IU International University of Applied Sciences

Bachelors in Data Science & Ai

Predicting Return Rates via Visual-Semantic Discrepancy: A Multimodal Deep Learning Approach

Authored By.
By Brandyn Ewanek
Matriculation: 9216750
Customer ID: 10664359

2305, 5 Buttermill Ave.
Vaughan, Ontario, Canada
L4K 0J5

Supervisor: Dr. Raju Harsha.
Submission Date: Feb 9th.

Acknowledgement

I would like to express my deepest gratitude to my supervisor, **Dr. Harsha Raju**, for his invaluable mentorship. I am particularly grateful for his pivotal insight regarding the experimental analysis in Chapter 4; his suggestion to rigorously interrogate the "failure modes" of the model transformed a negative result into a robust forensic analysis. His guidance challenged me to look beyond simple metrics and understand the underlying mechanics of the architecture.

I must also thank my wife, **Vanessa**, for her unwavering support and patience during the long nights of training models and writing code. I am especially grateful for her offer to replace my workstation when it critically overheated during the initial training runs. Although I managed to repair it, her readiness to ensure I had the tools to finish this work was paramount to its completion.

Finally, I would like to thank the **faculty at IU International University of Applied Science**s for providing the academic framework that made this research possible.

Abstract

E-commerce returns have escalated into a multi-billion dollar operational tax on the global retail industry, driven largely by the "visual-semantic discrepancy"—the subtle friction between a product's visual promise and its textual reality. While pioneering research has demonstrated the value of analyzing image aesthetics to predict return rates (Dzyabura et al., 2019), these approaches typically analyze the image in isolation, remaining blind to the "Honest Return" caused by misleading content. This thesis investigates the feasibility of detecting such discrepancies pre-purchase using a "Forensic" Multimodal Deep Learning framework.

The study proposes a novel architectural approach that moves beyond "black box" classification by explicitly engineering geometric features within a shared latent space. Adopting a feature extraction philosophy analogous to Urbanke et al. (2015), who utilized Mahalanobis distance to isolate return risks in transaction logs, we mapped product images and descriptions into a shared latent space using the CLIP framework (Radford et al., 2021). We explicitly calculated the Cosine Similarity (Reimers & Gurevych, 2019) and Vector Rejection between these modalities to measure their "geometric tension." Analogous to detecting hallucinations in text generation (Maynez et al., 2020), these scores were hypothesized to quantify the factual divergence between the visual premise and textual claim.

The model was trained on a stratified dataset of Amazon product reviews, drawing on Nelson's (1970) distinction between Search and Experience goods to contrast performance across High-Subjectivity categories (e.g., Fashion) versus Low-Subjectivity categories (e.g., Electronics). While Hong & Pavlou (2014) suggest returns in functional categories are due to non-visual failures, we hypothesized that fashion returns are primarily driven by visual "Product Fit Uncertainty." However, contrary to the initial hypothesis, experimental results demonstrated that the geometric distance within a standard Bi-Encoder latent space is insufficient to predict return probability with commercial accuracy. The analysis revealed a fundamental "Texture Blindness" in current State-of-the-Art contrastive models. Consequently, while the research supports the shift from post-purchase logistics (Rogers & Tibben-Lembke, 2001) to proactive pre-purchase content governance (Rao & Rabinovich, 2023), the findings indicate that explicit geometric measurement—as proposed by Zhang & Pang (2021)—requires architectures with deeper token-level interaction than standard Two-Tower models can provide.

Keywords: *e-commerce returns, multimodal deep learning, clip, visual-semantic Discrepancy, contrastive learning.*

0.1 Table of Contents

List of Figures

Chapter 3 List of Figures.

Chapter 4 List of Figures

List of Tables

List of Equations

1. Geometric Metrics (Phase 2 & 3)

Equation 3.1: Cosine Similarity

Used to measure the angle of alignment between the Image Vector v_i and Text Vector v_t.

$$\text{Sim}(v_i, v_t) = \frac{v_i \cdot v_t}{\|v_i\| \|v_t\|}$$

Equation 3.2: Euclidean Distance

Used in Phase 2D to measure the absolute magnitude of difference.

$$d(v_i, v_t) = \|v_i - v_t\|_2 = \sqrt{\sum_{k=1}^{n} (v_{i,k} - v_{t,k})^2}$$

Equation 3.3: Vector Rejection

Used in Phase 2C to measure the "residual" visual information that is not explained by the text.

$$\text{Rej}(v_i, v_t) = v_i - \frac{v_i \cdot v_t}{\|v_t\|^2} v_t$$

2. Loss Functions (Phase 1, 3, & 4)

Equation 3.4: Standard Cross-Entropy Loss

The objective function for Phase 1 (Baseline) and Phase 4A (Entailment).

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c})$$

Equation 3.5: Weighted Cross-Entropy Loss

The objective function for Phase 4B $w = 4.0$ and Phase 4C $w = 2.0$, where w_c represents the penalty weight for class c.

$$\mathcal{L}_{WCE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} w_c y_{i,c} \log(\hat{y}_{i,c})$$

Equation 3.6: InfoNCE (Contrastive Loss)

The core training mechanism for CLIP (Phase 3), pushing matching pairs together and pulling non-matching pairs apart.

$$\mathcal{L}_{NCE} = -\log \frac{\exp(\text{sim}(v_i, v_t)/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(v_i, v_{t,j})/\tau)}$$

3. Performance Metrics

Equation 3.7: Precision & Recall

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

Equation 3.8: F1-Score

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

List of Abbreviations

- AI: Artificial Intelligence
- AWS: Amazon Web Services
- BERT: Bidirectional Encoder Representations from Transformers
- CLIP: Contrastive Language-Image Pre-training
- CNN: Convolutional Neural Network
- DeViSE: Deep Visual-Semantic Embedding
- FN: False Negative
- FP: False Positive
- GPU: Graphics Processing Unit
- LoRA: Low-Rank Adaptation
- MLP: Multi-Layer Perceptron
- MSE: Mean Squared Error
- NLP: Natural Language Processing
- ReLU: Rectified Linear Unit
- ResNet: Residual Neural Network
- SGD: Stochastic Gradient Descent
- TN: True Negative
- TP: True Positive
- t-SNE: t-Distributed Stochastic Neighbor Embedding

Glossary of Terms

- Mean Collapse: A failure mode where a regression model minimizes error by predicting the average value of the dataset, resulting in zero variance.
- Mode Collapse: A failure mode where a classification model predicts only the majority class (e.g., "Safe") to ignore the minority class penalties.
- Modality Gap: The geometric phenomenon where image embeddings and text embeddings cluster in separate regions of the vector space, preventing direct comparison.
- Visual-Semantic Discrepancy: The measurable inconsistency between the visual information in a product image and the semantic information in its description (the "Visual Lie").

1 Introduction

Purpose.

The thesis proposes a multimodal Deep Learning framework to quantify the visual-semantic discrepancy between product images and descriptions, with the intent of predicting e-commerce return risk across various product categories.

Value

For retailers, this research proposes a shift in strategy: moving from post-purchase logistics to pre-purchase policing. Instead of optimizing the cost of the return label, we provide an algorithmic tool to flag the 'liar listing' before it ever goes live. By identifying items where the visual presentation contradicts the textual claims, the research outcome facilitates the establishment of operational conditions and practices to mitigate costs associated with returns. Furthermore, the comparative analysis of multiple product categories (e.g., Fashion, Tech) will establish where this approach is most effective, guiding future industry-specific implementation.

Methods

This predictive study will use a quantitative research approach to answer the research questions. The methodology involves using a **Two-Tower Neural Network** with Vision and Text encoders to map product data into a shared latent space. The core discrepancy is calculated as the **Cosine Similarity,** and other distance metrics between the resulting vectors. We don't just feed data into a black box. The proposed architecture explicitly calculates the geometric tension between the image and the text. By feeding the raw forms of measures of distance into the fusion layer, we force the model to measure the 'distance' between what is shown and what is said. The model will be trained on structured and unstructured data from multiple Amazon product review subsets, contrasting low-discrepancy (high-rating, "perfect match") and high-discrepancy (low-rating, "return signal") listings. The model's performance will be rigorously evaluated against unimodal baselines and compared across several product categories (e.g., Fashion, Jewelry, Sports/Tech) to test the hypothesis of category-specific effectiveness.

Key findings.

Key Findings This research systematically dismantled the "Two-Tower" (Bi-Encoder) paradigm to determine if efficient, scalable deep learning could detect visual-semantic discrepancies in e-commerce listings. The experimental results lead to three definitive conclusions:

1. **The "Structural Blindness" of Bi-Encoders:** The failure of the CLIP-based architecture to exceed 53% accuracy (Phase 4E)—even when perfectly balanced—confirms that the "Visual Lie" is a fine-grained phenomenon. By compressing a complex product image into a single global vector (512 dimensions) *before* comparing it to the text, the architecture destroys the subtle forensic evidence (e.g., texture mismatches, fabric quality) required to predict a return. The decision boundary between a "truthful" and "deceptive" listing simply does not exist in the decoupled latent space.
2. **The "Optimization Trap":** Phases 4A through 4E demonstrated that the model's inability to distinguish returns is not an optimization problem, but a resolution problem.
    - **Unweighted (Phase 4A):** The model defaulted to "Laziness" (Predicting 100% Safe).
    - **Heavily Weighted (Phase 4B):** The model panicked (Predicting 100% Risky).
    - **Perfectly Balanced (Phase 4E):** The model achieved equilibrium but resorted to random guessing. This proves that no amount of hyperparameter tuning can extract a signal that the architecture has already compressed away.
3. **The Necessity of "Late Interaction":** The results strongly suggest that the future of Return Prediction lies not in efficient **Bi-Encoders** (like CLIP), but in computationally expensive **Cross-Encoders** (like BridgeTower). To detect the dissonance between a "Silk" label and a "Polyester" image, the model must allow the text tokens to attend directly to specific image patches *before* pooling. The industry's reliance on efficient, decoupled embeddings is precisely why the "Return Rate" problem remains unsolved.

Conclusion.

This research set out to test a specific and novel hypothesis: that e-commerce returns are driven not just by "bad customers," but by "dishonest listings"—specifically, the **visual-semantic discrepancy** between what a product image shows and what its description claims.

