UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

FROM TRANSPARENCY TO AGENCY: THE MICA FRAMEWORK FOR ADDRESSING THE UNDERSTANDING-TRUST GAP IN AI-ASSISTED CAMOUFLAGE ANALYSIS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

BY

DEBRA L. HOGUE
Norman, Oklahoma
2026

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/342777

FROM TRANSPARENCY TO AGENCY: THE MICA FRAMEWORK FOR ADDRESSING THE UNDERSTANDING–TRUST GAP IN AI-ASSISTED CAMOUFLAGE ANALYSIS

A DISSERTATION APPROVED FOR THE
SCHOOL OF COMPUTER SCIENCE

BY THE COMMITTEE CONSISTING OF

Dr. Ghulam Jilani Quadri, Chair

Dr. Dean F. Hougen, Co-Chair

Dr. Sridhar Radhakrishnan

Dr. Justin G. Metcalf

© Copyright by DEBRA L. HOGUE 2026
All Rights Reserved.

Acknowledgements

This journey would not have been possible without the remarkable people who illuminated my path.

To my mother, who instilled courage to pursue impossible dreams; to my husband, who stood beside me through late nights and doubt; to my son, whose future inspired me to persevere—your love has been my foundation.

I am deeply thankful to the DoD SMART Scholarship program for making this pursuit possible, and to Mr. Rick McIver, Mr. Michael Gourley, Dr. Tom Turner, Dr. Wayne Jones, and Mr. David Renfroe for championing my journey. Dr. Justin Lewis and Mrs. Shannon Alfaro became advocates whose support reminded me why this research matters.

A special thank you to Lacey Schley for introducing me to Dr. Shane Connelly, who was instrumental to this work—helping me secure the participation my user study needed, providing facility space, and granting access to the SONA system.

To Dr. Dean Hougen and Dr. Ghulam Jilani Quadri, who graciously stepped in to guide this work to completion—your mentorship during this difficult time has been invaluable.

To everyone named and the many whose contributions are imprinted on my heart, this dissertation exists because you do.

In Memory of Dr. Chris E. Weaver

This dissertation is dedicated to my advisor and friend, Dr. Chris Weaver, who passed away suddenly during the final stages of this work. Chris transformed me from a student to a scientist. He helped me refine my hypotheses, challenged me to think more critically, and taught me how to be an effective scientist. He revealed the elegant connections between human perception and artificial intelligence, encouraging me to explore them with both rigor and creativity. More than an advisor, he was a mentor who believed in my potential and a friend who celebrated every victory along this journey. Although he cannot see this work completed, his wisdom and influence permeate every page. The questions he taught me to ask and the passion for discovery he shared are his lasting gifts. This accomplishment honors his memory.

"We honor his legacy by carrying forward the curiosity, compassion, and excellence that defined his life and career."

— Wong et al. [2025]

Abstract

Effective integration of Artificial Intelligence (AI) into high-stakes visual detection tasks requires not only technical accuracy but calibrated human trust. Explainable AI (XAI) approaches assume that increasing transparency improves user understanding and, in turn, increases trust and reliance. However, this assumption remains insufficiently validated in operationally relevant domains.

This dissertation investigates the relationships among explainability, user understanding, agency, and trust in Camouflaged Object Detection and Segmentation (CODS). A controlled between-subjects factorial study ($N = 150$) compared an interactive explainable CODS system (MURDOC) to a baseline system (FACE). Participants completed structured detection tasks followed by standardized assessments of understanding, trust, usability, and perceived interactivity.

Results revealed a persistent dissociation between understanding and trust. While interactive explainability significantly increased perceived interactivity and engagement, it did not significantly improve user understanding and produced only modest increases in trust. Participants demonstrated high comprehension of system capabilities but only moderate willingness to rely on system outputs. This finding—termed the understanding–trust gap—challenges the assumption that transparency alone produces calibrated trust.

In response, this dissertation proposes the Mixed-Initiative Camouflage Anal-

ysis (MICA) framework, which reconceptualizes explainability as an active human–AI partnership rather than passive explanation delivery. MICA introduces structured interactive controls, including region-of-interest selection and sensitivity adjustment, to operationalize user agency within the detection process. A functional prototype demonstrates the feasibility of mixed-initiative CODS and establishes infrastructure for future empirical validation.

Together, these contributions provide empirical clarification of the limits of transparency-based approaches and introduce a framework that positions structured user agency as a mechanism supporting trust calibration in AI-assisted visual detection.

Keywords: Camouflaged object detection, explainable AI, mixed-initiative systems, trust calibration, human–AI collaboration, computer vision

Contents

Acknowledgements **iv**

Abstract **vi**

List of Figures **xiii**

List of Tables **xv**

List of Algorithms **xvi**

1 Introduction **1**
1.1 Problem Context and Motivation 2
1.2 The Understanding–Trust Gap 2
1.2.1 Initial Research Plan 3
1.2.2 Scope Evolution 3
1.2.3 Final Dissertation Scope 4
1.3 Research Questions and Hypotheses 5
1.3.1 Research Questions 5
1.3.2 Hypotheses 5
1.4 Theoretical Framework 6
1.4.1 Computer Vision Challenges 6
1.4.2 Human Perception and Expertise 6
1.4.3 Psychology of Control and Trust 7
1.5 Methodology Overview 7
1.5.1 Phase 1: Empirical Study 7
1.5.2 Phase 2: Framework Design 7
1.5.3 Phase 3: Synthetic Camouflage Generation 8
1.5.4 Integration and Evaluation Strategy 8
1.6 Contributions and Scope 9
1.6.1 Empirical Contribution 9
1.6.2 Theoretical Contribution 9
1.6.3 Technical Contribution 9
1.6.4 Computing Infrastructure Contribution 10

1.7 Scope and Delimitations 11
1.8 Contribution Summary 11
1.9 Dissertation Structure 12

2 Literature Review **13**
2.1 Camouflage and Visual Perception 13
2.2 Camouflaged Object Detection and Segmentation 14
2.2.1 Traditional Approaches 15
2.2.2 Deep Learning Approaches 15
2.3 Explainable AI in Visual Detection 16
2.4 Trust in Automation and AI 17
2.4.1 Control and Locus of Control 17
2.4.2 Human Factors in High-Stakes Systems 18
2.5 Human–AI Teaming 18
2.6 Mixed-Initiative Interaction 19
2.7 Limitations of One-Way Explainability 19
2.8 Synthetic Camouflage Generation 20
2.9 Synthesis of Gaps and Design Requirements 20
2.10 Summary 21

3 Empirical Evaluation of Explainability and Trust in AI-Assisted CODS **23**
3.1 Research Framework and Hypotheses 24
3.2 Methods 24
3.2.1 Research Questions 24
3.2.2 Experimental Design 25
3.2.3 Materials 29
3.2.4 Procedure 31
3.2.5 Measures and Analysis 33
3.3 Results 33
3.3.1 Descriptive Statistics 34
3.3.2 Assumption Checks 35
3.3.3 ANOVA Results: Composite Scores 36
3.3.4 ANOVA Results: Individual Items 39
3.3.5 Cybersecurity Deception Effects 41
3.3.6 Effect Sizes 41
3.3.7 The Understanding–Trust Gap 42
3.3.8 Qualitative Observations 43
3.4 Discussion 44
3.4.1 The Insufficiency of Passive Explainability 45
3.4.2 Agency Without Confidence 46
3.4.3 Three Mechanisms Underlying the Gap 47

3.4.4 Implications for Operational Contexts 48
3.4.5 Limitations 49
3.5 Summary 51

4 From Transparency to Agency: The MICA Framework 53
4.1 Theoretical Foundations 53
4.1.1 Formalizing the Understanding–Trust Gap 54
4.1.2 Modeling Trust through Shared Agency 54
4.1.3 The Insufficiency of Passive Explainability 55
4.1.4 Control Theory and Trust Formation 56
4.1.5 Mixed-Initiative Principles 56
4.2 The MICA Framework Architecture 57
4.2.1 Core Components 57
4.2.2 The Trust-Building Cycle 57
4.3 Operationalizing Interactive Control 58
4.3.1 Region-of-Interest Mechanisms 58
4.3.2 Sensitivity Controls 59
4.4 Theoretical Contributions 60
4.4.1 Reconceptualizing Explainability 60
4.4.2 Agency as a Trust Mechanism 61
4.4.3 Expertise-Adaptive Collaboration 61
4.5 Design Principles for Mixed-Initiative CODS 62
4.6 Relationship to Existing Frameworks 63
4.6.1 Cognitive Systems Engineering 63
4.6.2 Distributed Cognition 63
4.6.3 Activity Theory 63
4.6.4 Sensemaking Theory 63
4.7 Validation Approach 64
4.8 Summary 64

5 Instantiating Mixed-Initiative Agency: Prototype Design and Implementation 66
5.1 System Architecture 67
5.2 Mixed-Initiative Interaction Loop 70
5.3 Framework Component Instantiation 71
5.3.1 Explainability Enhancement 71
5.3.2 Human–AI Iterative Control 73
5.3.3 Contextual Environment Interpretation 73
5.3.4 Adaptive Decision Support 74
5.4 Unified Editing Environment 74
5.4.1 Interface Layout and Collapsible Panels 74
5.4.2 Tool Palette 76

5.4.3 Polygon Mask Editor 78
5.4.4 Rank Brush Editor 78
5.5 Mixed-Initiative Interaction Workflow 79
5.6 Detection Feedback and Session Management 81
5.6.1 Detection Feedback System 82
5.6.2 Session Management and Data Serialization 82
5.7 Algorithmic Implementation 85
5.7.1 Localization Update Mechanism 85
5.7.2 Rank Map Adjustment 86
5.7.3 Session Data Serialization 87
5.7.4 Session-Based Model Refinement 87
5.8 Signal Detection Parameter Controls 88
5.9 Prototype Limitations 90
5.9.1 Failure Modes and Their Implications for Trust 90
5.10 Chapter Summary 93

6 Perceptual Loss Design for Synthetic Camouflage Generation 95
6.1 Motivation and Research Context 96
6.2 The Latent-Space Integration Challenge 97
6.3 HSV-Gestalt Loss Framework 99
6.3.1 Loss Formulation 99
6.3.2 Perceptual Weight Map Mechanism 100
6.3.3 Progressive Integration Design 102
6.4 Experimental Setup 103
6.4.1 Dataset and Preprocessing 104
6.4.2 Multi-Architecture Training 104
6.4.3 Multi-Metric Evaluation 106
6.5 Results 107
6.5.1 SD-1.5: Modest Metrics, Stable Realism 109
6.5.2 SD-2.0: Best Metrics, Wandering Realism 109
6.5.3 SD-2.1: Larger Metric Improvements, Degradation to Abstraction 111
6.5.4 The Optimization–Perception Gap 113
6.6 Discussion 115
6.6.1 Architecture as a Design Variable 117
6.6.2 Implications for the Latent Diffusion Pipeline 118
6.6.3 Relationship to CODS Evaluation Infrastructure 118
6.6.4 Limitations 119
6.7 Summary 120

7 Beyond Deliberation: The Crisis Adoption Problem and Future Research Directions 122
7.1 Summary of Contributions . . . 123
7.2 Theoretical Implications: The Crisis Adoption Problem . . . 123
7.2.1 From Deliberation to Crisis . . . 124
7.2.2 The Crisis Adoption Problem . . . 125
7.2.3 Theoretical Foundations . . . 126
7.2.4 Relationship to the Understanding–Trust Gap . . . 128
7.2.5 Scope and Status . . . 129
7.3 Proposed Empirical Program . . . 129
7.3.1 Research Questions and Hypotheses . . . 129
7.3.2 Experimental Paradigms . . . 130
7.3.3 Measurement Considerations . . . 132
7.3.4 Prerequisite Relationship to Phase 1 . . . 133
7.4 Extensions to the MICA Framework . . . 133
7.4.1 Adaptive Automation and Trust-Aware Interaction . . . 134
7.4.2 Multimodal Camouflage Detection . . . 134
7.4.3 Collaborative Multi-User Analysis . . . 135
7.4.4 Integration with Broader Surveillance Pipelines . . . 135
7.5 Synthetic Data Generation and Model Robustness . . . 135
7.5.1 Controlled Evaluation via Synthetic Camouflage . . . 136
7.5.2 Adversarial Robustness Testing . . . 136
7.5.3 Transfer Learning and Domain Adaptation . . . 136
7.5.4 Environment-Adaptive Generation via Conditioned LoRA 137
7.6 Broader Implications for Human–AI Teaming . . . 138
7.6.1 Medical Imaging . . . 138
7.6.2 Autonomous Vehicle Monitoring . . . 138
7.6.3 Cybersecurity Threat Detection . . . 139
7.6.4 The Common Thread . . . 139
7.7 Limitations and Open Questions . . . 139
7.7.1 Ecological Validity . . . 140
7.7.2 Expertise Effects . . . 140
7.7.3 Longitudinal Trust Dynamics . . . 140
7.7.4 Cultural and Organizational Context . . . 141
7.7.5 Generalization Across AI Systems . . . 141
7.8 Summary . . . 141

8 Toward Calibrated Human–AI Collaboration 143
8.1 From Transparency to Agency: Contributions and Field-Level Implications . . . 143
8.2 Concluding Reflections . . . 146

A Definition of Terms **156**

B CITI Training Documentation **160**
B.1 Certificates . . . 160
B.1.1 Human Research . . . 160
B.1.2 Responsible Conduct of Research . . . 162
B.2 CITI Completion Reports . . . 163
B.2.1 Human Research . . . 163
B.2.2 Responsible Conduct of Research . . . 165

C Institutional Review Board (IRB) Documentation **167**
C.1 Approval of Study Modification — Personnel Change . . . 167
C.2 Approval of Study Modification — Instrument Update . . . 169
C.3 Approval of Initial Submission . . . 169

D Survey Instruments **172**
D.1 Pre-Interaction Survey . . . 172
D.1.1 Informed Consent . . . 172
D.1.2 Demographics . . . 173
D.1.3 Technology Experience and Habits . . . 173
D.2 Post-Interaction Survey . . . 175
D.2.1 Trust Items . . . 175
D.2.2 Understanding Items . . . 176
D.2.3 Usability Items . . . 177
D.2.4 Interactivity Items . . . 178
D.2.5 Additional Items . . . 178

E Statistical Analysis Code **180**
E.1 Python Analysis Script . . . 180
E.2 SPSS Verification Syntax . . . 187

F Complete Loss Ablation Results **191**

G Technical Background: Latent Diffusion Models **193**
G.1 Variational Autoencoders . . . 193
G.2 The UNet Architecture . . . 195
G.3 Latent Diffusion Models and Stable Diffusion . . . 196
G.4 Low-Rank Adaptation (LoRA) . . . 197

List of Figures

3.1 MURDOC interface with full explainability pipeline 27
3.2 FACE baseline interface 27
3.3 Composite scores by condition 35
3.4 Main effect of Application Type on Interactivity 38
3.5 Main effect of Application Type on Q15 (adjust settings) 40
3.6 Main effect of CD condition on Q19 (privacy priority) 42

4.1 Trust-building cycle 58

5.1 MURDOC_2024 system architecture 68
5.2 MICA mixed-initiative interaction loop 72
5.3 MICA workspace layout 75
5.4 MICA tool palette 77
5.5 MICA interactive editing tools 80
5.6 MICA session management interface 84
5.7 Signal detection parameter controls in the MICA interface. Sensitivity (d') and response bias (β) are exposed as interactive sliders, enabling users to adjust the detection decision boundary to match operational cost asymmetries. 89

6.1 COD10K dataset categories and representative examples 105
6.2 SD-1.5 baseline training progression across epochs 110
6.3 SD-2.0 baseline training progression across epochs 112
6.4 SD-2.1 baseline training progression across epochs 114
6.5 Optimization–perception paradox 116

B.1 CITI Training Certificate for Human Research 161
B.2 CITI Training Certificate for Responsible Conduct of Research . . 162
B.3 CITI Training Completion Report for Human Research 164
B.4 CITI Training Completion Report for Responsible Conduct of Research 166

C.1 IRB Approval of Personnel Change (June 2025) 168
C.2 IRB Approval of Study Modification (February 20, 2025) 170

C.3 IRB Approval of Initial Submission (December 08, 2024) 171

List of Tables

3.1 System Feature Comparison: MURDOC vs. FACE 26
3.2 Descriptive Statistics for Composite Scores by Experimental Condition 34
3.3 Two-Way ANOVA Results for Composite Scores 36
3.4 Two-Way ANOVA Results for Individual Items with Significant Application Effects 39
3.5 Significant Cybersecurity Deception (CD) Effects on Individual Items 41
3.6 Effect Sizes (Cohen's d) for Main Effects on Composite Scores . . 43

5.1 Recurring failure modes in MURDOC_2024, their trigger conditions, observable symptoms, and the MICA interaction mechanisms available for operator recovery. 92

6.1 Progressive HSV-Gestalt Loss Function Integration 103
6.2 Six-preset perceptual-loss progression across three architectures . 108

D.1 Post-Interaction Survey Item–Composite Mapping 176

F.1 Complete 15-preset ablation across three architectures 192

List of Algorithms

1 Hierarchical ROI Selection 59
2 Polygon Mask Update via Contour Rasterization 85
3 Rank Map Adjustment via Gaussian Brush 86
4 Session-Based LoRA Model Refinement 88
5 HSV-Gestalt Total Loss Computation 100
6 Perceptual Weight Map Construction via CPU-Side VAE Decoding 102
7 LoRA Fine-Tuning with HSV-Gestalt Perceptual Loss 106

Chapter 1

Introduction

Artificial Intelligence (AI) systems are increasingly deployed for visual detection tasks, where human operators must interpret and act on model outputs under uncertainty. In high-stakes contexts, technical accuracy alone is insufficient; effective deployment depends on calibrated human trust.

Explainable Artificial Intelligence (XAI) approaches aim to improve trust by increasing transparency. The prevailing assumption is that greater transparency improves user understanding, and that greater understanding, in turn, increases trust and reliance. However, this relationship remains insufficiently validated in complex visual detection domains.

Camouflaged Object Detection and Segmentation (CODS) presents a uniquely demanding context for examining this assumption. Camouflage suppresses visual cues and produces ambiguous outputs that require interpretive judgment. In such environments, operators may understand system capabilities yet remain hesitant to rely on them.

This dissertation investigates whether transparency-based explainability is sufficient to produce calibrated trust in AI-assisted camouflage detection systems.

1.1 Problem Context and Motivation

Despite extensive investment in XAI, the relationship between explainability, understanding, and trust remains unresolved. Many systems assume a linear progression:

$$\text{Transparency} \rightarrow \text{Understanding} \rightarrow \text{Trust} \rightarrow \text{Reliance}$$

However, empirical validation of this progression remains limited. If understanding does not reliably translate into trust, then transparency-based approaches alone may be insufficient for operational adoption.

The central problem addressed in this dissertation is the unresolved relationship between explainability, user understanding, agency, and trust in AI-assisted camouflage detection.

1.2 The Understanding–Trust Gap

This dissertation identifies an important limitation in current explainable AI systems: increased transparency may improve user understanding, but it does not necessarily lead to operational trust or confident decision-making. This phenomenon is referred to as the *understanding–trust gap*. In operational environments, this gap frequently manifests as a broader *trust–action gap*, in which users understand system outputs yet remain hesitant to rely on them in decision-making contexts.

1.2.1 Initial Research Plan

The initial research plan investigated three complementary dimensions of camouflaged object detection and segmentation:

1. **Explainability and Trust:** Examining whether explainable AI features improve user understanding, trust, and willingness to rely on CODS systems.

2. **Multidimensional Segmentation:** Exploring segmentation approaches that integrate color, texture, and geometric cues.

3. **Color Space Optimization:** Evaluating whether alternative color spaces (e.g., HSV, LAB) improve separability between camouflaged objects and background textures.

Together, these dimensions were intended to advance both technical performance and operational usability.

1.2.2 Scope Evolution

Two factors shaped the final direction of this dissertation.

First, the passing of my advisor, Dr. Chris Weaver, in June 2025 required a transition in committee guidance and a refinement of research scope.

Second, early empirical investigations revealed a recurring pattern in user interactions with CODS systems: users generally understood system outputs, yet remained hesitant to trust or rely on them in decision-making contexts. This observation suggested that explainability alone may be insufficient for trust formation, particularly under short exposure and high uncertainty, conditions common in operational CODS use.

These insights motivated a shift in emphasis from purely explanatory interfaces toward interaction mechanisms that allow users to influence, guide, and interrogate system behavior during analysis.

1.2.3 Final Dissertation Scope

Based on this evolution, the dissertation focuses on three core components:

1. **Empirical Investigation (Chapter 3):** A controlled user study examining how explainability affects understanding, trust, usability, and engagement in CODS systems.

2. **Framework Development (Chapter 4):** Design of the Mixed-Initiative Camouflage Analysis (MICA) framework, which introduces interactive control mechanisms to support user agency and bidirectional human–AI collaboration.

3. **Theoretical Integration:** Analysis of how perceptual psychology, trust in automation, and mixed-initiative interaction theory explain the limitations of explainability-only approaches.

Multidimensional segmentation and color space optimization remain important research directions and are addressed as future work. The synthetic camouflage generation framework developed in Chapter 6 provides the infrastructure to support those investigations and constitutes a completed technical contribution in its own right.

1.3 Research Questions and Hypotheses

1.3.1 Research Questions

This dissertation addresses the following research questions:

- **RQ1:** How does explainability influence user understanding and trust in AI-assisted camouflage detection systems?
- **RQ2:** What structured interaction mechanisms enable users to meaningfully influence CODS system behavior under uncertainty?
- **RQ3:** How can synthetic camouflage generation support controlled evaluation of CODS systems across varied concealment strategies?

1.3.2 Hypotheses

The research is guided by three hypotheses:

H1: Explainability features improve user understanding of CODS systems but are insufficient on their own to produce calibrated trust.

H2: Interactive control mechanisms enable users to guide detection processes and address limitations that explainability-only approaches cannot resolve.

H3: Synthetic camouflage generation enables controlled evaluation of CODS systems across varied concealment strategies, and perceptual loss design fundamentally shapes generative behavior in ways that standard evaluation metrics do not capture.

Empirical evaluation of H1 is presented in Chapter 3. H2 motivates the design of the MICA framework introduced in Chapter 4. H3 is operationalized in

Chapter 6, which evaluates a perceptually guided generation framework and documents a systematic optimization–perception gap in latent diffusion fine-tuning.

1.4 Theoretical Framework

This research integrates three complementary theoretical foundations: limitations of computer vision, principles of human perception, and psychological models of control and trust.

1.4.1 Computer Vision Challenges

Camouflaged objects suppress visual cues such as edges, contrast, and distinctive texture features that many detection algorithms rely on. Although deep learning architectures improve detection accuracy, their internal representations remain opaque. This opacity limits interpretability and constrains operator trust in operational settings where understanding system behavior is critical.

1.4.2 Human Perception and Expertise

Human observers detect camouflage by leveraging contextual inconsistencies, perceptual grouping, and domain expertise. However, human perception is susceptible to cognitive overload, attentional limits, and bias. Effective human–AI collaboration, therefore, requires systems that combine human contextual sensitivity with algorithmic consistency rather than replacing one with the other.

1.4.3 Psychology of Control and Trust

Trust in automation is influenced not only by performance but by perceived control and agency. Theories of self-efficacy and locus of control [Bandura, 1977, Rotter, 1966] suggest that individuals are more likely to trust systems they can influence. These principles inform the MICA framework's emphasis on structured interactive control as a mechanism for trust calibration, formalized in Chapter 4 (Equations 4.1 and 4.2).

1.5 Methodology Overview

This dissertation proceeds in three primary phases: empirical investigation, framework development, and synthetic camouflage generation.

1.5.1 Phase 1: Empirical Study

Chapter 3 presents a controlled experimental study examining the relationship between explainability, user understanding, trust, usability, and engagement in Camouflaged Object Detection and Segmentation (CODS) systems. Participants interacted with either an explainable system (MURDOC) or a baseline system (FACE) under a between-subjects design. Quantitative measures assessed perceived understanding, trust, interactivity, and usability to evaluate whether explainability produces measurable gains in trust.

1.5.2 Phase 2: Framework Design

Chapter 4 introduces the Mixed-Initiative Camouflage Analysis (MICA) framework. Informed by Phase 1 findings and established theories of trust and agency,

MICA operationalizes structured interactive control mechanisms, such as region-of-interest selection and sensitivity adjustment, to position user agency as a central component of trust formation, detailed in Algorithm 1 and Equations 4.3 and 4.4. The prototype implementation operationalizes these mechanisms through rank map adjustment and mask editing formalized in Equations 5.1, 5.2, and 5.3, with session-based model refinement described in Algorithms 3 and 4.

1.5.3 Phase 3: Synthetic Camouflage Generation

Chapter 6 presents a framework for synthetic camouflage generation that integrates Hue-Saturation-Value (HSV) color analysis and Gestalt perceptual principles into the fine-tuning of Stable Diffusion via Low-Rank Adaptation (LoRA). The training procedure is formalized in Algorithm 7, with perceptual guidance provided through the CPU-side variational autoencoder (VAE) decoding mechanism described in Algorithm 6. This work establishes a controlled evaluation infrastructure for CODS research and documents a systematic optimization–perception gap in latent diffusion training.

1.5.4 Integration and Evaluation Strategy

The empirical findings from Phase 1 establish the understanding–trust gap, which motivates the design principles formalized in MICA. Chapter 5 implements a functional prototype to demonstrate the feasibility of mixed-initiative CODS. Comprehensive empirical evaluation of MICA's long-term effects on trust calibration, performance, and operational adoption is identified as future work.

1.6 Contributions and Scope

This dissertation makes four primary contributions:

1.6.1 Empirical Contribution

It demonstrates a measurable gap between users' understanding and trust in an AI-assisted camouflage detection system. A controlled study (N=150) shows that interactive explainability increases perceived interactivity but does not significantly improve understanding and produces only modest increases in trust. This finding empirically establishes the understanding–trust gap.

1.6.2 Theoretical Contribution

It proposes the Mixed-Initiative Camouflage Analysis (MICA) framework, which reconceptualizes explainability as a structured human–AI partnership. MICA introduces user agency through region-of-interest selection and sensitivity control to support trust calibration, with expertise-adaptive interaction modes formalized in Equation 4.5.

1.6.3 Technical Contribution

It implements a functional prototype that operationalizes the MICA framework and introduces a perceptually guided synthetic camouflage generation framework, documented in Chapter 6, which establishes a controlled evaluation infrastructure for CODS research. Framework effectiveness is evaluated using the composite metric formalized in Equation 4.6. The synthetic generation framework is operationalized through the LoRA fine-tuning and perceptual weight map procedures

formalized in Algorithms 6 and 7.

1.6.4 Computing Infrastructure Contribution

Implementing MICA required resolving a practical engineering problem that lacks a standard solution in the mixed-initiative AI literature: coordinating a responsive, event-driven graphical interface with a computationally intensive Python-based inference pipeline across a runtime boundary. This dissertation contributes NEXUS (*.NET–Python EXecution Unified System*), a bridging architecture that launches Python inference scripts as managed subprocesses from within the C# runtime, passing parameters and image file paths at invocation and ingesting structured JSON outputs upon completion. This design preserves the responsiveness of the Windows Presentation Foundation (WPF) interface while executing RankNet and EfficientDet-D7 asynchronously, without requiring the embedding overhead and dependency conflicts that arise from tighter cross-runtime integration strategies. The application follows the Model-View-ViewModel (MVVM) architectural pattern, which decouples detection engine outputs, user editing actions, and session analytics from the presentation layer, allowing any component to be modified or replaced independently. Because the frontend and backend are architecturally isolated, the prototype functions as a reusable experimental platform: alternative detection models can be substituted into the Python pipeline without modifying the interface, and new interaction mechanisms can be evaluated without retraining underlying AI components. This extensibility is a deliberate design choice that positions MURDOC_2024 as infrastructure for the future empirical program described in Chapter 7. The rank map adjustment and session-based retraining pipelines that operationalize this architecture are

formalized in Algorithms 3 and 4 and Equations 5.2 and 5.3.

1.7 Scope and Delimitations

This dissertation empirically evaluates trust and understanding within controlled laboratory conditions. It does not measure long-term operational reliance, crisis decision-making behavior, or field deployment outcomes.

The Crisis Adoption Problem introduced in Chapter 7 is presented as a theoretical extension motivated by the findings of this work and is not empirically validated within the scope of this study.

1.8 Contribution Summary

This dissertation provides four primary contributions:

1. **Empirical evidence of the understanding–trust gap** in AI-assisted camouflage detection, demonstrating that interactive and explainable features significantly improve perceived interactivity but do not proportionally increase user trust or understanding.

2. **The MICA framework** for mixed-initiative camouflage analysis, grounded in psychological theory, which reconceptualizes the human–AI relationship as a bidirectional partnership and introduces structured user agency as the operative mechanism for trust calibration (Equations 4.1 and 4.2; Algorithm 1).

3. **Synthetic generation infrastructure** that translates Gestalt perceptual principles and HSV color analysis into differentiable training objectives for

Stable Diffusion, with a documented optimization–perception gap that provides principled guidance for the design of perceptual losses in domain-specific generative modeling (Algorithms 6 and 7).

4. **NEXUS computing infrastructure** implementing the MICA prototype as a modular, extensible hybrid architecture that bridges a C#/WPF interactive interface with a Python-based AI detection pipeline, establishing a reusable experimental platform for future mixed-initiative interaction research (Algorithms 3 and 4; Equations 5.1, 5.2, and 5.3).

Together, these contributions advance the design of trustworthy AI systems for high-stakes applications by clarifying the structural relationship among transparency, agency, and trust, and by demonstrating that structured interactive control is a more reliable mechanism for calibrated reliance than passive explainability alone.

1.9 Dissertation Structure

Chapter 2 synthesizes relevant literature. Chapter 3 reports the empirical evaluation. Chapter 4 introduces the MICA framework, and Chapter 5 describes its implementation. Chapter 6 presents the synthetic camouflage generation framework. Chapter 7 discusses theoretical implications, the Crisis Adoption Problem, and future research directions. Chapter 8 summarizes conclusions and closing reflections.

Chapter 2

Literature Review

Camouflaged Object Detection and Segmentation (CODS) remains one of the most challenging problems in computer vision because camouflage suppresses or distorts the cues typically used for object identification. This chapter synthesizes prior work on camouflage theory and visual perception; computational approaches to CODS; explainable AI (XAI) for visual systems; and foundations in trust in automation, human–AI teaming, and mixed-initiative interaction. Together, these literatures motivate the need for CODS systems that support not only transparency, but also user agency and bidirectional interaction—design requirements that inform the Mixed-Initiative Camouflage Analysis (MICA) framework introduced in Chapter 4.

2.1 Camouflage and Visual Perception

Camouflage conceals objects by manipulating color, texture, shape, or motion to achieve background matching or disruptive patterning [Cott, 1940, Stevens and Merilaita, 2009, Price et al., 2019]. These strategies reduce contrast and disrupt

structural cues necessary for reliable detection. Human detection of camouflage can depend on contextual inconsistencies, subtle shadows, and texture discontinuities, but performance degrades in cluttered or unfamiliar environments. Classical perceptual work shows that humans use Gestalt principles—closure, similarity, proximity, and figure–ground segmentation—to infer structure from ambiguous scenes [Wagemans et al., 2012]. In the CODS context, these principles help explain why camouflaged scenes are cognitively demanding and why users may seek visual cues that align with perceptual grouping and boundary inference.

While humans are adaptable and context-sensitive, they are susceptible to fatigue, cognitive overload, and attentional biases. Automated systems offer speed and consistency, but often lack the high-level contextual inference humans apply when interpreting ambiguous scenes. These complementary strengths motivate hybrid approaches that combine computational detection with human perceptual expertise.

2.2 Camouflaged Object Detection and Segmentation

CODS aims to identify objects blended into their environment. Unlike conventional detection tasks that rely on salient features and strong boundaries, CODS must accommodate low-contrast, irregular patterns and suppressed edges. These characteristics create recurring technical challenges:

1. **Low feature visibility:** camouflage weakens color and edge contrast, complicating feature extraction.

2. **Background clutter:** complex natural scenes can produce false positives

and distract attention.

3. **Intra-class variability:** camouflage strategies vary widely across environments and object types.

2.2.1 Traditional Approaches

Early methods relied on grayscale conversion, edge detection, and region-based segmentation [Rosenfeld, 1988]. Techniques such as Canny edges, thresholding, and Gabor filters often struggle because camouflage is explicitly designed to obscure these cues, limiting robustness across realistic environments.

2.2.2 Deep Learning Approaches

Recent CODS methods use CNN- and transformer-based architectures to identify subtle patterns, often incorporating multi-scale feature extraction and attention mechanisms. Representative approaches include SINet [Fan et al., 2020a], PraNet [Fan et al., 2020b], TPRNet [Mei et al., 2021], and RankNet [Lv et al., 2023]. These methods have advanced benchmark performance, but several limitations persist in applied settings:

- **Interpretability limitations:** models are often difficult to interrogate, complicating user understanding and calibrated reliance.
- **Dataset constraints:** camouflage datasets remain limited in size, diversity, and environmental coverage.
- **Generalization challenges:** performance can degrade under clutter, low-light conditions, and domain shifts.

- **Static emphasis:** many pipelines prioritize single-image inference and may not generalize to motion-based camouflage cues.

These limitations motivate research directions that integrate explainability, user-centered interaction, and mechanisms for human guidance during analysis.

2.3 Explainable AI in Visual Detection

Explainable AI aims to make model behavior observable and interpretable. Common visual XAI techniques include Grad-CAM [Selvaraju et al., 2017], integrated gradients [Lucas et al., 2022], guided backpropagation [Springenberg et al., 2014], and occlusion sensitivity [Zeiler and Fergus, 2014, He et al., 2025]. These approaches frequently produce saliency maps or attribution heatmaps indicating regions that influence predictions.

However, prior work highlights recurring challenges when XAI is applied to visual detection systems:

- **Human alignment:** explanations may not correspond to features users consider meaningful.
- **Stability:** heatmaps can be sensitive to noise, model initialization, or minor input perturbations.
- **Post-hoc nature:** explanations often do not provide actionable pathways to change model behavior.
- **Misuse risk:** users may overinterpret explanations or infer reliability that the system has not earned.

In high-stakes settings (i.e., security, intelligence, search-and-rescue), transparency alone may be insufficient. Prior literature emphasizes that explainability does not reliably produce calibrated trust and can sometimes create misplaced confidence [Ashoori and Weisz, 2019, Akash et al., 2020]. This motivates interaction paradigms that support not only observation of model behavior, but also user agency and corrective influence when uncertainty is high.

2.4 Trust in Automation and AI

Trust influences whether users rely on, comply with, or override automated systems [Kohn et al., 2021, Chiou and Lee, 2023]. Over-trust can lead to complacency, while under-trust can result in disuse and lost operational benefits. Across the trust-in-automation literature, factors that shape trust include perceived reliability, transparency, user expertise, and perceived control [Lee and See, 2004, Kohn et al., 2021, Chiou and Lee, 2023].

2.4.1 Control and Locus of Control

Foundational psychological theories emphasize the role of agency. Bandura's self-efficacy theory [Bandura, 1977] and Rotter's locus of control [Rotter, 1966] suggest that people engage more productively with tools when they believe their actions can influence outcomes. In automation contexts, perceived controllability and recourse—being able to intervene, adjust, or correct—are repeatedly linked to trust calibration and reliance behaviors [Chiou et al., 2021, Kohn et al., 2021, Chiou and Lee, 2023]. Applied to CODS, these theories motivate interface and interaction mechanisms that allow users to shape analysis (i.e., prioritizing regions, adjusting sensitivity, or providing corrective input), especially when camouflage

uncertainty is unavoidable.