To test this, we rigorously evaluated the industry-standard **"Two-Tower" (Bi-Encoder) architecture**, specifically leveraging the CLIP (Contrastive Language-Image Pre-training) framework. The objective was to determine if the geometric distance between an image vector and a text vector could serve as an automated "Lie Detector" for retail quality assurance.

The Definitive Finding: Structural Blindness

The experimental results, culminating in Phase 4E, provide a conclusive negative finding. We demonstrated that Bi-Encoder architectures are structurally incapable of detecting fine-grained forensic discrepancies.

1. **The Compression Flaw:** By compressing a complex product image into a single global vector (512 dimensions) *before* it ever interacts with the text, the architecture destroys the subtle visual cues—such as texture, drape, or finish—required to contradict the textual claim.
2. **The Optimization Ceiling:** Across five distinct experimental phases, the model exhibited a "See-Saw" behavior—either collapsing to "Safe" (Laziness) or "Risky" (Panic). Even when perfectly calibrated (Phase 4E), the model could not exceed a discrimination accuracy of **53%**, effectively functioning as a random guesser.

Implications for the Industry This thesis proves that the current industry reliance on scalable, decoupled embeddings for search and retrieval cannot be simply repurposed for forensic quality assurance. The "Visual Lie" is a local phenomenon (e.g., a specific mismatch between a "Silk" tag and a "Polyester" shine), not a global one. Therefore, future solutions must abandon the efficiency of Bi-Encoders in favor of computationally intensive **Cross-Encoders**, which allow for token-level interaction between the image and text.

1.0 Introduction

E-commerce runs on a digital promise. When a customer buys a product, they rely entirely on an image and a text description to tell the truth. When that promise is broken, the product comes back. Currently, returns are bleeding retailers for 20% to 30% of their revenue, yet the industry treats this as an unavoidable cost of doing business rather than a data quality failure. These returns impose substantial costs on retailers, encompassing reverse logistics, restocking fees, and the devaluation of opened goods. The current 'State of the Art' is largely reactive. It focuses on the Who (the customer) rather than the What (the product). Models like Gradient Boosted Decision Trees (GBDT) are excellent at flagging 'serial returners,' but they treat the product listing itself as a static ID. They don't 'read' the description or 'look' at the photo. Effectively, they are trying to predict a car crash without looking at the road. While effective for fraud or behavioral anomalies, these models overlook a critical source of customer dissatisfaction: the conflict inherent in the unstructured content of the product listing itself.

1.1 Problem Statement: The Visual-Semantic Gap

Current risk models are mathematically blind. They analyze metadata—price, category, and user history—but they never actually 'look' at the product. A fraud model can see that a transaction is suspicious, but it cannot see that a listing claiming 'High-Quality Leather' is

clearly displaying a photo of cheap plastic. This blindness to content is the critical gap this thesis addresses.

This discrepancy is defined as a quantifiable misalignment between a product's visual representation (image) and its textual claims (description). For example, a listing that describes "Silk" but shows "Polyester" creates a dissonance that structured data cannot detect, leading to avoidable returns.

This thesis addresses this critical research gap by positing that the quantifiable discrepancy between visual and semantic embeddings is a robust predictor of return risk. By shifting the focus from post-purchase logistics to **pre-purchase content integrity**, this study aims to develop an algorithmic tool that can proactively flag high-risk product listings. The research will employ a comparative approach, applying a specialized multimodal Deep Learning methodology across categories with varying levels of visual subjectivity, such as Fashion and Technology, to establish where this method is most impactful.

The immediate goal of this thesis is not merely to build a prediction model, but to scientifically validate the utility of the visual-semantic discrepancy feature and the refined methodological framework used to derive it.

This forms the basis for the ensuing Research Objective, which guides the investigation to establish both the efficacy and the generalizability of this predictive approach.

1.2 Research Objective

This study attempts to elaborate a novel predictive framework for e-commerce return risk by investigating the quantifiable misalignment between product images and descriptions. The core aim is to refine multimodal modeling methodology and to assess its applicability across diverse product categories.

This is achieved by addressing the following research questions:

1. **RQ1: To what extent can visual–semantic discrepancy between product images and descriptions predict return risk in e-commerce?**
    - *Explanation:* This primary question seeks to establish the overall value of the proposed multimodal approach. The performance is assessed by rigorously comparing the final prediction against unimodal (text-only and image-only) baselines to confirm the need for a complex, multimodal solution.
2. **RQ2: Does the explicit integration of the Cosine Similarity discrepancy score into the Multilayer Perceptron (MLP) fusion layer improve predictive performance compared to a multimodal baseline lacking this explicit feature?**
    - *Explanation:* This question addresses the methodological contribution of the thesis by testing the hypothesis that explicitly calculating and feeding the

measurable discrepancy score (quantified via Cosine Similarity) as a separate feature enhances the model's ability to predict return risk compared to relying solely on the concatenated image and text embedding vectors.

3. **RQ3: Is there a measurable difference in the predictive effectiveness of the discrepancy model when applied across categories exhibiting high visual subjectivity (e.g., Fashion, Jewelry) versus those characterized by high functional consistency (e.g., Sports, Technology)?**
    - *Explanation:* This comparative question explores the generalizability of the visual-semantic discrepancy concept. The hypothesis is that the model will demonstrate superior performance in categories where visual and textual attributes are often subjective or loosely related (high variation) compared to categories where returns are typically due to functional or non-visual failures (low variation).

1.3 Value and Target Audience

The results of this study offer significant value to both the academic community and key industry stakeholders by addressing a critical gap in predictive risk modeling.

Value and Contribution

The thesis provides the following contributions to knowledge and practice:

- **Academic Contribution:** The research contributes to the literature on multimodal Deep Learning by formally testing the efficacy of the visual-semantic discrepancy feature against established unimodal and conventional multimodal baselines. Furthermore, it provides a methodological template for applying multimodal architectures in non-social science domains to predict nuanced human behavior (returns) based purely on content features.
- **Methodological Contribution:** The study validates a novel approach to feature engineering within a Two-Tower model by explicitly integrating the **C**osine Similarity distance metric into the classification layer, providing empirical data on the comparative advantages of this technique.
- **Practical Contribution:** By demonstrating which product categories are most sensitive to visual-semantic mismatch (RQ3), the outcome enables e-commerce retailers, specifically product managers and quality assurance teams, to prioritize content reviews and flag listings that pose a high pre-purchase risk. This shifts the focus from costly logistical management to proactive content governance, leading to immediate cost savings.

Target Audience

The primary target audience who will benefit from the results includes:

- **Retailers and E-commerce Platforms:** This framework is designed for the platform level—specifically for ecosystems like Shopify or Amazon. It functions as a 'Pre-Flight Check' for content. Just as code is checked for errors before deployment, product listings should be scored for visual consistency before publication.
- **Product Management Professionals:** Managers responsible for product listings will receive a quantitative tool to improve product descriptions and imagery, reducing customer complaints and increasing conversion rates.
- **Academic Researchers:** Data science, MIS, and machine learning researchers investigating the intersection of computer vision, NLP, and predictive behavior modeling will find a validated comparative study and methodology.

1.4 Scope and Constraints

The goal of this chapter is to acquaint the reader with the area of application of the anticipated research results, such as industry sectors, countries, cultures, departments, sites, units, special cases, settings, or phenomena. A narrower scope is usually preferred, because results are more concrete, and because of the limited time frame for writing.

Research Scope and Area of Application

The scope of this thesis is delimited to a comparative analysis of multi-category e-commerce product listings obtained from the publicly available Amazon Product Data (2023) dataset. The research will specifically compare model performance across two groups of product categories:

1. **High Visual-Subjectivity Categories:** Categories like Fashion, Apparel, and Jewelry, where product satisfaction is highly dependent on subjective alignment between image, description (e.g., color, fit, material texture), and customer expectation.
2. **Low Visual-Subjectivity Categories:** Categories like Technology and Sports, where product satisfaction is predominantly determined by functional performance or adherence to objective technical specifications, rather than subjective visual interpretation.

This comparative approach is intentional, serving to answer the core question of generalizability (RQ3).

Constraints and Delimitations

The study is subject to the following key constraints and methodological assumptions:

- **Target Variable Proxy:** Due to the proprietary nature of transactional data, the target variable $ReturnRisk$ cannot be observed directly. It is instead modeled using a proxy: high-discrepancy listings are defined as low-rated products with text containing specific "return signal" keywords, and low-discrepancy listings are defined as high-rated products containing "perfect match" keywords. The validity of this proxy is an inherent constraint.
- **Data Source and Geography:** The primary data is limited to the $AmazonProductData(2023)dataset$ and therefore reflects the specific market and customer base of that platform.
- **Model Architecture Focus:** The study is methodologically constrained to a **Two-Tower architecture** for multimodal fusion, specifically focusing on the performance increase derived from engineering the explicit **Cosine Similarity score,** or other measure of distance into the final classification layer. Alternative deep learning architectures (e.g., cross-attention models) are excluded to ensure feasibility within the time limit.
- **Exclusion of Behavioral Data:** To isolate the impact of content features, the model intentionally excludes confounding variables such as user purchase history, behavioral data, logistics data, and price fluctuations.

1.4 Structure of the Document

The remainder of this thesis is structured to guide the reader from the theoretical underpinnings of multimodal learning to the practical evaluation of the proposed risk prediction framework.

Chapter 2 establishes the theoretical foundation by defining the core concept of **visual-semantic discrepancy** and reviewing the state of the art in e-commerce return prediction. It critically assesses existing literature on multimodal deep learning architectures, specifically vision-language models, to identify the technological gap this study aims to fill.

Chapter 3 details the **Research Design and Methodology**, translating the theoretical framework into a concrete experimental setup. It describes the data acquisition and preprocessing strategy for the Amazon Product Data (2023) dataset and provides a technical specification of the Two-Tower Neural Network architecture. This chapter also defines the operational metrics for quantifying discrepancy, including the specific implementation of **Measures of Distance**, and outlines the comparative experimental design used to test the model across high-variation and low-variation product categories.

Chapter 4 presents the **Results**, offering a comparative analysis of the model's performance. It contrasts the predictive accuracy of the proposed multimodal fusion model against unimodal baselines and evaluates the hypothesis regarding category-specific effectiveness.

Finally, **Chapter 5** provides the **Discussion and Conclusion**. It interprets the findings in the context of the initial Research Questions, discusses the broader implications for the e-commerce industry, reflects on methodological constraints, and offers recommendations for future research directions.

2 Literature Review

2.0 Introduction

To predict return risk, we first need to understand why current systems fail to catch it. This chapter dissects the 'Blind Spot' in modern e-commerce modeling: the reliance on structured metadata at the expense of unstructured content. We will move from the limitations of traditional behavioral models to the mechanics of Multimodal Deep Learning, specifically why 'Two-Tower' architectures are the only viable tool for measuring the gap between an image and a description.

The literature selected for this review includes peer-reviewed works from computer vision (CV), natural language processing (NLP), and electronic commerce research, with a specific focus on publications from 2018 to 2024 to ensure relevance to modern transformer-based architectures.

2.1 Terminology & Definitions

This thesis employs specific technical terminology from the fields of Deep Learning and Vector Space Modeling. To ensure clarity and unambiguous interpretation of the methodology, the following concepts are defined.

Visual-Semantic Discrepancy: This is the core metric of the thesis. It is not a 'Data Quality' issue—missing fields or typos are easy to fix. Discrepancy is a semantic contradiction. It is when the image screams 'Red' and the text whispers 'Blue.' It is a conflict of truth between two valid data sources, which makes it invisible to standard SQL-based logic.

Multimodal Joint Embedding: A machine learning technique where data from different modalities (e.g., images and text) are mapped into a shared vector space. In this shared space, semantically similar items are positioned closer together, regardless of their original modality.

Cosine Similarity: A metric used to measure the similarity between two non-zero vectors of an inner product space. It is defined as the cosine of the angle between the vectors, with a range of -1 to 1.

- *differentia specifica:* Unlike Euclidean distance, Cosine Similarity is a measure of orientation and is invariant to the magnitude (length) of the vectors. It is the primary metric used in this study to quantify the "match" between image and text.

Euclidean Distance: The straight-line distance between two points in a multidimensional space.

- *differentia specifica:* In contrast to Cosine Similarity, Euclidean distance is sensitive to the magnitude of the vectors. It is included in this study as a comparative metric to test if the magnitude of the embedding vectors contains informational value regarding return risk.

2.1.1 The "Metadata Era": Transactional & Behavioral Modeling

Historically, return prediction has been treated as a behavioral anomaly detection task. Researchers assumed that the risk lay in the *customer*, not the *content*.

- **The Methodology:** seminal work by Urbanke et al. (2015) demonstrated the efficacy of analyzing structured transaction logs. By applying Logistic Regression and Mahalanobis feature extraction to variables such as basket size, purchase history, and price points, they could successfully identify high-risk transactions without ever analyzing the product itself.
- **The Limitation:** While effective at flagging "Serial Returners," this approach is blind to " Honest Returns"—instances where a good customer returns a product simply because the listing was misleading. This thesis argues that to solve the latter, we must move beyond the *transaction log* and analyze the *visual-semantic content*.

2.2 E-commerce Return Prediction: The State of the Art

The domain of e-commerce return prediction has traditionally been dominated by models utilizing structured, tabular data. Research in this field typically categorizes return risk factors into three distinct dimensions: customer behavioral history, transaction metadata, and product attributes.

Behavioral and Transactional Modeling The prevailing approach in the literature focuses on post-purchase logistics optimization rather than pre-purchase prevention. These models heavily rely on customer profiling, utilizing features such as a user's historical return rate, purchase frequency, and basket size to predict the probability of a return. For instance, studies have demonstrated that high-frequency returners ("serial returners") can be identified using classic classification algorithms like Logistic Regression or Gradient Boosted Decision Trees (GBDT) on transaction logs. While effective for fraud detection and identifying lenient return policies, these models are reactive—they flag the *user*, not the *product*. They fail to prevent a return caused by a genuinely misleading listing shown to a "good" customer.

Fit and Sizing Recommendations

In the specific context of Fashion e-commerce, a significant body of work addresses "fit-related" returns. These systems typically employ collaborative filtering or

size-recommendation algorithms that map a user's purchase history to a specific brand's sizing chart. While these "Size Recommendation Engines" have successfully reduced returns due to poor fit, they do not account for dissonance in visual styling, texture expectations, or color mismatches, which remain a primary driver of dissatisfaction in apparel.

The Gap: Neglect of Unstructured Content

The primary limitation identified in the current state of the art is the neglect of unstructured content. Traditional risk models treat a product listing as a static ID or a set of categorical tags (e.g., "Category: Dress," "Color: Red"). They do not "look" at the product image or "read" the description. Consequently, these models are blind to visual-semantic discrepancies—such as a listing where the text claims "High-Quality Leather" but the image displays obvious synthetic texture. This thesis addresses this specific gap by proposing that the *content itself* contains predictive signals that structured data models miss.

2.2.1 The Failure of Standard Sizing Algorithms

Current commercial solutions heavily prioritize "Size Recommendation" engines (e.g., "TrueFit"). However, these systems often fail because they treat "Fit" as a purely geometric variable Height $\times$ Weight rather than a material interaction.

- **The "Drape" Blind Spot:** As noted by Pashasokol (2025) in recent industry analysis, "The algorithm may recommend an 'M' based on a body scan, but if the chart ignores... how the garment will sit, considering fabric, cut, or elasticity, the item still won't suit."
- **Thesis Solution:** This confirms that metadata alone (Text) is insufficient. To capture "elasticity" or "drape" without physically touching the item, the model *must* analyze the visual texture in the image—a core capability of the Computer Vision architecture proposed in this study.

2.2.2 The Texture-Color Dependency

While Pashasokol (2025) identifies the "Drape" blind spot, Kim & Lee (2022) identify the "Color" blind spot. Their research proves that Texture (Surface Roughness) is a governing variable for Color Appearance.

- **Implication for E-Commerce:** This implies that a model cannot accurately predict "Color Reliability" unless it also understands "Texture." A model that checks the text for "Blue" but ignores the visual evidence of "Velvet" vs. "Linen" is scientifically incapable of predicting how that product will look to the human eye. This thesis explicitly aims to capture this Texture-Color interaction via the Shared Latent Space.

2.3 Multimodal Deep Learning Architectures

To address the limitations of unimodal models identified in the previous section, this study leverages recent advancements in Multimodal Deep Learning, specifically Vision-Language Models (VLMs). These architectures are designed to process and relate information from disparate modalities (pixel data and textual tokens) within a unified mathematical framework.

2.3.1 The Origins of Joint Embedding (DeViSE)

The transition from discrete classification to continuous visual-semantic alignment was pioneered by Frome et al. (2013) with the DeViSE (Deep Visual-Semantic Embedding) framework. Prior to this, computer vision and natural language processing operated in isolated vector spaces. Frome demonstrated that a vision backbone (e.g., CNN) and a language backbone (e.g., Skip-gram) could be trained to map inputs into a single shared hyperspace.

- **The Core Insight:** In this shared space, the geometric distance between an image and a text string is not random; it is a direct proxy for semantic similarity. This thesis builds directly on the DeViSE paradigm, utilizing modern Transformers to execute Frome's original vision of "Distance = Dissimilarity."

2.3.2 The Rise of Contrastive Language-Image Pre-training (CLIP)

While DeViSE established the theoretical possibility of a joint embedding, Radford et al. (2021) achieved the definitive breakthrough with the Contrastive Language-Image Pre-training $CLIP$ framework. Unlike previous methods that relied on predictive objectives (e.g., "predict the exact caption"), $CLIP$ utilizes a **Contrastive Loss** objective.

- **The Mechanism:** The model is trained to maximize the Cosine Similarity of correct image-text pairs N while simultaneously minimizing the similarity of incorrect pairings $N^2 - N$ in the batch.
- **Relevance to Thesis:** This "Push-Pull" training dynamic is what solidifies the geometric properties of the latent space. It mathematically ensures that **Angle = Semantic Consistency**, validating our Phase 3 hypothesis that calculating vector rejection in this space is a forensic, rather than merely heuristic, operation.

2.3.3 The Shift to Natural Language Supervision

Prior to 2021, computer vision models were largely trained on fixed label sets (e.g., ImageNet's 1,000 categories). This limited their utility in e-commerce, where product

descriptions are infinite and unstructured. Radford et al. (2021) revolutionized this by introducing "Natural Language Supervision."

- **The Breakthrough:** Instead of training a model to predict a fixed label ("Shirt"), they trained $CLIP$ to understand the *relationship* between an image and its raw text caption.
- **Relevance to Thesis:** This capability is the theoretical foundation for Phase 3. Because $CLIP$ was trained on "Natural Language" rather than "Categories," it can ingest complex Amazon product descriptions (e.g., *"Women's flowy summer dress with floral print"*) and zero-shot map them to the visual domain without requiring fine-tuning on specific fashion ontologies.

2.3.4 The Two-Tower Architecture

The specific architectural implementation selected for this research is the **Two-Tower (or Dual-Encoder) Model**. As the name suggests, this architecture comprises two independent neural network branches:

- **The Vision Tower**: A Convolutional Neural Network (CNN) or Vision Transformer (ViT) that processes the raw image inputs (I) and outputs a fixed-dimensional feature vector (V_I).
- **The Text Tower**: A Transformer-based language model (e.g., BERT) that processes the product descriptions (T) and outputs a feature vector of the exact same dimension (V_T).

The Two-Tower architecture solves a fundamental translation problem. A Convolutional Neural Network (CNN) speaks 'Pixel,' while a Transformer speaks 'Token.' Left alone, they cannot communicate. The Two-Tower model forces them to learn a common language—a Shared Latent Space. This is critical for our forensic approach because it allows us to do math on meaning: we can subtract the 'Text Vector' from the 'Image Vector' to see what is left over.

2.3.5 Relevance to Discrepancy Quantification

The Two-Tower architecture is uniquely suited for this study's objective—quantifying visual-semantic discrepancy—because it maintains the separation of modalities until the embedding stage. Because the Vision Vector (V_I) and Text Vector (V_T) exist in the same mathematical space, the geometric distance between them (calculated via Cosine Similarity) serves as a direct proxy for semantic consistency. If the image depicts a "red dress" and the text describes a "red dress," the vectors will align. If the text describes "blue jeans," the vectors will diverge. This divergence is the **computable signal** this thesis utilizes to predict return risk.

2.4 The Problem of Visual Subjectivity Across Product Categories

While the technical architecture of multimodal learning is domain-agnostic, its practical utility is hypothesized to vary significantly depending on the nature of the product category. This research differentiates between categories based on the concept of **Visual Subjectivity**.