2.4.2 Human Factors in High-Stakes Systems

Research in aviation, defense, and medical domains emphasizes that trust deteriorates when errors are unexplained and when users cannot predict or influence system behavior. These findings support a shift from one-way "explanation delivery" toward designs that incorporate bidirectional interaction and user control mechanisms.

2.5 Human–AI Teaming

Human–AI teaming focuses on coordination between human expertise and machine capabilities. Effective teams support mutual awareness of goals and constraints, leverage complementary strengths (human contextual inference; AI consistency and speed), and provide communication channels for feedback and adjustment.

In medical imaging workflows, human-in-the-loop systems allow experts to guide, correct, or refine AI-generated segmentations using interactive tools [Marinov et al., 2024, Zhu and Wu, 2024]. These patterns suggest a generalizable lesson for CODS: when uncertainty is inherent, effective systems often provide structured pathways for users to steer model behavior, resolve ambiguity, and calibrate reliance rather than merely observe outputs.

2.6 Mixed-Initiative Interaction

Mixed-initiative systems allow both humans and AI to contribute proactively to task execution [Horvitz, 1999, Allen et al., 1999]. These systems support shared control, dynamic task allocation, continuous feedback, and adaptation to user expertise and preferences. Mixed-initiative interaction is especially appropriate in domains where perfect automation is unattainable or undesirable—conditions that characterize CODS due to its ambiguity, variable visibility, and sensitivity to context.

In medical image analysis, clinicians frequently incorporate AI suggestions while retaining decision authority, supported by interactive annotation and adjustment tools [Wang et al., 2021, Rezazade Mehrizi et al., 2023, Bernstein et al., 2023]. Analogously, CODS systems may benefit from interaction techniques that preserve user agency while leveraging AI for systematic scanning.

2.7 Limitations of One-Way Explainability

Many XAI interfaces are fundamentally unidirectional: the system explains and the user receives information without a structured mechanism to respond. Prior literature has identified several shortcomings of this model in applied settings: transparency does not guarantee calibrated trust; users cannot express intent, challenge outputs, or supply corrective feedback; and XAI is treated as an endpoint rather than an interactive process.

These limitations motivate mixed-initiative alternatives in which explanations are coupled with mechanisms for intervention, refinement, and verification—particularly in high-stakes contexts where users need recourse under un-

certainty.

2.8 Synthetic Camouflage Generation

Synthetic data generation can enable controlled evaluation of model performance across environments, textures, lighting conditions, and concealment strategies. Advances in generative modeling—including diffusion models and parameter-efficient fine-tuning such as LoRA [Hu et al., 2022]—support high-quality synthesis that preserves content while varying camouflage strength and texture.

Dataset constraints remain a bottleneck for CODS. Examples of natural camouflage can be difficult to collect at scale and may lack precise ground-truth segmentation masks. Synthetic generation can support systematic manipulation of camouflage properties (i.e., color relationships, texture patterns, blending intensity) while maintaining known labels, enabling controlled testing of model behaviors under targeted camouflage conditions.

Because synthetic generation can support evaluation across controlled conditions, it is relevant to downstream research questions involving robustness, feature separability, and comparative analysis across model variants. (Specific tool implementations and dissertation contributions are described in Chapter 1 and later technical chapters, rather than in this literature review.)

2.9 Synthesis of Gaps and Design Requirements

Across the CODS, XAI, trust-in-automation, human–AI teaming, and mixed-initiative literatures, several gaps emerge that motivate the design of mixed-initiative CODS systems.

Gap 1: Transparency without recourse. Visual explanations can increase observability, but often do not provide users with structured means to guide, correct, or contest outputs, limiting their utility under uncertainty.

Gap 2: Trust requires more than understanding. Trust is shaped by perceived reliability, risk, and the ability to intervene when needed [Ashoori and Weisz, 2019, Akash et al., 2020, Kohn et al., 2021]. Thus, even when users understand system goals and outputs, they may remain unwilling to rely on the system without mechanisms for control, verification, or correction.

Gap 3: CODS uncertainty motivates shared control. Because camouflage is designed to suppress cues, CODS systems face unavoidable ambiguity and context sensitivity. Teaming and mixed-initiative paradigms suggest that shared control and bidirectional feedback are appropriate when automation cannot be perfectly dependable.

Gap 4: Dataset limitations constrain robustness claims. Limited and unevenly representative datasets complicate generalization across environments and camouflage strategies. Synthetic generation is a promising approach for controlled stress testing and targeted evaluation.

These gaps motivate a mixed-initiative approach in which CODS systems provide (a) interpretable visual cues aligned with user perception, and (b) interaction mechanisms that support user intent expression, adjustable operating points (i.e., sensitivity), and corrective feedback during analysis.

2.10 Summary

Camouflaged object detection lies at the intersection of computer vision, human perception, and trust in automation. While deep CODS models have advanced

benchmark performance, interpretability limitations and unavoidable ambiguity create challenges for calibrated reliance in applied settings. Prior work in XAI, trust, human–AI teaming, and mixed-initiative interaction motivates designs that move beyond one-way explanation toward bidirectional interaction and user agency.

Chapter 3 reports an empirical study evaluating how explainability and interactive controls influence trust, understanding, usability, and perceived interactivity in CODS interfaces. Chapter 4 then introduces the Mixed-Initiative Camouflage Analysis (MICA) framework, which operationalizes user agency and bidirectional interaction for CODS analysis.

Chapter 3

Empirical Evaluation of Explainability and Trust in AI-Assisted CODS

This chapter reports the empirical study examining interactive explainability, user understanding, trust, and usability in AI-assisted CODS. A 2×2 factorial design compared an interactive system (MURDOC) with a baseline system (FACE) and assessed whether cybersecurity deception awareness moderates user perceptions.

Methods and statistical procedures are detailed in Section 3.2, with results presented in Section 3.3. The findings establish a measurable dissociation between understanding and trust, providing the empirical foundation for the MICA framework introduced in Chapter 4.

Portions of this work have been submitted for publication [Hogue et al., 2026b]; this chapter presents the complete analysis.

3.1 Research Framework and Hypotheses

This dissertation advances three hypotheses spanning algorithmic, perceptual, and interactive dimensions of camouflaged object detection. The present chapter focuses on the first:

H1 (Explainability and Interactive Features): Integrating explainability and interactive features into CODS systems improves user trust, understanding, and perceived usability.

H2 (Interactive Control and Mixed-Initiative Design) and **H3 (Synthetic Camouflage Generation)** address the framework and infrastructure contributions, respectively. H2 motivates the MICA framework introduced in Chapter 4. H3 is operationalized in Chapter 6.

3.2 Methods

This section details the methodology used to empirically evaluate the impact of interactive explainability on user understanding and trust. It first outlines the primary research questions driving the study, then describes the 2×2 between-subjects factorial experimental design. It then details the materials used, including the MURDOC and FACE system interfaces, and outlines the standardized experimental procedure. Finally, it describes the quantitative and qualitative measures utilized for data analysis.

3.2.1 Research Questions

The empirical study addressed:

- **RQ1.1**: Does interactive explainability increase user trust and engagement

relative to a baseline system?

- **RQ1.2**: Does cybersecurity deception awareness moderate user perceptions of AI systems?

- **RQ1.3**: What is the relationship between system understanding and trust in high-stakes visual detection contexts?

3.2.2 Experimental Design

A 2×2 between-subjects factorial design manipulated:

Application Type:

- *MURDOC* (Interactive Explainable): Grad-CAM visualizations, confidence overlays, detection rationale, and interactive image adjustment controls.

- *FACE* (Baseline): Identical detection model providing segmentation output and confidence scores without explainability or interactive features.

Cybersecurity Deception Condition:

- *CD Present*: Participants received a brief orientation on cybersecurity deception techniques and how camouflage relates to adversarial attacks on AI systems.

- *CD Absent*: Participants received standard task orientation without cybersecurity framing.

This yielded four groups with balanced cell sizes of $n = 37$–38 (total $N = 150$): MURDOC–CD ($n = 37$), MURDOC–No CD ($n = 38$), FACE–CD ($n = 37$), FACE–No CD ($n = 38$).

Table 3.1: System Feature Comparison: MURDOC vs. FACE

Feature Category	**MURDOC (Interactive Explainable)**	**FACE (Baseline)**
Visual Explanations	Grad-CAM maps and confidence heatmaps	None (hidden from user)
Textual Rationale	Natural language detection rationale	Numerical confidence scores only
Image Controls	Active sliders for brightness, contrast, saturation	None (static viewing)
Feedback Loop	Real-time preview and dual-pane comparison	Static segmentation mask output

Note: Both systems used identical PyTorch-based SINet detection models. MURDOC added interactive controls and explainability features; FACE provided only the core detection output.

Table 3.1 summarizes the feature differences between the two systems.

Figures 3.1 and 3.2 illustrate the visual difference between the two interfaces as presented to participants.

Design Note: Interactive and explainable features were jointly manipulated: MURDOC included both, while FACE included neither. Therefore, the independent effects of interactivity and explainability could not be isolated in this design. This choice was deliberate. Estimating the combined effect of interactive explainability represents a necessary first step before future studies isolate the independent contributions of interactivity and explanation features [Hogue et al., 2026b].

Power Analysis and Sample Size Justification

The target sample size of $N = 150$ was determined through *a priori* power analysis using G*Power 3.1 [Faul et al., 2009], based on a hierarchical analysis strategy

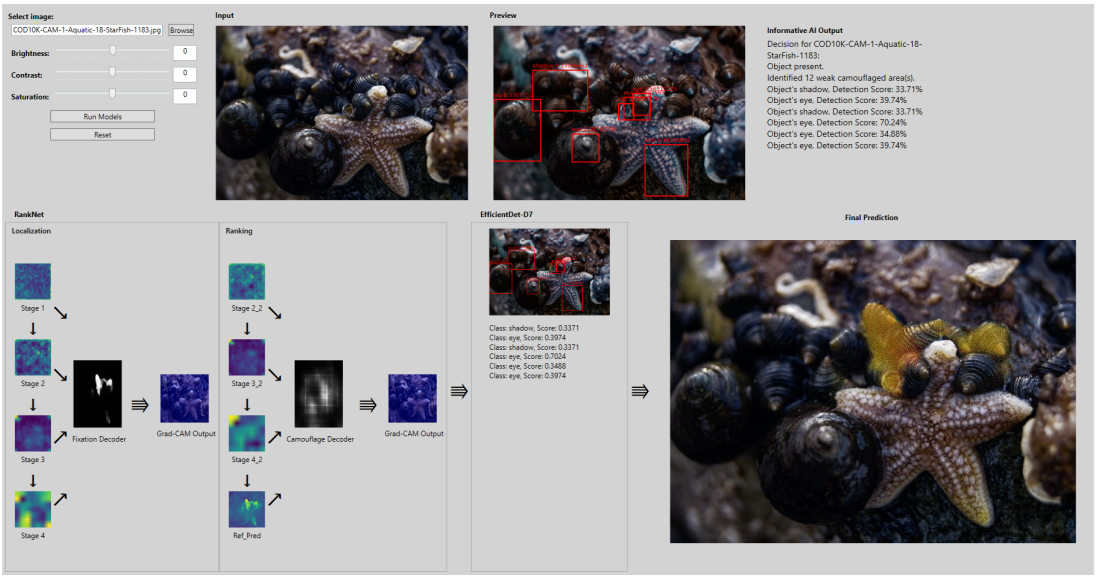

Figure 3.1: The MURDOC interface displaying the full explainability pipeline for a starfish camouflage image. The system shows the input image with preprocessing controls (left), a preview with EfficientDet-D7 bounding boxes and detection scores (top center), RankNet localization and ranking stage outputs with Grad-CAM visualizations (bottom left and center), and the final prediction with Informative AI output (right). This interface constitutes the interactive explainable condition in the study.

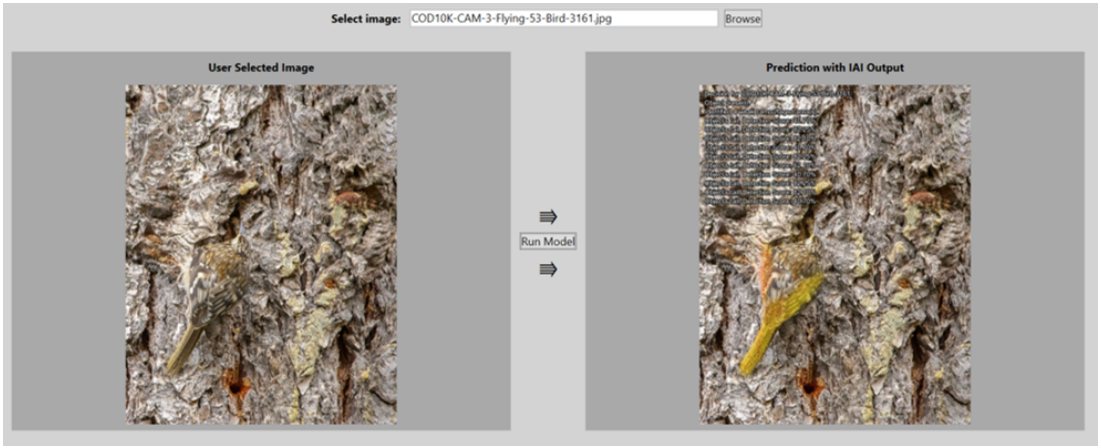

Figure 3.2: The FACE baseline interface displaying a camouflaged bird image. The system provides only the input image (left) and the segmentation prediction with basic IAI output (right), without Grad-CAM visualizations, confidence heatmaps, or interactive preprocessing controls. This interface constitutes the baseline condition in the study.

that prioritizes the primary research question while maintaining adequate power for secondary analyses.

Hierarchical Analysis Approach: The study employs two levels of analysis:

- **Level 1 (Primary)**: Comparison of MURDOC vs. FACE (2 groups), addressing the core research question of whether interactive explainability enhances trust, understanding, and usability.
- **Level 2 (Secondary)**: Full 2×2 factorial analysis (Application $\times$ CD), examining whether cybersecurity deception awareness moderates user perceptions.

Primary Analysis Power: For the main comparison of MURDOC vs. FACE, treating this as an independent-samples t-test (or equivalently, a one-way ANOVA with two groups), power calculations assumed:

- Effect size: $d = 0.5$ (medium effect, based on prior automation trust literature [Kohn et al., 2021])
- Significance level: $\alpha = 0.05$ (two-tailed)
- Desired power: $1 - \beta = 0.80$

The final sample size provides 80% power to detect medium effects ($d = 0.5$) and substantially higher power for larger effects.

Secondary Analysis Power: For the full 2×2 factorial ANOVA with approximately 37–38 participants per cell, the design achieves 62.2% power for medium effects ($f = 0.25$), 81.1% power for medium-large effects ($f = 0.30$), and 94.6% power for large effects ($f = 0.40$).

The final sample met the target with perfectly balanced cell sizes ($n = 37$–38), providing high confidence in the adequacy of power to detect meaningful effects. Observed effect sizes ranged from negligible to large ($\eta_p^2 < 0.001$ to $\eta_p^2 = 0.507$), consistent with adequate sample size.

3.2.3 Materials

Systems: Both applications used identical PyTorch-based SINet [Fan et al., 2020a] detection models wrapped in a C#/.NET interface. MURDOC included interactive features (brightness, contrast, and saturation controls with real-time preview) and explainability enhancements (Grad-CAM activation maps, confidence visualizations, detection rationale text); FACE omitted these features, providing only the segmentation output with basic confidence scores.

Images: Test set images from the COD10K dataset [Fan et al., 2020a] were selected to span easy, medium, and hard camouflage difficulty levels based on model confidence and visual complexity.

Surveys:

- *Pre-interaction Survey*: Demographics, technology experience, and baseline attitudes toward automation and AI.
- *Post-interaction Survey*: 20 items measuring multiple constructs on 5-point Likert scales (1=Strongly Disagree, 5=Strongly Agree):
  - Q1: How confident are you in the accuracy of the application's camouflage detection results?
  - Q2: I understand what the camouflage detection application should do.

- Q3: How likely are you to rely on this application's outputs in real-world scenarios?
- Q4: I understand the application's limitations.
- Q5: I understand the application's capabilities.
- Q6: How comfortable would you feel using this application without human oversight?
- Q7: I understand how the application executes its tasks.
- Q8: How would you rate the overall reliability of the application?
- Q9: The application performs the way that it should.
- Q10: How likely are you to recommend this application to colleagues for camouflage detection tasks?
- Q11: The application helps me achieve my goals in camouflage detection.
- Q12: The application performs consistently.
- Q13: I feel comfortable relying on the information that the application provided.
- Q14: I am rarely surprised by how the application responds.
- Q15: The application allowed me to adjust settings to refine its output.
- Q16: I felt engaged with the application while using it.
- Q17: The application provided real-time feedback as I used it.
- Q18: I feel that the application protects and handles data responsibly.
- Q19: The application seems to prioritize user privacy.

- Q20: I am confident that unauthorized parties cannot easily manipulate the application's output.

Composite Scores: Calculated as the mean of constituent items:

- *Trust*: Q1 (confidence in accuracy), Q3 (likelihood to rely), Q8 (overall reliability), Q13 (comfort relying on information)
- *Understanding*: Q2 (understand purpose), Q4 (understand limitations), Q5 (understand capabilities), Q7 (understand execution)
- *Usability*: Q9 (performs as expected), Q10 (likelihood to recommend), Q11 (achievement of goals), Q12 (consistency)
- *Interactivity*: Q15 (adjust settings), Q16 (engagement), Q17 (real-time feedback)

All items used 5-point Likert scales to support composite calculation.

3.2.4 Procedure

Each session lasted approximately 20 minutes and followed a standardized protocol:

1. **Pre-interaction survey** (2 min): Demographics and baseline technology attitudes.
2. **Cybersecurity deception orientation** (3 min, CD condition only): Brief explanation of how camouflage techniques relate to adversarial attacks on AI systems and cybersecurity deception.
3. **System interaction** (12 min):

- Standardized introduction to assigned application with user guide.
- Guided walkthrough of system features.
- Two structured detection tasks with think-aloud commentary encouraged.
- Free exploration period with additional test images.

4. **Post-interaction survey** (2 min): Trust, understanding, usability, and interactivity assessments.

5. **Debriefing**: Full disclosure of study purpose and experimental manipulations.

Randomization: Participants were randomly assigned to one of four conditions using a computer-generated randomization sequence with block sizes of 8, ensuring balanced cell sizes throughout data collection. Image presentation order within the free exploration phase was counterbalanced using a Latin square design to control for potential order effects. Experimenters were blind to the study hypotheses but not to the condition assignment, given the visible differences between the MURDOC and FACE interfaces.

Participants were recruited from two sources: non-expert participants ($n = 92$), undergraduate students recruited through the university's SONA system who received 1.5 research credits, and domain experts ($n = 58$), students with educational backgrounds in Computer Science, Software Engineering, or related technical fields recruited through university departmental channels, who were entered into a raffle for a Nintendo Switch Lite (valued at $160). No participants withdrew data following debriefing.

3.2.5 Measures and Analysis

The primary dependent variables were the Trust, Understanding, Usability, and Interactivity composites, each calculated as the mean of its constituent Likert items.

Separate two-way ANOVAs (Application $\times$ CD) were conducted for each dependent variable and individual item. Type III sums of squares were used to accommodate potential unbalanced designs. Effect sizes are reported as partial η_p^2 ($small = 0.01 - 0.06, medium = 0.06 - 0.14, large \geq 0.14$) and Cohen's d.

Qualitative Analysis: Participant comments from the post-interaction survey open-ended questions were reviewed for recurring themes. Comments were iteratively reviewed and organized into thematic categories following an inductive approach informed by Braun and Clarke [Braun and Clarke, 2006]. The analysis focused on identifying patterns that illuminated or contextualized the quantitative findings, particularly regarding user preferences for interactive features and the relationship between understanding and trust.

3.3 Results

This section presents the statistical and qualitative findings from the empirical study. It begins by reporting descriptive statistics and assumption checks, followed by the primary ANOVA results for both the composite scores and the individual survey items. It then examines the specific effects of the cybersecurity deception framing and reports the standardized effect sizes. Finally, it analyzes the empirical dissociation between understanding and trust, the understanding–trust gap, and contextualizes the quantitative data with qualitative observations derived from participant feedback.

3.3.1 Descriptive Statistics

Table 3.2 presents means and standard deviations for composite scores by Application and CD condition. The sample was perfectly balanced, with cell sizes ranging from $n = 37$ to $n = 38$.

Table 3.2: Descriptive Statistics for Composite Scores by Experimental Condition

Variable	**Condition**	M	SD	n
Trust	MURDOC–CD	3.78	0.51	37
	MURDOC–No CD	3.67	0.74	38
	FACE–CD	3.47	0.91	37
	FACE–No CD	3.47	0.89	38
	Overall	*3.59*	*0.78*	*150*
Understanding	MURDOC–CD	4.25	0.54	37
	MURDOC–No CD	4.13	0.54	38
	FACE–CD	4.25	0.51	37
	FACE–No CD	4.15	0.61	38
	Overall	*4.19*	*0.55*	*150*
Usability	MURDOC–CD	4.12	0.57	37
	MURDOC–No CD	4.05	0.74	38
	FACE–CD	3.87	0.94	37
	FACE–No CD	3.82	0.92	38
	Overall	*3.96*	*0.81*	*150*
Interactivity	MURDOC–CD	4.24	0.73	37
	MURDOC–No CD	4.14	0.51	38
	FACE–CD	3.55	0.76	37
	FACE–No CD	3.04	0.87	38
	Overall	*3.74*	*0.87*	*150*

Note. All scores on a 5-point Likert scale. Trust = mean of Q1, Q3, Q8, Q13; Understanding = mean of Q2, Q4, Q5, Q7; Usability = mean of Q9, Q10, Q11, Q12; Interactivity = mean of Q15, Q16, Q17.

Descriptive statistics indicate a persistent gap between Understanding ($M = 4.19$) and Trust ($M = 3.59$), a difference of 0.60 points. Understanding showed

virtually no difference between applications (MURDOC: $M = 4.19$; FACE: $M = 4.20$), while Trust showed a meaningful difference (MURDOC: $M = 3.72$; FACE: $M = 3.47$).

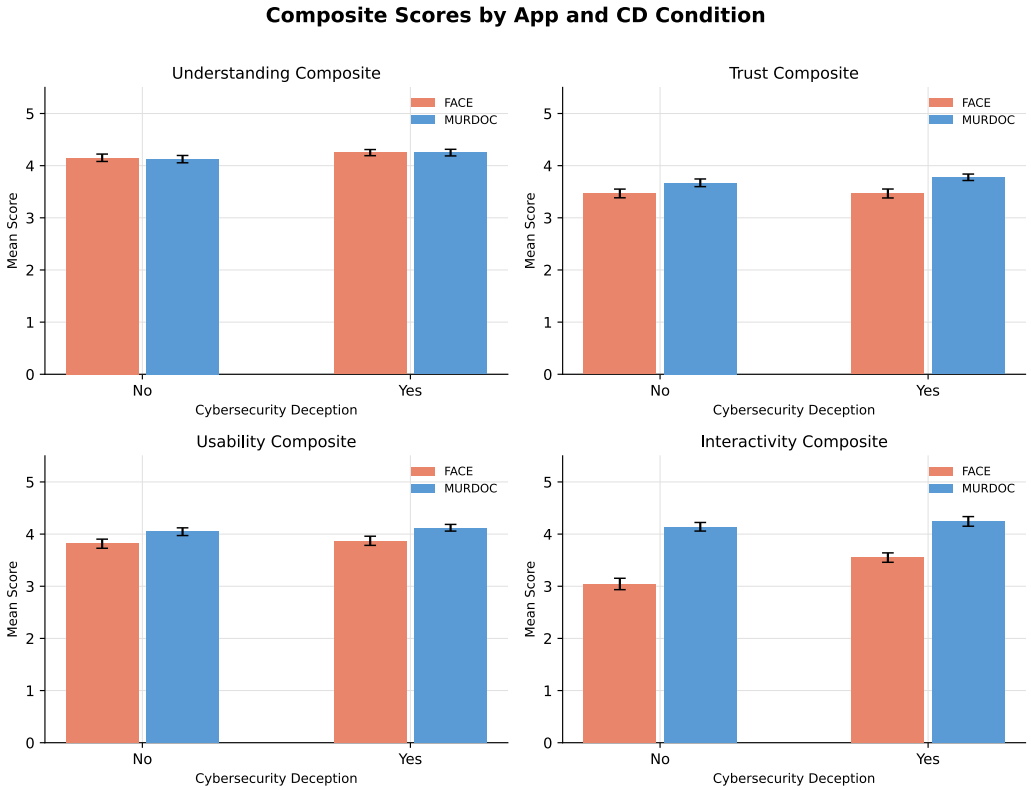

Figure 3.3: Composite scores by Application Type and Cybersecurity Deception condition. Understanding scores were uniformly high ($M > 4.1$) across conditions, while Interactivity showed the largest differences between MURDOC and FACE. The persistent gap between Understanding and Trust is visible across all conditions. Error bars represent standard errors.

3.3.2 Assumption Checks

Normality and homogeneity of variance assumptions were evaluated for all dependent variables. While some violations were detected (common with Likert-scale data and balanced designs), ANOVA is robust to moderate departures with sam-

ples of this size ($N = 150$) [Schmider et al., 2010]. All variance ratios were within acceptable limits (ratio < 4.0), and the balanced design further ensures robustness [Box, 1954].

3.3.3 ANOVA Results: Composite Scores

Table 3.3 presents two-way ANOVA results for the four composite scores.

Table 3.3: Two-Way ANOVA Results for Composite Scores

Source	**DV**	*df*	*F*	*p*	η_p^2
Application	Trust	1, 146	4.09	0.045*	0.027
	Understanding	1, 146	0.02	0.883	< 0.001
	Usability	1, 146	3.32	0.070	0.022
	Interactivity	1, 146	56.47	< 0.001***	0.279
CD	Trust	1, 146	0.17	0.680	0.001
	Understanding	1, 146	1.53	0.218	0.010
	Usability	1, 146	0.25	0.618	0.002
	Interactivity	1, 146	6.49	0.012*	0.043
Application × CD	Trust	1, 146	0.18	0.675	0.001
	Understanding	1, 146	0.02	0.884	< 0.001
	Usability	1, 146	0.01	0.940	< 0.001
	Interactivity	1, 146	2.84	0.094	0.019

Note. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

Trust

The main effect of Application on Trust reached significance, $F(1, 146) = 4.09$, $p = 0.045$, $\eta_p^2 = 0.027$, with a small effect size. MURDOC produced higher Trust scores ($M = 3.72$, $SD = 0.63$) than FACE ($M = 3.47$, $SD = 0.89$), representing a mean difference of 0.25 points (Cohen's $d = 0.33$). Neither the main effect of CD, $F(1, 146) = 0.17$, $p = 0.680$, $\eta_p^2 = 0.001$, nor the Application × CD interaction, $F(1, 146) = 0.18$, $p = 0.675$, $\eta_p^2 = 0.001$, reached significance.

Understanding

The main effect of Application on Understanding was not significant, $F(1, 146) = 0.02$, $p = 0.883$, $\eta_p^2 < 0.001$. MURDOC ($M = 4.19$, $SD = 0.54$) and FACE ($M = 4.20$, $SD = 0.56$) produced virtually identical Understanding scores, with a negligible mean difference of 0.01 points (Cohen's $d = -0.02$). Both systems achieved high Understanding scores ($M > 4.1$ on a 5-point scale). Neither the main effect of CD, $F(1, 146) = 1.53$, $p = 0.218$, $\eta_p^2 = 0.010$, nor the Application $\times$ CD interaction, $F(1, 146) = 0.02$, $p = 0.884$, $\eta_p^2 < 0.001$, reached significance.

Usability

The main effect of Application on Usability approached but did not reach conventional significance, $F(1, 146) = 3.32$, $p = 0.070$, $\eta_p^2 = 0.022$. MURDOC ($M = 4.08$, $SD = 0.66$) was rated slightly higher than FACE ($M = 3.84$, $SD = 0.92$), with a mean difference of 0.24 points and a small effect size (Cohen's $d = 0.30$). Neither the main effect of CD, $F(1, 146) = 0.25$, $p = 0.618$, $\eta_p^2 = 0.002$, nor the Application $\times$ CD interaction, $F(1, 146) = 0.01$, $p = 0.940$, $\eta_p^2 < 0.001$, reached significance.

Interactivity

The main effect of Application on Interactivity was highly significant and represented the largest observed effect, $F(1, 146) = 56.47$, $p < 0.001$, $\eta_p^2 = 0.279$. MURDOC ($M = 4.19$, $SD = 0.63$) was rated dramatically higher than FACE ($M = 3.29$, $SD = 0.85$), representing a mean difference of 0.90 points with a very large effect size (Cohen's $d = 1.19$). This effect accounted for 27.9% of the variance in Interactivity scores.

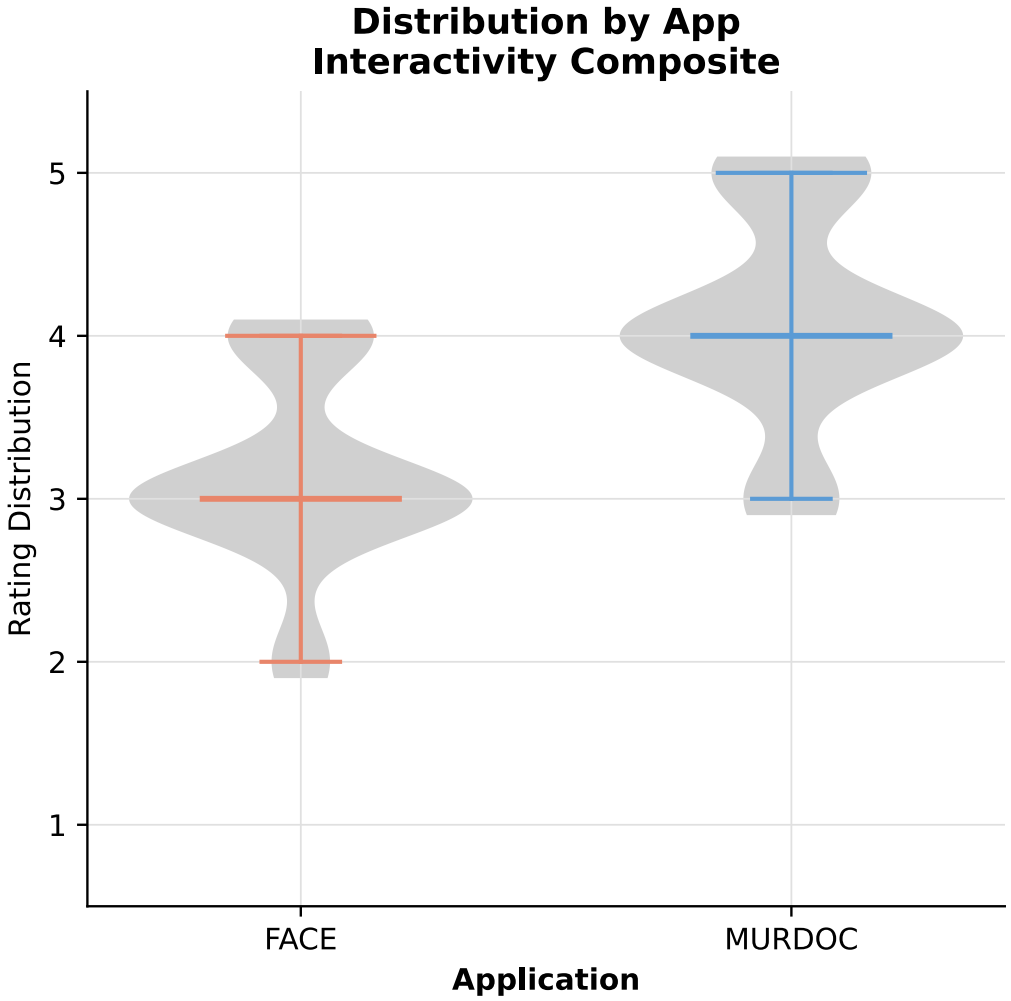

Figure 3.4: Distribution of Interactivity Composite ratings by Application Type. MURDOC users (median = 4, IQR = 3–5) rated interactivity substantially higher than FACE users (median = 3, IQR = 2–4), consistent with the large main effect, $F(1, 146) = 56.47$, $p < 0.001$, $\eta_p^2 = 0.279$.

There was also a significant main effect of CD on Interactivity, $F(1, 146) = 6.49$, $p = 0.012$, $\eta_p^2 = 0.043$, with a small effect size. Participants in the CD condition ($M = 3.90$, $SD = 0.82$) rated interactivity higher than those in the No CD condition ($M = 3.59$, $SD = 0.90$). The Application × CD interaction approached but did not reach conventional significance, $F(1, 146) = 2.84$, $p = 0.094$, $\eta_p^2 = 0.019$. Cell means showed that the CD effect was slightly larger for

FACE users (No CD: $M = 3.04$; CD: $M = 3.55$; $\Delta = 0.51$) than for MURDOC users (No CD: $M = 4.14$; CD: $M = 4.24$; $\Delta = 0.10$), but this pattern was not statistically reliable.

3.3.4 ANOVA Results: Individual Items

Table 3.4 presents ANOVA results for individual questionnaire items showing significant or noteworthy Application effects.

Table 3.4: Two-Way ANOVA Results for Individual Items with Significant Application Effects

Item	**Question Text**	$F(1, 146)$	p	η_p^2
Significant Application Effects (MURDOC M vs. FACE M)				
Q15	The application allowed me to adjust settings to refine its output	150.10	< 0.001***	0.507
			4.49 vs. 2.71	
Q16	I felt engaged with the application while using it	17.68	< 0.001***	0.108
			4.23 vs. 3.55	
Q11	The application helps me achieve my goals in camouflage detection	6.86	0.010**	0.045
			4.24 vs. 3.88	
Q10	How likely are you to recommend this application to colleagues?	6.82	0.010**	0.045
			4.15 vs. 3.73	
Marginal Application Effects				
Q8	How would you rate the overall reliability of the application?	3.80	0.053	0.025
			3.76 vs. 3.51	
Q3	How likely are you to rely on this application's outputs in real-world scenarios?	2.81	0.096	0.019
			3.53 vs. 3.25	

Note. All items on a 5-point Likert scale. Means shown as MURDOC vs. FACE. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

Item Q15 ("The application allowed me to adjust settings to refine its output") produced the study's most dramatic finding, $F(1, 146) = 150.10$, $p < 0.001$, $\eta_p^2 = 0.507$, accounting for 50.7% of the variance. MURDOC users ($M = 4.49$, $SD = 0.78$) rated adjustment ability far higher than FACE users ($M = 2.71$, $SD = 1.00$), with a mean difference of 1.78 points (Cohen's $d = 2.00$).