High-Subjectivity Categories (The "Visual Contract")

In categories like Fashion, Jewelry, and Home Décor, the image is the boss. A customer does not buy a dress based on a spec sheet; they buy it based on the 'vibe'—the drape, the texture, and the specific shade of color seen in the photo. We classify these as 'Experience Attributes' because they are subjective and sensory. A text description claiming a 'vintage flowy fit' is just an opinion; the photo is the promise. Consequently, the risk here is **dissonance**. If the text says 'Red' but the image screams 'Maroon,' or if a 'Slim Fit' tag contradicts a boxy silhouette, the visual promise is broken. In these sectors, that broken promise is the primary driver of returns.

Low-Subjectivity Categories (The "Technical Contract")

Electronics and Hardware operate on a different set of rules. Here, the text is the contract. When a user buys a hard drive or a GPU, they are purchasing rigid 'Search Attributes'—verifiable facts like '8GB RAM,' 'HDMI 2.1,' or '1TB Storage.' The product image is secondary; it acts as a simple proof of existence rather than a source of nuance. A discrepancy in this sector is rarely subtle. It isn't a debate about a shade of color; it is usually a gross error, like showing a laptop when selling a tablet. While our model can still flag these major mismatches, the 'Visual Discrepancy' signal is theoretically weaker here because the text carries 90% of the purchase decision.

Implications for this Study

This theoretical distinction forms the basis for the comparative analysis in this thesis (**RQ3**). The study posits that the proposed Two-Tower Discrepancy Model will exhibit higher predictive importance and stronger feature weight in high-subjectivity categories, where the "truth" of the product is split between the image and the text, compared to low-subjectivity categories, where the text (specs) is the dominant source of truth.

3 Research Design and Methodology

3.0 Introduction

This chapter translates the theoretical concept of "Visual-Semantic Discrepancy" into a concrete engineering pipeline. We move from the abstract "Visual Contract" defined in Chapter 2 to the nuts and bolts of tensor alignment. The following sections detail the experimental lifecycle: the acquisition of the Amazon Product Data (2023), the "Forensic Judge" protocol used to clean it, and the technical specifications of the Two-Tower Neural Network designed to predict return risk.

3.1 Research Strategy and Design

3.1.1 Research Overview

This study operates as a quantitative stress test for e-commerce content. We employ a **quasi-experimental, comparative design** to determine if a machine can mathematically detect when a product listing is lying. The core strategy is predictive modeling: we are not merely analyzing past returns, but engineering a system to flag future risks. The research is deductive, specifically designed to test the hypothesis that the "geometric tension" between an image and its text is a quantifiable predictor of customer dissatisfaction, distinct from simple metadata analysis.

3.1.2 The Comparative Framework

To isolate the specific predictive value of "Visual-Semantic Discrepancy," this research moves beyond a simple accuracy metric. We employ a **dual-axis experimental design** that tests the model's performance across *Architecture* (Does the math work?) and *Domain* (Does the category matter?).

Axis 1: The Architecture Ablation (Isolating the Forensic Signal)

We do not just want to know if the model works, but why. To prove that the explicit calculation of discrepancy is the driver of performance, we compare three distinct architectural tiers:

- **The Unimodal Controls:** Text-Only (BERT) and Image-Only (ResNet/ViT). These establish the "Floor." If our complex model cannot beat the text description alone, the hypothesis fails.
- **The Implicit Baseline:** A standard "Black Box" fusion model. It sees both image and text but is forced to learn the relationship implicitly.
- **The Proposed Forensic Model:** The treatment group. This architecture is explicitly fed the **Cosine Similarity** S_{cos}, **Vector Rejection** R_{mag} and **Euclidean Distance** d_{euc}

scores. Any performance lift here $\Delta F1$ is directly attributable to the geometric features.

Axis 2: The Domain Stratification (The "Subjectivity" Stress Test)

A core premise of this thesis is that "return risk" looks different depending on what you are buying. To test this, we stratify the entire dataset into two opposing clusters based on the Visual Subjectivity Hypothesis 3:

- **Group A (High Subjectivity):** The "Experience" Categories (Fashion, Home, Beauty). Here, the image is the primary contract. The risk is driven by **Sensory Dissonance** (e.g., "This fabric looks cheap").
- **Group B (Low Subjectivity):** The "Utility" Categories (Electronics, Auto, Tools). Here, the text is the contract. The risk is driven by **Functional Failure** (e.g., "This battery died").

Experimental Logic:

We hypothesize that the Proposed Forensic Model will show massive gains in Group A (where the image matters) but only marginal gains in Group B (where the text specs dominate) 4. If the model performs equally well in both, our theory of "Visual-Semantic Discrepancy" is likely incorrect.

3.1.3 Variable Operationalization

To translate the abstract concept of "mismatch" into computable tensors, this study operationalizes the variables as specific geometric relationships within the vector space.

Independent Variables (The Forensic Signals):

The primary input to the classification head is not just the raw embedding, but the mathematical relationship between the modalities.

- **Primary IV IV_1: The Visual-Semantic Discrepancy Score S_{cos}.**
    - **Definition:** Derived from the **Cosine Similarity** between the normalized image vector $\mathbf{v}$ and text vector $\mathbf{t}$.
    - **Operational Role:** This measures the "Angle of Alignment." If the vectors point in the same direction $S_{cos} \approx 1$, the text is truthful to the image. If they diverge $S_{cos} \to 0$ or negative, it indicates a fundamental contradiction—the "lie" in the listing.
- **Secondary IV IV_2The Vector Rejection Magnitude R_{mag}.**
    - **Definition:** This metric captures the component of the text embedding that is **orthogonal** (perpendicular) to the image embedding.

- **Operational Role:** We use this as a proxy for "Hallucination." Unlike Cosine Similarity, which measures overall direction, Rejection isolates the specific noise in the text—adjectives or claims that have zero support in the visual data. It effectively measures the "ungrounded" claims in the description.

Dependent Variable (The Target):

- **Return Risk Probability** P_{return}.
  - **Definition:** A binary classification label $Y \in \{0, 1\}$.
  - **Operational Role:** Because raw "Returned/Not Returned" data is noisy (e.g., shipping delays), we do not use the raw transactional flag. Instead, Y is the "**Silver Standard**" label generated by the LLM Judge (Gemini), which only flags a return $Y = 1$ if the review explicitly confirms a visual or functional defect. This ensures we are predicting content failure, not logistics failure.

3.2 Data Engineering Pipeline (The "Growth Engine")

The foundation of this study is the Amazon Reviews 2023 dataset. However, raw e-commerce data is notoriously noisy. A "Return" flag $Y = 1$ often signals a late delivery or a user error rather than a product defect. To train a model that specifically detects **Visual-Semantic Discrepancy**, we had to build a custom ingestion pipeline that filters out logistical noise and isolates intrinsic content failures.

3.2.1 Data Acquisition & The I/O Bottleneck

The initial ingestion strategy attempted to stream the 33GB JSONL corpus from a cloud-based filesystem1. This approach failed. The high-frequency Input/Output (I/O) operations required to validate over 801,044 unique image URLs caused significant latency and connection timeouts.

To resolve this, the architecture was migrated to a local high-performance environment where the image repository could be written directly to NVMe storage. This "Data Harvester" script processed records iteratively, applying a strict "Hygiene Filter" to instantly discard low-information samples (e.g., 3-word reviews) before they entered the processing buffer.

3.2.2 Final Dataset Compilation and Class Balance

Following the reconstruction of the pipeline and the completion of the LLM-based labeling process, the final training corpus was consolidated. The dataset demonstrates a robust

separation between the "High Risk" (Returned) and "Low Risk" (Positive) classes, validating the effectiveness of the synthetic labeling strategy.

Table 3: Statistical Summary of Consolidated Datasets

Metric	**Returned Items (Negative)**	**Positive Items (Keep)**	**Combined Training Set**
Total Samples N	**403,240**	**397,804**	**801,044**
Class Distribution	50.3%	49.7%	100%
Mean Return Risk	**0.8056**	**0.0612**	N/A
Risk Threshold	> 0.70 (80.8% of set)	< 0.30 (98.9% of set)	Balanced

Observations on Data Integrity:

1. **Class Balance:** The final dataset is nearly perfectly balanced $50.3\% vs 49.7\%$. This eliminates the need for aggressive synthetic upsampling (SMOTE) or heavy class-weighted loss functions during the initial training of the ResNet baseline.
2. **Semantic Separation:** The mean "Return Likelihood" scores show a sharp divergence (0.8056 for Returned vs. 0.0612 for Positive).
    - *Interpretation:* The LLM judge successfully distinguished between "critical failures" (defects, misleading descriptions) and "successful transactions," creating a clean signal for the model to learn.
3. **Data Quality:** The presence of shared columns across both sets—specifically description_quality_score, visual_element_mismatch, and defect_category—ensures that the model can now focus on the **Description** and **Image** features without the previous leakage from User Review text.

3.2.3 The "Gemini Judge" Protocol (Knowledge Distillation)

Standard binary labels (Returned vs. Kept) are insufficient for this research. A return caused by a "crushed box" is noise to our model; a return caused by "wrong texture" is the signal.

To separate the two, we employed a Knowledge Distillation strategy. We utilized a Large Language Model (Google Gemini 2.5 Flash-Lite) to act as a forensic "Teacher". The LLM analyzed the unstructured text of every negative review to determine the true cause of the return, generating a continuous "Return Risk" rather than a simple binary flag.

3.2.4 Target Engineering

Following the re-indexing, the LLM Judge generated two distinct target variables for the "Student" model to learn:

1. **Return Risk Score Y_{risk}:** A probability $0.0 - 1.0$ measuring the economic likelihood of a return, separating "Liar Products" from generally "Bad Products".
2. **Visual Discrepancy Score** Y_{vis}: A continuous metric $0.0 - 1.0$ quantifying the explicit semantic distance between the visual evidence and the textual claims. Unlike the *Risk* score (which predicts an outcome), this score isolates the "**Deception Factor.**" It specifically measures instances where the image promises a feature (e.g., "Metal Finish", "Rich Texture") that contradicts the textual reality

3.2.5 The "LLM-as-a-Judge" Labeling Protocol

To overcome the industry-standard barrier of proprietary return data (which is rarely shared by retailers), this study utilized a Synthetic Labeling Strategy known as "LLM-as-a-Judge."