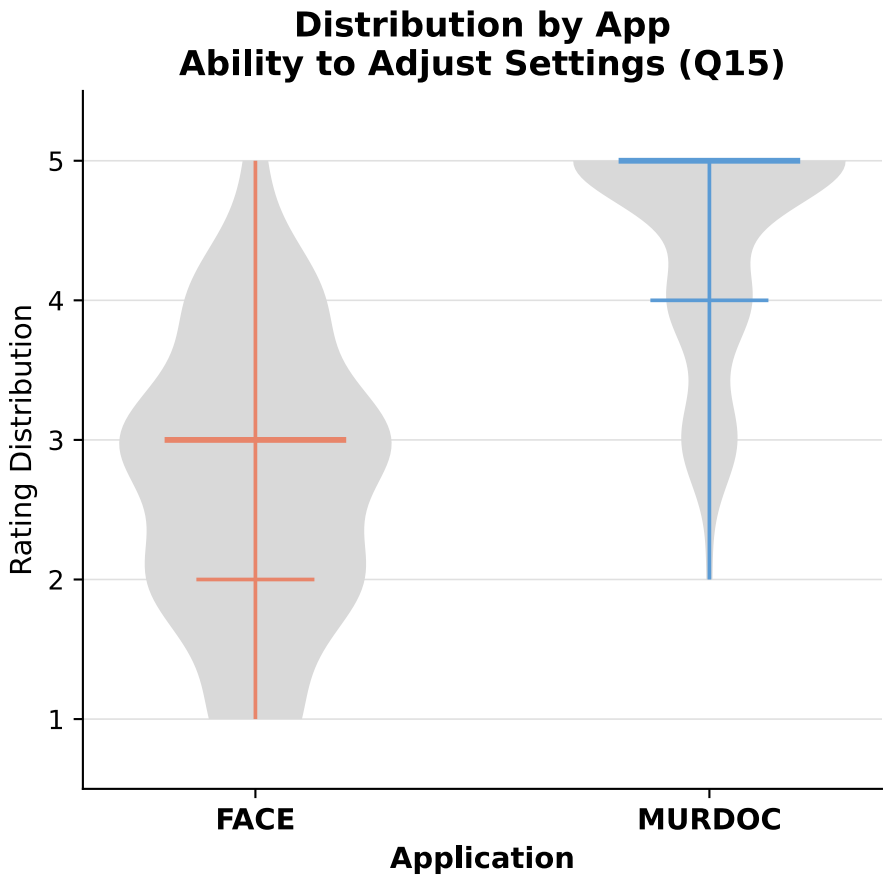

Figure 3.5: Main effect of Application Type on ability to adjust settings (Q15). This item produced the study's largest effect ($\eta_p^2 = 0.507$), with MURDOC users rating adjustment capabilities nearly 2 points higher than FACE users on the 5-point scale. $F(1, 146) = 150.10$, $p < 0.001$, $\eta_p^2 = 0.507$. $^{***}p < 0.001$.

Q16 (engagement) showed a large Application effect, $F(1, 146) = 17.68$, $p < 0.001$, $\eta_p^2 = 0.108$, with MURDOC ($M = 4.23$, $SD = 0.89$) rated substantially

higher than FACE ($M = 3.55$, $SD = 1.12$). Q11 (achievement of goals) and Q10 (likelihood to recommend) both showed significant Application effects with small-to-medium effect sizes ($\eta_p^2 = 0.045$ for both, $p = 0.010$).

3.3.5 Cybersecurity Deception Effects

While Application Type dominated the results, the CD manipulation produced selective effects on privacy awareness and engagement (see Table 3.5).

Table 3.5: Significant Cybersecurity Deception (CD) Effects on Individual Items

Item	**Description**	$F(1, 146)$	p	η_p^2	**CD** M
Q19	User privacy priority	17.59	< 0.001***	0.108	4.24 vs. 3.66
Q16	Engagement	7.30	0.008**	0.048	4.11 vs. 3.67
Q7	Understanding execution	4.03	0.047*	0.027	4.12 vs. 3.79

Note. CD M = mean for CD Present condition; values after "vs." are No CD means. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

The strongest CD effect appeared on Q19 ("The application seems to prioritize user privacy"), $F(1, 146) = 17.59$, $p < 0.001$, $\eta_p^2 = 0.108$. Participants who received cybersecurity framing ($M = 4.24$, $SD = 0.86$) rated privacy concerns greater than those without such framing ($M = 3.66$, $SD = 0.84$).

3.3.6 Effect Sizes

Table 3.6 summarizes standardized effect sizes for all main effects on composite scores.

The pattern of effect sizes reveals a striking dissociation: the interactive explainable system produced a massive improvement in perceived Interactivity

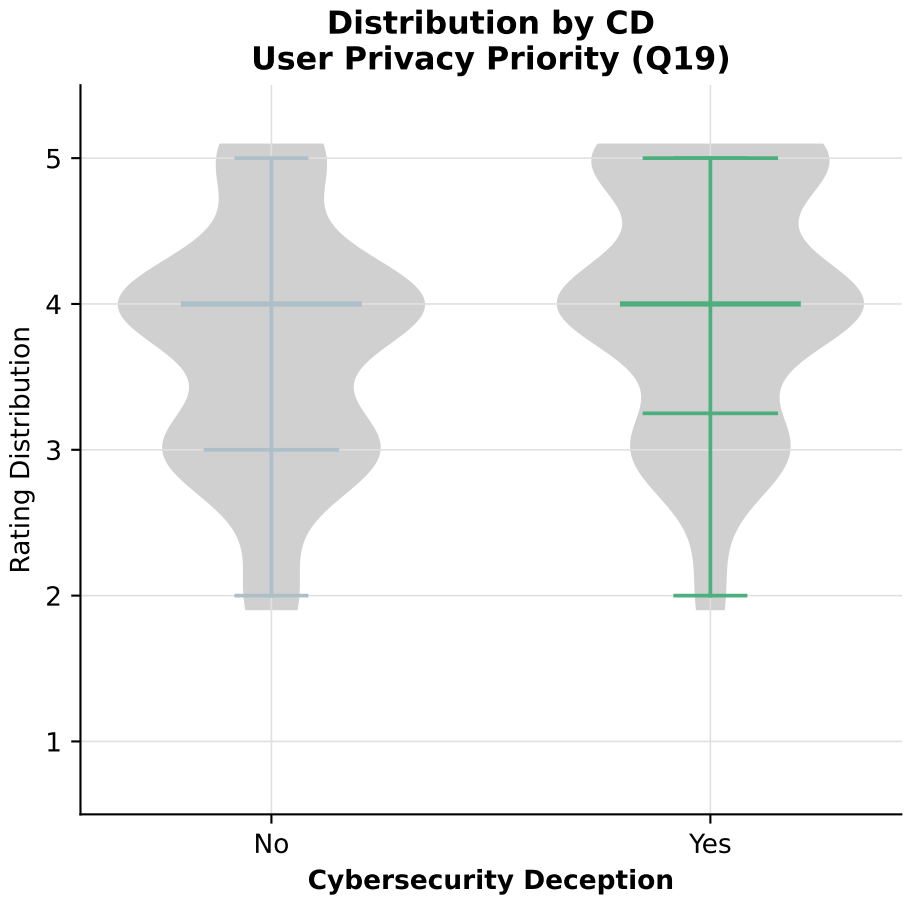

Figure 3.6: Main effect of Cybersecurity Deception condition on user privacy priority (Q19). Participants who received cybersecurity deception orientation rated privacy concerns significantly higher than those without such framing. $F(1, 146) = 17.59$, $p < 0.001$, $\eta_p^2 = 0.108$. $^{***}p < 0.001$.

($d = 1.19$), but only negligible differences in Understanding ($d = -0.02$) and small improvements in Trust ($d = 0.33$) and Usability ($d = 0.30$).

3.3.7 The Understanding–Trust Gap

The persistent gap between Understanding and Trust scores is visible across all experimental conditions. Despite achieving high Understanding (overall $M = 4.19/5.0$), participants demonstrated only moderate Trust ($M = 3.59/5.0$), re-

Table 3.6: Effect Sizes (Cohen's d) for Main Effects on Composite Scores

Composite	Application (MURDOC vs. FACE)	CD (CD vs. No CD)
Trust	+0.33 (small)	+0.07 (negligible)
Understanding	−0.02 (negligible)	+0.20 (small)
Usability	+0.30 (small)	+0.08 (negligible)
Interactivity	+1.19 (very large)	+0.36 (small)

Note. Positive values favor MURDOC over FACE (Application) and CD over No CD. Effect size interpretation: $|d| < 0.20$ = negligible; 0.20–0.50 = small; 0.50–0.80 = medium; > 0.80 = large [Cohen, 1988].

sulting in an average gap of 0.60 points (15% of the scale range). MURDOC users exhibited a 0.47-point gap (Understanding: $M = 4.19$; Trust: $M = 3.72$), while FACE users showed a larger 0.73-point gap (Understanding: $M = 4.20$; Trust: $M = 3.47$). The gap persisted regardless of CD condition.

This dissociation is further demonstrated by the ANOVA results: Application had no effect on Understanding ($F = 0.02$, $p = 0.883$, $\eta_p^2 < 0.001$) but a significant effect on Trust ($F = 4.09$, $p = 0.045$, $\eta_p^2 = 0.027$). Understanding and trust were empirically dissociated.

3.3.8 Qualitative Observations

Thematic analysis of participant comments revealed several consistent patterns:

- **Interactive controls universally valued**: Participants across all conditions praised MURDOC's brightness, contrast, and saturation controls. These features were described as "intuitive," "helpful for exploring images," and "making it feel like I have more control."

- **Explainability appreciated but insufficient for trust**: Participants

acknowledged that Grad-CAM visualizations showed *what* the system detected, but expressed that these did not address *why* regions indicated camouflage or *how reliable* detections were. Representative comment: "I can see where it's looking, but I don't know if it's looking at the right things."

- **Desire for bidirectional influence**: Participants consistently requested features to "tell the system where to focus," "adjust how sensitive it is," or "give feedback when it's wrong." These requests appeared across both CD conditions.

- **Trust requires performance evidence**: Multiple participants stated they would need to see "consistent results across many images" or "proof it works in different scenarios" before fully trusting outputs.

- **CD condition increased security awareness**: Participants in the CD condition made more comments about "adversarial attacks," "system vulnerabilities," and "data security," aligning with the significant effect on Q19 (privacy priority). However, this heightened awareness did not translate into a measurable difference in trust levels.

- **Recognition of the understanding–trust dissociation**: Several participants explicitly articulated the gap: "I understand how it works, but I'm not sure I'd bet on it being right" and "It makes sense what it's doing, but I wouldn't use it for anything important yet."

3.4 Discussion

This section interprets the empirical results and explores their broader implications for human-AI interaction in camouflage detection. It examines the insuffi-

ciency of passive explainability and the phenomenon of users exercising agency without necessarily developing performance-based confidence. It then proposes three potential mechanisms that may underlie the observed understanding–trust gap, discusses the implications of these findings for operational deployment contexts, and concludes by acknowledging the study's methodological limitations.

3.4.1 The Insufficiency of Passive Explainability

The central finding of this study is that an interactive explainable system, despite producing exceptionally large effects on user experience, failed to close the understanding–trust gap. Users achieved high comprehension of system capabilities ($M = 4.19/5.0$) but withheld operational confidence ($M = 3.59/5.0$) pending performance evidence they could not obtain in a single session. This dissociation aligns with Jacovi *et al.*'s [Jacovi et al., 2021] distinction between *understanding-based trust* and *performance-based trust*, and extends it by demonstrating that even combined interactive and explainable features, not just passive explanations alone, are insufficient to bridge the gap.

The null effect on Understanding is theoretically informative: it challenges the core assumption of XAI research that providing explanations improves comprehension. Both MURDOC and FACE users achieved virtually identical Understanding scores ($M \approx 4.2$), suggesting that users formed adequate mental models through interaction alone, regardless of whether explicit explainability features were present. This pattern is consistent with Ehsan et al.'s [Ehsan et al., 2021] finding that users readily form mental models through direct system interaction, and raises the possibility that self-reported understanding measures may be insufficiently sensitive to detect differences in mental model depth or accuracy.

3.4.2 Agency Without Confidence

The magnitude of the interactivity effect warrants examination. When users can adjust system behavior and observe consequences directly, they are exercising the kind of control that self-determination theory identifies as a fundamental psychological need [Deci and Ryan, 2000]. That MURDOC's interactive controls produced a Cohen's $d = 2.00$ on adjustment ability while generating only $d = 0.33$ on trust reveals that agency and confidence are not the same construct. Users valued the ability to manipulate the system without necessarily believing its outputs.

This finding suggests that interactivity satisfies the need for autonomy but, by itself, does not generate the performance evidence that trust requires. The feature gap between MURDOC and FACE was experienced primarily as an *agency gap*: users responded powerfully to the opportunity for control, independent of whether that control translated into greater confidence in the AI's reliability. Because MURDOC bundles interactive controls with explainability features, we cannot attribute this effect to either feature type individually, but the magnitude ($\eta_p^2 = 0.507$) is exceptional by HCI standards and suggests that the interactive component is likely a substantial contributor, given its direct relationship to the adjustment ability item.

That the interactivity effect was consistent across both CD conditions further indicates that the desire for control is not moderated by contextual threat framing but reflects a general human preference in human–AI collaboration. This generalizability strengthens the case for incorporating interactive control mechanisms into AI systems deployed across diverse operational contexts.

3.4.3 Three Mechanisms Underlying the Gap

The understanding–trust gap reflects an asymmetry in cognitive versus evidential evaluation. Understanding is achieved through observation of system behavior; trust requires accumulated evidence of reliability [Lee and See, 2004, Hoff and Bashir, 2015]. Three mechanisms may explain why understanding-based trust does not convert to performance-based trust.

Temporal asymmetry appears to be the primary driver. Understanding can be achieved in a single session through demonstration and explanation, while trust develops over extended interaction histories. The 12-minute interaction period was sufficient for understanding but insufficient for trust calibration. Multiple participants explicitly stated they would need "consistent results across many images" or "proof it works in different scenarios" before fully trusting outputs. This aligns with Muir's [Muir, 1994] finding that trust in automation develops through repeated exposure and observed reliability rather than initial explanation.

Epistemic versus performance judgments represent a second mechanism. Understanding addresses "What does this system do?" while trust addresses "Will it do it correctly when I need it to?" These are distinct epistemic judgments. A user can fully understand that a system detects camouflaged objects while remaining uncertain whether it will correctly detect *this particular* camouflaged object in *this particular* scene.

Asymmetric risk perception may also contribute. Errors of misplaced trust (relying on an incorrect output) may be perceived as more costly than errors of misplaced distrust (failing to rely on a correct output), particularly in high-stakes domains such as defense applications. This asymmetry would rationally

produce lower trust than understanding, as users hedge against potential false reliance [Parasuraman and Riley, 1997]. Future research employing longitudinal designs could distinguish the relative contributions of these mechanisms.

3.4.4 Implications for Operational Contexts

A critical observation is that these findings were obtained under *ideal deliberative conditions*: no time pressure, no consequences for errors, no competing task demands, and a structured 12-minute exploration period. Participants could examine the system at their own pace, explore features freely, and reflect carefully before completing the survey. If the understanding–trust gap persists under conditions that maximize the opportunity for careful evaluation, then the assumption underlying most XAI research, that users will deliberatively assess explanations before deciding whether to rely on AI outputs, may be fundamentally misaligned with operational reality.

Defense analysts, medical professionals, and security operators rarely encounter AI systems for the first time under calm, exploratory conditions. The persistence of the gap even under ideal conditions suggests that the problem is not merely one of providing *better* explanations or *more* interactive features, but of addressing a structural mismatch between how XAI systems are designed and how adoption decisions actually unfold. This observation motivates a broader reexamination of how AI trust is conceptualized, which is developed in Chapter 7.

The design response to the empirical findings, shifting from passive explanation to active mixed-initiative collaboration, is presented in Chapter 4, which introduces the MICA framework.

3.4.5 Limitations

Confounded design: The most consequential limitation is that MURDOC and FACE differed on both interactivity and explainability simultaneously. Because these features were never separated, effects cannot be attributed to either feature type individually, and potential interactions between them cannot be assessed. A three-condition design, baseline, explainability-only, and interactive-only, would be required to isolate individual contributions. The current results establish what the combined approach achieves; future controlled studies are needed to determine the mechanism [Hogue et al., 2026b]. The design estimates the combined effect; isolating mechanisms requires follow-up studies.

Single-session exposure: The 12-minute interaction period, while sufficient to establish understanding, may have been too brief for meaningful trust development. The temporal asymmetry mechanism discussed above suggests that trust requires extended observation of system performance. Longitudinal studies tracking trust development over multiple sessions would provide stronger tests of whether interactive features eventually close the understanding–trust gap.

Sample characteristics: Participants included undergraduate students recruited through the university's SONA system (non-experts, $n = 92$) and students with technical backgrounds in Computer Science or Software Engineering recruited through departmental channels (domain experts, $n = 58$). While this sample is appropriate for establishing basic psychological phenomena, generalization to operational contexts (i.e., military analysts, security professionals) requires caution. Expert populations with specialized camouflage detection training may show different trust calibration patterns.

Self-report measures: All dependent variables were measured through self-

report questionnaires. While these measures have established validity for assessing subjective perceptions, they may not perfectly predict behavioral trust, that is, actual reliance on system outputs in consequential decisions. Future research should incorporate behavioral measures such as decision-making tasks with real consequences.

Ceiling effects on Understanding: Understanding scores approached the scale maximum ($M = 4.19$ on a 5-point scale), which may have attenuated potential differences between conditions. This ceiling may reflect either that the CODS task is inherently easy to comprehend or that self-reported understanding measures lack sufficient sensitivity to detect differences in mental model depth or accuracy.

Ecological validity: The laboratory setting and artificial task framing differ from operational deployment contexts. Real-world camouflage detection involves time pressure, consequences for errors, and integration with other information sources. These contextual factors may moderate the effects observed here, a possibility that the operational implications discussed above make particularly salient.

Despite these limitations, the study provides robust evidence for the central theoretical contribution: the dissociation between understanding and trust in AI-assisted decision-making systems. The confounded design, while limiting mechanistic conclusions, establishes the empirical baseline necessary for future controlled studies isolating the independent effects of interactivity and explainability on trust formation [Hogue et al., 2026b].

3.5 Summary

This chapter reported an empirical evaluation of an interactive explainable approach to AI-assisted camouflaged object detection using a controlled 2×2 factorial design ($N = 150$). The principal findings are:

- **Interactivity**: Application Type produced a large effect on perceived interactivity, with adjustment capability (Q15) showing the strongest observed effect in the study.
- **Understanding**: Application Type had no effect on understanding. MURDOC and FACE produced virtually identical comprehension scores.
- **Trust**: Application Type produced a small but significant increase in trust; however, this improvement did not eliminate the gap between understanding and trust.
- **Understanding–Trust Dissociation**: A consistent 0.60-point gap separated understanding and trust across conditions, indicating that comprehension and confidence are empirically distinct constructs [Jacovi et al., 2021].
- **Cybersecurity Deception**: CD framing selectively influenced privacy awareness and engagement but did not significantly affect trust or understanding.
- **Design Constraint**: Because interactivity and explainability were jointly manipulated, independent feature effects cannot be isolated. The study estimates the combined impact of interactive explainability; mechanism-level decomposition remains future work [Hogue et al., 2026b].

Collectively, the results indicate that enhancing interactivity and providing explanations substantially improve user experience but do not resolve the dissociation between understanding and trust. Agency alone does not generate performance-based confidence. Notably, this gap persisted under ideal deliberative laboratory conditions, suggesting structural limitations in current explainability paradigms when translated to operational contexts.

These findings motivate the development of a mixed-initiative alternative. Chapter 4 introduces the Mixed-Initiative Camouflage Analysis (MICA) framework, which reconceptualizes explainability as bidirectional collaboration. Chapter 7 extends the implications to broader questions of AI adoption under high-stakes conditions.

Chapter 4

From Transparency to Agency: The MICA Framework

Chapter 3 demonstrated a dissociation between understanding and trust: explainability improved user experience but produced negligible gains in comprehension and only modest increases in trust. This persistent gap motivates a structural reconsideration of human–AI interaction in camouflage detection.

4.1 Theoretical Foundations

This section establishes the theoretical basis for MICA. It first formalizes the understanding–trust gap as a relationship among understanding, control, and performance, then models trust calibration as a temporal process driven by agency, examines why passive explainability is insufficient, connects control theory to trust formation, and introduces the mixed-initiative principles that ground the framework.

4.1.1 Formalizing the Understanding–Trust Gap

Trust T is modeled as a function of understanding U, control C, and perceived performance P [Lee and See, 2004, Hoff and Bashir, 2015, Kohn et al., 2021]:

$$T = f(U, C, P) \tag{4.1}$$

where each term is normalized to $[0, 1]$. For trust, $T = 0$ denotes complete distrust and $T = 1$ denotes full reliance. Understanding U ranges from no comprehension of the model's reasoning (0) to complete comprehension (1); control C ranges from no interactive control (0) to full control over the inference process (1); and perceived performance P ranges from consistently incorrect (0) to consistently correct (1) observed behavior. Calibrated trust corresponds to alignment between T and P.

The findings suggest that, in this context, $\frac{\partial T}{\partial C} > \frac{\partial T}{\partial U}$ [Rotter, 1966, Bandura, 1977]. That is, understanding is necessary but insufficient, whereas control more strongly shapes operational confidence.

4.1.2 Modeling Trust through Shared Agency

To address this gap, the MICA framework operationalizes trust-building as a temporal process influenced by the level of perceived agency $A(t)$ and observed performance $P(t)$ [Lee and See, 2004, Hoff and Bashir, 2015, Kohn et al., 2021]. The rate of trust calibration is modeled as:

$$\frac{dT}{dt} = \alpha \cdot A(t) \cdot (P(t) - T(t)) \tag{4.2}$$

where:

- $T(t)$ represents the calibrated trust at time t [Muir, 1994, Lee and See, 2004].
- $A(t)$ is the degree of interactive agency afforded by the system [Rotter, 1966, Bandura, 1977].
- $P(t)$ is the empirical performance of the model as witnessed by the user [Dzindolet et al., 2003, Lee and See, 2004].
- α is a learning rate parameter reflecting individual user characteristics [Hoff and Bashir, 2015].

When $A(t) \to 0$, $\frac{dT}{dt} \to 0$, formalizing the limitation of passive explainability [Parasuraman and Riley, 1997]. Increasing agency accelerates trust calibration by allowing users to iteratively align perceived performance with observed system behavior.

4.1.3 The Insufficiency of Passive Explainability

The Phase 1 empirical findings (Chapter 3) align with emerging critiques of the limits of contemporary XAI systems. Grad-CAM heatmaps and confidence overlays illustrate *what* the model attends to, but provide limited insight into *why* those regions constitute evidence of camouflage or *how reliable* those inferences are under varying environmental conditions. This creates an explanatory asymmetry: the system transmits information, but the user cannot interrogate or influence the explanation.

Such asymmetry is consistent with theoretical concerns that explanations must be interactive and context-sensitive to support human reasoning [Doshi-Velez and Kim, 2017, Norton and Miller, 2019] meaningfully. The persistent

understanding–trust gap observed in Phase 1 indicates that comprehension alone does not generate calibrated trust or appropriate reliance.

4.1.4 Control Theory and Trust Formation

Research in psychology and automation suggests that trust is influenced not only by understanding but also by perceived control. Bandura's theory of self-efficacy [1977] and Rotter's locus of control [1966] indicate that agency enhances confidence in one's ability to act effectively. In automation research, controllability is repeatedly linked to calibrated trust and appropriate reliance [Chiou et al., 2021, Kohn et al., 2021].

4.1.5 Mixed-Initiative Principles

MICA is grounded in established mixed-initiative interaction principles [Horvitz, 1999], including:

- **Uncertainty awareness**: Represent uncertainty continuously.
- **Initiative negotiation**: Dynamically allocate control.
- **Value-directed dialogue**: Prioritize high-value human input.
- **Attention management**: Guide focus without restricting autonomy.

These principles provide the conceptual foundation for a collaborative, adaptive approach to CODS.

4.2 The MICA Framework Architecture

The MICA framework is designed to address the trust–action gap (Section 1.2) by enabling mixed-initiative collaboration between human operators and AI-based camouflage detection systems. The framework operationalizes design implications derived from the empirical findings of Chapter 3, which demonstrated that explainable interfaces alone do not reliably produce trust or reliance in camouflage detection workflows.

4.2.1 Core Components

MICA comprises four integrated architectural components:

Explainability Enhancement: Transforms static explanations into manipulable inference components.

Human–AI Iterative Control: Enables Region-of-Interest (ROI) specification, sensitivity adjustment, and real-time parameter manipulation to operationalize user agency.

Contextual Environment Interpretation: Grounds detection decisions in scene-level perception, domain knowledge, and temporal coherence.

Adaptive Decision Support: Adjusts explanation depth and recommendations based on user expertise and uncertainty.

4.2.2 The Trust-Building Cycle

The trust-building cycle consists of iterative user interaction, model prediction, system refinement, and trust reassessment, thereby framing trust as a dynamic calibration process rather than a static judgment.

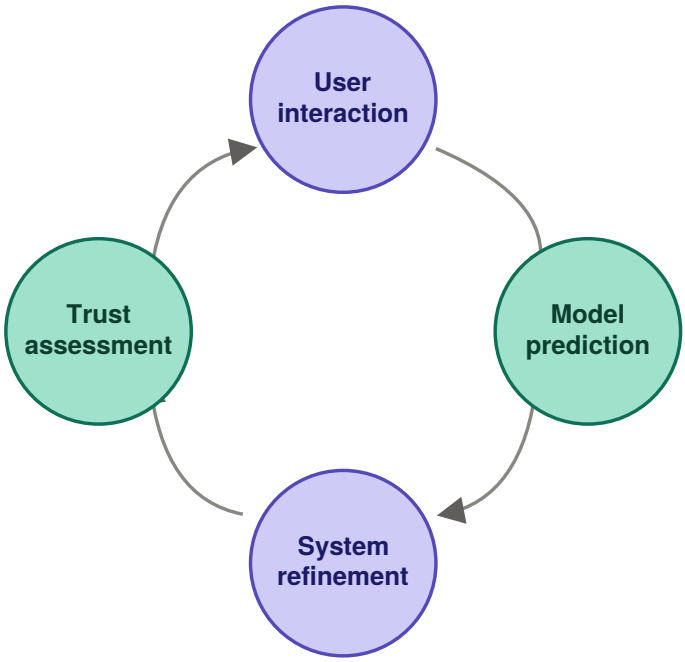

Figure 4.1: Trust-building cycle: a continuous loop of user interaction, model prediction, system refinement, and trust evaluation.

4.3 Operationalizing Interactive Control

This section describes how MICA turns the principle of interactive control into concrete mechanisms. It presents the hierarchical region-of-interest refinement procedure that lets operators direct the system's attention, followed by the sensitivity controls that let operators tune detection behavior through signal detection theory parameters.

4.3.1 Region-of-Interest Mechanisms

Drawing on visual analytics and information foraging theory [Pirolli and Card, 2005], MICA implements hierarchical ROI refinement:

Algorithm 1 Hierarchical ROI Selection

Require: Image I, user-selected region S
Ensure: Weighted attention map Ψ

1: $R_{\text{coarse}} \leftarrow \textsc{BoundingBox}(S)$
2: $R_{\text{fine}} \leftarrow \textsc{RefineContour}(R_{\text{coarse}}, I)$
3: $w \leftarrow \textsc{UserDefinedWeight}(R_{\text{fine}})$
4: $\Psi \leftarrow \textsc{GaussianWeight}(R_{\text{fine}}, w, \sigma)$
5: **return** Ψ

Here σ is the width of the Gaussian kernel applied around the refined region R_{fine}: a larger σ spreads attentional weight more broadly around the user-selected region, while a smaller σ concentrates it. The user-defined weight w scales the peak attention assigned to the region.

4.3.2 Sensitivity Controls

Building on signal detection theory, MICA provides user-adjustable detection thresholds:

$$d' = Z(H) - Z(FA) \tag{4.3}$$

where d' is the detection sensitivity index from signal detection theory and $Z(\cdot)$ is the inverse of the standard normal cumulative distribution function, which converts a proportion into a z-score. In this context, the hit rate H is the proportion of truly camouflaged regions that the system flags as camouflage, and the false alarm rate FA is the proportion of background regions that the system incorrectly flags as camouflage. A larger d' indicates better separation between camouflaged targets and background.

Users may adjust three controls. Detection sensitivity is governed by the decision threshold on d', which trades hits against false alarms. The response bias

β shifts the decision criterion toward either liberal reporting (more detections, more false alarms) or conservative reporting (fewer detections, fewer false alarms) without changing d'. Feature weighting $\mathbf{w}$ controls cue prioritization, that is, how strongly each visual feature (such as texture, edge disruption, or color contrast) contributes to the detection score, allowing operators to emphasize the cues most diagnostic for a given scene.

The decision function is formalized as:

$$y = \phi\left(\sum_{i=1}^{n} w_i \cdot f_i(x) + b\right) \tag{4.4}$$

where x is the input image patch under evaluation, $f_i(x)$ is the i-th feature extractor returning the response of visual feature i for patch x, w_i is the corresponding user-adjustable weight, b is a bias term that sets the baseline decision threshold, ϕ is the activation function mapping the weighted sum to a bounded detection score, and $y \in [0, 1]$ is the resulting per-patch camouflage likelihood. This formulation preserves interpretability while enabling user-adjustable inference.

4.4 Theoretical Contributions

MICA does not replace existing frameworks; rather, it operationalizes their principles within CODS.

4.4.1 Reconceptualizing Explainability

MICA introduces the concept of *interactive explainability*, defined as follows:

Definition 4.1 (Interactive Explainability)**.** An explanation E is interactively explainable if:

1. E can be modified by user input u through a transformation g: $E' = g(E, u)$, where g maps an explanation and a user action to a revised explanation, with $g(E, u) \neq E$ for at least some u, so that user input is capable of changing the explanation.

2. The modification influences model behavior: $M(x \mid E') \neq M(x \mid E)$, where $M(x \mid E)$ denotes the model's output for input x conditioned on explanation E.

3. The influence is interpretable: $\Delta M = h(u)$, where ΔM is the resulting change in model output and h maps the user action u to that change in a human-understandable form.

This definition reframes explanations as manipulable components of inference rather than static outputs.

4.4.2 Agency as a Trust Mechanism

In this work, perceived interactivity is treated as a proxy for perceived agency. Agency functions as a structural mechanism for trust calibration. By enabling users to influence model behavior, the system transforms passive evaluation into active hypothesis testing. This shifts trust formation from static belief to iterative evidence alignment.

4.4.3 Expertise-Adaptive Collaboration

MICA adapts interaction modes based on inferred expertise:

$$I_{\text{mode}} = \begin{cases} I_{\text{guided}} & \text{if } L_{\text{exp}} < \theta_{\text{expert}} \\ I_{\text{exploratory}} & \text{if } L_{\text{exp}} \geq \theta_{\text{expert}} \end{cases} \tag{4.5}$$

where I_{mode} is the interaction mode, L_{exp} is the user expertise level, and θ_{expert} is the expertise threshold. This supports both novice users who require scaffolding and expert users who require flexible control.

4.5 Design Principles for Mixed-Initiative CODS

The theoretical commitments above translate into five design principles that guide the construction of mixed-initiative CODS. Each principle operationalizes one facet of the framework: shared agency between human and AI, transparent communication of uncertainty, iterative trust development, grounding of detection in context, and adaptation of explanation depth to the user.

Principle 1 (Shared Agency)**.** *Human and AI contributions should be balanced based on the level of uncertainty and expertise.*

Principle 2 (Transparent Uncertainty)**.** *Confidence should be communicated through continuous gradients $p(y \mid x) \in [0, 1]$ rather than binary classifications.*

Principle 3 (Iterative Refinement)**.** *Trust develops through cycles of interaction and observation, following the trust-building cycle in Figure 4.1.*

Principle 4 (Contextual Grounding)**.** *Detection must incorporate domain knowledge K and environmental factors $\mathcal{E}$: $y = f(x, K, \mathcal{E})$.*

Principle 5 (Adaptive Depth)**.** *Explanation complexity should scale with user expertise and task difficulty: $D_{expl} \propto L_{exp} \times D_{task}$.*

4.6 Relationship to Existing Frameworks

MICA integrates principles from joint cognitive systems, distributed cognition, activity theory, and sensemaking theory by treating human and AI elements as a coupled analytical system in which expertise, artifacts, and contextual reasoning interact dynamically [Hutchins, 1995, Hollnagel and Woods, 2005, Klein et al., 2006, Engeström, 2014].

4.6.1 Cognitive Systems Engineering

MICA aligns with joint cognitive systems principles by treating human and AI elements as components of a coupled analytical system [Hollnagel and Woods, 2005].

4.6.2 Distributed Cognition

Consistent with Hutchins [1995], MICA distributes cognitive processes across human expertise and computational capabilities.

4.6.3 Activity Theory

The AI system functions as both a tool and a collaborator, consistent with the mediating role of artifacts in activity theory [Engeström, 2014].

4.6.4 Sensemaking Theory

MICA supports both bottom-up and top-down processing, aligning with the data-frame model of Klein et al. [2006].

4.7 Validation Approach

Evaluation focuses on three dimensions. Trust calibration T_{calib} measures the alignment between a user's reported trust and the system's objective performance, so that a well-calibrated user trusts the system when it is correct and withholds trust when it is not. Joint detection performance P_{joint} measures the accuracy of the combined human and AI team rather than either agent alone. Interaction efficiency η_{interact} captures the cost of reaching a decision, combining time-to-decision with cognitive load (for example, as measured by NASA-TLX), so that higher η_{interact} reflects accurate decisions reached with less time and effort.

These dimensions define a composite effectiveness metric:

$$V_{\text{MICA}} = \lambda_1 \cdot T_{\text{calib}} + \lambda_2 \cdot P_{\text{joint}} + \lambda_3 \cdot \eta_{\text{interact}} \tag{4.6}$$

where the weighting factors $\lambda_i \geq 0$ with $\sum_i \lambda_i = 1$ encode the relative priority of each dimension for a given deployment. Operational requirements set these weights: a time-critical triage setting might increase λ_3 to prioritize speed, whereas a high-stakes analysis setting might increase λ_1 to prioritize calibrated trust. Each component is normalized to $[0, 1]$ so the terms are commensurable.

4.8 Summary

The MICA framework addresses the central limitation identified in Chapter 3: explainability alone does not bridge the understanding–trust gap. MICA reframes explainability as mixed-initiative collaboration, integrating agency, contextual reasoning, and adaptive interaction to support calibrated human–AI teaming.

Chapter 5 describes how these theoretical principles are instantiated in a

functional prototype system.

Chapter 5

Instantiating Mixed-Initiative Agency: Prototype Design and Implementation

This chapter documents the design and implementation of the MICA (Mixed-Initiative Camouflage Analysis) prototype, designated MURDOC_2024. The prototype operationalizes the MICA framework described in Chapter 4 by instantiating mixed-initiative interaction as a mechanism for trust calibration within a functional camouflaged object detection and segmentation (CODS) workflow. Rather than limiting users to passive observation of AI outputs, MICA places them in an active collaborative role, enabling iterative refinement of both detection boundaries and camouflage salience rankings within a unified editing environment.

The implementation serves two purposes within this dissertation. First, it provides a concrete technical realization of the mixed-initiative principles derived from the empirical findings of Chapter 3. Second, it establishes a controlled

experimental platform to evaluate how interactive agency influences user trust, decision confidence, and analytical performance in camouflage detection tasks.

The chapter proceeds as follows. Section 5.1 describes the system architecture and technology stack. Section 5.2 details the mixed-initiative interaction loop. Section 5.3 describes how each MICA framework component is instantiated in the prototype. Section 5.4 presents the unified editing environment, including the interface layout, tool palette, polygon mask editor, and rank brush editor. Section 5.5 describes the mixed-initiative interaction workflow. Section 5.6 covers the detection feedback and session management subsystems. Section 5.7 documents the algorithmic implementation of key operations. Section 5.8 describes the controls for the signal-detection parameters. Section 5.9 addresses prototype limitations, and Section 5.10 summarizes the chapter.

5.1 System Architecture

MICA is implemented as a desktop application targeting Windows 10 and later, built on WPF with the .NET Framework using C#. As illustrated in Figure 5.1, the system separates user interaction, model inference, explanation generation, and session management into modular components coordinated through the Model-View-ViewModel (MVVM) application architecture. This separation supports the modularity required for mixed-initiative interaction, where detection engine outputs, user editing actions, and session analytics remain independently maintainable and not coupled to presentation concerns.

The AI backend is implemented in Python 3.9, executing RankNet [Lv et al., 2021] (with a ResNet50 backbone) for camouflage salience ranking and EfficientDet-D7 [Tan et al., 2020] for object detection. RankNet addresses two tasks simulta-

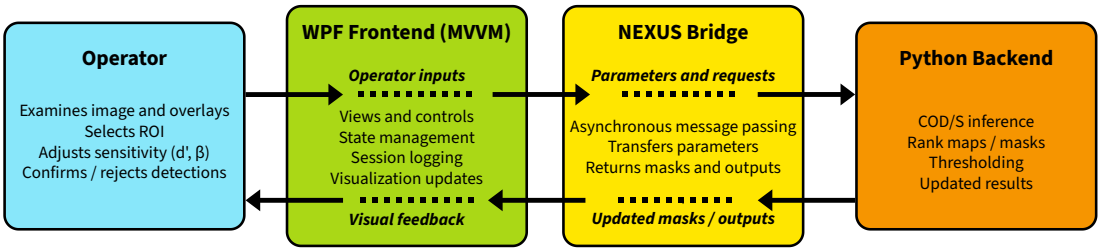

Figure 5.1: System architecture of the MURDOC_2024 prototype implementing the MICA framework. The WPF/MVVM frontend manages user interaction and visualization, while the NEXUS bridge coordinates communication with the Python-based camouflage detection pipeline.

neously using a joint learning strategy. The first task, camouflaged object ranking (COR), estimates the perceptual difficulty of detecting a given camouflaged region by assigning a continuous salience score that reflects how effectively the object blends into its background. The second task, camouflaged object localization (COL), identifies the discriminative regions where concealment is weakest, that is, the areas of highest local contrast between the camouflaged object and its surroundings. A reverse attention module drives the localization decoder to progressively suppress already-identified regions and refocus on remaining concealment boundaries, enabling the model to simultaneously segment the object and regress the regions that make it perceptually apparent. In the MICA prototype, RankNet's COR output is rendered as a JET-colormap rank map overlay (a rainbow pseudocolor scale ranging from blue for low values to red for high values) in the prediction pane, where warm tones indicate regions of weak camouflage (high detectability) and cool tones indicate regions of strong camouflage (low detectability), providing the operator with a spatially grounded representation of the system's uncertainty across the image.

Communication between the C# frontend and the Python backend is managed through NEXUS (the .NET–Python EXecution Unified System), a bridging

layer that launches Python scripts as managed subprocesses and handles parameter serialization, image file handoff, and result ingestion. This architecture preserves the responsiveness of the WPF interface while allowing the computationally intensive detection pipeline to execute asynchronously in a separate process.

The hybrid architecture was selected to combine the rapid interface development capabilities of the .NET WPF ecosystem with the extensive machine learning libraries available in Python.

The principal ViewModel components are organized as follows.
`MainWindowViewModel` serves as the application orchestrator, managing the session lifecycle, coordinating model execution, and routing events between child ViewModels. `MICAControlViewModel` exposes the detection parameter controls, including sensitivity (d') and response bias (β), and triggers model runs.
`EditorControlsPaneViewModel` manages the unified edit mode state, the tool palette, detection feedback commands, and rank brush parameter bindings.
`FinalPredictionPaneViewModel` owns the image layer state (original image, binary mask, and rank map) and implements the pixel-level editing operations.
`SessionInfoPaneViewModel` tracks session duration, images analyzed, and aggregate modification counts, and fires the end-session event that initiates data serialization.

The modular architecture also allows the prototype to function as a reusable experimental research platform, enabling controlled evaluation of mixed-initiative interaction strategies, trust calibration mechanisms, and alternative camouflage detection models without requiring modifications to the core user interface.

5.2 Mixed-Initiative Interaction Loop

The mixed-initiative loop in MICA is structured around a four-stage cycle that alternates control between the user and the AI model. In the first stage, the user selects an image from the file system and optionally adjusts image-level parameters (brightness, contrast, saturation) to enhance camouflage visibility. In the second stage, the user triggers model execution, causing NEXUS to invoke RankNet and EfficientDet-D7, retrieve their outputs, and render a layered visualization in the prediction pane: the original image, a JET-colormap rank map overlay, and the binary detection mask. In the third stage, the user evaluates the AI output and exercises agency through two complementary channels. Structural corrections are made through the unified edit mode, where the user selects either the Polygon Tool to refine detection boundaries or the Rank Brush Tool to adjust camouflage salience weights. Epistemic corrections are made through the detection feedback panel, where the user confirms correct detections, rejects false positives, or flags missed detections for correction. In the fourth stage, the user saves the session, triggering serialization of all modifications, parameters, and feedback to structured JSON summaries and optionally queuing the session for Low-Rank Adaptation (LoRA) model retraining.

This loop instantiates the core theoretical claim of the MICA framework: that interactive control over AI outputs, rather than passive transparency about AI internals, is the operative mechanism for trust calibration. Each pass through the loop produces a durable record of the human–AI collaboration that supports both retrospective analysis and prospective model adaptation.

The mixed-initiative interaction loop (see Figure 5.2) illustrates how user actions and system responses form a continuous analytical cycle. The opera-

tor provides input through region-of-interest selection, sensitivity adjustments, and corrective feedback. The AI system processes these inputs to refine detection boundaries, generate updated probability maps, and produce explanations. These outputs then inform subsequent user adjustments, creating an iterative collaboration between human expertise and algorithmic inference.

5.3 Framework Component Instantiation

The MICA prototype instantiates the four architectural components defined in Chapter 4 through concrete interface mechanisms, model integrations, and session-level persistence logic. This section describes how each component is realized in the implementation.

5.3.1 Explainability Enhancement

Explainability is operationalized through integration of Grad-CAM visualizations [Selvaraju et al., 2017] and model confidence outputs generated by the underlying PyTorch-based detection models. Grad-CAM heatmaps are computed in the Python backend and returned to the interface via the NEXUS inter-process communication bridge. These visualizations may be overlaid on the input image to reveal the salient regions that influence detection decisions.

Confidence values and decision summaries are displayed within the Informative AI (IAI) pane. Together, these elements transform static model outputs into inspectable artifacts that support diagnostic interpretation. Importantly, explainability artifacts are user-invoked through collapsible panes, preserving primary workspace focus while enabling on-demand inspection, a design choice discussed further in Section 5.4.1.

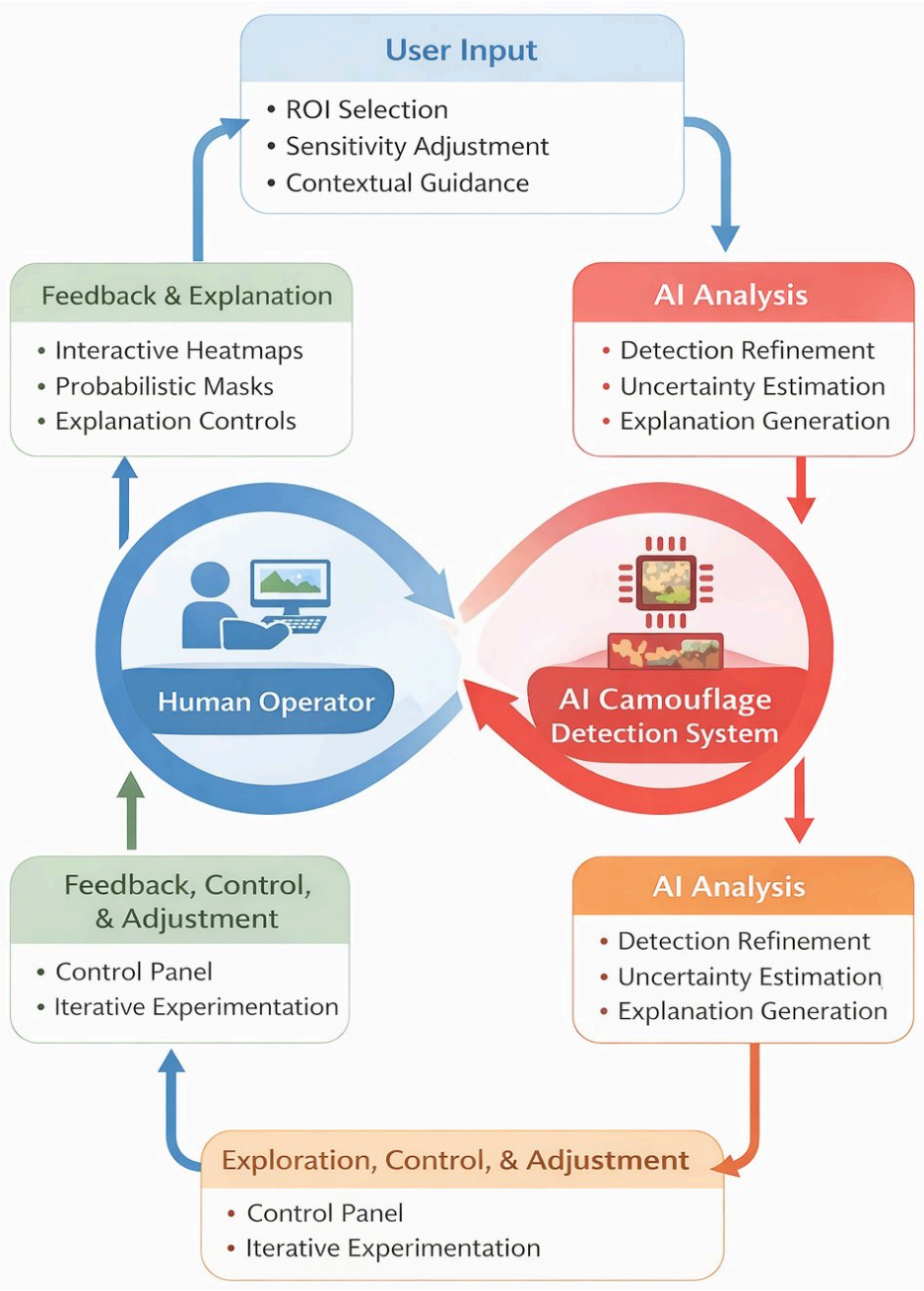

Figure 5.2: The MICA mixed-initiative interaction loop. User agency is exercised through two channels in Stage 3: structural editing (Polygon Tool / Rank Brush Tool) and epistemic feedback (Confirm / Reject).

5.3.2 Human–AI Iterative Control

User agency is implemented through interactive editing mechanisms that directly modify model outputs. The localization mask is rendered as an editable polygon boundary, allowing users to refine segmentation by adjusting control points. Modifications to the polygon update the underlying binary mask representation in real time.

Ranking adjustments are implemented through brush-based interaction with the camouflage rank map. Users may increase or decrease local rank values using adjustable brush size and intensity parameters. These edits provide structured corrective supervision rather than binary acceptance or rejection of model predictions.

Inference may be re-executed following parameter changes, enabling iterative refinement within a single analysis session.

5.3.3 Contextual Environment Interpretation

Contextual reasoning is supported through the integration of image metadata, session parameters, and user-defined adjustments. While the current prototype does not incorporate external environmental sensors, it enables contextual reinterpretation through dynamic parameter modification and iterative correction of both localization and ranking outputs.

This approach allows users to incorporate domain knowledge implicitly by reshaping model outputs in accordance with perceived environmental conditions, operationalizing expert judgment as a first-class input to the detection process.

5.3.4 Adaptive Decision Support

Adaptive decision support is partially instantiated through structured feedback mechanisms within the IAI pane. Users may confirm detections, reject outputs, or indicate missed objects. These actions are logged as structured session events and incorporated into the session summary.

Although real-time adaptive interface restructuring is not yet implemented, session-level adaptation is achieved by queuing retraining of the RankNet model on edited artifacts. This mechanism enables incremental refinement across sessions, establishing a pathway from individual user corrections to durable improvements in model behavior.

5.4 Unified Editing Environment

A central design contribution of the MICA prototype is the unified editing environment, which integrates polygon mask editing and rank map brush editing within a single modal interaction space. Prior to this implementation, these two editing surfaces (structural boundary correction and confidence weight adjustment) required separate workflows. The unified environment allows users to move between them within a single edit session without exiting and re-entering the editing context, reducing cognitive overhead and preserving spatial orientation in the image.

5.4.1 Interface Layout and Collapsible Panels

The MICA workspace is organized into three functional regions. The input panel on the left provides image selection and preprocessing controls, including bright-

ness, contrast, and saturation adjustments. The central visualization area renders the layered detection output: the original image, JET-colormap rank map overlay, and binary mask boundary. The editor controls panel on the right houses the tool palette, editing sub-controls, detection feedback actions, and session management. Figure 5.3 illustrates the three-panel workspace layout and the placement of interactive editing and feedback controls.

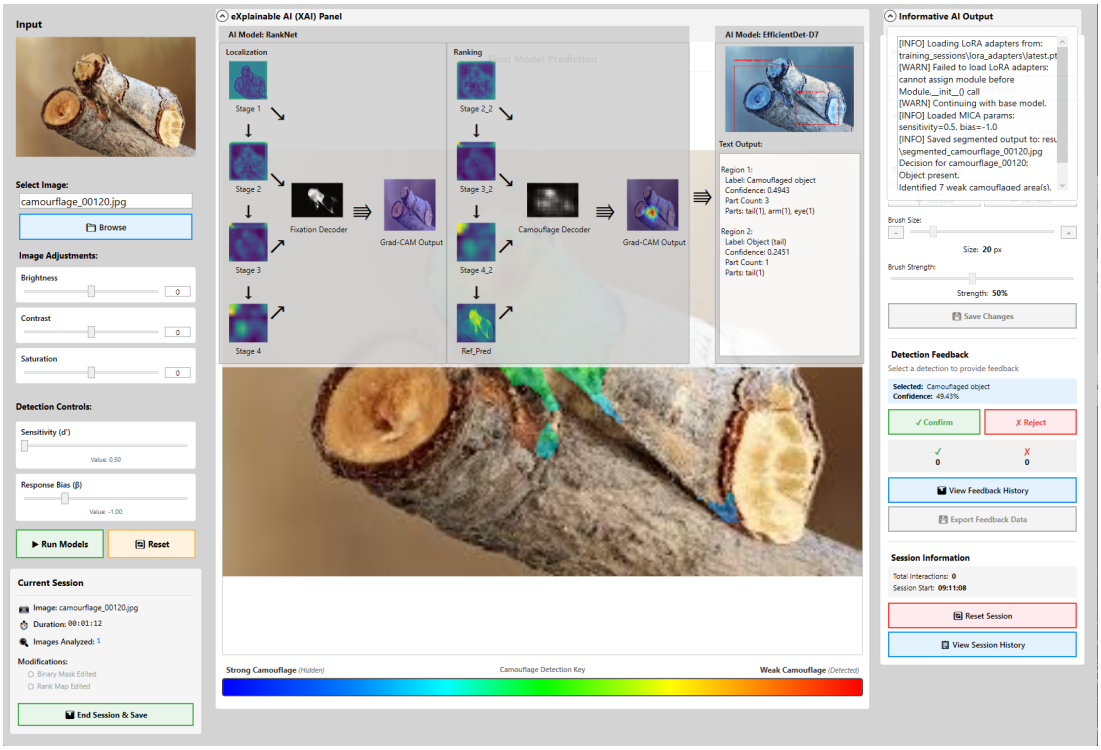

Figure 5.3: The MICA workspace layout. The three-panel design separates preprocessing controls, the primary detection canvas, and the interactive editing and feedback controls.

Both the Explainable AI (XAI) and Informative AI (IAI) components are implemented as collapsible interface panes. This design reflects MICA's emphasis on mixed-initiative interaction and attention management: by making both XAI and IAI user-invoked rather than persistently visible, the interface reframes explanation from presentation to negotiation, consistent with the mixed-initiative

principles outlined in Chapter 4.

This structure reduces cognitive load and supports adaptive depth of engagement. Users may expand the XAI pane to inspect Grad-CAM visualizations or confidence heatmaps as needed, and collapse it during iterative editing. Similarly, the IAI pane provides structured feedback, confidence summaries, and session information without constraining the primary workspace. Unlike MURDOC, where explanation artifacts were persistently displayed, MICA treats explanation and informational output as negotiable resources within the interaction loop.

5.4.2 Tool Palette

The editing environment is governed by a tool palette that appears in the editor controls panel whenever edit mode is active. The palette exposes two mutually exclusive tools: the Polygon Tool and the Rank Brush Tool. Tool selection is modeled by the `EditTool` enumeration, defined in `Model/PointEditMode.cs`, with values `PolygonTool` and `RankBrushTool`. The active tool is maintained as a bindable property, `ActiveEditTool`, in `EditorControlsPaneViewModel`, with derived boolean properties `IsPolygonToolActive` and `IsRankBrushToolActive` serving as XAML binding targets for visual state and conditional visibility.

Selecting the Polygon Tool fires the existing `PointEditModeChanged` event, which propagates through `MainWindowViewModel` to `FinalPredictionPane`, where it sets the internal `EditSubMode` to `PolygonEditing` and configures the canvas cursor to a hand pointer. Selecting the Rank Brush Tool fires the existing `RankBrushChanged` event, which causes `FinalPredictionPane` to set `EditSubMode` to `RankBrushing` and configure the canvas cursor to a crosshair. This design routes tool selection through the pre-existing event infrastructure without intro-

ducing new wiring, ensuring architectural consistency across the application.

The tool palette UI renders each tool as a large toggle button with descriptive iconography. An active tool is visually distinguished by a blue border and tinted background, implemented via `DataTrigger` bindings in the XAML. The polygon sub-controls (Add Vertex / Remove Vertex toggle buttons) are visible only when `IsPolygonToolActive` is true; the rank brush sub-controls (direction, brush size, brush strength) are visible only when `IsRankBrushToolActive` is true. This conditional visibility collapses irrelevant controls, reducing interface clutter without removing functionality from the workflow. Figure 5.4 illustrates the tool palette interface and the context-sensitive sub-controls that appear when each editing tool is active.

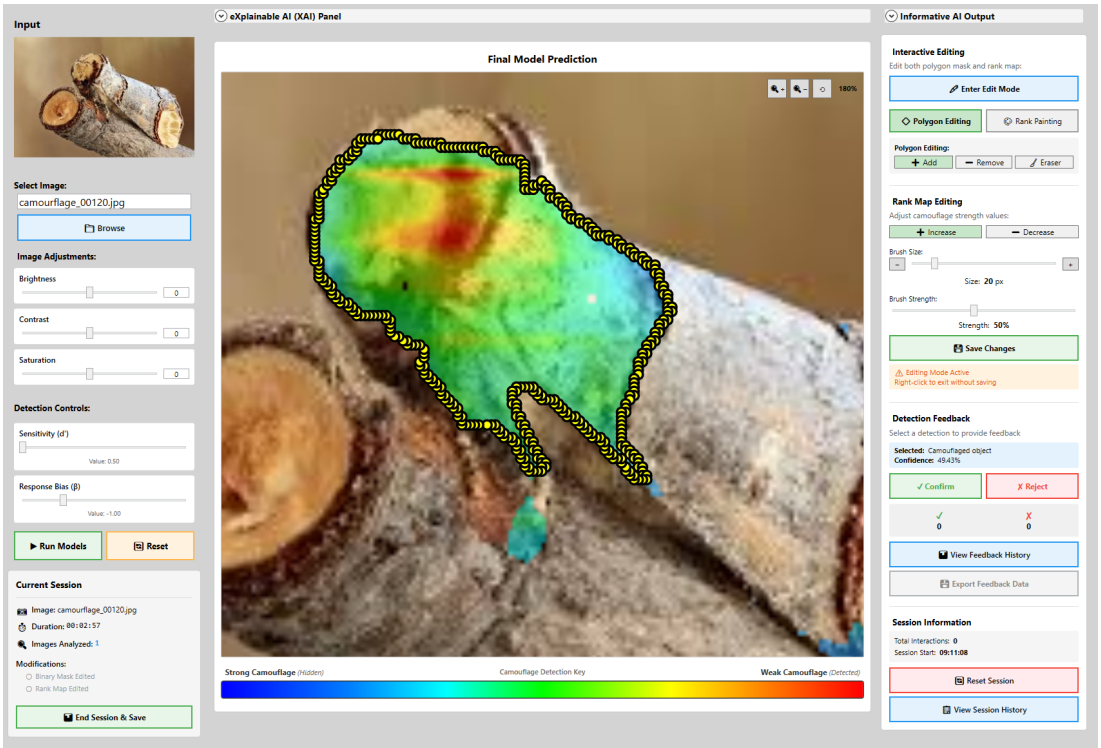

Figure 5.4: The MICA tool palette in the editor controls panel. Sub-controls adapt dynamically to the active tool, collapsing irrelevant options to minimize interface complexity.

5.4.3 Polygon Mask Editor

The Polygon Tool enables users to correct the spatial boundary of the AI's binary detection mask by manipulating a polygon overlay rendered on the editing canvas. When the user enters edit mode, `FinalPredictionPane` initializes the `PolygonDrawingService` and renders the current mask boundary as a movable polygon with draggable vertex markers. The user may operate in one of two sub-modes: Add Vertex, which inserts a new polygon point at the clicked canvas location, and Remove Vertex, which removes the nearest existing point. Both sub-modes support real-time drag editing of any vertex, with the polygon outline and marker positions updating continuously as the user drags.

Zoom and pan are available throughout polygon editing via mouse wheel scaling and space-drag panning, allowing users to work at high precision on large imagery without switching contexts. The zoom and pan state is maintained in a `TransformGroup` composed of a `ScaleTransform` and a `TranslateTransform`, applied to the `LayeredImageGrid` so that both the image and the polygon overlay transform in concert. On save, the edited polygon is rasterized into an updated binary mask using contour fill, replacing the AI-generated mask and producing a corrected localization that is persisted to disk.

5.4.4 Rank Brush Editor

The Rank Brush Tool enables users to locally adjust the AI's camouflage salience rank map by painting over regions with a configurable circular brush. The brush operates in two directions: Increase, which raises local rank values to strengthen camouflage attribution in a region, and Decrease, which lowers local rank values to attenuate spurious salience. Three parameters govern brush behavior and are

exposed as slider controls in the editor panel: direction (Increase/Decrease toggle), brush size (radius in pixels, range 5–100), and brush strength (proportional intensity, range 0.0–1.0).

During a brush stroke, mouse movement events on the editing canvas trigger `ApplyRankBrush` in `FinalPredictionPaneViewModel`, which computes the set of rank map pixels within the brush radius of the current mouse position and applies a weighted delta proportional to brush strength. The modified rank data accumulates in a working buffer that is separately tracked from the original AI output, enabling non-destructive editing throughout the session. The rank map overlay is updated on each stroke to provide immediate visual feedback. On save, the modified rank buffer is serialized to disk in the LoRA training folder structure alongside the corrected binary mask. These editing tools operationalize the mixed-initiative design principle central to the MICA framework: rather than passively observing AI outputs, users can directly intervene in model interpretations and detection boundaries, transforming explanations into actionable control and helping bridge the understanding–trust gap identified in Chapter 3.

5.5 Mixed-Initiative Interaction Workflow

The interaction workflow reflects a structured mixed-initiative loop that transforms explanation inspection into active model interrogation. Rather than passively accepting outputs, users iteratively reshape predictions, thereby operationalizing shared control between human and model. A complete analysis session proceeds as follows:

1. The user selects an image and initiates model inference.

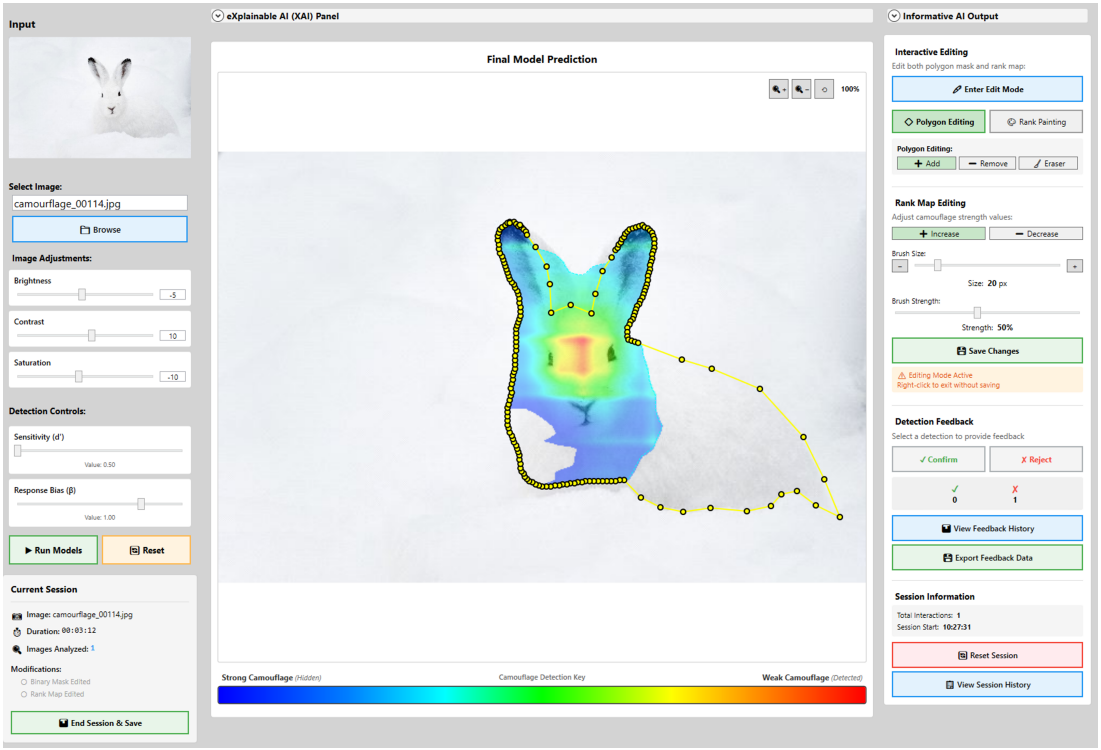

Figure 5.5: Interactive editing tools in the MICA prototype. The Polygon Tool allows users to refine binary detection mask boundaries by adjusting polygon vertices, while the Rank Brush Tool enables localized modification of camouflage salience values in the rank map.

2. The system generates a localization mask and camouflage rank map.

3. The user inspects outputs and optionally expands the XAI pane to review Grad-CAM explanations or confidence metrics.

4. The user refines localization by editing polygon boundaries via the Polygon Tool.

5. The user adjusts camouflage ranking using brush-based modifications via the Rank Brush Tool.

6. Feedback actions (confirm, reject, add missed object) are recorded in the IAI pane.

7. At session termination, all edited artifacts are consolidated, serialized, and optionally queued for LoRA retraining.

Steps 4 through 6 may be performed in any order and repeated as needed within a session. The workflow is non-prescriptive regarding the depth of engagement at each stage: users may skip polygon editing entirely and rely solely on feedback actions, or perform detailed rank map corrections without confirming individual detections. This flexibility preserves operator autonomy while ensuring that all interaction, regardless of type or depth, is captured in the session record.

5.6 Detection Feedback and Session Management

This section describes the two subsystems that close the mixed-initiative loop. Section 5.6.1 presents the structured detection feedback system through which

operators explicitly validate or challenge AI detection outputs. Section 5.6.2 describes the session management and data serialization infrastructure that persists interaction data for retrospective analysis and LoRA model retraining.

5.6.1 Detection Feedback System

MICA implements a structured detection feedback system that enables users to explicitly validate or challenge the AI's detection outputs. The feedback panel, managed by `EditorControlsPaneViewModel`, exposes three feedback actions for each detected region: Confirm, which records the detection as correct; Reject, which flags it as a false positive; and Correction, which marks that the system failed to detect a target the user has identified. Each action is logged as a `DetectionFeedback` record carrying a feedback type, detection identifier, and timestamp.

Aggregate feedback counts—confirmed detections, rejected detections, and user corrections—are tracked in real time and displayed in the session summary panel. These counts are included in the per-image data records written to the session JSON at session end, providing a quantitative record of the human–AI agreement state for each analyzed image. This data supports retrospective analysis of where the model's confidence was well-calibrated relative to expert judgment and where systematic disagreements occurred.

5.6.2 Session Management and Data Serialization

When the user ends a session, the application presents a confirmation dialog offering three options: save and queue for LoRA retraining, save only the summary, or cancel. This design gives the user explicit control over whether their

session data will be used for model adaptation, deliberately instantiating the MICA framework's agency principle at the data governance level.

Session data is serialized to a structured JSON file in the `training_sessions` directory, named using a timestamp-based session identifier (format: `yyyyMMdd_HHmmss`). Logged artifacts include:

- Session metadata (ID, start and end timestamps, formatted duration)
- Aggregate statistics (images analyzed, binary mask edits, rank map edits, confirmed detections, rejections, and corrections)
- Per-image records, each capturing:
  - Image filename and analysis timestamp
  - Image adjustment parameters (brightness, contrast, saturation)
  - Detection parameters (sensitivity d', response bias β)
  - Modification flags (binary mask edited, rank map edited)
  - Detection feedback counts (confirmed, rejected, corrections)

When retraining is requested, the session identifier is appended to a `retrain_queue.txt` file in the application directory. The LoRA retraining pipeline (`lora_retrain.py`) consumes this queue to identify which session data to incorporate into the next round of model fine-tuning. The session data folder structure organizes artifacts into three subdirectories: `img/` (original images), `bi_gt/` (corrected binary masks), and `fix/` (modified rank maps), following the format expected by the LoRA training pipeline. This persistence mechanism operationalizes mixed-initiative collaboration by transforming interaction into incremental learning signals.

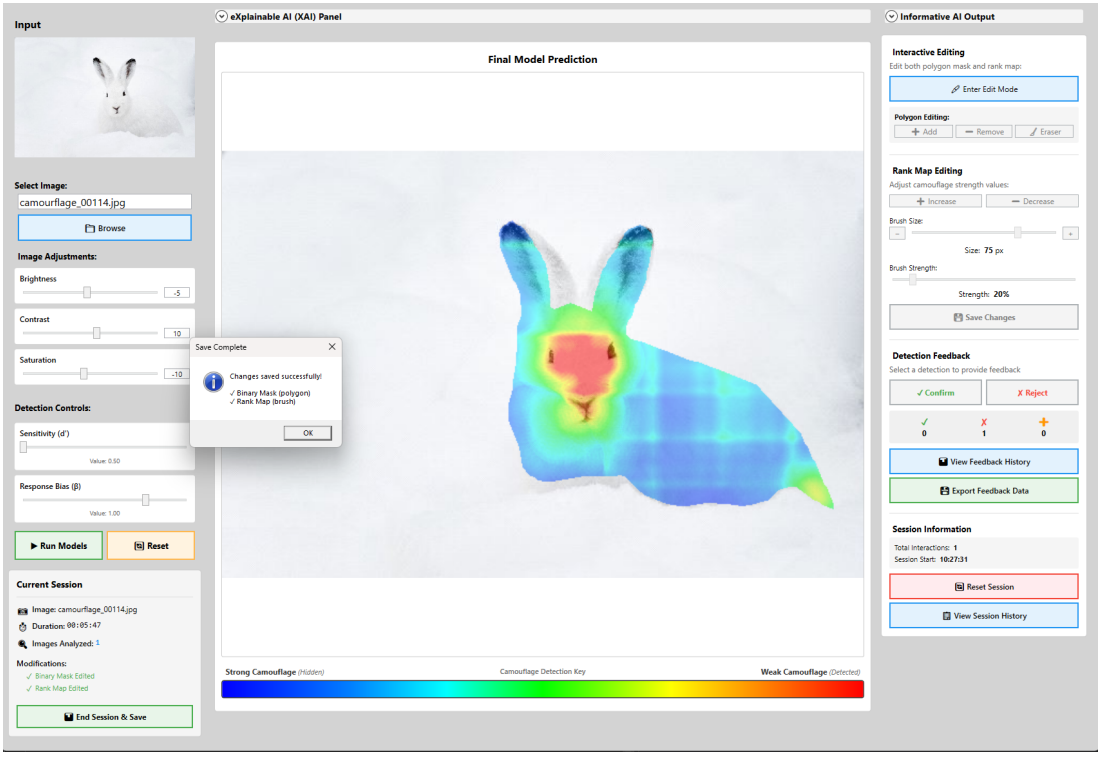

Figure 5.6: The MICA session management interface and representative session JSON output. Per-image records capture the full parameter and feedback state at the time of analysis.

5.7 Algorithmic Implementation

Several key operations within the MICA prototype require algorithmic processing beyond standard user-interface interactions. These operations include polygon mask updates, rank map modification, session data serialization, and session-based model refinement.

5.7.1 Localization Update Mechanism

Polygon boundary corrections are translated into updated binary masks through contour rasterization. Let M_0 denote the original binary mask produced by the AI pipeline, and let P denote the user-edited polygon. The updated mask M' is generated by:

$$M'(x, y) = \begin{cases} 1 & \text{if } (x, y) \in P \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

Algorithm 2 Polygon Mask Update via Contour Rasterization

Require: User-edited polygon P, image dimensions (W, H)
Ensure: Updated binary mask M'
Initialize $M' \leftarrow \mathbf{0}_{W \times H}$ ▷ Empty mask
$C \leftarrow \textsc{RasterizeContour}(P, W, H)$ ▷ Polygon boundary pixels
$F \leftarrow \textsc{FloodFill}(C)$ ▷ Interior pixel set
for all $(x, y) \in F$ **do**
    $M'(x, y) \leftarrow 1$
end for
$M' \leftarrow \textsc{Clip}(M', 0, 1)$ ▷ Enforce binary values
$\textsc{PersistToDisk}(M', \texttt{bi_gt/})$
return M'

This updated mask replaces the previous localization state for the current image and is used to constrain subsequent rank map persistence, ensuring spatial

consistency between boundary and salience representations.