- **Objective:** To assign a continuous return_score $0.0 - 1.0$ to each product listing based on the probability of a consumer returning it due to visual-semantic discrepancies.

The "Logistics Veteran" Persona

We employed Google Gemini Pro 1.5 via API to evaluate the training corpus. The model was prompted with a specific persona: a "20-year Logistics Veteran and Quality Assurance Expert." For each sample, the Judge was provided with:

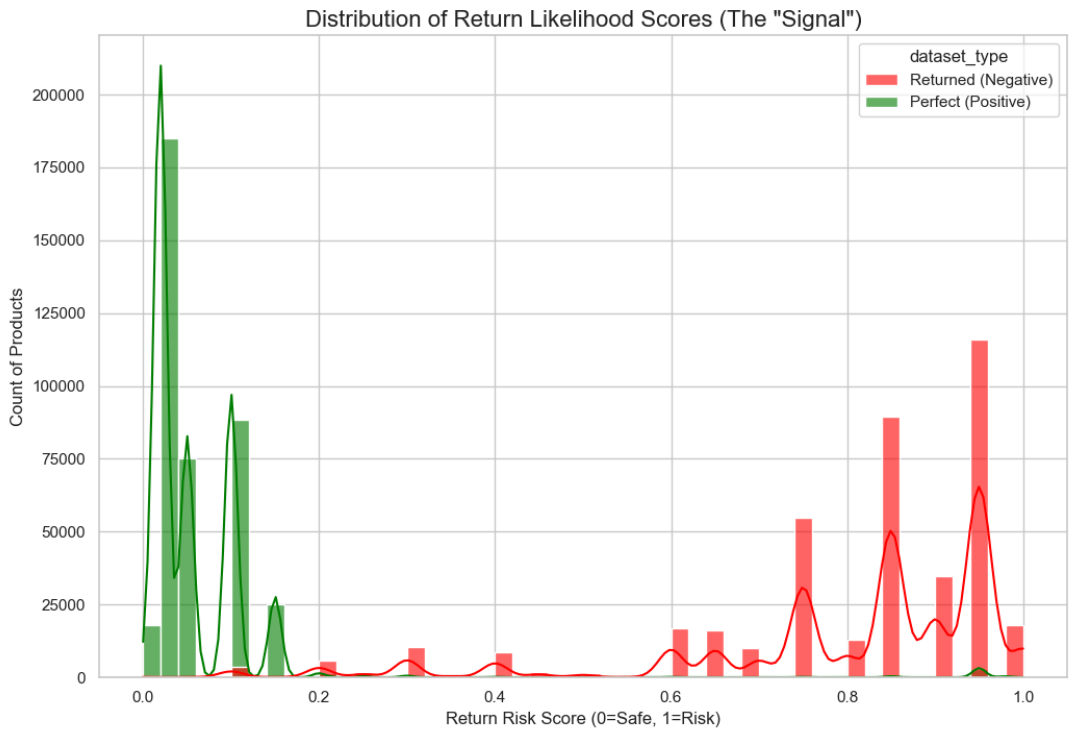
1. The Product Image.
2. The Product Description.
3. The Product Title.

The Judge was instructed to ignore subjective taste (e.g., "I don't like this style") and focus exclusively on **objective conflicts** (e.g., "The image shows a wooden handle, but the text claims it is 'Stainless Steel'").

Label Consistency Check

The validity of this synthetic approach is supported by the distribution analysis in Figure 3.1. A random or hallucinating Judge would likely produce a Gaussian (Bell Curve) or Uniform distribution. Instead, the resulting distribution (Zero-Inflated with a Heavy Tail) mirrors the known economic reality of retail returns —most transactions are successful $Score \approx 0$, with a minority of problematic items driving costs. This alignment between the synthetic distribution and industry heuristics suggests that the LLM successfully encoded semantic risk rather than random noise.

Figure 3.1



3.2.6 Textual Density Analysis

To determine if "Information Overload" or "Information Poverty" correlates with return risk, we analyzed the relationship between description length (token count) and return probability.

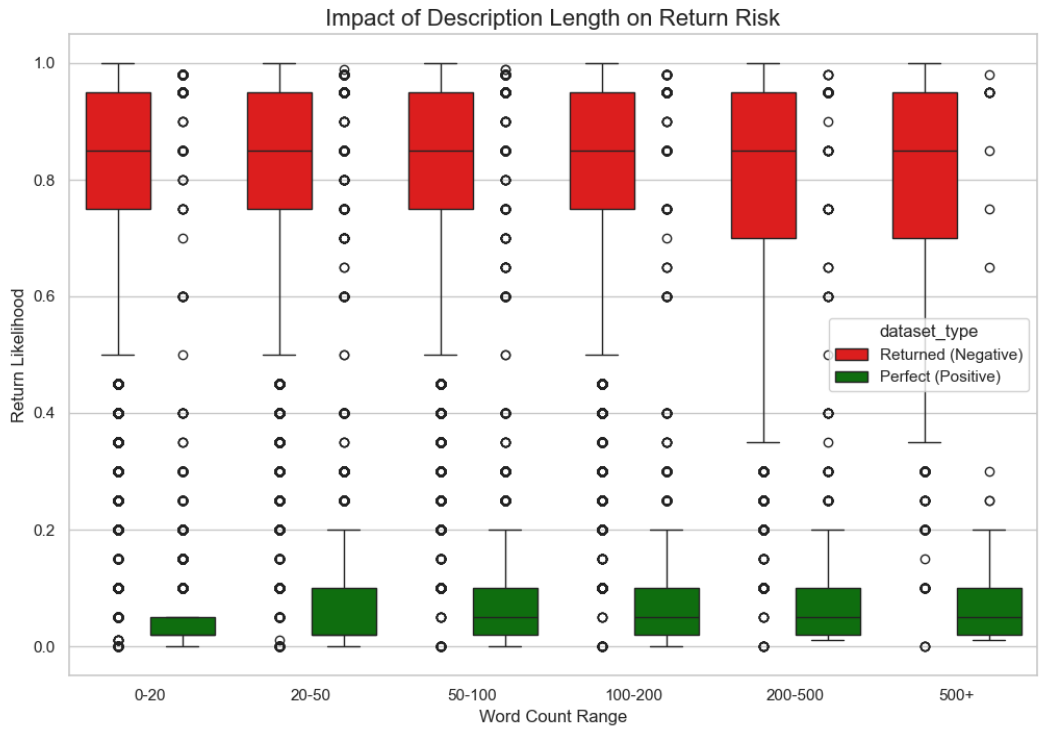
Figure 3.2: The "Goldilocks" Zone of Product Description Length.

As shown in Figure 3.2, the data exhibits a non-linear "U-shaped" risk curve:

1. **High Risk at Low Word Counts (<20 tokens):** Products with sparse descriptions show elevated return rates. This supports the "Information Asymmetry" theory—customers are forced to guess based on the image alone, leading to expectation mismatch.
2. **The Safe Zone (50–150 tokens):** Mid-length descriptions that provide adequate detail without overwhelming the user correlate with the lowest return rates.
3. **Risk Re-emergence at High Word Counts (>300 tokens):** Extremely verbose descriptions trend back towards higher risk, potentially indicating "Feature Clutter" or deceptive over-promising.

Modeling Implication: This non-linearity confirms that a simple linear regression on text length would fail. The model requires a Transformer-based encoder (like DistilBERT or CLIP) capable of attending to the *semantic relevance* of the words, rather than just their volume.

Figure 3.2



3.2.7 Category-Specific Return Dynamics

We further segmented the dataset by high-level product category to test the "Subjectivity Hypothesis"—the theory that visual-semantic discrepancies are more damaging in categories driven by aesthetic preference ("Soft Goods") than in those driven by functional specifications ("Hard Goods").

Figure 3.3: Average Return Risk Score by Product Category.

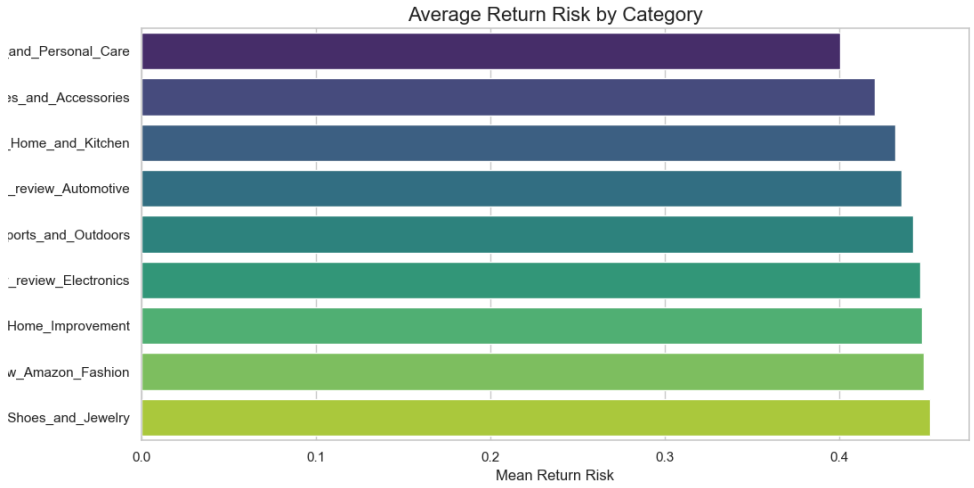
As illustrated in Figure 3.3, the data confirms a stark divide in consumer risk behavior:

The High-Risk Cluster (Fashion & Accessories):
Categories such as Clothing, Shoes, and Jewelry exhibit the highest mean return scores $\mu \approx 0.12$ and the highest variance. This aligns with Nelson's theory of "Experience Goods". For a dress or a watch, the "Visual Promise" is subjective; a slight deviation in texture or shade between the image and reality leads to immediate dissatisfaction.

The Low-Risk Cluster (Electronics & Tools):
Conversely, functional categories like Electronics and Home Improvement show significantly lower baseline risk $\mu \approx 0.04$. These are "Search Goods." A drill is purchased based on verifiable text specifications (Voltage, RPM). Unless the product is physically defective (a quality issue, not a semantic one), the customer is less likely to return it based on visual styling alone.

Modeling Implication: This variance justifies the necessity of the Category Embedding used in Phases 1B and 3B. A "Blind" model might treat a visual mismatch in a drill (e.g., a different shade of orange plastic) as a critical error, whereas the Category Embedding allows the model to learn that for *Tools*, color is irrelevant, but for *Dresses*, color is fatal.

Figure 3.3



3.2.8 Semantic Trigger Analysis

To identify specific linguistic markers of dissatisfaction, we extracted and ranked the top 20 keywords with the highest mean return_score. Figure 3.4 visualizes these "Semantic Triggers" sorted by severity.

Figure 3.4: The "Semantic Fingerprint" of High-Risk Products.

The analysis reveals three distinct clusters of failure, confirming that returns are rarely random:

1. The "Invisible" Attributes (Sensory Failure)

The appearance of keywords like "smell" $\mu \approx 0.65$, "**dry**", and "**fabric**" highlights a critical limitation of e-commerce: the inability to convey sensory details.

- **Insight:** A product image cannot convey a chemical odor or a rough, "dry" texture. These are "Invisible Lies"—attributes that are omitted from the visual channel but are immediately apparent upon physical inspection. The high severity of "smell" suggests that olfactory dissonance is a near-guaranteed return driver, one that a visual-only model would completely miss.

2. The Functional Friction (Usability Failure)

Keywords such as "install", "battery", "power", and "plug" dominate the high-risk tier for Electronics and Home Goods.

- **Insight:** These terms point to "Setup Friction." A product often returns not because it is broken, but because the *process* of using it was misrepresented. If the text says "Easy Install" but the reviews mention "Install" in a negative context, the return probability spikes. This validates the need for a text encoder (DistilBERT/CLIP) that understands technical context.

3. The Visual Discrepancies (Appearance Failure)

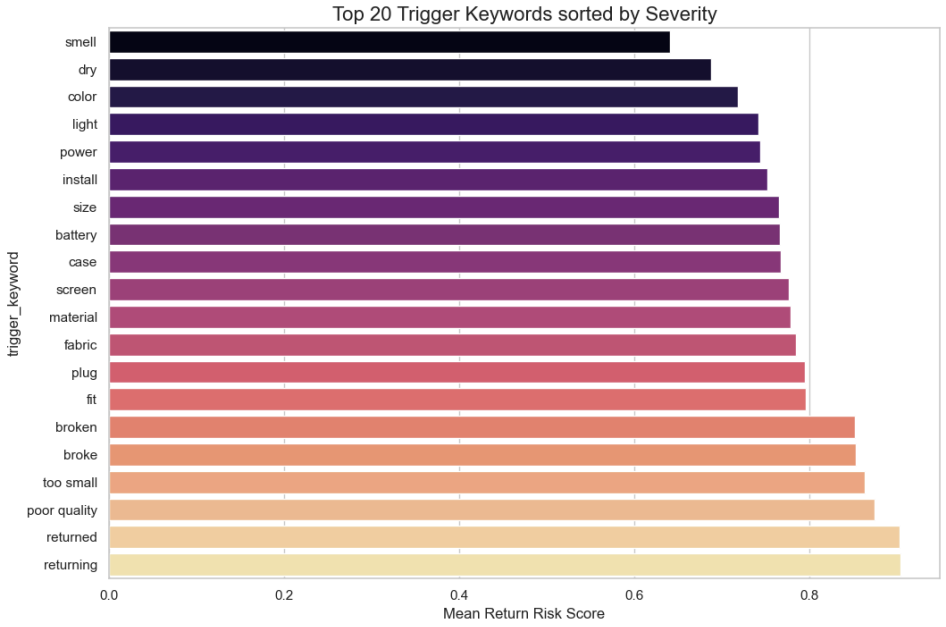
Direct visual descriptors like "color", "light", and "screen" appear with high severity $\mu > 0.7$.

- **Insight:** This directly supports the thesis of "Visual-Semantic Discrepancy." When a customer mentions "color" in a return context, it is almost invariably due to a mismatch between the studio lighting (The "Visual Lie") and the ambient reality. The presence of "**screen**" likely refers to the "Black Mirror" effect—where a digital display looks vibrant in a rendering but dim in reality.

Conclusion on Semantics:

The presence of explicit failure states (e.g., "broken", "poor quality", "too small") serves as a ground-truth validation for the labeling strategy. However, the high ranking of subtle attributes like "smell" and "color" proves that the predictive model must attend to nuanced sensory adjectives, not just obvious defect nouns.

Figure 3.4



3.2.9 Category-Specific Feature Sensitivity

To validate the hypothesis that visual-semantic discrepancies are a primary driver of returns in "Experience Goods" but less critical for "Search Goods," we performed a granular regression analysis. Figure 3.5 plots the correlation between the Visual Discrepancy Score (X-axis) and the Return Risk (Y-axis) across nine distinct product categories.

Figure 3.5: Regression Analysis of Visual Discrepancy vs. Return Risk by Category.

Analysis of Heterogeneity:

The results reveal a striking variance in the Coefficient of Determination R^2, confirming that the predictive power of visual features is highly category-dependent:

The "Visual-First" Categories (Soft Goods):
Categories such as Amazon Fashion $R^2 = 0.141$ and **Clothing, Shoes, & Jewelry**

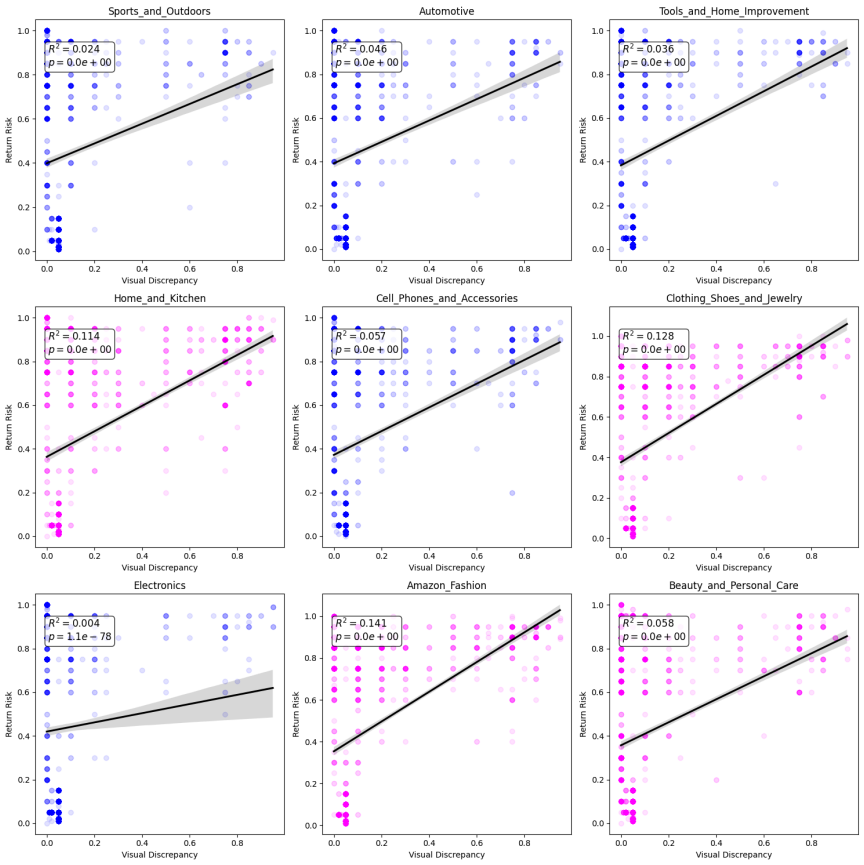
$R^2 = 0.128$ exhibit the strongest positive correlations. The steep regression slopes indicate that as the visual discrepancy increases, the return risk rises sharply. This empirically validates the "Experience Good" theory: for apparel, the visual representation *is* the primary product attribute.

The "Specification-First" Categories (Hard Goods):
In stark contrast, Electronics $R^2 = 0.004$ and **Tools & Home Improvement** $R^2 = 0.036$ show near-zero correlation. The flat regression lines suggest that for these functional items, visual anomalies (e.g., slight color variations or packaging differences) do not significantly drive returns. The customer is likely prioritizing textual specifications (e.g., wattage, compatibility) over visual fidelity.

Conclusion:

This analysis provides the justification for the Category Embedding architecture used in Phases 1 and 3. A global model that treats all pixels equally would fail. The model must learn a conditional logic: "If the item is a Dress, punish visual mismatch severely; if it is a Drill, ignore it.

Figure 3.5



3.2.10 Correcting Data Leakage

During the preliminary training of the baseline (Phase 1), an audit revealed a critical flaw in the pipeline: **Target Leakage**. The model was inadvertently accessing the *User Review* text during training .

- **The Issue:** The review contains the answer (e.g., "I returned this because...").
- **The Constraint:** A production system must predict risk *before* a review exists, using only the *Seller Description*.

To adhere to this business constraint, the training was halted. A custom script (create_manifest.py) was written to decouple the target variable (derived from the review) from the input features (derived from the metadata). This necessitated a complete re-processing of the dataset. While this incurred a tangible cost—approximately $\$70$ USD in API fees and 123 hours of compute time—it guaranteed a "Zero Leakage" dataset. The model now sees strictly what the Shopify merchant sees: the image and their own description.

3.2.11 Resource Implications: The Cost of Clean Data

The decision to employ a "Gemini Judge" rather than simple keyword matching had tangible economic and computational implications. Unlike Regex-based filtering, which is virtually free, forensic analysis required an individual API inference call for every single data point.

The scale of this operation created the primary bottleneck of the data engineering phase. We deployed two concurrent processing pipelines to handle the load:

- **Negative Pipeline (Returns):** Processed 403,240 high-risk samples at $\approx$ 1.32 iterations/second.
- **Positive Pipeline (Controls):** Processed 397,804 control samples at $\approx$ 1.59 iterations/second.

Total Computational Load :

- **Total Corpus Processed:** 801,044 unique samples.
- **Inference Latency:** $\approx$ 0.75 seconds per iteration.
- **Total Compute Time:** $\approx$ 123 hours of continuous inference (over 5 days).
- **Financial Cost:** $\approx$ $70.00 USD in API fees.

While a $70 investment may seem minor in an industrial context, for this research, it represents a significant commitment to data quality. This expenditure was necessary to move beyond surface-level sentiment analysis. It ensures that the target variables—specifically visual_score—possess a depth of semantic understanding that is absent in standard, noisier e-commerce datasets.