5.7.2 Rank Map Adjustment

Let $R(x, y)$ denote the original rank map produced by RankNet. Brush-based interaction modifies local regions according to:

$$R'(x, y) = R(x, y) + \Delta(x, y) \tag{5.2}$$

where $\Delta(x, y)$ is determined by the brush direction, strength parameter $s \in [0, 1]$, and a Gaussian spatial weighting over the brush radius r:

$$\Delta(x, y) = \pm s \cdot \exp\left(\frac{-d(x, y)^2}{2r^2}\right) \tag{5.3}$$

Algorithm 3 Rank Map Adjustment via Gaussian Brush

Require: Original rank map R, brush center (x_b, y_b), radius r, strength $s \in [0, 1]$, direction $\delta \in \{+1, -1\}$
Ensure: Updated rank map R'

$R' \leftarrow R$ ▷ Initialize working buffer
for all (x, y) such that $d(x, y) \leq r$ **do**
    $d \leftarrow \sqrt{(x - x_b)^2 + (y - y_b)^2}$
    $\Delta(x, y) \leftarrow \delta \cdot s \cdot \exp\left(\frac{-d^2}{2r^2}\right)$
    $R'(x, y) \leftarrow R'(x, y) + \Delta(x, y)$
end for
$R' \leftarrow \textsc{Clip}(R', 0, 255)$ ▷ Enforce valid intensity range
return R'

The sign is positive for the Increase direction and negative for the Decrease direction, where $d(x, y)$ is the Euclidean distance from the brush center to pixel (x, y). Modifications are accumulated in a working buffer across the editing session. The modified rank map R' is bound to valid intensity values prior to

serialization.

5.7.3 Session Data Serialization

All user interactions are recorded as structured session events. When the session is saved, these events are serialized into JSON format containing the image identifier, edited mask coordinates, rank map modification parameters, detection confirmation or rejection labels, and signal detection parameters (d' and β). These artifacts provide both an audit trail of human–AI collaboration and training data for subsequent model refinement experiments.

5.7.4 Session-Based Model Refinement

At session termination, edited masks and rank maps are serialized into the structured JSON summaries described in Section 5.6.2. Session identifiers are appended to the retraining queue, which the LoRA fine-tuning pipeline consumes to update RankNet parameters using the edited rank maps as supervisory signals. Retraining is deferred to the post-session context rather than executed during interaction, preserving interface responsiveness while enabling incremental model adaptation across sessions.

Algorithm 4 Session-Based LoRA Model Refinement

Require: Session identifier sid, edited artifacts in `training_sessions/`sid`/`, retraining queue file Q

Ensure: Updated RankNet parameters θ'

1: AppendToQueue(Q, sid) ▷ Deferred: executed by `lora_retrain.py` at next run
2: $S \leftarrow$ ReadQueue(Q)
3: $D \leftarrow \emptyset$
4: **for** all $s \in S$ **do**
5: $M \leftarrow$ LoadMasks(s, `bi_gt/`)
6: $R \leftarrow$ LoadRankMaps(s, `fix/`)
7: $I \leftarrow$ LoadImages(s, `img/`)
8: $D \leftarrow D \cup \{(I, M, R)\}$
9: **end for**
10: $\theta' \leftarrow$ LoRAFineTune(θ, D)
11: ClearQueue(Q)
12: **return** θ'

5.8 Signal Detection Parameter Controls

The MICA interface exposes two parameters derived from signal detection theory: sensitivity d' and response bias β. These controls allow the operator to adjust the detection threshold used during inference, enabling dynamic calibration of model behavior.

The sensitivity parameter influences the degree to which the detection model distinguishes between target and background signals. Increasing sensitivity increases the likelihood of detecting subtle camouflage features but may also increase false positive detections. Decreasing sensitivity produces more conservative detections.

The response bias parameter controls the decision threshold applied to the model's confidence scores. Higher β values bias the system toward conservative decisions, favoring precision over recall, while lower values encourage more

permissive detection behavior.

These parameters are exposed through slider controls within the detection controls panel and are bound directly to the model execution pipeline through `MICAControlViewModel`. When modified, the updated values are passed to the Python inference pipeline via the NEXUS communication layer. Figure 5.7 illustrates the sensitivity and response bias controls used to dynamically adjust the detection decision boundary during inference.

Providing users with direct control over these parameters enables operators to adapt the detection process to varying environmental conditions and mission objectives. This capability reflects the mixed-initiative philosophy underlying the MICA framework, where human expertise complements automated inference.

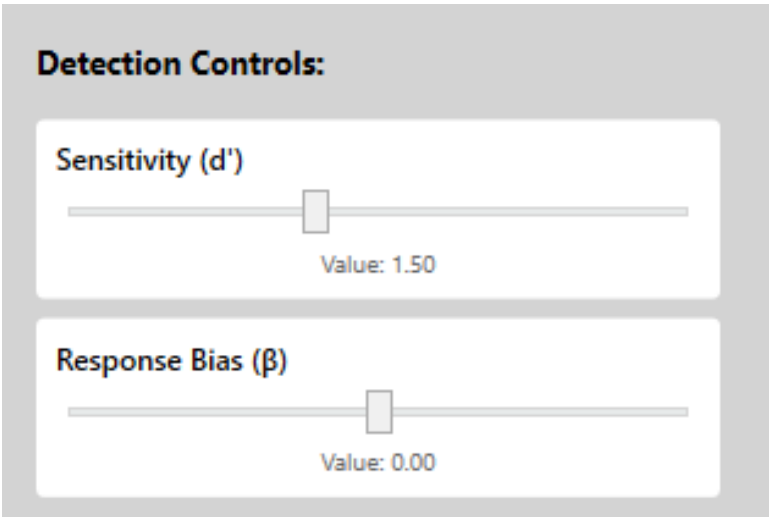

Figure 5.7: Signal detection parameter controls in the MICA interface. Sensitivity (d') and response bias (β) are exposed as interactive sliders, enabling users to adjust the detection decision boundary to match operational cost asymmetries.

5.9 Prototype Limitations

Although the MICA prototype demonstrates the feasibility of mixed-initiative camouflage analysis, several limitations remain.

First, the current implementation operates as a research prototype rather than a production-grade system. Model inference is performed via subprocess invocation rather than through an optimized inference server, which may limit scalability when processing large datasets.

Second, adaptive interface behavior is currently limited to session-level retraining rather than real-time model adaptation. While edited artifacts may be incorporated into future training cycles, the system does not yet support online learning during a single analysis session.

Third, the prototype relies primarily on static imagery and does not yet incorporate temporal information from video streams or external contextual data sources such as environmental sensors or mission metadata. Integrating these inputs could further enhance contextual reasoning during camouflage analysis.

Finally, the evaluation of the system's impact on trust calibration and analytical performance remains an open research question, to be addressed in the future empirical program outlined in Chapter 7. The prototype demonstrates feasibility but does not constitute a full empirical evaluation of MICA.

5.9.1 Failure Modes and Their Implications for Trust

Understanding where the system fails is as important as documenting what it achieves, particularly in the context of trust calibration. The understanding–trust gap identified in Chapter 3 is not an abstract statistical finding; it is the perceptual consequence of users encountering detection outputs they cannot fully

verify. Characterizing the specific conditions under which MURDOC_2024 produces incorrect outputs is therefore a prerequisite for designing interfaces that support appropriate reliance rather than either blind acceptance or blanket rejection of AI outputs.

RankNet's salience ranking is most reliable when a single camouflaged object occupies a well-defined foreground region against a visually homogeneous background. Performance degrades in three recurring conditions. First, in highly textured scenes where background complexity approaches the spatial frequency of the camouflaged object, the rank map assigns elevated salience scores to background regions, producing false positive heat concentrations that may mislead the operator into examining areas that contain no target. Second, when multiple camouflaged objects appear at different depths or overlap spatially, the ranking function collapses their salience into a single undifferentiated region, obscuring the boundaries between distinct targets. Third, in images captured under low-contrast lighting conditions, such as overcast environments where shadow cues are absent, RankNet's localization decoder produces diffuse rank maps with low spatial specificity, making it difficult to determine where the system believes the camouflaged object boundary lies.

EfficientDet-D7 exhibits a complementary failure pattern. The object detection pipeline was trained on annotated features that compromise camouflage, including eyes, limbs, and cast shadows. When these features are partially occluded or absent, the detection model tends toward false negatives, returning empty bounding box outputs for images that RankNet correctly identifies as containing a camouflaged target. The inverse condition also occurs: EfficientDet occasionally generates false positive bounding boxes in high-clutter scenes by latching onto background textures that share feature responses with the train-

ing annotations. In both cases, the discrepancy between the rank map output and the detection bounding box is itself informative, signaling to an experienced operator that the system is operating near the boundary of its competence.

Table 5.1 summarizes the recurring failure conditions identified in MURDOC_2024, the observable symptoms each produces, and the corresponding MICA recovery mechanism available to the operator.

Table 5.1: Recurring failure modes in MURDOC_2024, their trigger conditions, observable symptoms, and the MICA interaction mechanisms available for operator recovery.

Component	**Failure Mode**	**Trigger Condition**	**Observable Symptom**	**MICA Recovery**
RankNet	False positive salience	High-texture background matches target spatial frequency	Rank map heat concentrated on background; no target present	Rank Brush Tool (Decrease)
RankNet	Salience collapse	Multiple overlapping or depth-separated targets	Single undifferentiated region obscures target boundaries	Rank Brush Tool (Increase / Decrease)
RankNet	Diffuse localization	Low-contrast lighting; absent shadow cues	Rank map spatially diffuse; boundary unresolvable	Rank Brush Tool + Polygon Tool
EfficientDet-D7	False negative	Discriminating features (eyes, limbs, shadows) occluded	Empty bounding box despite RankNet detecting the target	Correction feedback
EfficientDet-D7	False positive	High-clutter scene; background matches training annotations	Spurious bounding box on non-target region	Reject feedback

The failure modes summarized in Table 5.1 have direct implications for trust calibration. A user who observes a confident rank map overlaying the wrong region, with no mechanism to correct it, is likely to experience a sharp decrease in performance-based trust even if their understanding of the system's general capabilities remains intact. This is precisely the dynamic that motivates MICA's mixed-initiative design. The Polygon Tool allows structural corrections to false-positive detection boundaries; the Rank Brush Tool allows salience corrections to mislocalized rank map concentrations; and the detection feedback panel allows operators to explicitly reject false positives and flag missed detections for subsequent model refinement. Failure is therefore not terminal in the MICA workflow. It is a recoverable event that the system is designed to absorb, record, and eventually learn from through the LoRA retraining pipeline. This recovery mechanism is the engineering operationalization of the framework's core theoretical claim: that agency over system outputs, rather than transparency into system internals, is the operative mechanism for sustained trust calibration under conditions of imperfect AI performance.

5.10 Chapter Summary

This chapter presented the design and implementation of the MICA prototype system, demonstrating how the mixed-initiative framework described in Chapter 4 can be instantiated within an operational camouflage detection workflow. The implementation integrates a WPF-based interactive interface with a Python-based computer-vision backend via the NEXUS bridging architecture, enabling responsive interaction with AI detection outputs.

Key interface components, including the unified editing environment, poly-

gon mask editor, rank brush editor, and detection feedback mechanisms, provide operators with structured control over both detection boundaries and camouflage salience rankings. Together, these mechanisms transform AI analysis from a passive explanatory process into an interactive collaborative workflow.

The prototype establishes the technical foundation for evaluating the central hypothesis of this dissertation: that interactive agency in AI-assisted analysis can improve user trust and analytical effectiveness in camouflage detection tasks. The following chapter presents the synthetic camouflage generation framework, which establishes the controlled evaluation infrastructure required for future empirical validation of MICA.

Chapter 6

Perceptual Loss Design for Synthetic Camouflage Generation

This chapter presents the third primary contribution of this dissertation: a framework for synthetic camouflage generation that integrates human perceptual principles into the training objectives of latent diffusion models. Motivated by the dataset limitations identified in Chapter 2 and the evaluation infrastructure requirements of the MICA framework introduced in Chapter 4, this work investigates whether optimizing generative models to better satisfy perceptual criteria actually produces outputs that more effectively simulate natural camouflage.

The central finding is a systematic optimization–perception gap. Across a fifteen-preset loss ablation evaluated on three architectures, the standard evaluation metrics change little, yet the visual character of the generated images varies substantially, and the architecture whose metrics respond most strongly to perceptual loss guidance is not the architecture whose images are most useful for camouflage. This paradox reveals that standard evaluation metrics can shift independently of the perceptual dimensions most relevant to domain-specific

generation tasks, a finding with implications that extend beyond camouflage synthesis to the design of perceptual loss for generative models more broadly.

Realizing this finding required identifying and resolving a previously unaddressed technical obstacle: naive perceptual loss formulations used in latent diffusion training collapse to mean-squared-error equivalents because the learned latent channels of diffusion models do not encode RGB color semantics. The perceptual weight map mechanism introduced here addresses this by performing CPU-side decoding via a VAE, which provides pixel-space perceptual guidance at no additional GPU memory cost. Portions of this chapter have been submitted for publication [Hogue et al., 2026a].

6.1 Motivation and Research Context

As established in Section 2.8, dataset constraints represent a persistent bottleneck for CODS research. Natural camouflage examples are difficult to collect at scale, ground-truth segmentation masks require expert annotation, and existing benchmarks such as CAMO [Le et al., 2019] and COD10K [Fan et al., 2020a] offer limited coverage across environments, camouflage strategies, and difficulty levels. Synthetic generation addresses these constraints by enabling systematic manipulation of camouflage properties: color relationships, texture patterns, and blending intensity, while preserving known labels and maintaining controlled variation across conditions.

Hypothesis H3, introduced in Chapter 1, posits that synthetic camouflage generation enables controlled evaluation of CODS systems across varied concealment strategies. This chapter operationalizes that hypothesis by developing and evaluating a generation framework grounded in the perceptual mechanisms that

govern camouflage effectiveness.

Effective camouflage exploits specific perceptual vulnerabilities rather than achieving generic visual similarity [Lv et al., 2023]. Biological and designed camouflage disrupt the Gestalt mechanisms that enable figure-ground segregation: specifically, background matching and disruptive coloration that suppress the edge grouping and continuity cues on which detection depends [Troscianko et al., 2009, Stevens and Merilaita, 2009]. A perceptually grounded generation framework should therefore optimize for these specific disruption properties rather than for general image-quality criteria designed for different tasks.

Existing synthetic camouflage methods are constrained by three limitations. First, most generative pipelines operate in RGB color space, which does not separate chromatic information from luminance as human vision does. Second, training objectives typically optimize pixel-level metrics (MSE and binary cross-entropy), which do not encode the perceptual grouping principles that determine whether camouflage succeeds or fails. Third, the field lacks systematic comparisons across generative architectures, making it difficult to disentangle the effects of loss function design from those of architectural capacity. This work addresses all three limitations while introducing a technical solution to a deeper obstacle: integrating perceptual objectives into the latent spaces in which modern diffusion models operate.

6.2 The Latent-Space Integration Challenge

Modern diffusion models such as Stable Diffusion [Rombach et al., 2022] operate on learned latent representations rather than directly on pixel space (see Appendix G for technical background on latent diffusion models). A UNet, a

convolutional encoder-decoder with skip connections, predicts the noise $\hat{\epsilon}$ added to latent encodings $z = E(x)$, where E is a learned VAE encoder. The resulting latent space has four channels with no fixed correspondence to RGB color: the encoder learns an abstract, compressed representation optimized for reconstruction fidelity rather than semantic interpretability.

This design creates a representational mismatch for perceptual loss functions. HSV color conversion requires RGB inputs with well-defined chromatic semantics: hue encoding angular color position, saturation encoding chromatic purity, and value encoding luminance. Gestalt edge detection requires spatial features that correspond to visual boundaries in pixel space. Neither operation produces meaningful results when applied to latent-space tensors, regardless of normalization.

Three approaches and their failure modes were identified through systematic experimentation. First, applying an HSV conversion to sigmoid-normalized noise predictions treats the first three latent channels as pseudo-RGB channels. Because these channels encode learned features rather than color, the resulting "HSV" values are arbitrary nonlinear transformations that reduce to MSE variants. Second, estimating clean latents via the diffusion reverse formula yields spatially coherent representations, but the same channel-interpretation problem persists. Third, full VAE decoding with backpropagation through the decoder is theoretically sound, but computationally prohibitive at 512×512 resolution, because storing intermediate activations for gradient computation alongside the UNet's own gradient graph exceeds available VRAM.

The latent-space collapse phenomenon was confirmed empirically. An SD-1.5 model trained with naive perceptual components applied to latent tensors produced outputs that were metrically indistinguishable from an MSE-only base-

line: FID differed by less than 1 point, LPIPS by 0.0006, and CLIP Score by 0.01 (these metrics are defined in Section 6.4.3). The perceptual components computed non-zero gradients but encoded no meaningful color or edge information, resulting in training dynamics functionally equivalent to pure MSE. This validation establishes that the differences observed with the corrected implementation reflect genuine perceptual guidance rather than incidental variation.

6.3 HSV-Gestalt Loss Framework

This section describes the HSV-Gestalt loss framework in three parts. Section 6.3.1 presents the loss formulation that integrates noise prediction and weighted reconstruction objectives. Section 6.3.2 details the perceptual weight map mechanism that resolves the latent-space integration challenge. Section 6.3.3 describes the progressive integration design and the configurable loss-weight presets used in this study.

6.3.1 Loss Formulation

The HSV-Gestalt framework integrates a standard noise-prediction loss with spatial guidance derived from real pixel-space analysis. The total training objective is:

$$\mathcal{L}_{\text{total}} = w_{\text{MSE}}\,\mathcal{L}_{\text{MSE}}(\hat{\epsilon}, \epsilon) + (w_{\text{HSV}} + w_{\text{Gestalt}})\,\mathcal{L}_{\text{weighted}}(x_0, \hat{x}_0) \tag{6.1}$$

where $\mathcal{L}_{\text{MSE}}$ operates on noise predictions in latent space (the standard diffusion training objective), and $\mathcal{L}_{\text{weighted}}$ is a spatially-weighted x_0 reconstruction loss whose weights are informed by perceptual analysis of decoded RGB images.

Algorithm 5 HSV-Gestalt Total Loss Computation

Require: Noise prediction $\hat{\epsilon}$, ground-truth noise ϵ, estimated clean latents $\hat{x}_0$, target clean latents x_0, spatial weight map W, loss weights w_{MSE}, w_{HSV}, w_{Gestalt}

Ensure: Total training loss $\mathcal{L}_{\text{total}}$

1: $\mathcal{L}_{\text{MSE}} \leftarrow \|\hat{\epsilon} - \epsilon\|^2$ ▷ Standard latent-space noise prediction loss

2: $\mathcal{L}_{\text{weighted}} \leftarrow \frac{1}{N}\sum_i (\hat{x}_{0,i} - x_{0,i})^2 \cdot W_i$ ▷ Perceptually-weighted x_0 reconstruction loss

3: $\mathcal{L}_{\text{total}} \leftarrow w_{\text{MSE}} \cdot \mathcal{L}_{\text{MSE}} + (w_{\text{HSV}} + w_{\text{Gestalt}}) \cdot \mathcal{L}_{\text{weighted}}$

4: **return** $\mathcal{L}_{\text{total}}$

6.3.2 Perceptual Weight Map Mechanism

The perceptual weight map mechanism addresses the challenge of latent-space integration through a three-stage design that provides pixel-space perceptual guidance at zero additional GPU memory cost.

On every fourth training step, clean latents $\hat{x}_0$ are estimated from the noise prediction $\hat{\epsilon}$ using the diffusion reverse formula:

$$\hat{x}_0 = \frac{x_t - \sqrt{1 - \bar{\alpha}_t}\,\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \tag{6.2}$$

This estimate is differentiable: gradients flow from $\hat{x}_0$ through $\hat{\epsilon}$ to the UNet parameters. Both $\hat{x}_0$ and the target clean latents x_0 are then decoded using a permanent CPU-resident copy of the VAE decoder with `torch.no_grad()`, yielding real RGB images at zero GPU memory cost. On these decoded images, HSV color difference maps are computed (weighted 0.4 hue, 0.3 saturation, 0.3 value, following the perceptual primacy of chromatic over luminance channels [Cheng et al., 2001]), and Gestalt edge difference maps are computed via local gradient analysis [Wagemans et al., 2012]. These maps are combined into a spatial weight map W at the latent resolution (64×64), normalized to have a unit mean

so that the overall loss magnitude is preserved. The weighted reconstruction loss is:

$$\mathcal{L}_{\text{weighted}} = \frac{1}{N} \sum_{i} (\hat{x}_{0,i} - x_{0,i})^2 \cdot W_i. \tag{6.3}$$

The weight map is detached from the computation graph; no gradients flow through the VAE, but it spatially concentrates the differentiable x_0 reconstruction loss in regions with the largest perceptual errors, including areas with incorrect hue, broken edges, or disrupted figure-ground boundaries. A uniform weight map would be mathematically equivalent to standard MSE on x_0; the non-uniform weights derived from HSV and Gestalt analysis are what make the loss perceptually informed. This mechanism ensures that perceptual principles guide training through real pixel-space analysis while maintaining a clean gradient path to the UNet.

Algorithm 6 Perceptual Weight Map Construction via CPU-Side VAE Decoding

Require: Noisy latents x_t, noise prediction $\hat{\epsilon}$, diffusion schedule $\bar{\alpha}_t$, CPU-resident VAE decoder D_{CPU}, HSV weights $(w_H, w_S, w_V) = (0.4, 0.3, 0.3)$

Ensure: Spatial weight map W, estimated clean latents $\hat{x}_0$

if training step $\equiv 0 \pmod 4$ **then**

  $\hat{x}_0 \leftarrow \dfrac{x_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}}$ ▷ Differentiable x_0 estimate; gradients flow to UNet

  $\hat{I} \leftarrow D_{\text{CPU}}(\hat{x}_0\texttt{.detach()})$ ▷ Decode predicted latents to RGB; no GPU memory cost

  $I_0 \leftarrow D_{\text{CPU}}(x_0\texttt{.detach()})$ ▷ Decode target latents to RGB

  $W_{\text{HSV}} \leftarrow w_H \cdot \Delta H(\hat{I}, I_0) + w_S \cdot \Delta S(\hat{I}, I_0) + w_V \cdot \Delta V(\hat{I}, I_0)$ ▷ HSV channel difference map

  $W_{\text{Gestalt}} \leftarrow \|\nabla \hat{I} - \nabla I_0\|$ ▷ Gestalt edge difference map via local gradient analysis

  $W_{\text{raw}} \leftarrow W_{\text{HSV}} + W_{\text{Gestalt}}$

  $W \leftarrow \textsc{Resize}(W_{\text{raw}},\ 64 \times 64)$ ▷ Downsample to latent resolution

  $W \leftarrow W \,/\, \texttt{mean}(W)$ ▷ Normalize to unit mean; preserves overall loss magnitude

  $W \leftarrow W\texttt{.detach()}$ ▷ Detach from graph; no gradients flow through VAE

else

  $W \leftarrow \mathbf{1}^{64\times 64}$ ▷ Use uniform weights on non-update steps (equivalent to MSE)

  $\hat{x}_0 \leftarrow \dfrac{x_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}}$

end if

return $W,\ \hat{x}_0$

6.3.3 Progressive Integration Design

The framework supports progressive integration of perceptual components via configurable loss-weight presets (Table 6.1), ranging from an MSE-only baseline through increasing perceptual emphasis. The six presets reported in the body of this chapter trace a staged progression, from the `mse_only` baseline ($w_{\text{MSE}} = 1.0$, all perceptual components zero) to the `theoretical` configuration ($w_{\text{MSE}} = 0.4$, $w_{\text{HSV}} = 0.4$, $w_{\text{Gestalt}} = 0.2$), in which the perceptual terms carry the largest combined weight. The intermediate `full_gestalt` preset ($w_{\text{MSE}} = 0.5$, $w_{\text{HSV}} = 0.3$,

$w_{\text{Gestalt}} = 0.2$) serves as a balanced reference point that activates both perceptual terms without fully subordinating the noise-prediction objective. These weights were selected to sample the design space at roughly even intervals of perceptual emphasis rather than to optimize any single metric, so that the progression isolates the effect of increasing perceptual guidance. A further nine single-component and tuned presets, exercising the HSV and Gestalt terms in isolation, complete the full fifteen-preset ablation reported in Appendix F.1.

Table 6.1: Progressive HSV-Gestalt Loss Function Integration

Preset	w_{MSE}	w_{HSV}	w_{Gestalt}
Stage 1: Baseline Establishment			
`mse_only`	1.0	0.0	0.0
`conservative_hsv`	0.9	0.1	0.0
Stage 2: Balanced Integration			
`moderate_hsv`	0.8	0.2	0.0
`hsv_enhanced`	0.7	0.2	0.1
Stage 3: Theoretical Limits			
`full_gestalt`	0.5	0.3	0.2
`theoretical`	0.4	0.4	0.2

6.4 Experimental Setup

This section describes the experimental configuration in three parts. Section 6.4.1 describes the dataset and preprocessing pipeline. Section 6.4.2 details the multi-architecture LoRA fine-tuning procedure. Section 6.4.3 defines the five complementary evaluation metrics used to detect optimization–perception discrepancies.

6.4.1 Dataset and Preprocessing

Experiments draw 1,000 samples per configuration from a processed subset of COD10K [Fan et al., 2020a], encompassing camouflaged objects across six categories: amphibians, aquatic creatures, flying animals, humans, terrestrial animals, and miscellaneous objects. When the subset exceeds 1,000 samples, deterministic sampling (seed = 42) selects the first 1,000 after shuffling. Dataset splits follow an 80/20 train/validation protocol (800/200 samples), using a fixed random seed. The original dataset was augmented with descriptive text prompts and negative prompt specifications to enable fine-tuning of Stable Diffusion. Figure 6.1 illustrates representative examples from each category, highlighting the diversity of camouflage strategies and environmental contexts covered by the training set.

6.4.2 Multi-Architecture Training

Fine-tuning evaluated three Stable Diffusion architectures: SD-1.5, SD-2.0, and SD-2.1 [Rombach et al., 2022]. All three belong to the same latent diffusion family but differ in their pretrained UNet and VAE weights, which makes them a controlled testbed for isolating the effect of architecture from that of loss design. SDXL was excluded for principled architectural reasons: its VAE produces a different latent distribution (scaling factor $\approx$0.13025 vs. $\approx$0.18215 for SD-1.5) with distinct channel semantics, meaning the perceptual weight map mechanism cannot transfer across architectures with non-aligned latent spaces.

LoRA fine-tuning [Hu et al., 2022] targeted attention mechanism parameters (the query `to_q`, key `to_k`, value `to_v`, and output projection `to_out.0`), with rank $r = 16$, alpha scaling 32, and 0.1 dropout at 512×512 resolution. SD-2.1 employs a 0.5× reduction in learning rate relative to the base rate to improve

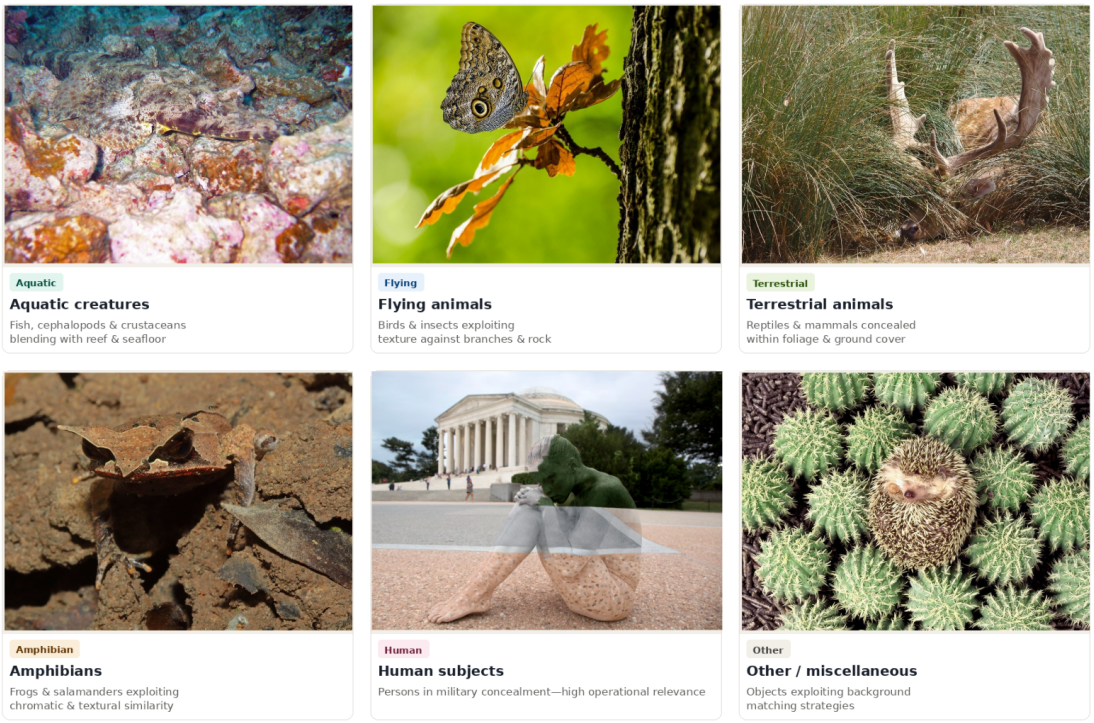

Figure 6.1: Representative examples from the six COD10K categories used in this study. Each category presents distinct camouflage challenges: aquatic creatures exploit reef and seafloor texture similarity; flying animals blend with branches and rock surfaces; terrestrial animals conceal within foliage and ground cover; amphibians exploit chromatic and textural similarity; human subjects represent high-operational-relevance concealment scenarios; and miscellaneous objects exploit background matching strategies. Training and evaluation drew 1,000 samples per configuration spanning all six categories.

training stability. Training spans 4–12 epochs with checkpoint saving at each interval. Standardized test images were generated at each checkpoint using fixed prompts (rainforest amphibian, military woodland, and aquatic camouflage scenarios) with consistent inference parameters (50 steps, guidance scale 7.5). All experiments employed deterministic seeding (seed = 42) across sampling, splitting, and initialization, with gradient clipping (max_norm = 1.0) and AdamW optimization with weight decay 0.01.

Algorithm 7 LoRA Fine-Tuning with HSV-Gestalt Perceptual Loss

Require: Pretrained UNet θ, VAE encoder E, CPU VAE decoder D_{CPU}, dataset $\mathcal{D}$, loss weights $w_{\text{MSE}}, w_{\text{HSV}}, w_{\text{Gestalt}}$, LoRA rank $r = 16$, epochs $T \in \{4, \ldots, 12\}$

Ensure: Fine-tuned LoRA parameters $\Delta\theta$

Initialize LoRA adapters $\Delta\theta$ on `to_q`, `to_k`, `to_v`, `to_out.0`
for each epoch $t = 1, \ldots, T$ **do**
  for each batch $(x_0, \text{prompt}) \in \mathcal{D}$ **do**
    $z_0 \leftarrow E(x_0)$ ▷ Encode image to latent space
    Sample timestep $t \sim \mathcal{U}[1, T_{\max}]$, noise $\epsilon \sim \mathcal{N}(0, I)$
    $x_t \leftarrow \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$ ▷ Forward diffusion
    $\hat{\epsilon} \leftarrow \text{UNet}_{\theta+\Delta\theta}(x_t, t, \text{prompt})$ ▷ Noise prediction with LoRA
    $W, \hat{x}_0 \leftarrow \textsc{PerceptualWeightMap}(x_t, \hat{\epsilon}, \bar{\alpha}_t, D_{\text{CPU}})$ ▷ Algorithm 6
    $\mathcal{L} \leftarrow \textsc{TotalLoss}(\hat{\epsilon}, \epsilon, \hat{x}_0, z_0, W, w_{\text{MSE}}, w_{\text{HSV}}, w_{\text{Gestalt}})$ ▷ Algorithm 5
    $\Delta\theta \leftarrow \Delta\theta - \eta \nabla_{\Delta\theta} \mathcal{L}$ ▷ AdamW update; only LoRA parameters trained
  end for
  $\textsc{SaveCheckpoint}(\Delta\theta, t)$
  $\textsc{GenerateTestImages}(\theta + \Delta\theta, \text{fixed prompts})$
end for
return $\Delta\theta$

6.4.3 Multi-Metric Evaluation

A central methodological decision in this work is the use of five complementary evaluation metrics rather than relying on FID alone, because no single metric captures the full spectrum of perceptual quality. Fréchet Inception Distance

(FID) [Heusel et al., 2017] measures distributional distance between generated and reference image features. Inception Score (IS) [Salimans et al., 2016] evaluates both quality and diversity of generated images. The Structural Similarity Index (SSIM) [Wang et al., 2004] quantifies structural fidelity by comparing luminance, contrast, and structure. Learned Perceptual Image Patch Similarity (LPIPS) [Zhang et al., 2018] computes perceptual distance using deep features that are aligned with human visual judgments. CLIP Score [Radford et al., 2021] measures text-image alignment between generated images and their conditioning prompts.

This multi-metric approach is essential for detecting optimization–perception discrepancies: a configuration may achieve a strong FID while producing structurally dissimilar outputs, or generate perceptually distant images despite a low training loss. Evaluation employed deterministic sampling (seed = 42) with up to 500 paired comparisons for SSIM and LPIPS metrics.

6.5 Results

Results are presented in four parts. Section 6.5.1 reports SD-1.5, Section 6.5.2 reports SD-2.0, and Section 6.5.3 reports SD-2.1. Section 6.5.4 then synthesizes the three architectures and characterizes the optimization–perception gap. Table 6.2 reports the five-metric evaluation of the six-preset progression for all three architectures and is referenced throughout; the complete fifteen-preset ablation appears in Appendix F.1. The headline observation, developed below, is that the metrics span a narrow range within each architecture while the generated images vary substantially in character, and that the ranking of metric responsiveness across architectures runs nearly opposite to the ranking of perceptual quality.

Table 6.2: Five-metric evaluation of the six-preset HSV-Gestalt progression across SD-1.5, SD-2.0, and SD-2.1. Within each architecture, the best value per metric is shown in bold.

Architecture	Preset	FID↓	IS↑	SSIM↑	LPIPS↓	CLIP↑
SD-1.5	`mse_only`	149.97	**2.87**	**0.0903**	0.7091	21.81
	`conservative_hsv`	149.40	2.76	0.0903	0.7090	21.77
	`moderate_hsv`	**148.35**	2.80	0.0900	0.7086	21.79
	`hsv_enhanced`	149.30	2.81	0.0890	0.7082	21.80
	`full_gestalt`	149.05	2.79	0.0881	0.7082	21.82
	`theoretical`	149.02	2.79	0.0882	**0.7072**	**21.82**
SD-2.0	`mse_only`	129.59	3.03	0.0990	0.6834	22.48
	`conservative_hsv`	**129.58**	3.00	0.0997	0.6836	22.49
	`moderate_hsv`	129.84	3.00	0.1006	0.6838	22.50
	`hsv_enhanced`	129.84	3.04	0.1011	0.6838	22.49
	`full_gestalt`	129.72	3.00	0.1010	0.6831	22.54
	`theoretical`	129.91	**3.04**	**0.1023**	**0.6830**	**22.57**
SD-2.1	`mse_only`	162.75	**3.95**	0.2222	0.8533	22.39
	`conservative_hsv`	164.25	3.86	0.2223	0.8542	22.46
	`moderate_hsv`	163.02	3.92	**0.2227**	0.8549	22.42
	`hsv_enhanced`	163.21	3.93	0.2223	0.8533	22.42
	`full_gestalt`	158.37	3.86	0.2204	0.8426	22.47
	`theoretical`	**155.15**	3.91	0.2173	**0.8272**	**22.48**

6.5.1 SD-1.5: Modest Metrics, Stable Realism

For SD-1.5, the metrics move only slightly across the perceptual progression. Relative to the `mse_only` baseline (FID 149.97), FID changes by less than one point through the staged presets, reaching 149.05 at `full_gestalt` and 149.02 at `theoretical`, with the lowest value in the full ablation, 148.32, attained by a single-component HSV preset (Appendix F.1). LPIPS improves marginally from 0.7091 to 0.7072, and the remaining metrics shift within comparably narrow bands. These numerical differences are modest. Yet the generated images exhibit substantial visual divergence across the evaluation prompts, with the perceptually guided presets producing markedly different color palettes, texture distributions, and compositional structures relative to the baseline. Training loss dynamics also differ substantially: perceptually guided runs peak at approximately 0.5 against 0.8–1.0 for the baseline, indicating that the perceptual weight maps reshape the loss landscape even when metric shifts are small. Figure 6.2 traces SD-1.5's generative character across the full training progression, showing that the model selected by validation loss (Epoch 2) produces outputs consistent with the photorealistic, scene-integrated quality the architecture maintains throughout training.

6.5.2 SD-2.0: Best Metrics, Wandering Realism

SD-2.0 records the strongest standard metrics of the three architectures, with the lowest FID (a baseline of 129.59, against roughly 150 for SD-1.5 and 163 for SD-2.1) and the lowest LPIPS (near 0.683). It is also the architecture whose metrics respond least to perceptual guidance. Across the progression, FID moves by a few hundredths to a few tenths of a point and does not improve monotonically:



Figure 6.2: SD-1.5 baseline training progression across 12 epochs for the underwater aquatic prompt. SD-1.5 maintains photorealistic, scene-integrated outputs throughout training. The best model checkpoint at Epoch 2 (highlighted, selected by validation loss) produces a visually coherent and perceptually compelling output, consistent with the model's other checkpoints. This stable generative character contrasts with the degradation observed for SD-2.1 in Figure 6.4.

`full_gestalt` sits at 129.72 and `theoretical` at 129.91, both marginally above the baseline, while the lowest FID in the full ablation, 129.53, differs from the baseline by less than one tenth of a point (Appendix F.1). By the logic of the metrics alone, SD-2.0 is the best model in the study and perceptual loss does essentially nothing to it.

The images tell a more complicated story. Figure 6.3 traces the SD-2.0 baseline across its twelve epochs. The architecture holds photorealistic, scene-integrated rendering through most of training, but it does not hold it stably: Epoch 2 departs into a painterly, illustration-like register before the run returns to photorealism, and later epochs exhibit isolated coherence failures, including an implausible bright-green head covering at Epoch 5 and a semantic breakdown at Epoch 11 in which the subject is turned away from the viewer beside two large, unexplained hollows bored into the tree trunks. The checkpoint selected by validation loss is exactly this Epoch 11, one of the least coherent frames in the run. As with SD-2.1 below, the metric-optimal checkpoint is not the perceptually preferable one. SD-2.0 thus occupies a distinct position in the study: its metrics are the best and the most inert, while its visual output is generally realistic but unstable from epoch to epoch, a combination the metrics give no indication of.

6.5.3 SD-2.1: Larger Metric Improvements, Degradation to Abstraction

SD-2.1 is the architecture whose metrics respond most strongly to perceptual guidance. Relative to its baseline (FID 162.75), `full_gestalt` lowers FID by 4.38 points to 158.37 and `theoretical` by 7.60 points to 155.15, while LPIPS improves from 0.8533 to 0.8272, an order of magnitude larger than the LPIPS



Figure 6.3: SD-2.0 baseline training progression across 12 epochs for the military woodland prompt. SD-2.0 holds photorealistic rendering through most of training but wanders, with a painterly excursion at Epoch 2 and isolated coherence failures at Epochs 5 and 11. The best model checkpoint at Epoch 11 (highlighted, selected by validation loss) shows a semantic breakdown, the subject turned away beside unexplained hollows in the trunks, despite achieving the lowest validation loss. The metric-selected checkpoint is again not the perceptually preferable one.

movement seen for SD-1.5 or SD-2.0. On the metrics, SD-2.1 is the clear beneficiary of the perceptual loss. Its images, however, occupy a fundamentally different generative register from SD-1.5's photorealistic scene integration: across the evaluation prompts the outputs are abstract and desaturated, with limited recognizable subject detail, and they change little in character between the baseline and the perceptually guided presets. Figure 6.4 traces this behavior across the full training progression, showing that SD-2.1 begins with photorealistic early epochs and degrades progressively toward desaturated abstraction, and that the checkpoint selected by validation loss (Epoch 11) is among the visually weakest in the run despite achieving the lowest validation loss. SD-2.1 therefore presents the sharpest within-architecture form of the gap: the largest metric gains accompany the poorest perceptual outcomes.

6.5.4 The Optimization–Perception Gap

Taken together, the three architectures expose the optimization–perception gap as a systematic pattern rather than an isolated observation. Two regularities stand out in Table 6.2.

The first concerns responsiveness. Ranked by how much the metrics move under perceptual loss guidance, the architectures order as SD-2.1 (theoretical ΔFID -7.60, ΔLPIPS -0.0261), then SD-1.5 (ΔFID -0.96, ΔLPIPS -0.0019), then SD-2.0 (ΔFID $+0.32$, essentially inert or marginally worse). This ranking runs nearly opposite to the ranking of perceptual quality established by the progression figures, where SD-1.5 is the most stable and scene-integrated, SD-2.0 is realistic but unstable, and SD-2.1 degrades to abstraction. The architecture that the metrics reward most for perceptual guidance is the one whose images

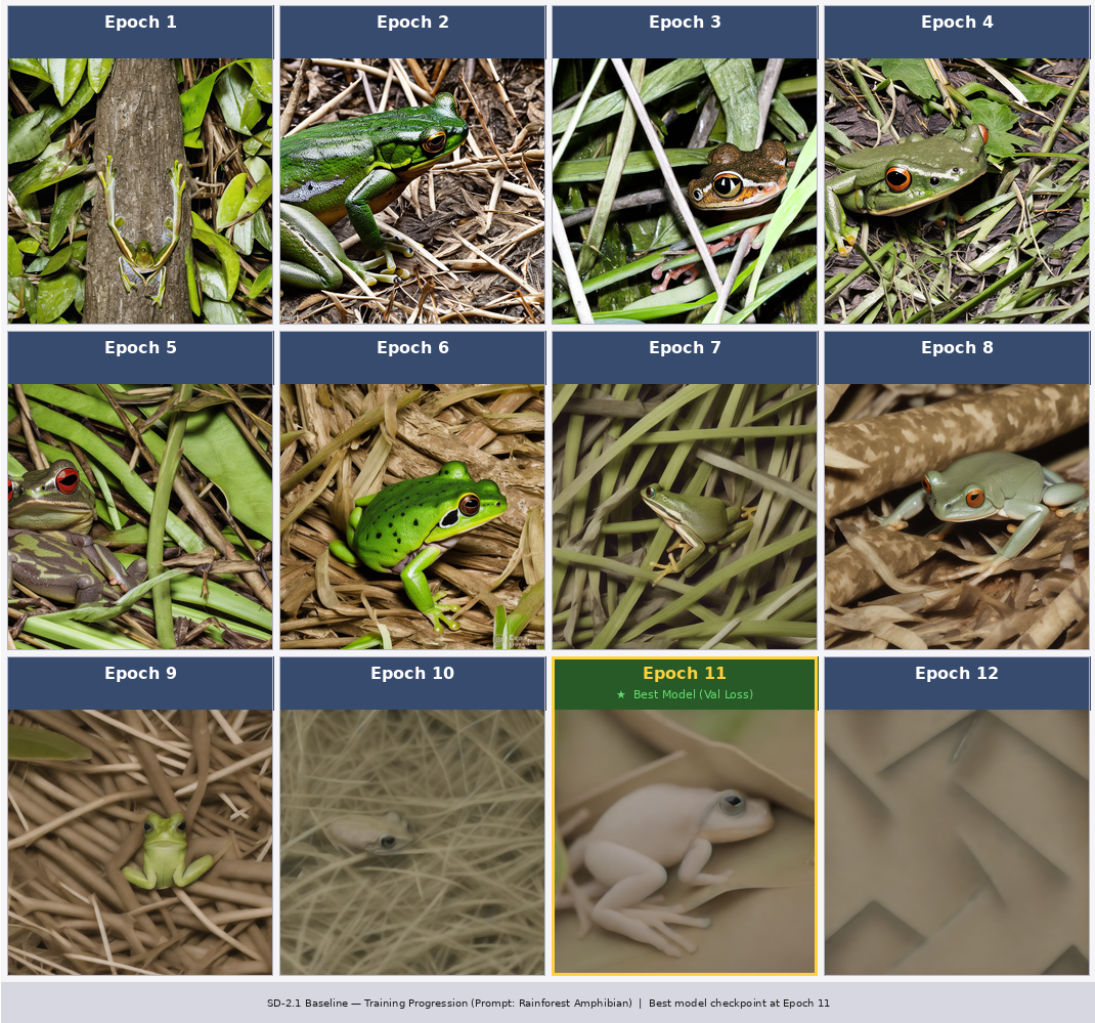

Figure 6.4: SD-2.1 baseline training progression across 12 epochs for the rainforest amphibian prompt. Outputs are photorealistic in early epochs but progressively desaturate and lose subject detail. The best model checkpoint at Epoch 11 (highlighted, selected by validation loss) produces an abstract, nearly featureless output, visually inferior despite achieving the lowest validation loss. This progression illustrates how standard metric optimization can diverge from perceptual quality over the course of training.

are least suited to camouflage, and the architecture the metrics leave untouched is among the most realistic. A researcher selecting a model or a loss configuration on metric movement alone would be led directly away from the perceptually preferable outcome.

The second concerns absolute metric standing. SD-2.0 attains the best FID and LPIPS in the study, and SD-2.1 attains the best IS and SSIM; however, based on the progression figures, neither architecture is one a human observer would choose for photorealistic camouflage. Abstract, desaturated outputs achieve higher Inception Scores through classifiable diversity, higher SSIM through low-frequency structural similarity to the muted tones of real camouflage, and competitive CLIP Scores through coarse text-image alignment, all without producing a recognizable camouflaged subject. Metric superiority does not imply perceptual superiority, and in this study, the two diverge in both directions: better metrics with worse images across architectures, and larger metric gains with no perceptual benefit. Figure 6.5 illustrates the effect directly, pairing baseline and perceptually guided outputs with their metric deltas for the two architectures that bracket the gap.

6.6 Discussion

This section interprets the experimental findings across four dimensions. Section 6.6.1 examines architecture as a design variable. Section 6.6.2 discusses implications for the latent diffusion pipeline. Section 6.6.3 addresses the relationship to CODS evaluation infrastructure. Section 6.6.4 acknowledges study limitations.

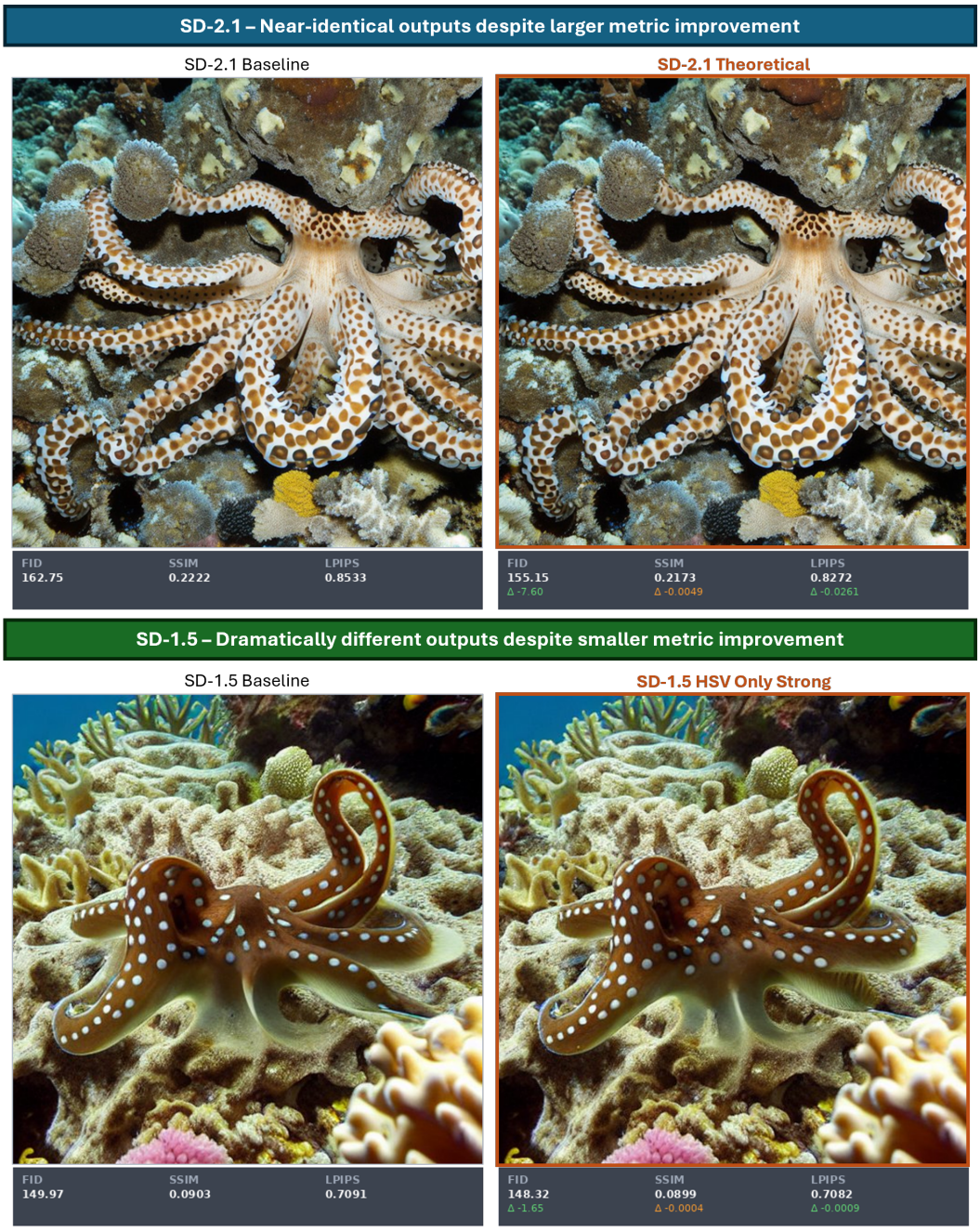

Figure 6.5: The optimization–perception paradox across the two architectures that bracket the gap. *Top*: SD-2.1 produces near-identical outputs despite substantial metric improvements ($\Delta\text{FID} = -7.60$, $\Delta\text{LPIPS} = -0.0261$). *Bottom*: SD-1.5 produces a substantially different color palette, texture, and subject rendering despite modest metric shifts ($\Delta\text{FID} = -1.65$, $\Delta\text{LPIPS} = -0.0009$). Metric improvement does not predict perceptual change.

6.6.1 Architecture as a Design Variable

The three architectures, trained on identical datasets with identical LoRA configurations and loss presets, produce outputs in distinct generative registers and respond to perceptual guidance in distinct ways. SD-1.5 is perceptually responsive in the sense that matters for the task: its images change substantially under the weight maps while remaining photorealistic and scene-integrated. SD-2.1 is perceptually resistant in the way that matters least: it produces the largest metric shifts but negligible visual improvement, suggesting that its abstract generative tendencies are embedded in architectural priors that loss-level intervention cannot override. SD-2.0 sits between them on visual quality but at the extreme on metric inertia, posting the best FID and LPIPS in the study while barely moving under perceptual loss, and wandering stylistically across epochs in a way the metrics do not register. The three cases make the same point from different directions: architecture, not loss design alone, determines whether perceptual guidance helps, and metric movement is a poor proxy for whether it does.

For camouflage generation that requires photorealistic environmental integration, SD-1.5 is the appropriate architecture, not because it achieves the best metrics (it does not), but because its generative character aligns with the task and it responds meaningfully to perceptual guidance. Architecture selection must therefore be informed by qualitative domain requirements rather than by metric rankings alone. This finding challenges the assumption that perceptual loss functions can be designed once and applied universally across architectures, and it reinforces the importance of multi-metric, perceptually grounded evaluation in any domain-specific generative modeling work.

6.6.2 Implications for the Latent Diffusion Pipeline

The latent-space collapse finding demonstrates that implementation matters as much as mathematical formulation: the same HSV and Gestalt principles produce either no effect (when applied to latent tensors) or substantial generative changes (when applied through pixel-space weight maps), depending solely on where in the pipeline the analysis occurs. This has implications beyond camouflage generation. Any researcher integrating domain-specific perceptual objectives into latent diffusion architectures will encounter the same representational mismatch: the learned latent channels are not RGB, and operations that assume color or spatial semantics produce meaningless values. The perceptual weight map mechanism, decoupling pixel-space analysis from the gradient path via CPU-side VAE decoding, provides an extensible solution applicable to any perceptual objective that requires real image content.

This representational mismatch extends to architectural compatibility. Each diffusion model is tightly coupled to its VAE: latent distributions, scaling factors, and channel semantics are architecture-specific and non-interchangeable. The perceptual weight map mechanism sidesteps this coupling by always decoding through the matched VAE, but the constraint implies that any perceptual loss framework must be validated on a per-architecture basis rather than assumed to transfer. The divergent responses of SD-1.5, SD-2.0, and SD-2.1 to an identical loss specification are direct evidence of this.

6.6.3 Relationship to CODS Evaluation Infrastructure

This generation framework addresses the dataset constraints identified in Chapter 2 and supports the controlled evaluation infrastructure that the MICA frame-

work's future empirical program will require. Synthetic camouflage data generated via this pipeline enables systematic stress-testing of detection models across texture complexity, environmental clutter, and lighting conditions. The COD10K categorical organization provides a natural scaffold for environment-specific generation, and future work can extend the present six-preset progression and full ablation to additional domain-specific architectures. Most promisingly, leveraging detection saliency maps as adversarial evaluators offers a path toward closing the loop between generation and detection, optimizing camouflage effectiveness directly against the perceptual vulnerabilities that detection models exploit.

6.6.4 Limitations

Visual assessment in this study relies on the researcher's evaluation rather than crowdsourced human perceptual judgment, limiting the generalizability of perceptual quality claims. The progression figures and Figure 6.5 present representative outputs with inline metric deltas, enabling readers to assess the observed differences independently and evaluate whether the researcher's perceptual judgments are warranted.

A rigorous future evaluation would recruit naive observers via a crowdsourcing platform such as Amazon Mechanical Turk or Prolific to perform forced-choice comparisons between baseline and perceptually-guided outputs across matched image pairs, using a standardized protocol that controls for order effects and presentation duration. Such a study would establish whether the visual differences reported here are reliably detectable by independent human judges, enabling calibration of the optimization–perception gap against an empirically validated perceptual standard. Until such validation is conducted, perceptual quality claims

in this chapter should be understood as researcher-observed rather than independently confirmed.

The evaluation covers three architectures within the same model family, and conclusions may not generalize to substantially different architectures or training paradigms. Each experiment uses 1,000 training samples from a single dataset, and the perceptual weight map is computed every fourth training step from a single decoded sample, thereby introducing stochasticity in the guidance signal. The progression analyses focus on the baseline run of each architecture; per-preset progression studies and isolated component ablation beyond the metrics reported in Appendix F.1 remain ongoing work.

6.7 Summary

This chapter introduced a framework for synthetic camouflage generation that translates Gestalt perceptual principles and HSV color analysis into differentiable training objectives for latent diffusion models. Three primary findings emerged.

First, naive perceptual loss formulations applied in latent diffusion training collapse to MSE equivalents because diffusion model latent channels lack RGB color semantics; the perceptual weight map mechanism resolves this by decoupling pixel-space analysis from the gradient path via CPU-side VAE decoding.

Second, a systematic optimization–perception gap characterizes both within- and between-architecture comparisons. Across a fifteen-preset ablation on three architectures, metric movement is small and does not track visual quality, and the ranking of metric responsiveness runs nearly opposite to the ranking of perceptual quality, so that the architecture metrics reward most for perceptual guidance is the one least suited to the task.

Third, SD-1.5 is the appropriate architecture for photorealistic camouflage generation given its perceptual responsiveness and stable, scene-integrated generative character, despite SD-2.0 and SD-2.1 posting superior standard evaluation metrics.

Together, these findings operationalize Hypothesis H3, establish infrastructure for controlled evaluation of CODS systems, and contribute principled guidance for integrating perceptual objectives into latent diffusion training pipelines.

Chapter 7

Beyond Deliberation: The Crisis Adoption Problem and Future Research Directions

This chapter synthesizes the implications of the empirical findings (Chapter 3), the framework design (Chapter 4), the implementation (Chapter 5), and the synthetic camouflage generation framework (Chapter 6), and identifies directions for future research. Section 7.1 reviews the dissertation's contributions. Section 7.2 develops the theoretical implications of the understanding–trust gap, proposing an initial formulation of the Crisis Adoption Problem as a lens for understanding AI trust formation under operational conditions. Sections 7.3–7.6 outline future empirical, technical, and applied research directions.

7.1 Summary of Contributions

This dissertation identified a persistent understanding–trust gap in AI-assisted CODS systems, showing that explainability and interactivity improve perceived engagement without proportionally increasing trust. The MICA framework was proposed to address this structural limitation, and a prototype was implemented to demonstrate the feasibility of the architecture. The synthetic camouflage generation framework (Chapter 6) additionally established a controlled evaluation infrastructure and documented a systematic optimization–perception gap in latent diffusion training, providing principled guidance for perceptual loss design in domain-specific generative modeling.

The NEXUS computing infrastructure also provides a reusable experimental platform for future mixed-initiative interaction research, isolating the WPF interface from the Python detection pipeline via a subprocess-based bridging architecture that enables alternative models and interaction mechanisms to be evaluated independently.

Together, these contributions clarify the structural relationship between understanding, agency, and trust in AI-assisted detection systems and provide a framework for designing interaction mechanisms that support calibrated reliance in high-stakes visual tasks.

7.2 Theoretical Implications: The Crisis Adoption Problem

This section develops the theoretical implications of the understanding–trust gap in five parts. Section 7.2.1 contextualizes the finding within operational

AI deployment. Section 7.2.2 states the Crisis Adoption Problem. Section 7.2.3 grounds the problem in established theory. Section 7.2.4 reframes the understanding–trust gap through this lens. Section 7.2.5 describes the scope and status of the formulation.

7.2.1 From Deliberation to Crisis

Chapter 3 demonstrated that the understanding–trust gap persists under ideal deliberative conditions: no time pressure, no consequences for errors, no competing task demands, and a structured exploration period in which participants could examine the system at their own pace. If passive explainability does not eliminate the gap under ideal deliberative conditions, its effectiveness under operational constraints warrants systematic examination.

This question exposes a structural assumption in XAI research that has received insufficient scrutiny. The dominant paradigm assumes that users will encounter AI systems under conditions conducive to deliberative reasoning, adequate time, low cognitive load, minimal operational stress, and will engage thoughtfully with the explanations provided. Under these conditions, the reasoning goes, transparency enables comprehension, comprehension enables calibrated trust, and calibrated trust enables appropriate reliance.

The Phase 1 findings challenge this pipeline at its first link: transparency did not differentially improve comprehension ($d = -0.02$), and comprehension did not translate into trust in proportion (a 0.60-point gap). If the deliberative pathway fails under ideal conditions, the problem is not that users need better explanations or more time. The problem is that the relationship between understanding and trust is not the linear pipeline that the XAI paradigm assumes.

This observation motivates what we term the *Crisis Adoption Problem*: the recognition that AI adoption decisions in operational contexts occur under conditions fundamentally different from those assumed in system design and evaluation.

7.2.2 The Crisis Adoption Problem

The Crisis Adoption Problem can be stated concisely: *we design for deliberation, but adoption happens in crisis.* This formulation is not intended to imply that all operational decisions occur in crisis, but rather that high-stakes AI adoption often unfolds under constraints not reflected in laboratory evaluation paradigms.

In domains where camouflaged object detection is deployed, surveillance, defense, search-and-rescue, intelligence analysis, operators rarely encounter AI systems for the first time under calm, exploratory conditions. Initial adoption decisions and ongoing reliance behaviors occur under time pressure, high cognitive load, cascading task demands, and uncertain consequences Klein [1989], Endsley [1995]. These are precisely the conditions least conducive to the deliberative evaluation that explainability features are designed to support.

The Crisis Adoption Problem is not simply a matter of ecological validity, the familiar concern that laboratory findings may not generalize to operational settings. It is a deeper structural claim: that the *mechanisms* by which users develop trust in AI systems may differ qualitatively between deliberative and crisis contexts. Under deliberation, users may engage in systematic evaluation of explanations, building trust through comprehension. In a crisis, users may rely on heuristic strategies such as pattern matching, satisficing, automation offloading, that bypass the explanatory pathway entirely.

7.2.3 Theoretical Foundations

The Crisis Adoption Problem draws on several established research traditions that, taken together, suggest the deliberation-crisis mismatch is not merely a deployment inconvenience but a fundamental challenge for XAI.

Bounded Rationality and Satisficing

The concept of bounded rationality [Simon, 1956] provides the intellectual foundation. Human decision-makers do not optimize; they satisfice, accepting solutions that meet threshold criteria ("is this good enough?") rather than conducting exhaustive evaluation. Under operational constraints, satisficing becomes the dominant strategy. Applied to AI-assisted detection, this suggests that operators may rely on heuristic reliance rules ("trust the system when confidence is high") rather than systematically evaluate the explanations that XAI systems are designed to provide.

The understanding–trust gap documented in Chapter 3 can be interpreted through this lens: even under ideal conditions, participants may have satisficed on understanding ("I get what this does") without engaging in the deeper evaluative process required for calibrated trust ("I have sufficient evidence to rely on this"). Under operational pressure, this tendency toward satisficing would only intensify, further widening the gap between comprehension and confidence.

Naturalistic Decision-Making

Klein [1989]'s Recognition-Primed Decision (RPD) model demonstrates that experienced operators in time-pressured environments rely on pattern recognition and rapid situation assessment rather than exhaustive option comparison Zsam-

bok [1997]. In the RPD framework, decision-makers recognize situations as typical or atypical, mentally simulate a course of action, and execute if the simulation succeeds, without systematically comparing alternatives or consulting detailed analytical outputs.

Applied to AI-assisted CODS, the RPD model predicts that operators under time pressure will attend to high-level system cues (confidence scores, alert salience, spatial patterns) rather than engaging with detailed explanations (Grad-CAM visualizations, textual rationale, uncertainty decompositions). This does not mean explanations are valueless; they may support training, post-hoc review, and expertise development, but it does mean their role in real-time adoption decisions may be marginal.

Cognitive Load and Resource Competition

Sweller's cognitive load framework Sweller [1988], Sweller et al. [2011] predicts that extraneous cognitive demands reduce capacity for deliberative reasoning. In operational CODS contexts, concurrent tasks, including multiple feeds, coordinating with teammates, managing cascading alerts, and maintaining situational awareness, impose substantial extraneous load. Detailed XAI outputs, however well-designed, add to this load rather than reducing it. Under high cognitive load, users may rationally choose to ignore explanations and rely on simpler decision heuristics, effectively neutralizing the XAI features that designers intended to support trust calibration.

Automation Reliance Under Stress

Parasuraman and Manzey's [2010] complacency-potential model and Onnasch et al.'s meta-analysis Onnasch et al. [2014] demonstrate that time pressure and

high task demands increase reliance on automation, sometimes beyond what system reliability warrants. This suggests a paradoxical risk: the conditions under which users most need calibrated trust (high-stakes, time-pressured operations) are precisely those that promote uncalibrated overreliance. Interactive control mechanisms that maintain user engagement with the detection process may mitigate this complacency risk, but this hypothesis remains to be tested empirically.

7.2.4 Relationship to the Understanding–Trust Gap

The Crisis Adoption Problem reframes the understanding–trust gap identified in Chapter 3. Under the standard XAI account, the gap represents a failure of explanation quality: if explanations were better, users would develop appropriate trust. Under the Crisis Adoption account, the gap reflects a structural mismatch between how trust is *designed to develop* (through deliberative evaluation of explanations) and how trust *actually develops* under operational conditions (through accumulated performance observation, heuristic reliance, and satisficing under constraint).

This reframing has a practical implication: closing the understanding–trust gap may require not better explanations but different *mechanisms*, mechanisms that support trust calibration through the heuristic, recognition-primed processes that operators actually use under pressure. MICA's interactive control features (Chapter 4) represent one such mechanism: they provide rapid-access agency that does not depend on sustained engagement with detailed explanations. Whether these mechanisms prove sufficient under crisis conditions is an empirical question that motivates the research program proposed in Section 7.3.

7.2.5 Scope and Status

The Crisis Adoption Problem is presented here as an emerging theoretical direction motivated by this dissertation's empirical findings, rather than as a validated framework. The core claim, that the deliberation-crisis mismatch undermines XAI effectiveness in operational contexts, is grounded in established theory (bounded rationality, NDM, cognitive load) and consistent with the Phase 1 data, but it has not been directly tested. The experimental paradigms proposed in Section 7.3 outline the empirical program necessary to evaluate and refine this initial formulation.

7.3 Proposed Empirical Program

Testing the Crisis Adoption Problem requires controlled manipulations of the contextual factors that distinguish deliberative from operational conditions. This section outlines five experimental paradigms, associated hypotheses, and measurement considerations.

7.3.1 Research Questions and Hypotheses

The proposed empirical program addresses four research questions:

RQ1: How does time pressure influence reliance on explainability versus control features?

Hypothesis 1a: Under time constraints, users will engage less frequently with passive explanation features (saliency maps, attention visualizations) compared to low-pressure conditions.

Hypothesis 1b: Under time constraints, users will increase reliance on inter-

active control features as rapid-access influence mechanisms.

RQ2: Does cognitive load moderate the understanding–trust gap?

Hypothesis 2: The understanding–trust gap will widen under high cognitive load, as users lack cognitive resources to translate comprehension into calibrated reliance.

RQ3: How do operational stressors influence trust calibration and error detection?

Hypothesis 3a: Time pressure will increase reliance on AI outputs, including inappropriate reliance on incorrect detections (complacency).

Hypothesis 3b: Interactive control mechanisms will mitigate complacency by preserving user agency and situational awareness under stress.

RQ4: Does training context influence crisis performance?

Hypothesis 4: Users trained to operate MICA under simulated time pressure and cognitive load will demonstrate more appropriate reliance and better error detection than users trained under low-pressure conditions, when both groups are tested under crisis conditions.

7.3.2 Experimental Paradigms

Paradigm 1: Time-Constrained Detection Tasks

Participants perform CODS analysis under countdown timers (30 seconds per image vs. unlimited time). This paradigm directly tests whether time pressure shifts feature engagement from passive explanations to interactive controls. Dependent measures include explanation feature engagement (frequency and duration), control feature usage, detection accuracy, and appropriateness of reliance (correlation between system confidence and user acceptance).

Paradigm 2: Dual-Task Cognitive Load

Participants monitor multiple surveillance feeds simultaneously or respond to secondary alerting tasks while performing primary CODS analysis. This paradigm simulates operational multitasking and tests whether interactive controls preserve performance under divided attention. The dual-task methodology allows parametric manipulation of cognitive load while maintaining ecological relevance.

Paradigm 3: Simulated Operational Environment

Participants operate within a scenario-based simulation (perimeter monitoring, threat assessment) with cascading task demands, uncertain consequences, and dynamic prioritization requirements. This paradigm tests ecological validity and examines how users allocate attention between consuming explanations and executing tasks directly. Unlike Paradigms 1 and 2, which isolate individual stressors, this paradigm combines multiple operational demands to approximate real-world conditions.

Paradigm 4: Consequence Framing

Participants are informed that detection errors have operational consequences (missed threats incur security costs; false alarms incur resource costs). This manipulation tests whether perceived stakes influence reliance strategies without introducing physical stressors, isolating the cognitive effects of consequence awareness from the physiological effects of stress.

Paradigm 5: Longitudinal Training Study

Participants are trained under either low-pressure or crisis-simulated conditions, then tested under crisis conditions. This design examines whether crisis-adapted training improves performance and whether deliberative training transfers to crisis contexts. This paradigm is particularly important for the Crisis Adoption Problem: if deliberative training fails to transfer, it would provide strong evidence that the design-deployment mismatch requires training interventions, not merely interface improvements.

7.3.3 Measurement Considerations

Testing the Crisis Adoption Problem requires measures sensitive to both process (how users interact with AI) and outcome (detection performance and appropriateness of reliance).

Process measures should capture feature engagement patterns: explanation engagement frequency and duration (via eye-tracking or interaction logs), control feature utilization (adjustment frequency and parameter ranges), decision time distributions, and cognitive load indicators, including self-report (NASA-TLX Hart and Staveland [1988]) and physiological measures (heart rate variability, pupil diameter).

Outcome measures should capture decision quality: detection accuracy (hits, misses, false alarms, correct rejections), trust calibration indices Lee and See [2004] (correlation between user reliance and system reliability), compliance and reliance rates under varying system confidence levels, and post-task trust, workload, and situational awareness assessments.

Analytical approaches should include mixed-effects models to assess how

time pressure, cognitive load, and task conditions moderate the relationships between interface features and trust outcomes. Signal detection theory analyses can decompose performance into perceptual sensitivity (d') and response bias (β), testing whether stress conditions shift users toward liberal or conservative decision criteria. This decomposition is particularly relevant for Hypothesis 3a, which predicts that time pressure will increase reliance (liberal bias) without improving sensitivity.

7.3.4 Prerequisite Relationship to Phase 1

The Phase 1 study (Chapter 3) was a necessary prerequisite for this empirical program. Establishing the understanding–trust gap under controlled, low-pressure conditions provides a baseline against which crisis effects can be measured. Without knowing that the gap exists at $\Delta = 0.60$ under ideal conditions, Hypothesis 2 (that the gap widens under load) would have no reference point. Similarly, the Phase 1 finding that interactive features produce substantial engagement effects ($\eta_p^2 = 0.507$) while generating only modest trust improvements ($d = 0.33$) provides the specific empirical puzzle—why does agency not translate to confidence—that the Crisis Adoption Problem is designed to explain.

7.4 Extensions to the MICA Framework

Beyond crisis adoption testing, several technical and interaction design extensions can strengthen the MICA framework's operational relevance.

7.4.1 Adaptive Automation and Trust-Aware Interaction

Current MICA implementations provide static control options: users can adjust sensitivity and select regions of interest, but the system does not adapt its behavior based on inferred user state. Future work can incorporate adaptive automation that adjusts system autonomy based on user trust state, task difficulty, and environmental conditions Akash et al. [2020], Zhang et al. [2021]. Trust-aware systems might suppress low-confidence detections when users exhibit under-trust (to avoid overwhelming a skeptical operator) or highlight uncertain regions when over-trust is detected (to counteract complacency).

This extension connects directly to the Crisis Adoption Problem: adaptive automation that responds to operator stress indicators could dynamically shift the balance between explanatory and control features, foregrounding rapid-access controls under high load while providing richer explanations when cognitive resources are available.

7.4.2 Multimodal Camouflage Detection

Extending MICA to incorporate thermal imaging, motion cues, and temporal analysis would address the limitations of static single-image analysis and support the detection of dynamic camouflage strategies, including behavioral mimicry and motion camouflage Cott [1940]. Multimodal integration also introduces new interaction design challenges: operators must manage information from multiple sensor modalities, increasing cognitive load and making the crisis adoption considerations discussed above even more pressing.