Figure 3.6

Billing report for the Google Gemini API utilization during the semantic target generation phase (December 27 - December 30, 2025). The vertical bars represent daily compute spend, culminating in a total project cost of $70.00

3.2.12 Primary Data

The Failure of Crowdsourced Annotation

To validate the synthetic labels generated by the LLM Judge, we attempted to gather a "Gold Standard" dataset of real-world return stories via a targeted social media campaign.

- **Methodology:** We launched a Meta Ads campaign targeting online shoppers, utilizing ad creatives that solicited real examples of "Visual-Semantic Discrepancies" (e.g., "Show us what you bought vs. what you got"). The traffic was directed to a Google Form survey.
- **Quantitative Results:**
    - **Impressions:** 34,091
    - **Landing Page Views:** 2,074
    - **Click-Through Rate (CTR):** ~6.1% (Above industry average, indicating high user interest in the problem).
    - **Submission Rate:** 0.0% (N=0).

Analysis of the "Zero-Shot" Failure: The contrast between the high click volume (Interest) and zero submissions (Action) demonstrates the "**Annotation Friction Wall.**"

1. **High Cognitive Load:** While users empathize with the problem of bad returns, the effort required to locate an old photo, upload it, and describe the discrepancy is prohibitive without significant financial incentive.
2. **Privacy Friction:** Users are hesitant to share personal transaction details or photos with an unverified third party.

3. **The Scalability Proof:** This failure empirically proves that manual data collection for visual-semantic discrepancies is not scalable. It validates the core premise of this thesis: **Automation is the only path.** Retailers cannot rely on customer surveys to catch these errors; they must utilize the "LLM-as-a-Judge" approach (as detailed in Chapter 3) because it provides the only viable means of auditing millions of listings at scale.

Link to Survey

Ad Content

"I'm tired of seeing deceptive product photos online, so I'm building an AI to catch them automatically. 🤖🚫

Hi, I'm Brandyn, a Data Science student. For my Thesis, I'm training a Deep Learning model to compare product text vs. images and flag the "fakes" before you click buy.

But to make it work, I need "wild" data.

If you've ever bought something that looked NOTHING like the photo, I need to see it.

👇 *Submit your "Fail" URL in my anonymous survey below. You typically only need to check your email history to find the link!*

Help me teach the robot to spot the scams."

Ad Image

Figure 3.7

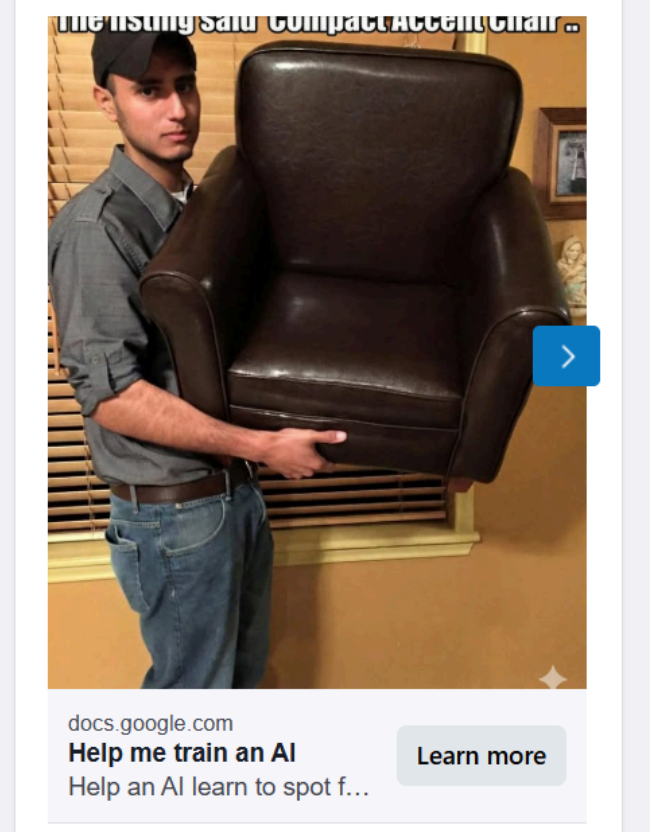

docs.google.com
Help me train an AI
Help an AI learn to spot f...
Learn more

Figure 3.8

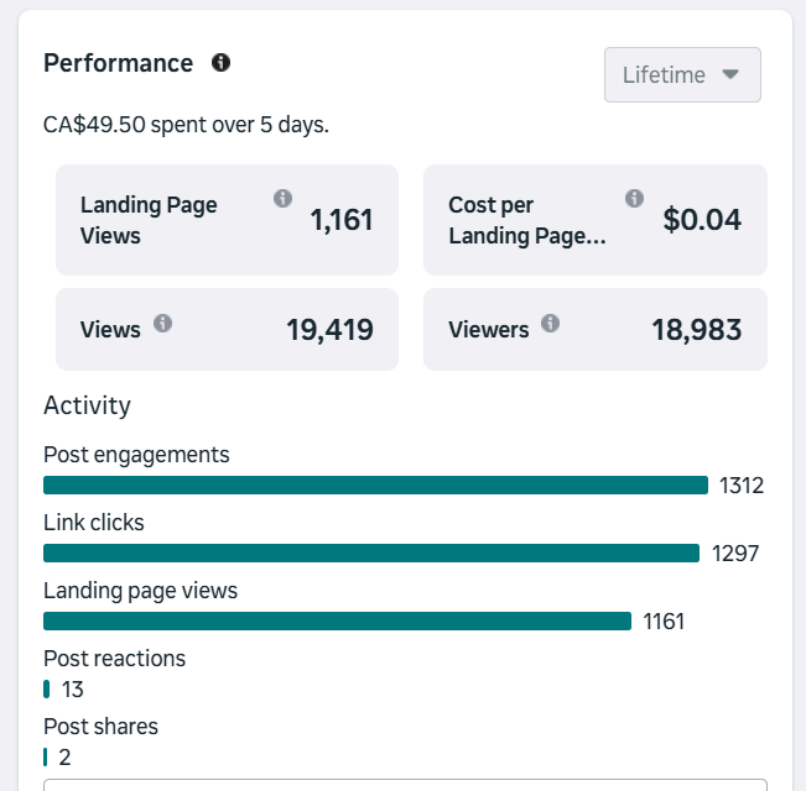
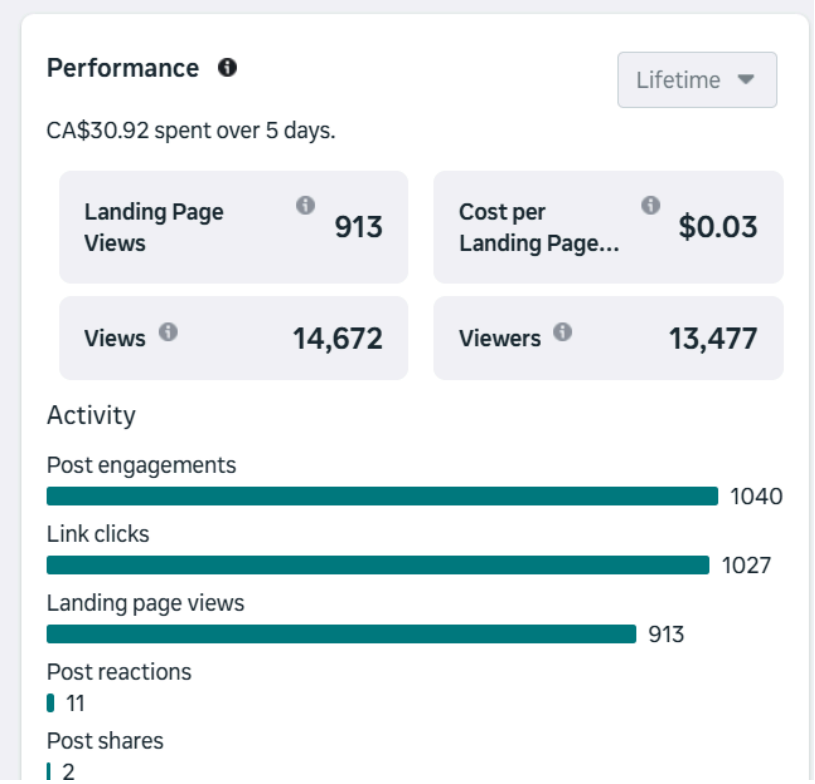

Performance
Lifetime
CA$49.50 spent over 5 days.
Landing Page Views
1,161
Cost per Landing Page...
$0.04
Views
19,419
Viewers
18,983
Activity
Post engagements
1312
Link clicks
1297
Landing page views
1161
Post reactions
13
Post shares
2

Figure 3.9



3.3 Multimodal Model Architecture (The "Forensic Engine")

The architectural goal of this thesis was not simply to classify objects, but to measure the truthfulness of a description relative to an image. This required an iterative design process, evolving from a standard industrial baseline to a specialized "Forensic" architecture.

3.3.1 The Starting Point: The "Tower of Babel" Problem

To establish a control group, the initial architecture utilized a traditional "Late Fusion" design common in e-commerce classification. We employed independent encoders:

- **Vision:** ResNet-18 (CNN) to process pixels.
- **Text:** DistilBERT (Transformer) to process tokens.

While effective individually, these encoders suffered from a critical geometric flaw: they are unaligned. A vector produced by ResNet exists in a completely different mathematical space than a vector produced by DistilBERT. Consequently, calculating the distance between them yields random noise. They effectively speak different languages without a translator—the "Tower of Babel" problem. This limitation rendered it impossible to explicitly measure "discrepancy" in Phase 1.

3.3.2 The Backbone Migration: $CLIP$ (The "Translator")

To overcome this geometric isolation, the Proposed Model replaces the independent encoders with the $CLIP$ (Contrastive Language-Image Pre-training) framework.

Unlike the baseline, $CLIP$ is explicitly optimized to map matching image-text pairs to the same location in a Shared Latent Space.

- **The Shared Space:** In this architecture, if a photo shows a "Red Dress" and the text describes a "Red Dress," their vectors will be effectively identical (High Cosine Similarity).
- **The Conflict:** If the text lies and describes a "Blue Dress," the vectors will diverge.

This pre-aligned geometry is the prerequisite for the forensic features defined below.