7.4.3 Collaborative Multi-User Analysis

Current implementations assume single-user operation. Extending MICA to support team-based analysis, where multiple analysts contribute ROI selections, validate detections, and aggregate confidence assessments, would align with operational intelligence workflows and enable studies of collective trust calibration. Team-based analysis introduces additional research questions about how individual trust states interact with group decision-making processes and whether mixed-initiative mechanisms that support individual operators scale to team contexts.

7.4.4 Integration with Broader Surveillance Pipelines

Standalone CODS analysis represents only one component of operational surveillance systems. Integrating MICA with alerting systems, case management tools, and decision support frameworks would enable end-to-end evaluation of mixed-initiative interaction in realistic operational contexts and provide the ecological validity necessary to test the Crisis Adoption Problem at scale.

7.5 Synthetic Data Generation and Model Robustness

This section identifies four directions for extending the synthetic camouflage generation framework. Section 7.5.1 describes controlled evaluation via synthetic camouflage. Section 7.5.2 addresses adversarial robustness testing. Section 7.5.3 discusses transfer learning and domain adaptation. Section 7.5.4 proposes environment-adaptive generation via conditioned LoRA adapters.

7.5.1 Controlled Evaluation via Synthetic Camouflage

Chapter 6 describes the LoRA-enhanced diffusion infrastructure for synthetic camouflage generation. Future work can systematically vary camouflage strength, texture complexity, environmental clutter, and lighting conditions to stress-test detection models across a controlled parameter space. This approach addresses dataset limitations and supports fine-grained analysis of model failure modes that are difficult to isolate with naturally occurring camouflage examples.

7.5.2 Adversarial Robustness Testing

Camouflage represents a naturalistic form of adversarial perturbation. Extending synthetic generation to produce adversarial camouflage patterns, designed specifically to fool detection models while remaining perceptually plausible, can evaluate robustness and inform defensive training strategies. This direction connects to the broader adversarial ML literature while maintaining the ecological grounding that distinguishes camouflage detection from abstract adversarial attack research.

7.5.3 Transfer Learning and Domain Adaptation

Synthetic data can support few-shot learning and domain adaptation studies, testing whether models trained on synthetic camouflage generalize to novel environments and whether human-in-the-loop correction via MICA accelerates adaptation. This is particularly relevant for operational deployment, where target environments may not be known in advance and labeled training data are scarce.

7.5.4 Environment-Adaptive Generation via Conditioned LoRA

Preliminary observations of architectural specialization across Stable Diffusion variants in Chapter 6 suggest that different model configurations develop distinct generative biases suited to different aspects of camouflage synthesis. Future work can extend this finding by training environment-conditioned LoRA adapters, separate low-rank update sets optimized for specific deployment contexts such as woodland, desert, urban, and aquatic environments, and implementing a routing mechanism that selects or interpolates between adapters based on the perceptual characteristics of the input scene.

The COD10K dataset's categorical organization (amphibian, aquatic, flying, terrestrial, and human subcategories) provides a natural scaffold for initial environment-specific adapter training. A more advanced formulation would adopt a meta-learning approach, training adapters to rapidly specialize to novel environments given a small number of reference images. Coupling environment-adaptive generation with MICA's interactive detection framework would enable closed-loop evaluation: synthetically generated camouflage patterns could be assessed by human analysts within MICA, providing ecologically grounded feedback on perceptual effectiveness across environments and informing iterative adapter refinement.

7.6 Broader Implications for Human–AI Teaming

The understanding–trust gap and the Crisis Adoption Problem extend beyond camouflaged object detection to other high-stakes, high-uncertainty AI applications where the same structural tension, systems designed for deliberation, deployed under pressure, is likely to arise.

7.6.1 Medical Imaging

Radiologists operate under time pressure and high case volume, conditions that parallel the crisis adoption scenario. Testing whether interactive control mechanisms (adjustable sensitivity, region prioritization) improve diagnostic accuracy and appropriate reliance under clinical workload conditions could generalize MICA principles to healthcare AI. The understanding–trust dissociation observed in this work may be particularly relevant in radiology, where physicians may comprehend how a detection algorithm works without trusting its outputs for individual patient decisions.

7.6.2 Autonomous Vehicle Monitoring

Operators supervising semi-autonomous vehicles face similar challenges: understanding system capabilities does not guarantee appropriate takeover decisions under time-critical failure scenarios. Crisis-adapted interfaces that foreground rapid intervention controls may improve intervention timing and reduce mode confusion. The satisficing framework is directly applicable: operators satisfice on monitoring ("the system seems fine") until a crisis demands rapid, high-stakes

intervention.

7.6.3 Cybersecurity Threat Detection

Security analysts triage alerts under cognitive overload and time constraints. Mixed-initiative systems that support rapid tuning of detection thresholds and alert prioritization may reduce false alarm fatigue and improve threat detection. The Phase 1 finding that cybersecurity deception framing increased privacy awareness ($\eta_p^2 = 0.108$) without affecting trust suggests that contextual threat awareness interacts with but does not substitute for trust calibration mechanisms.

7.6.4 The Common Thread

Across these domains, the core tension persists: AI systems are designed and evaluated under ideal conditions (low pressure, single-task focus, ample time for deliberation), yet deployed in environments characterized by stress, multitasking, and operational urgency. The Crisis Adoption Problem provides a theoretical structure for addressing this design–deployment mismatch, and the empirical program proposed in Section 7.3 outlines one approach to testing whether the mismatch has the practical consequences that theory predicts.

7.7 Limitations and Open Questions

Several limitations constrain current findings and motivate continued investigation.

7.7.1 Ecological Validity

Laboratory studies, even with time-pressure manipulations, cannot fully replicate operational stress, uncertainty about consequences, and team coordination demands. Field studies and operational deployments are necessary to validate whether the Crisis Adoption Problem manifests as predicted in real-world settings. The proposed experimental paradigms (Section 7.3) represent intermediate steps between laboratory control and operational realism.

7.7.2 Expertise Effects

The Phase 1 study recruited university students without specialized training in camouflage detection. Professional analysts (military intelligence, wildlife biologists, surveillance operators) may exhibit different trust calibration and interaction patterns. Expertise may moderate crisis adoption behaviors: experienced operators may have developed stable reliance heuristics that are less susceptible to contextual stressors, or they may be more sensitive to situational cues that novices miss. Future work should examine expertise as both a moderator and a dependent variable (i.e., how crisis-adapted training develops expertise over time).

7.7.3 Longitudinal Trust Dynamics

Current studies capture trust at a single time point following a brief interaction session. Long-term use may reveal trust recalibration, habit formation, automation complacency, or skill transfer effects not observable in short-term studies. The temporal asymmetry mechanism identified in Chapter 3's Discussion, that trust requires accumulated performance observation, makes longitudinal designs

essential for understanding whether the understanding–trust gap eventually closes with extended experience.

7.7.4 Cultural and Organizational Context

Trust in automation is shaped by organizational culture, training protocols, error consequences, and institutional risk tolerance Hoffman et al. [2018]. Future work should examine how these factors moderate individual-level trust and interaction patterns. Military organizations, healthcare institutions, and technology companies may create different baseline conditions for crisis adoption.

7.7.5 Generalization Across AI Systems

The empirical findings reported in this dissertation are specific to camouflaged object detection systems. Whether the understanding–trust gap and the Crisis Adoption Problem generalize to other high-uncertainty AI tasks, natural language inference, anomaly detection, predictive maintenance, remains an open empirical question. The theoretical arguments (bounded rationality, NDM, cognitive load) suggest generalizability, but domain-specific factors (task structure, error consequences, user expertise) may substantially moderate the effects.

7.8 Summary

This dissertation establishes that explainability alone does not guarantee appropriate trust in AI-assisted camouflaged object detection. The persistent gap between understanding and trust, documented under ideal deliberative conditions with $N = 150$ participants, challenges the foundational assumption of XAI

research and motivates a fundamental reexamination of how AI trust is conceptualized.

The MICA framework (Chapter 4) provides one design response: shifting from passive explanation to active mixed-initiative collaboration. The Crisis Adoption Problem, proposed here as an emerging theoretical direction, offers a complementary analytical response: recognizing that the conditions under which AI systems are designed and evaluated may differ fundamentally from those under which adoption decisions are made. If operators satisfice under pressure, rely on recognition-primed heuristics rather than deliberative evaluation, and lack the cognitive resources to engage with detailed explanations, then the XAI paradigm's emphasis on transparency may be insufficient to address the mechanisms by which trust forms under operational pressure.

Testing this claim requires the empirical program outlined in Section 7.3, including controlled manipulations of time pressure, cognitive load, and operational demands, to determine whether the deliberation-crisis mismatch has the practical consequences predicted by theory. Extensions to MICA (adaptive automation, multimodal detection, team-based analysis) and synthetic generation infrastructure (environment-conditioned LoRA adapters, adversarial robustness testing) provide the technical foundations for this research.

The broader objective is to encourage a shift from systems optimized solely for comprehension to systems designed to support calibrated reliance under operational pressure, a direction necessary to realize the potential of AI-assisted decision-making in high-stakes domains.

Chapter 8

Toward Calibrated Human–AI Collaboration

This dissertation examined the intersection of computer vision, human perception, and trust in automation within the domain of camouflaged object detection and segmentation (CODS). By identifying and analyzing the persistent understanding–trust gap, this work provides empirical clarification and a structured framework for advancing calibrated human–AI collaboration.

8.1 From Transparency to Agency: Contributions and Field-Level Implications

The central argument of this dissertation is that transparency is necessary but insufficient for calibrated trust in AI-assisted detection systems. Explainability improves engagement and comprehension, but the gap between understanding and willingness to rely on AI outputs persists under controlled conditions and is likely to widen under operational pressure. Bridging that gap requires a structural

shift: from systems designed to explain themselves to systems designed to share control with their operators.

This reframing has implications beyond camouflage detection. Any high-stakes domain in which AI outputs must be acted upon under uncertainty—medical imaging, intelligence analysis, autonomous vehicle supervision—faces a structurally identical problem. The understanding–trust gap is not a CODS-specific artifact; it is a general feature of human–AI interaction in which performance ambiguity exceeds what transparency alone can resolve. The lesson this dissertation offers to those domains is concrete: interactive agency, operationalized as the ability to influence, correct, and redirect system outputs, is a more reliable mechanism for trust calibration than passive explanation delivery.

Four contributions substantiate this argument.

The **empirical contribution** provides the evidentiary foundation. A controlled study (N=150) demonstrated a persistent dissociation between user understanding of and willingness to trust an AI-assisted CODS system. Explainability features significantly improved perceived interactivity and engagement, but did not eliminate the gap between comprehension and reliance. This finding empirically establishes the understanding–trust gap and challenges the foundational assumption of XAI research that transparency produces calibrated trust. For other researchers, this result is a methodological caution: self-reported understanding and behavioral reliance are not interchangeable dependent variables, and conflating them obscures the mechanism that actually governs adoption.

The **theoretical and design contribution** provides the structural response. The Mixed-Initiative Camouflage Analysis (MICA) framework reconceptualizes the human–AI relationship as a bidirectional partnership in which shared control, rather than shared information, is the operative mechanism for calibrating trust.

By shifting from passive explanation delivery to structured user agency through region-of-interest selection and sensitivity adjustment, MICA provides a design vocabulary that other mixed-initiative systems can adopt. The trust calibration model, formalized in Equations 4.1 and 4.2, provides a theoretical framework for future empirical work examining how agency accelerates or stabilizes trust formation over time.

The **technical infrastructure contribution** extends the argument into generative modeling. A synthetic camouflage generation framework translating Gestalt perceptual principles and HSV color analysis into differentiable training objectives for Stable Diffusion establishes a controlled evaluation infrastructure for CODS research. The discovery of a systematic optimization–perception gap—in which standard evaluation metrics and human perceptual assessments diverge as a function of architecture—provides principled guidance for the design of perceptual loss functions in domain-specific generative modeling. The broader lesson: evaluation metric selection is not a neutral choice, and in perceptual domains, metric improvement and perceptual improvement can move in opposite directions.

The **computing infrastructure contribution** addresses the engineering precondition that makes mixed-initiative AI research tractable. The NEXUS bridging architecture resolves a recurring practical problem in human–AI interaction research: coordinating a responsive, event-driven interface with a computationally intensive inference pipeline without sacrificing either interactivity or model flexibility. By launching Python inference scripts as managed subprocesses from within the C# runtime and ingesting structured JSON results upon completion, NEXUS eliminates the embedding overhead and runtime conflicts that arise from tighter integration strategies. The MVVM architectural pattern fur-

ther decouples detection outputs, user editing actions, and session analytics from presentation concerns, allowing any component to be replaced independently. For researchers building interactive AI evaluation platforms, NEXUS demonstrates that architectural isolation between interface and inference pipeline is not merely a software engineering preference—it is what makes controlled experimentation with alternative models and interaction mechanisms feasible at all.

Together, these contributions clarify the structural relationship among transparency, agency, and trust in AI-assisted detection and demonstrate that interactive control is a more reliable mechanism for calibrated reliance than passive explainability alone. The empirical, theoretical, technical, and infrastructure dimensions of this work are designed to be separable: future researchers may build on any one layer without requiring the others, and the MICA prototype is explicitly architected to support that modularity.

8.2 Concluding Reflections

The trajectory of this dissertation reflects both empirical evolution and personal transformation. The research direction adapted in response to emerging findings and changing circumstances, including the passing of my advisor, Dr. Chris Weaver, in June 2025. His emphasis on the intellectual connections between human perception and artificial intelligence shaped the foundation of this work.

This dissertation argues that trust in AI systems is not achieved solely through transparency, but through structured interaction. The findings indicate that while explainability improves engagement and comprehension, meaningful agency plays a central role in shaping reliance decisions. The MICA prototype demonstrates the feasibility of embedding this agency into a CODS workflow, laying the

foundation for future empirical validation.

Importantly, the results reported here were obtained under deliberative laboratory conditions. These conditions represent a controlled baseline rather than operational reality. The next stage of this research program will investigate the Crisis Adoption Problem, examining how trust and reliance form under time pressure, cognitive load, and consequence-driven decision-making.

By distinguishing between deliberative evaluation and crisis-driven adoption, this work reframes the challenge of trustworthy AI as an interaction design problem under constraint. The broader objective is to develop systems that support calibrated human autonomy when it is most needed.

Bibliography

Kumar Akash, Griffon McMahon, Tahira Reid, and Neera Jain. Human trust-based feedback control: Dynamically varying automation transparency to optimize human-machine interactions. *IEEE Control Systems Magazine*, 40(6): 98–116, 2020.

James E. Allen, Curry I. Guinn, and Eric Horvtz. Mixed-initiative interaction. *IEEE Intelligent Systems and Their Applications*, 14(5):14–23, 1999.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

Maryam Ashoori and Justin D. Weisz. In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. *arXiv preprint arXiv:1912.02675*, 2019.

Albert Bandura. Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84, 1977.

Michael H. Bernstein, Michael K. Atalay, Elizabeth H. Dibble, Aaron W. P. Maxwell, Adib R. Karam, Saurabh Agarwal, Robert C. Ward, Terrance T. Healey, and Grayson L. Baird. Can incorrect artificial intelligence (AI) results impact radiologists, and if so, what can we do about it? A multi-reader pilot study of lung cancer detection with chest radiography. *European Radiology*, 33 (11):8263–8269, 2023. doi: 10.1007/s00330-023-09747-1.

George E. P. Box. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, pages 290–302, 1954.

Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

Heng-Da Cheng, X. H. Jiang, Ying Sun, and Jingli Wang. Color image segmentation: advances and prospects. *Pattern Recognition*, 34(12):2259–2281, 2001.

Erin K. Chiou and John D. Lee. Trusting automation: Designing for responsivity and resilience. *Human Factors*, 65(1):137–165, 2023.

Manolis Chiou, Faye McCabe, Markella Grigoriou, and Rustam Stolkin. Trust, shared understanding and locus of control in mixed-initiative robotic systems. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 684–691. IEEE, 2021.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2 edition, 1988.

Hugh B. Cott. *Adaptive Coloration in Animals*. Methuen & Co., Ltd., 1940.

Edward L. Deci and Richard M. Ryan. The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11 (4):227–268, 2000.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6):697–718, 2003.

Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

Mica R. Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1):32–64, 1995.

Yrjö Engeström. Activity theory and learning at work. In *Tätigkeit-Aneignung-Bildung: Positionierungen zwischen Virtualität und Gegenständlichkeit*, pages 67–96. Springer, 2014.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2777–2787, 2020a.

Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. PraNet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 263–273. Springer, 2020b.

Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4):1149–1160, 2009.

Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, 52:139–183, 1988.

Lingfei He, Siyu Wang, Cheng Chen, Yaping Wang, Qingcheng Fan, Congying Chu, Lingzhong Fan, and Junhai Xu. Network occlusion sensitivity analysis identifies regional contributions to brain age prediction. *Human Brain Mapping*, 46(8):e70239, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf`.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Kevin Anthony Hoff and Masooda Bashir. Trust in automation: An integrative review of the determinants and dynamics of trust. *Human Factors*, 57(3): 407–450, 2015.

Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Michael L. Littman. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

Debra Hogue, Shane Connelly, and Ghulam Jilani Quadri. Multi-Architecture Evaluation of Stable Diffusion with HSV–Gestalt Perceptual Loss. *IEEE Computer Graphics and Applications*, 2026a. Submitted for publication.

Debra Hogue, Shane Connelly, and Ghulam Jilani Quadri. Interactive Features and Trust in AI-Assisted Camouflaged Object Detection: Evidence for the Understanding-Trust Gap. *ACM Transactions on Computer-Human Interaction*, 2026b. Submitted for publication.

Erik Hollnagel and David D. Woods. *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering*. CRC Press, 2005.

Eric Horvitz. Uncertainty, action, and interaction: In pursuit of mixed-initiative computing. *IEEE Intelligent Systems*, 14(5):17–20, 1999.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022.

Edwin Hutchins. *Cognition in the Wild*. MIT Press, 1995.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 624–635, 2021.

Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Gary Klein, Brian Moon, and Robert R. Hoffman. Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems*, 21(4):70–73, 2006.

Gary A. Klein. Recognition-primed decisions. *Advances in Man-Machine Systems Research*, 5(1):47–92, 1989.

Spencer C. Kohn, Ewart J. De Visser, Eva Wiese, Yi-Ching Lee, and Tyler H. Shaw. Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, 12:604977, 2021.

Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding*, 184:45–56, 2019.

John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.

Mirtha Lucas, Miguel Lerma, Jacob Furst, and Daniela Raicu. Visual explanations from deep networks via Riemann-Stieltjes integrated gradient-based localization. *arXiv preprint arXiv:2205.10900*, 2022.

Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously Localize, Segment and Rank the Camouflaged Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11591–11601, 2021.

Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Nick Barnes, and Deng-Ping Fan. Towards deeper understanding of camouflaged object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

Zdravko Marinov, Paul F. Jäger, Jan Egger, Jens Kleesiek, and Rainer Stiefelhagen. Deep interactive segmentation of medical images: A systematic review and taxonomy. *arXiv preprint arXiv:2311.13964*, 2024.

Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3044–3053, 2021.

Bonnie M. Muir. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11): 1905–1922, 1994.

James Norton and Kristie Miller. A psychologistic theory of metaphysical explanation. *Synthese*, 196(7):2777–2802, 2019.

Linda Onnasch, Christopher D. Wickens, Huizhong Li, and Dietrich Manzey. Human performance consequences of stages and levels of automation: A meta-analysis. *Human Factors*, 56(3):476–488, 2014.

Raja Parasuraman and Dietrich Manzey. Complacency and bias in human-automation interaction. *Human Factors*, 52(3):381–410, 2010.

Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997.

Peter Pirolli and Stuart Card. The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. In *Proceedings of International Conference on Intelligence Analysis*, volume 5, pages 2–4. McLean, VA, USA, 2005.

Natasha Price, Simon Green, Jolyon Troscianko, Tom Tregenza, and Martin Stevens. Background matching and disruptive coloration as habitat-specific strategies for camouflage. *Scientific Reports*, 9(1), 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. Proceedings of Machine Learning Research (PMLR), 18–24 Jul 2021.

Mohammad H. Rezazade Mehrizi, Ferdinand Mol, Marcel Peter, Erik Ranschaert, Daniel Pinto Dos Santos, Ramin Shahidi, Mansoor Fatehi, and Thomas Dratsch. The impact of AI suggestions on radiologists' decisions: a pilot study of explainability and attitudinal priming interventions in mammography examination. *Scientific Reports*, 13:9230, 2023. doi: 10.1038/s41598-023-36435-3.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

Azriel Rosenfeld. Computer vision: Basic principles. *Proceedings of the IEEE*, 76(8):863–868, 1988. doi: 10.1109/5.5961.

Julian B. Rotter. *Generalized expectancies for internal versus external control of reinforcement*, volume 80. Psychological Monographs, 1966.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. *Advances in Neural Information Processing Systems*, 29, 2016. doi: 10.48550/arxiv.1606.03498.

Emanuel Schmider, Matthias Ziegler, Erik Danay, Luzi Beyer, and Markus Bühner. Is it really robust? *Methodology*, 2010.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

Herbert A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63, 1956.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Martin Stevens and Sami Merilaita. Animal camouflage: current issues and new perspectives. *Philosophical Transactions of the Royal Society B*, 364, 2009.

John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988.

John Sweller, Paul Ayres, and Slava Kalyuga. *Cognitive Load Theory*. Springer Science & Business Media, 2011.

Mingxing Tan, Ruoming Pang, and Quoc V. Le. EfficientDet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10781–10790, 2020.

Tom Troscianko, Christopher P. Benton, P. George Lovell, David J. Tolhurst, and Zygmunt Pizlo. Camouflage and visual perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1516):449–461, 2009.

Johan Wagemans, James H. Elder, Michael Kubovy, Stephen E. Palmer, Mary A. Peterson, Manish Singh, and Rüdiger von der Heydt. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychological Review*, 119(4):1141–1178, 2012. doi: 10.1037/a0029333.

Shuo Wang, Chenhao Li, Rang Wang, Ziyuan Liu, Meng Wang, Hui Tan, Yida Wu, Xia Liu, Hao Sun, Rui Yang, Xinfeng Zhang, Renzhi Zhou, Tingli Hou, Rong Jiang, Jiangyuan Li, Li Liu, Dinggang Shen, and Manning Liu. Annotation-efficient deep learning for automatic medical image segmentation. *Nature Communications*, 12, 2021.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Pak Chung Wong, June Abbas, Chaomei Chen, Christopher Collins, Danyel Fisher, Chi-Wing Fu, Samuel J. Huskey, Andreas Kerren, Alan M. MacEachren, Mike Potel, Anthony C. Robinson, and Michael Sedlmair. Gone Too Soon, Remembered Always: Chris Weaver and the Power of Visual Thinking. *IEEE Computer Graphics and Applications*, 45(05):8–11, 2025. doi: 10.1109/MCG.2025.3598531.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

Ruibo Zhang, Nathan J. McNeese, Guo Freeman, and Geoffrey Musick. The effect of adaptive explanations on human–AI teaming. *International Journal of Human–Computer Interaction*, 37(17):1630–1643, 2021.

Jiayuan Zhu and Junde Wu. MedUHIP: Towards human-in-the-loop medical segmentation. *arXiv preprint arXiv:2408.01620*, 2024.

Caroline E. Zsambok. Naturalistic decision making: Where are we now? In Caroline E. Zsambok and Gary Klein, editors, *Naturalistic Decision Making*, pages 3–16. Lawrence Erlbaum Associates, 1997.

Appendix A

Definition of Terms

The terms below are provided to help the reader understand the context of each term in this study.

Camouflaged Object Detection and Segmentation (CODS): The computational task of enabling machines to detect, locate, and segment one or more concealed objects within images or video, where objects blend with their surroundings through mechanisms such as background matching and disruptive coloration.

eXplainable Artificial Intelligence (XAI): An approach to enhance transparency, trust, and accountability in AI systems by providing interpretable explanations of model decision-making processes [Arrieta et al., 2020].

Find and Acquire Camouflage Explainability (FACE): The baseline camouflage detection system used in this research provides core detection capabilities with basic confidence scores but without enhanced explainability features.

Mixed Understanding and Recognition for Detection of Camouflage (MURDOC): The explainable camouflage detection system used in this research incorporates Grad-CAM visualizations, confidence displays, detection rationale, and interactive image adjustment controls.

Gradient-weighted Class Activation Mapping (Grad-CAM): A visualization technique that produces heatmaps highlighting which regions of an input image most influenced a neural network's predictions [Selvaraju et al., 2017].

Interpretability: The degree to which a human can understand and explain how a machine learning model makes predictions or decisions.

Low-Rank Adaptation (LoRA): A parameter-efficient fine-tuning technique for large models that introduces trainable low-rank matrices into model layers, enabling adaptation to specific tasks while minimizing computational requirements [Hu et al., 2022]. In this dissertation, LoRA is used to fine-tune diffusion models for synthetic camouflage generation.

Mixed-Initiative Camouflage Analysis (MICA): The framework proposed in this dissertation enables structured human-AI collaboration through interactive control mechanisms, including region-of-interest selection, sensitivity adjustment, feedback integration, and adaptive explanations.

Mixed-Initiative Interaction: A collaborative paradigm where both humans and AI systems can proactively contribute to task execution, with dynamic task allocation based on the relative strengths of each agent [Horvitz, 1999].

Region-of-Interest (ROI): A user-specified area within an image that guides the AI system's detection focus, enabling human expertise to direct computational analysis toward tactically or contextually significant regions.

Satisficing: A decision-making strategy characterized by accepting adequate solutions rather than seeking optimal ones [Simon, 1956]. Under operational constraints—time pressure, cognitive load, incomplete information—decision-makers satisfice by applying threshold criteria ("is this good enough?") rather than exhaustive evaluation. This concept is central to naturalistic decision-making research and explains why operators may adopt heuristic reliance strategies with AI systems under crisis conditions.

Trust (in Automation): The attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability [Kohn et al., 2021]. In this dissertation, trust is operationalized using the TOAST (Trust of Automated Systems Test) scale, which measures perceived reliability, understanding, and willingness to rely on system outputs.

Understanding: In the context of this research, the user's comprehension of the AI system's capabilities, limitations, and decision-making processes is measured through self-report assessments of system comprehension.

Understanding-Trust Gap: The empirical finding from this research shows a persistent dissociation between high user understanding of system capabilities (M=4.25/5.0) and moderate trust in system outputs (M=3.67/5.0), representing approximately 0.6 points on a 5-point scale.

Usability: The extent to which specified users can use a system to achieve specified goals with effectiveness, efficiency, and satisfaction. In this research, usability encompasses ease of use, interface quality, and overall user satisfaction with the detection system.

Appendix B

CITI Training Documentation

B.1 Certificates

B.1.1 Human Research

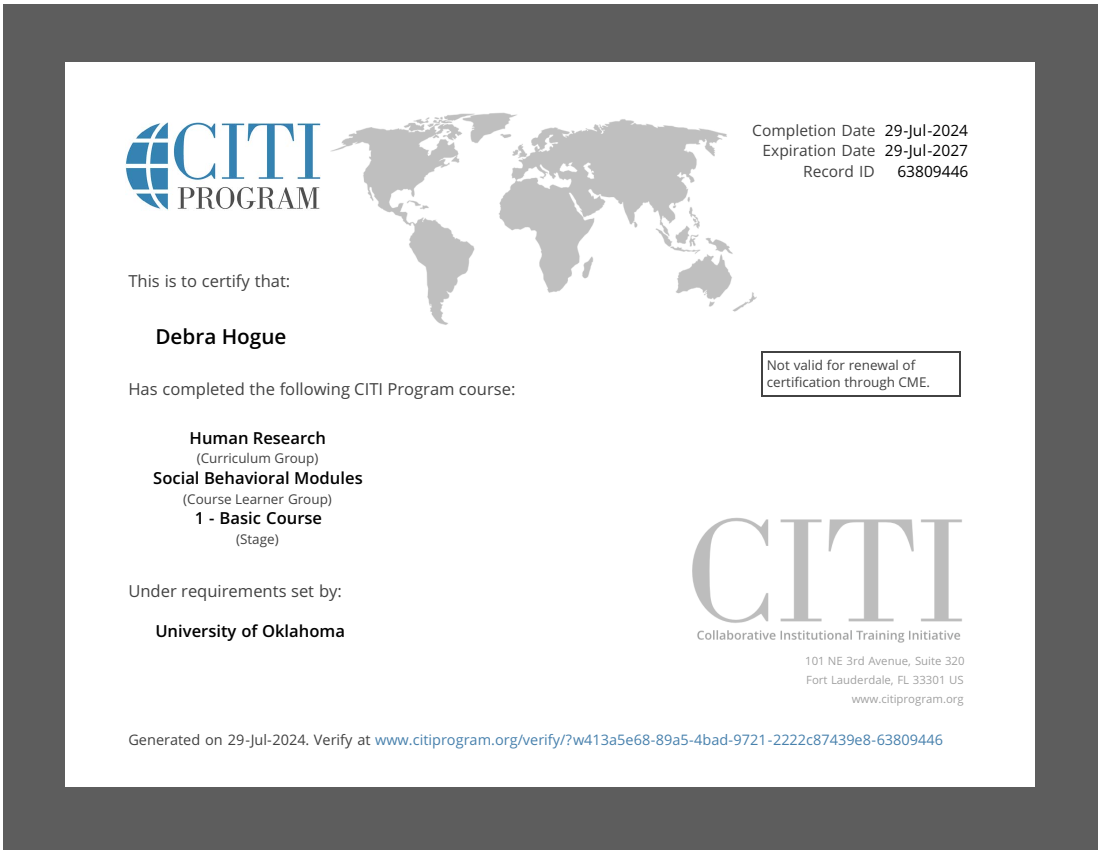

Figure B.1: CITI Training Certificate for Human Research

B.1.2 Responsible Conduct of Research

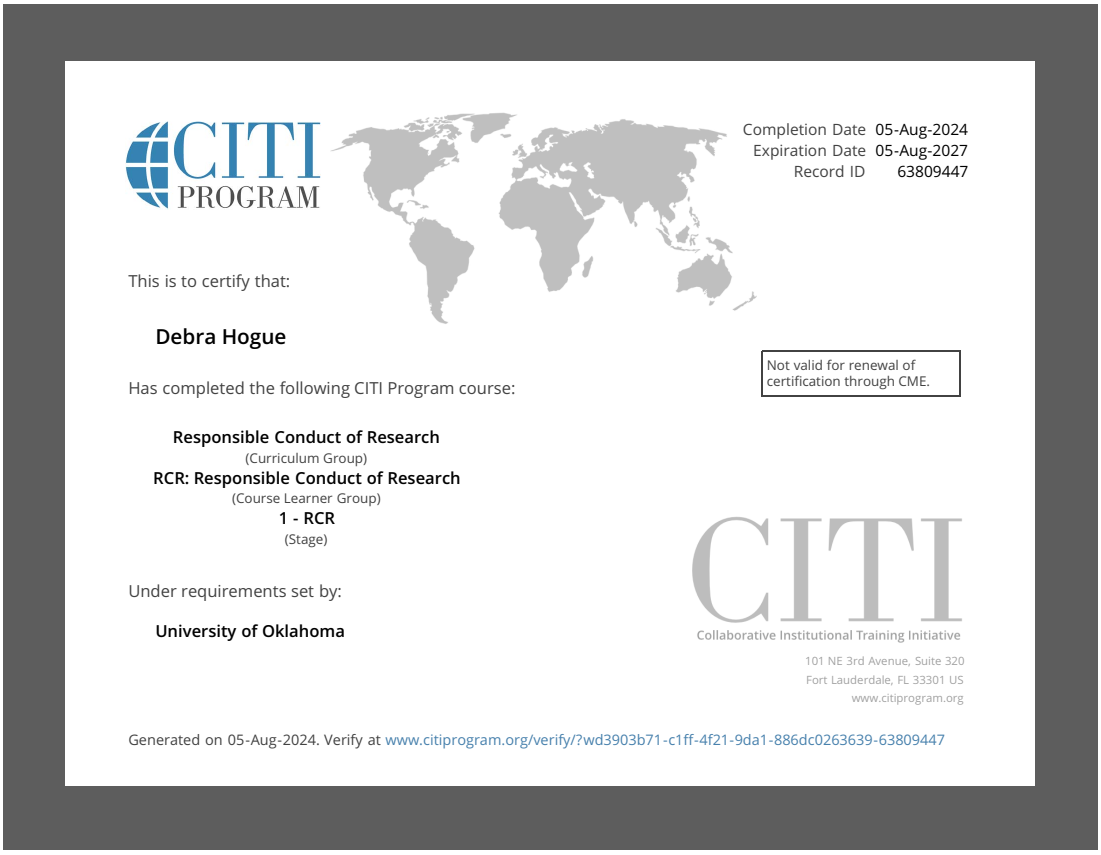

Figure B.2: CITI Training Certificate for Responsible Conduct of Research

B.2 CITI Completion Reports

B.2.1 Human Research

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 1 OF 2
COURSEWORK REQUIREMENTS*

* Scores on this Requirements Report (Part 1) reflect quiz completions at the time all requirements for the course were met. The Transcript Report (Part 2) lists more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** Debra Hogue (ID: 13463764)
- **Institution Affiliation:** University of Oklahoma (ID: 577)
- **Institution Email:** debra.l.hogue-1@ou.edu
- **Institution Unit:** Computer Science
- **Phone:** 4056154496

- **Curriculum Group:** Human Research
- **Course Learner Group:** Social Behavioral Modules
- **Stage:** Stage 1 - Basic Course

- **Record ID:** 63809446
- **Completion Date:** 29-Jul-2024
- **Expiration Date:** 29-Jul-2027
- **Minimum Passing:** 70
- **Reported Score*:** 88

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
Unanticipated Problems and Reporting Requirements in Social and Behavioral Research (ID: 14928)	23-Jul-2024	4/5 (80%)
Belmont Report and Its Principles (ID: 1127)	23-Jul-2024	3/3 (100%)
Conflicts of Interest in Human Subjects Research (ID: 17464)	23-Jul-2024	4/5 (80%)
Cultural Competence in Research (ID: 15166)	23-Jul-2024	4/5 (80%)
History and Ethical Principles - SBE (ID: 490)	24-Jul-2024	4/5 (80%)
Defining Research with Human Subjects - SBE (ID: 491)	24-Jul-2024	4/5 (80%)
The Federal Regulations - SBE (ID: 502)	24-Jul-2024	5/5 (100%)
Assessing Risk - SBE (ID: 503)	24-Jul-2024	5/5 (100%)
Informed Consent - SBE (ID: 504)	24-Jul-2024	4/5 (80%)
Privacy and Confidentiality - SBE (ID: 505)	24-Jul-2024	4/5 (80%)
Research with Prisoners - SBE (ID: 506)	24-Jul-2024	4/5 (80%)
Research with Children - SBE (ID: 507)	24-Jul-2024	5/5 (100%)
Research in Public Elementary and Secondary Schools - SBE (ID: 508)	24-Jul-2024	5/5 (100%)
International Research - SBE (ID: 509)	24-Jul-2024	5/5 (100%)
Internet-Based Research - SBE (ID: 510)	25-Jul-2024	4/5 (80%)
Research and HIPAA Privacy Protections (ID: 14)	25-Jul-2024	5/5 (100%)
Vulnerable Subjects - Research Involving Workers/Employees (ID: 483)	25-Jul-2024	3/4 (75%)
Hot Topics (ID: 487)	29-Jul-2024	No Quiz
University of Oklahoma (ID: 882)	29-Jul-2024	No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

This document was generated on 29-Jul-2024. Verify at:
www.citiprogram.org/verify/?k9a745c73-aa13-4c62-94ef-2d7be6c9cf89-63809446

Collaborative Institutional Training Initiative (CITI Program)
101 NE 3rd Avenue
Suite 320
Fort Lauderdale, FL 33301 US

Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

Figure B.3: CITI Training Completion Report for Human Research

B.2.2 Responsible Conduct of Research

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)
COMPLETION REPORT - PART 1 OF 2
COURSEWORK REQUIREMENTS*

* Scores on this Requirements Report (Part 1) reflect quiz completions at the time all requirements for the course were met. The Transcript Report (Part 2) lists more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** Debra Hogue (ID: 13463764)
- **Institution Affiliation:** University of Oklahoma (ID: 577)
- **Institution Email:** debra.l.hogue-1@ou.edu
- **Institution Unit:** Computer Science
- **Phone:** 4056154496

- **Curriculum Group:** Responsible Conduct of Research
- **Course Learner Group:** RCR: Responsible Conduct of Research
- **Stage:** Stage 1 - RCR

- **Record ID:** 63809447
- **Completion Date:** 05-Aug-2024
- **Expiration Date:** 05-Aug-2027
- **Minimum Passing:** 70
- **Reported Score*:** 97

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
Introduction to RCR (RCR-Basic) (ID: 17009)	29-Jul-2024	3/3 (100%)
Conflicts of Interest and Commitment (RCR-Basic) (ID: 16599)	29-Jul-2024	5/5 (100%)
Data Management (RCR-Basic) (ID: 20896)	29-Jul-2024	5/5 (100%)
Collaborative Research (RCR-Basic) (ID: 16598)	02-Aug-2024	5/5 (100%)
Mentoring and Healthy Research Environments (RCR-Basic) (ID: 20983)	02-Aug-2024	5/5 (100%)
Research Misconduct (RCR-Basic) (ID: 16604)	05-Aug-2024	5/5 (100%)
Financial Responsibility (ID: 16601)	05-Aug-2024	4/5 (80%)

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

This document was generated on 05-Aug-2024. Verify at:
www.citiprogram.org/verify/?k4901adb8-03b4-4a5a-9b4c-42606c2979c8-63809447

Collaborative Institutional Training Initiative (CITI Program)
101 NE 3rd Avenue
Suite 320
Fort Lauderdale, FL 33301 US

Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

Figure B.4: CITI Training Completion Report for Responsible Conduct of Research

Appendix C

Institutional Review Board (IRB) Documentation

The following documents represent the formal ethical oversight and approvals granted by the University of Oklahoma Institutional Review Board (IRB) for Protocol #17945, "Evaluating Trust and Explainability in Camouflaged Object Detection AI Systems." These documents establish the research life cycle from initial exempt determination through necessary modifications for personnel and instrument updates.

C.1 Approval of Study Modification — Personnel Change

Following the sudden passing of the original faculty advisor, Dr. Chris Weaver, in June 2025, this modification officially updated the study personnel to include Dr. Dean Hougen and Dr. Ghulam Jilani Quadri as guiding faculty C.1.



Institutional Review Board for the Protection of Human Subjects

Approval of Study Modification – Expedited Review – AP0

Date: September 15, 2025 **IRB#:** 17945

Principal Investigator: Debra L Hogue **Reference No: 779881**

Study Title: Evaluating Trust and Explainability in Camouflaged Object Detection AI Systems

Approval Date: 09/15/2025

Modification Description:
Update faculty sponsor from Christopher Weaver to Ghulam Quadri.

The review and approval of this submission is based on the determination that the study, as amended, will continue to be conducted in a manner consistent with the requirements of 45 CFR 46.

To view the approved documents for this submission, open this study from the My Studies option, go to Submission History, go to Completed Submissions tab and then click the Details icon.

If the consent form(s) were revised as a part of this modification, discontinue use of all previous versions of the consent form.

If you have questions about this notification or using iRIS, contact the HRPP office at (405) 325-8110 or irb@ou.edu. The HRPP Administrator assigned for this submission: Nicole A Cunningham.

Cordially,



Marshall Cheney, Ph.D.
Vice Chair, Institutional Review Board

Eric Pennello, Ph.D.
Vice Chair, Institutional Review Board

Figure C.1: IRB Approval of Personnel Change (June 2025)

C.2 Approval of Study Modification — Instrument Update

This modification, approved on February 19, 2025, authorized updates to the Psychology (PSY) Consent Form and the Pre-Evaluation Survey used in the final data collection phase C.2.

C.3 Approval of Initial Submission

The study was initially determined to be exempt from IRB review under Categories 1, 2, and 3 on December 7, 2024 C.3.

The UNIVERSITY of OKLAHOMA

Institutional Review Board for the Protection of Human Subjects

Approval of Study Modification – Expedited Review – AP0

Date: February 20, 2025 **IRB#:** 17945

Principal Investigator: Debra L Hogue **Reference No: 773139**

Study Title: Evaluating Trust and Explainability in Camouflaged Object Detection AI Systems

Approval Date: 02/19/2025

Modification Description: Mod PSY Consent & Pre-Eval Svy

The review and approval of this submission is based on the determination that the study, as amended, will continue to be conducted in a manner consistent with the requirements of 45 CFR 46.

To view the approved documents for this submission, open this study from the My Studies option, go to Submission History, go to Completed Submissions tab and then click the Details icon.

If the consent form(s) were revised as a part of this modification, discontinue use of all previous versions of the consent form.

If you have questions about this notification or using iRIS, contact the HRPP office at (405) 325-8110 or irb@ou.edu. The HRPP Administrator assigned for this submission: Kat L Braswell.

Cordially,

Aimee Franklin, Ph.D.
Chair, Institutional Review Board

Figure C.2: IRB Approval of Study Modification (February 20, 2025)



Institutional Review Board for the Protection of Human Subjects

Approval of Initial Submission – Exempt from IRB Review – AP01

Date: December 08, 2024 **IRB#:** 17945

Principal Investigator: Debra L Hogue

Approval Date: 12/07/2024

Exempt Category: 1, 2, & 3

Study Title: Evaluating Trust and Explainability in Camouflaged Object Detection AI Systems

On behalf of the Institutional Review Board (IRB), I have reviewed the above-referenced research study and determined that it meets the criteria for exemption from IRB review. To view the documents approved for this submission, open this study from the *My Studies* option, go to *Submission History*, go to *Completed Submissions* tab and then click the *Details* icon.

As principal investigator of this research study, you are responsible to:

- Conduct the research study in a manner consistent with the requirements of the IRB and federal regulations 45 CFR 46.
- Request approval from the IRB prior to implementing any/all modifications as changes could affect the exempt status determination.
- Maintain accurate and complete study records for evaluation by the HRPP Quality Improvement Program and, if applicable, inspection by regulatory agencies and/or the study sponsor.
- Notify the IRB at the completion of the project.

If you have questions about this notification or using iRIS, contact the IRB @ 405-325-8110 or irb@ou.edu.

Cordially,

Aimee Franklin, Ph.D.
Chair, Institutional Review Board

Figure C.3: IRB Approval of Initial Submission (December 08, 2024)

Appendix D

Survey Instruments

This appendix reproduces the pre-interaction and post-interaction survey instruments administered to participants in the Phase 1 empirical study (Chapter 3). Both surveys were delivered via Qualtrics under IRB protocol #17945. Separate consent forms were used for non-expert participants (Psychology/Communication students recruited through SONA) and expert participants (Computer Science/-Software Engineering students and professionals recruited through departmental outreach); the survey items themselves were identical across populations.

D.1 Pre-Interaction Survey

D.1.1 Informed Consent

Participants reviewed and acknowledged the following elements before proceeding:

- Confirmation of age (18 years or older)
- Consent to participate in the study

- Eligibility confirmation:
  - *Non-expert version*: Current enrollment in Psychology or Communication courses with SONA eligibility
  - *Expert version*: Educational or professional background in Computer Science, Software Engineering, AI/ML, or related fields

D.1.2 Demographics

1. **What is your age range?**

   Response options: 18–24 / 25–34 / 35–44 / 45–54 / 55–64 / 65+

2. **What gender do you identify as?**

   Response options: Male / Female / Non-binary or third gender / Prefer not to say

3. **What is your education level?**

   Response options: High school or equivalent / Some college / Bachelor's degree / Master's degree / Doctoral degree / Professional degree

4. **What is your employment status?**

   Response options: Full-time / Part-time / Self-employed / Unemployed / Student / Retired / Other

D.1.3 Technology Experience and Habits

All items used 5-point scales as indicated.

1. **How often do you rely on technology to assist with decision-making in your daily life?**

Scale: Never / Rarely / Sometimes / Often / Very frequently

2. **When faced with a complex task, how likely are you to seek advice from experts or specialized tools?**

   Scale: Extremely unlikely / Somewhat unlikely / Neither likely nor unlikely / Somewhat likely / Extremely likely

3. **How comfortable are you with using new technologies or applications?**

   Scale: Extremely uncomfortable / Somewhat uncomfortable / Neither comfortable nor uncomfortable / Somewhat comfortable / Extremely comfortable

4. **I generally trust new technologies to perform their intended functions.**

   Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

5. **I believe that automated systems are usually designed with the user's best interests in mind.**

   Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

6. **How would you rate your ability to work collaboratively with automated systems?**

   Scale: Poor / Fair / Good / Very Good / Excellent

7. **I am comfortable relying on automated systems for important tasks.**

Scale: Extremely uncomfortable / Somewhat uncomfortable / Neither comfortable nor uncomfortable / Somewhat comfortable / Extremely comfortable

8. **I feel that most automated systems perform consistently.**
   Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

9. **I typically understand the capabilities and limitations of automated systems I use.**
   Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

10. **How would you rate your overall openness to adopting new technological solutions in your work?**
    Scale: Not open at all / Slightly open / Moderately open / Quite open / Very open

D.2 Post-Interaction Survey

The post-interaction survey consisted of 20 items measuring perceptions of Trust, Understanding, Usability, Interactivity, and Cybersecurity. Table D.1 shows how items map to composite scores. All items used 5-point scales.

D.2.1 Trust Items

Q1. How confident are you in the accuracy of the application's camouflage detection results?

Table D.1: Post-Interaction Survey Item–Composite Mapping

Composite Score	**Constituent Items**
Trust ($\alpha = .82$)	Q1, Q3, Q8, Q13
Understanding ($\alpha = .79$)	Q2, Q4, Q5, Q7
Usability ($\alpha = .85$)	Q9, Q10, Q11, Q12
Interactivity ($\alpha = .81$)	Q15, Q16, Q17
Items not in composites: Q6, Q14, Q18, Q19, Q20	

Scale: Not confident at all / Slightly confident / Moderately confident / Confident / Very confident

Q3. How likely are you to rely on this application's outputs in real-world scenarios?

Scale: Extremely unlikely / Somewhat unlikely / Neither likely nor unlikely / Somewhat likely / Extremely likely

Q8. How would you rate the overall reliability of the application?

Scale: Very unreliable / Unreliable / Neither reliable nor unreliable / Reliable / Very reliable

Q13. I feel comfortable relying on the information that the application provided.

Scale: Extremely uncomfortable / Somewhat uncomfortable / Neither comfortable nor uncomfortable / Somewhat comfortable / Extremely comfortable

D.2.2 Understanding Items

Q2. I understand what the camouflage detection application should do.

Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree

/ Somewhat agree / Strongly agree

Q4. I understand the application's limitations.

Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

Q5. I understand the application's capabilities.

Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

Q7. I understand how the application executes its tasks.

Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

D.2.3 Usability Items

Q9. The application performs the way that it should.

Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

Q10. How likely are you to recommend this application to colleagues for camouflage detection tasks?

Scale: Extremely unlikely / Somewhat unlikely / Neither likely nor unlikely / Somewhat likely / Extremely likely

Q11. The application helps me achieve my goals in camouflage detection.

Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

Q12. The application performs consistently.

Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

D.2.4 Interactivity Items

Q15. The application allowed me to adjust settings to refine its output.

Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

Q16. I felt engaged with the application while using it.

Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

Q17. The application provided real-time feedback as I used it.

Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

D.2.5 Additional Items

The following items were analyzed individually but were not included in any composite score.

Q6. How comfortable would you feel using this application without human oversight?

Scale: Extremely uncomfortable / Somewhat uncomfortable / Neither comfortable nor uncomfortable / Somewhat comfortable / Extremely comfortable

Q14. I am rarely surprised by how the application responds.

Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

Q18. I feel that the application protects and handles data responsibly.

Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

Q19. The application seems to prioritize user privacy.

Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

Q20. I am confident that unauthorized parties cannot easily manipulate the application's output.

Scale: Strongly disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Strongly agree

Appendix E

Statistical Analysis Code

This appendix provides the analysis scripts used to produce the ANOVA results reported in Chapter 3 (Tables 3.3–3.5) and the descriptive statistics in Table 3.2. Both scripts implement identical two-way factorial ANOVAs (Application × CD) with Type III sums of squares and partial η_p^2 effect sizes. The Python script served as the primary analysis tool; the SPSS syntax was used for independent verification of all results.

E.1 Python Analysis Script

The script below reads the Qualtrics CSV export, computes composite scores (Trust, Understanding, Usability, Interactivity), runs two-way ANOVAs for all 20 individual items and the four composites, and exports the results tables. It requires Python 3.10+ with `pandas`, `numpy`, `scipy`, and `statsmodels`.

```
"""
Two-Way ANOVA Analysis: Interactive Features and Trust in AI-Assisted
Camouflaged Object Detection
```

```
Supplemental Analysis Script for:
  Hogue, D., Connelly, S., & Quadri, G. J. (2026). Interactive Features
  and Trust in AI-Assisted Camouflaged Object Detection: Evidence for the
  Understanding-Trust Gap. IEEE Access.

Design: 2 x 2 between-subjects factorial
  Factor 1 - Application Type: MURDOC (interactive explainable) vs.
             FACE (baseline)
  Factor 2 - Cybersecurity Deception (CD): Present vs. Absent
  Sample: N = 150 (37-38 per cell)

Dependent Variables:
  20 individual survey items (Q1-Q20) on 5-point Likert scales
  4 composite scores: Trust, Understanding, Usability, Interactivity

Statistical Method: Type III Sum of Squares ANOVA with partial eta-squared
  Equivalent to SPSS UNIANOVA with /METHOD=SSTYPE(3)

Requirements: Python 3.10+, pandas, numpy, scipy, statsmodels
  pip install pandas numpy scipy statsmodels

Usage: python anova_analysis.py <path_to_study_results.csv>

Author: Debra Hogue, University of Oklahoma
"""

import sys
import pandas as pd
import numpy as np
from scipy import stats
from statsmodels.formula.api import ols
```

```
from statsmodels.stats.anova import anova_lm
import warnings
warnings.filterwarnings('ignore')



# ======================================================================
# CONFIGURATION
# ======================================================================

ITEM_LABELS = {
    'Q1':  'Confidence in accuracy',
    'Q2':  'Understanding of application purpose',
    'Q3':  'Likelihood to rely on application',
    'Q4':  'Understanding of limitations',
    'Q5':  'Understanding of capabilities',
    'Q6':  'Comfort using without oversight',
    'Q7':  'Understanding of execution',
    'Q8':  'Overall reliability rating',
    'Q9':  'Performance as expected',
    'Q10': 'Likelihood to recommend',
    'Q11': 'Achievement of goals',
    'Q12': 'Consistency of performance',
    'Q13': 'Comfort relying on information',
    'Q14': 'Rarely surprised by responses',
    'Q15': 'Ability to adjust settings',
    'Q16': 'Engagement with application',
    'Q17': 'Real-time feedback',
    'Q18': 'Data responsibility',
    'Q19': 'User privacy priority',
    'Q20': 'Protection from manipulation',
}
```

```
COMPOSITES = {
    'Trust':          ['Q1', 'Q3', 'Q8', 'Q13'],
    'Understanding':  ['Q2', 'Q4', 'Q5', 'Q7'],
    'Usability':      ['Q9', 'Q10', 'Q11', 'Q12'],
    'Interactivity':  ['Q15', 'Q16', 'Q17'],
}

ETA_SQ_THRESHOLDS = {'small': 0.01, 'medium': 0.06, 'large': 0.14}



# ======================================================================
# DATA LOADING AND PREPARATION
# ======================================================================

def load_data(filepath):
    """Load and prepare study data from CSV export."""
    for enc in ['utf-8', 'latin-1', 'cp1252']:
        try:
            df = pd.read_csv(filepath, skiprows=[0], encoding=enc)
            break
        except (UnicodeDecodeError, pd.errors.ParserError):
            continue
    else:
        raise RuntimeError(f"Could not read {filepath}")

    df = df[df['Finished'] == 1].copy()
    df['App'] = df['App'].astype(str).str.strip()
    df['CD'] = df['CD?'].astype(str).str.strip()

    item_cols = list(ITEM_LABELS.keys())
```

```
    for col in item_cols:
        df[col] = pd.to_numeric(df[col], errors='coerce')
    df = df.dropna(subset=item_cols)

    for name, items in COMPOSITES.items():
        df[name] = df[items].mean(axis=1)

    return df


# ======================================================================
# STATISTICAL ANALYSIS
# ======================================================================

def run_anova(df, dv, label=""):
    """Run two-way ANOVA (Type III SS) for a single DV."""
    formula = f'{dv} ~ C(App) + C(CD) + C(App):C(CD)'
    model = ols(formula, data=df).fit()
    table = anova_lm(model, typ=3)

    results = []
    for source in ['C(App)', 'C(CD)', 'C(App):C(CD)']:
        ss_effect = table.loc[source, 'sum_sq']
        ss_resid = table.loc['Residual', 'sum_sq']
        df_effect = table.loc[source, 'df']
        ms_effect = ss_effect / df_effect
        f_val = table.loc[source, 'F']
        p_val = table.loc[source, 'PR(>F)']
        eta_sq_partial = ss_effect / (ss_effect + ss_resid)

        if p_val < 0.001:   sig = '***'
```

```
        elif p_val < 0.01:  sig = '**'
        elif p_val < 0.05:  sig = '*'
        else:               sig = 'ns'

        results.append({
            'DV': dv, 'Description': label, 'Source': source,
            'SS': ss_effect, 'df': df_effect, 'MS': ms_effect,
            'F': f_val, 'p': p_val, 'eta_sq': eta_sq_partial,
            'Significance': sig,
        })
    return results


def compute_descriptives(df, dv):
    """Compute cell-level descriptive statistics."""
    rows = []
    for app in ['FACE', 'MURDOC']:
        for cd in ['No', 'Yes']:
            subset = df[(df['App'] == app) & (df['CD'] == cd)][dv]
            rows.append({
                'DV': dv, 'App': app, 'CD': cd,
                'N': len(subset), 'Mean': subset.mean(),
                'SD': subset.std(ddof=1), 'SE': subset.sem(),
                'Min': subset.min(), 'Max': subset.max(),
            })
    return rows


def cohens_d(group1, group2):
    """Compute Cohen's d (pooled SD) for independent samples."""
    n1, n2 = len(group1), len(group2)
```

```
    var1, var2 = group1.var(ddof=1), group2.var(ddof=1)
    pooled = np.sqrt(((n1-1)*var1 + (n2-1)*var2) / (n1+n2-2))
    return (group1.mean() - group2.mean()) / pooled


# ======================================================================
# MAIN
# ======================================================================

def main():
    if len(sys.argv) > 1:
        filepath = sys.argv[1]
    else:
        filepath = '2025_MURDOC_v_FACE_Study_Results.csv'

    df = load_data(filepath)

    all_anova, all_desc = [], []

    for item, label in ITEM_LABELS.items():
        all_anova.extend(run_anova(df, item, label))
        all_desc.extend(compute_descriptives(df, item))

    composite_labels = {
        'Trust': 'Trust Composite (Q1, Q3, Q8, Q13)',
        'Understanding': 'Understanding Composite (Q2, Q4, Q5, Q7)',
        'Usability': 'Usability Composite (Q9, Q10, Q11, Q12)',
        'Interactivity': 'Interactivity Composite (Q15, Q16, Q17)',
    }
    for comp, label in composite_labels.items():
        all_anova.extend(run_anova(df, comp, label))
```

```
        all_desc.extend(compute_descriptives(df, comp))

    pd.DataFrame(all_anova).to_csv('ANOVA_Results_Complete.csv',
                                   index=False)
    pd.DataFrame(all_desc).to_csv('Descriptive_Statistics.csv',
                                  index=False)
    print("Analysis complete.")


if __name__ == '__main__':
    main()
```

E.2 SPSS Verification Syntax

The SPSS syntax below was used to independently verify all ANOVA results. Each `UNIANOVA` command corresponds to one dependent variable and specifies Type III sums of squares with descriptive statistics, partial η_p^2, and Levene's test for homogeneity of variance.

```
* SPSS Syntax for Two-Way ANOVA Analysis
* Independent Variables: App (MURDOC/FACE) and CD (Yes/No)
* ================================================================


* Filter for completed responses only.
SELECT IF (Finished = 1).
EXECUTE.


* Create composite scores.
COMPUTE Understanding = MEAN(Q2, Q4, Q5, Q7).
COMPUTE Trust = MEAN(Q1, Q3, Q8, Q13).
```

```
COMPUTE Usability = MEAN(Q9, Q10, Q11, Q12).
COMPUTE Interactivity = MEAN(Q15, Q16, Q17).
EXECUTE.

* Label variables.
VARIABLE LABELS
    Understanding 'Understanding Composite (Q2,Q4,Q5,Q7)'
    Trust 'Trust Composite (Q1,Q3,Q8,Q13)'
    Usability 'Usability Composite (Q9,Q10,Q11,Q12)'
    Interactivity 'Interactivity Composite (Q15,Q16,Q17)'.

* ================================================================
* TWO-WAY ANOVA FOR INDIVIDUAL QUESTIONS (Q1-Q20)
* ================================================================
* Each item follows the same template; Q1 shown as example.

* Q1 - Confidence in accuracy.
UNIANOVA Q1 BY App CD
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=DESCRIPTIVE ETASQ HOMOGENEITY
  /CRITERIA=ALPHA(.05)
  /DESIGN=App CD App*CD.

* Q2 through Q20 follow the identical pattern:
*   UNIANOVA Qn BY App CD
*     /METHOD=SSTYPE(3)
*     /INTERCEPT=INCLUDE
*     /PRINT=DESCRIPTIVE ETASQ HOMOGENEITY
*     /CRITERIA=ALPHA(.05)
*     /DESIGN=App CD App*CD.
```

```
* ================================================================
* TWO-WAY ANOVA FOR COMPOSITE SCORES
* ================================================================

UNIANOVA Understanding BY App CD
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=DESCRIPTIVE ETASQ HOMOGENEITY
  /CRITERIA=ALPHA(.05)
  /DESIGN=App CD App*CD.

UNIANOVA Trust BY App CD
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=DESCRIPTIVE ETASQ HOMOGENEITY
  /CRITERIA=ALPHA(.05)
  /DESIGN=App CD App*CD.

UNIANOVA Usability BY App CD
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=DESCRIPTIVE ETASQ HOMOGENEITY
  /CRITERIA=ALPHA(.05)
  /DESIGN=App CD App*CD.

UNIANOVA Interactivity BY App CD
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=DESCRIPTIVE ETASQ HOMOGENEITY
  /CRITERIA=ALPHA(.05)
```

```
  /DESIGN=App CD App*CD.

* ================================================================
* DESCRIPTIVE STATISTICS BY GROUP
* ================================================================

MEANS TABLES=Q1 TO Q20
    Understanding Trust Usability Interactivity BY App BY CD
  /CELLS=MEAN COUNT STDDEV.
```

Appendix F

Complete Loss Ablation Results

This appendix reports the complete HSV-Gestalt loss ablation summarized in Chapter 6. All fifteen loss-weight presets were evaluated on each of the three architectures (SD-1.5, SD-2.0, and SD-2.1) using the five-metric protocol described in Section 6.4.3. The six-preset progression reported in the body of Chapter 6 (Table 6.2) is a subset of the rows below. Within each architecture, the best value per metric is shown in bold.

Table F.1: Complete HSV-Gestalt loss ablation: all fifteen loss-weight presets evaluated on SD-1.5, SD-2.0, and SD-2.1. Within each architecture, the best value per metric is shown in bold.

Architecture	Preset	FID↓	IS↑	SSIM↑	LPIPS↓	CLIP↑
SD-1.5	mse_only	149.97	2.87	**0.0903**	0.7091	21.81
	conservative_hsv	149.40	2.76	0.0903	0.7090	21.77
	moderate_hsv	148.35	2.80	0.0900	0.7086	21.79
	hsv_enhanced	149.30	2.81	0.0890	0.7082	21.80
	full_gestalt	149.05	2.79	0.0881	0.7082	21.82
	theoretical	149.02	2.79	0.0882	**0.7072**	21.82
	hsv_only_light	149.40	2.76	0.0903	0.7090	21.77
	hsv_only_moderate	148.35	2.80	0.0900	0.7086	21.79
	hsv_only_strong	**148.32**	2.74	0.0899	0.7082	21.81
	hsv_only_maximum	148.69	2.76	0.0897	0.7077	21.78
	gestalt_only_light	149.89	2.81	0.0896	0.7087	21.80
	gestalt_only_moderate	149.79	**2.88**	0.0887	0.7081	21.80
	gestalt_only_strong	149.98	2.83	0.0878	0.7086	21.82
	gestalt_only_maximum	148.95	2.84	0.0870	0.7083	**21.85**
	optimized	148.93	2.73	0.0894	0.7081	21.79
SD-2.0	mse_only	129.59	3.03	0.0990	0.6834	22.48
	conservative_hsv	129.58	3.00	0.0997	0.6836	22.49
	moderate_hsv	129.84	3.00	0.1006	0.6838	22.50
	hsv_enhanced	129.84	3.04	0.1011	0.6838	22.49
	full_gestalt	129.72	3.00	0.1010	0.6831	22.54
	theoretical	129.91	3.04	0.1023	**0.6830**	**22.57**
	hsv_only_light	129.58	3.00	0.0997	0.6836	22.49
	hsv_only_moderate	129.84	3.00	0.1006	0.6838	22.50
	hsv_only_strong	129.81	3.02	0.1004	0.6835	22.50
	hsv_only_maximum	130.06	3.01	0.1015	0.6836	22.52
	gestalt_only_light	129.62	3.00	0.0999	0.6835	22.50
	gestalt_only_moderate	**129.53**	3.01	0.1007	0.6838	22.49
	gestalt_only_strong	129.77	3.05	0.1002	0.6835	22.49
	gestalt_only_maximum	129.86	3.00	0.1008	0.6833	22.52
	optimized	129.57	**3.05**	**0.1035**	0.6845	22.48
SD-2.1	mse_only	162.75	3.95	0.2222	0.8533	22.39
	conservative_hsv	164.25	3.86	0.2223	0.8542	22.46
	moderate_hsv	163.02	3.92	0.2227	0.8549	22.42
	hsv_enhanced	163.21	3.93	0.2223	0.8533	22.42
	full_gestalt	158.37	3.86	0.2204	0.8426	22.47
	theoretical	**155.15**	3.91	0.2173	**0.8272**	**22.48**
	hsv_only_light	163.10	3.92	0.2220	0.8533	22.42
	hsv_only_moderate	163.37	3.98	0.2225	0.8552	22.47
	hsv_only_strong	163.44	3.95	**0.2228**	0.8545	22.42
	hsv_only_maximum	161.92	3.79	0.2218	0.8499	22.46
	gestalt_only_light	162.85	3.90	0.2222	0.8541	22.42
	gestalt_only_moderate	162.44	3.95	0.2222	0.8532	22.42
	gestalt_only_strong	162.86	**4.00**	0.2220	0.8520	22.38
	gestalt_only_maximum	160.20	3.93	0.2211	0.8471	22.45
	optimized	163.69	3.91	0.2225	0.8547	22.42

Appendix G

Technical Background: Latent Diffusion Models

Chapter 6 employs several architectural components from the latent diffusion modeling literature that may not be familiar to all CS PhD readers, as they are taught primarily in elective courses and the primary research literature rather than in required coursework. This appendix provides self-contained explanations of the four components essential to understanding the contributions in Chapter 6: variational autoencoders (VAEs), the UNet architecture, the latent diffusion framework, and Low-Rank Adaptation (LoRA). Readers already familiar with these components may proceed directly to Chapter 6.

G.1 Variational Autoencoders

A *variational autoencoder* (VAE) [Kingma and Welling, 2014] is a generative model that learns a compressed, continuous representation of input data. It comprises two learned networks: an *encoder* $E(\cdot)$ that maps an input image $x \in$

$\mathbb{R}^{H\times W\times 3}$ to a lower-dimensional latent code, and a *decoder* $D(\cdot)$ that reconstructs the image from that code.

Unlike a standard autoencoder, which maps each input to a fixed point in latent space, a VAE maps each input to a *distribution*—a mean μ and a log-variance $\log \sigma^2$—and samples a latent code $z \sim \mathcal{N}(\mu, \sigma^2 I)$ during training. The training objective is:

$$\mathcal{L}_{\text{VAE}} = \underbrace{\|x - D(z)\|^2}_{\text{reconstruction}} + \underbrace{D_{\text{KL}}\big(\mathcal{N}(\mu, \sigma^2) \,\|\, \mathcal{N}(0, I)\big)}_{\text{regularization}} \tag{G.1}$$

The reconstruction term encourages the decoder to recover the original image; the KL divergence term regularizes the latent space to be approximately standard normal, which makes sampling at generation time well-defined.

The VAE used in Stable Diffusion compresses 512×512 RGB images (three channels) into 64×64 latent tensors with four channels, a compression factor of 48× in spatial resolution. Crucially, the four latent channels do not correspond to color channels or any human-interpretable feature: they are learned abstract representations whose structure is determined entirely by the reconstruction objective. This property creates the representational mismatch described in Section 6.2: operations that assume RGB or spatially grounded semantics (such as HSV conversion or Gestalt edge detection) yield meaningless values when applied directly to these latent tensors.

The VAE encoder and CPU-resident decoder are used concretely in Algorithm 6 (Perceptual Weight Map Construction), where decoded RGB images provide the pixel-space perceptual guidance described in Section 6.3.2.

G.2 The UNet Architecture

The UNet [Ronneberger et al., 2015] is a convolutional neural network architecture originally developed for biomedical image segmentation. Its defining characteristic is an encoder–decoder structure connected by *skip connections*:

- The **encoder** (contracting path) applies successive convolution and downsampling operations to produce a hierarchy of feature maps at decreasing spatial resolutions and increasing channel depths.
- The **decoder** (expanding path) applies successive upsampling and convolution operations to recover spatial resolution, producing an output of the same size as the input.
- **Skip connections** concatenate feature maps from each encoder level to the corresponding decoder level, allowing the network to combine fine-grained spatial detail from early layers with high-level semantic content from deeper layers.

This architecture is well-suited to tasks requiring dense per-pixel prediction, because the skip connections preserve spatial information that would otherwise be lost during downsampling.

In latent diffusion models, the UNet's role is to predict noise rather than perform segmentation. Given a noisy latent x_t at diffusion timestep t and an optional text prompt, the UNet predicts the noise component $\hat{\epsilon}$ that was added to the clean latent x_0 to produce x_t. The UNet in Stable Diffusion incorporates cross-attention layers that condition the noise prediction on text embeddings, which is the mechanism by which text prompts guide image generation.

The UNet's role in noise prediction is formalized in Algorithm 7 (LoRA Fine-Tuning), where $\text{UNet}_{\theta+\Delta\theta}(x_t, t, \text{prompt})$ denotes the conditioned forward pass at each training step.

G.3 Latent Diffusion Models and Stable Diffusion

Diffusion models [Ho et al., 2020] are a class of generative models that learn to reverse a gradual noise-addition process. Training proceeds in two stages: a fixed *forward process* progressively corrupts a clean data sample x_0 by adding Gaussian noise over T timesteps, producing a sequence $x_1, x_2, \ldots, x_T$; the model then learns to reverse this process—predicting and removing noise at each step—so that sampling from random Gaussian noise x_T and iteratively denoising produces a new data sample.

The forward process is defined by:

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I) \tag{G.2}$$

Where $\bar{\alpha}_t$ is a noise schedule parameter that controls how much of the original signal is retained at step t. Standard diffusion models [Ho et al., 2020] operate directly on pixel-space images, which is computationally expensive at high resolutions.

Latent diffusion models (LDMs) [Rombach et al., 2022] resolve this computational bottleneck by applying the diffusion process in the compressed latent space of a pretrained VAE rather than in pixel space. Concretely, a clean image x_0 is first encoded to a latent $z_0 = E(x_0)$; the forward diffusion process then operates

on z_0 to produce noisy latents z_t; and the UNet learns to predict the noise ϵ from z_t. At generation time, the denoising process produces a clean latent $\hat{z}_0$, which the VAE decoder maps back to a full-resolution image: $\hat{x} = D(\hat{z}_0)$.

Stable Diffusion [Rombach et al., 2022] is a widely used open-source LDM that adds text conditioning through a cross-attention mechanism, allowing the UNet's noise predictions to be guided by natural language prompts. The architecture couples three components:

1. A pretrained VAE that encodes and decodes images.
2. A text encoder (CLIP [Radford et al., 2021]) that converts prompts to dense embeddings.
3. A conditional UNet that accepts a noisy latent, a timestep embedding, and the text embeddings, and predicts the noise.

Stable Diffusion comes in several architecture variants (SD-1.5, SD-2.0, SD-2.1) that differ in their UNet and VAE weights, training data, and internal latent scaling factors. These differences produce distinct generative behaviors and distinct responses to fine-tuning, which is why Chapter 6 evaluates all three variants rather than treating them as interchangeable.

The forward diffusion equation above corresponds directly to line 6 of Algorithm 7, and the x_0 estimate in Equation 6.2 is derived from the same noise schedule parameter $\bar{\alpha}_t$.

G.4 Low-Rank Adaptation (LoRA)

Low-Rank Adaptation [Hu et al., 2022] is a parameter-efficient method for fine-tuning large pretrained models. Full fine-tuning updates all model parameters,

which for a model with billions of parameters is computationally prohibitive and prone to catastrophic forgetting of pretrained knowledge. LoRA avoids this by freezing the pretrained weights and instead learning a small set of additional low-rank weight updates.

Formally, for a pretrained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, LoRA introduces a decomposed update:

$$W = W_0 + \Delta W = W_0 + BA \tag{G.3}$$

Where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ are learned matrices with rank $r \ll \min(d, k)$. During training, W_0 is frozen, and only B and A are updated. Because r is small (the experiments in Chapter 6 use $r = 16$), the number of trainable parameters is reduced by several orders of magnitude relative to full fine-tuning, while the learned update $\Delta W = BA$ can still express meaningful task-specific adaptations.

At inference time, the LoRA update is applied by computing $W_0 + BA$; this can either be evaluated on the fly (keeping W_0 and ΔW separate) or merged into a single weight matrix at zero additional inference cost.

In the context of Chapter 6, LoRA is applied to the attention mechanism parameters of the Stable Diffusion UNet—specifically, the query (`to_q`), key (`to_k`), value (`to_v`), and output projection (`to_out.0`) matrices in each attention layer—to adapt the pretrained model to generate camouflage-consistent imagery while preserving the general visual capabilities acquired during pretraining. The full fine-tuning procedure is described in Algorithm 7 and Section 6.4.2.

The complete fine-tuning procedure incorporating all four components is formalized in Algorithm 7.