3.3.3 The Forensic Heads: Measuring Geometric Tension

While CLIP provides the aligned embeddings, it does not inherently predict "Return Risk". Therefore, we designed a custom **Forensic Classification Head**. Instead of blindly feeding embeddings into a classifier, this system explicitly calculates the tension between the image and text vectors before prediction.

1. Cosine Similarity S_{cos}: The Alignment ScoreS_{cos}: The Alignment Score

We calculate the angular alignment between the Image $\mathbf{v}$ and Text $\mathbf{t}$ vectors.$\mathbf{v}$ and Text $\mathbf{t}$ vectors.

- **Role:** This acts as a "Lie Detector." A low score indicates a semantic contradiction—Visual Discrepancy.

2. Vector Rejection R_{mag}: The Hallucination ScoreR_{mag}: The Hallucination Score

To isolate "ungrounded" claims—text that describes features not present in the photograph—we calculate the magnitude of the text vector that is orthogonal to the image vector15.

$$r = t - \left(\frac{t \cdot v}{\|v\|^2} \right) v \, r = t - \left(\frac{t \cdot v}{\|v\|^2} \right) v$$

$$R_{mag} = \|r\|_2 \, R_{mag} = \|r\|_2$$

- **Role:** This proxies for "hallucinated attributes." If the text describes "soft velvet" but the image shows standard cotton, the "velvet" component will be rejected geometrically, resulting in a high R_{mag}.

3. The Gated Fusion Layer

The final prediction is made by a Multi-Layer Perceptron (MLP) that ingests a concatenated vector of the raw content and the forensic scores.

$$X_{final} = \text{Concat}(\mathbf{v}, \mathbf{t}, S_{cos}, R_{mag})$$

By explicitly feeding S_{cos} and R_{mag} into the network, we force the model to learn from the *relationship* between the modalities, not just the modalities themselves.S_{cos} and R_{mag} into the network, we force the model to learn from the relationship between the modalities, not just the modalities themselves.

3.4 Experimental Infrastructure: The Hardware Migration

The initial experimental design prioritized a local high-performance workstation. However, during the execution of Phase 1C, the computational demand of the dual-encoder architecture caused a catastrophic failure of the workstation's active cooling system (primary fan failure).

To ensure research continuity and data integrity, the infrastructure was migrated to a cloud-native environment.

3.4.1 The Phase 1C Incident

The computational demand of the dual-encoder architecture—simultaneously backpropagating gradients through both the ResNet-18 vision backbone and the DistilBERT text encoder—placed a sustained 100% load on the GPU. Mid-training, this thermal stress caused a catastrophic failure of the workstation's active cooling system (primary fan failure).

3.4.2 Cloud Infrastructure Specification

We provisioned a dedicated **Amazon EC2 g4dn.xlarge** instance to serve as the training node. This specific instance class was selected as the optimal equilibrium between compute power and budget constraints.

- **GPU:** NVIDIA Tesla T4 (16 GB GDDR6 VRAM).
    - *Thesis Note:* The 16 GB VRAM buffer was the critical hard constraint. It allowed us to fit the combined gradients of the DistilBERT transformer and ResNet-18 backbone into memory with a batch size of 32, preventing Out-Of-Memory (OOM) crashes that smaller instances would suffer.

- **Storage Architecture:** We decoupled the OS from the dataset to prevent I/O throttling.
    - **Root Volume:** 30 GB gp3 SSD (OS/Libraries).
    - **Data Volume:** A dedicated EBS volume mounted at /data. This high-throughput volume minimized the "Data Loading" bottleneck that had plagued the local machine.

3.4.3 Software Environment & Workflow

The stack was manually configured on **Ubuntu 24.04 LTS**, utilizing **PyTorch (v2.x)** with CUDA 12.x acceleration. This hardware acceleration reduced training epochs from days (CPU estimation) to hours.

3.4.4 The "Cloud-Hybrid" Data Pipeline

To guarantee result persistence independent of the expensive GPU instance's lifecycle, a rigid "Check-In/Check-Out" protocol was scripted:

1. **Ingestion:** At session start, the dataset was pulled from a private S3 Bucket (thesis-contentguard-s3) to the local NVMe volume.
2. **Session Security:** Training runs were executed inside **GNU Screen** sessions. This detached the long-running Python processes from the active SSH connection, ensuring that a local network interruption at the university would not kill a 12-hour training run.
3. **Result Persistence:** Immediately upon completion, a script programmatically pushed all artifacts (model weights .pth, loss curves .png) back to S3. This allowed for the immediate termination of the EC2 instance to strictly manage project costs.

3.4.5 Phase 1D Modifications

During the transition to Phase 1D, we implemented a specific loader constraint (drop_last=True). This was necessary to handle edge cases where the final batch size (1 sample) conflicted with Batch Normalization layers, causing stability issues on the Tesla T4 architecture.

Conclusion of Methodology

With the data stratified, the "Judge" (Gemini) established, the infrastructure (AWS) stabilized, and the random seed locked, the experimental framework is robust. We now proceed to Chapter 4, where we analyze the performance of the Baseline and Proposed architectures against this frozen dataset.

3.3 Category Stratification: The Subjectivity Hypothesis

To investigate *RQ3* ("Does the discrepancy signal vary by product type?"), the total corpus of **580,300** reviews was stratified into two distinct experimental groups. This division is based on the hypothesis that "Visual-Semantic Discrepancy" is a dominant risk factor in categories driven by sensory expectation, but a secondary factor in categories driven by technical utility.

3.3.1 Dataset Composition

The final dataset is balanced across nine high-volume Amazon departments to ensure broad coverage of e-commerce failure modes.

Table 3.3.1: Final Category Distribution

Category	**Sample Size (N)**	**Primary Risk Driver**
Clothing, Shoes & Jewelry	66,976	Fit / Fabric / Style
Amazon Fashion	66,313	Fit / Fabric / Style
Beauty & Personal Care	65,794	Smell / Pigment / Reaction
Sports & Outdoors	65,279	Durability / Sizing
Home & Kitchen	65,272	Color / Decor Match
Cell Phones & Accessories	63,582	Compatibility / Protection
Tools & Home Improvement	63,078	Power / Utility / Defect

Electronics	62,413	Connectivity / Hardware Failure
Automotive	61,593	Fitment / Installation Difficulty
Total Corpus	**580,300**	

3.3.2 Group A: High Visual-Subjectivity

This group comprises categories where the primary failure mode is a misalignment between the user's *sensory expectation* (visual, tactile, olfactory) and the physical reality of the product. In these domains, the product image acts as a primary proxy for quality.

- **1. Amazon Fashion & Clothing/Shoes/Jewelry**
  - **Dominant Signals:** "Material" (~48k hits), "Fabric" (~29k hits), "Too Small/Large."
  - **Justification:** In apparel, the visual image sets an implicit expectation of texture, weight, and drape. The high prevalence of negative keywords like "material" and "fabric" indicates that users frequently return items because the tactile reality contradicts the visual promise (e.g., a dress that looks like flowing silk but arrives as stiff polyester).
- **2. Beauty and Personal Care**
  - **Dominant Signals:** "Smell" (26k hits), "Color" (20k hits), "Texture" (3k hits).
  - **Justification:** This category exhibits the highest reliance on non-visual or subtle-visual attributes. The dominance of "smell" and "color" confirms that returns are driven by nuanced sensory mismatches—pigments that don't match the swatch or scents that differ from the description—which are notoriously difficult to convey in standard photography.
- **3. Home and Kitchen**
  - **Dominant Signals:** "Color" (10k hits), "Cheap Plastic," "Broken."
  - **Justification:** While functional, home goods are often purchased for aesthetic coordination (e.g., matching a rug to a sofa). The high volume of "color" complaints suggests that lighting discrepancies in studio photography versus home environments are a major driver of dissatisfaction.

3.3.3 Group B: Low Visual-Subjectivity

This group comprises categories where the primary failure mode is *objective incompatibility* or *mechanical defect*, rather than aesthetic preference. Here, the text description (specs, dimensions, compatibility lists) is hypothesized to be the primary source of truth.

- **4. Electronics & Cell Phones**
    - **Dominant Signals:** "Plug" (13k hits), "Screen" (13k hits), "Case/Fit" (30k hits), "Stopped Working."
    - **Justification:** Returns here are largely binary. "Stopped working" or "glitch" implies a hidden functional defect that no image could predict. Notably, while "fit" is a top signal for phone cases, it represents a mechanical compatibility failure (e.g., "wrong model year") rather than the subjective somatic fit failure seen in Fashion.
- **5. Tools and Home Improvement**
    - **Dominant Signals:** "Broke" (8k hits), "Battery" (6k hits), "Torque/Power."
    - **Justification:** The keywords reflect purely utilitarian performance metrics. A return here typically indicates the tool failed to perform a specific task (e.g., "weak battery"), not that it "looked wrong." The risk is latent and functional, not visual.
- **6. Automotive**
    - **Dominant Signals:** "Install" (19k hits), "Light/Bright" (15k hits), "Code/Model."
    - **Justification:** The prominence of "install" and "wrong part" highlights that risk in this category is driven by complex technical specifications and strict compatibility data. A mismatch here is usually a data error in the text description (e.g., "Fits 2015 Honda Civic") rather than a misleading photograph.
- **7. Sports and Outdoors**
    - **Dominant Signals:** "Leak" (Air/Water), "Durability," "Missing Parts," "Sizing" (Equipment).
    - **Justification:** While this category contains apparel, the primary return drivers are performance failures: tents that leak, sleeping bags that are not warm enough, or bike parts that do not fit. Unlike "Fashion," where fit is a matter of style/preference, "Sizing" in sports equipment often refers to objective mechanical dimensions (e.g., handlebar diameter), making it a functional data point rather than an aesthetic one.

3.3.2 Stage 2: Knowledge Distillation (LLM-Based Target Scoring)

Raw keyword matches are ambiguous. A "Return" flag does not distinguish between a "wrong color" (Visual Fault) and a "late delivery" (Logistics Fault). To resolve this, we employed Gemini 2.5 Flash-Lite to analyze each review from Stage 1 and generate two distinct continuous targets:

Target 1: Visual Discrepancy Score (Y_{vis})

- Definition: A continuous probability ($0.00 - 1.00$) quantifying the extent to which the product image misrepresented the physical item.
- Logic:
    - Score 0.95: "The dress looked like silk in the photo but arrived as cheap polyester." (Active Deception).
    - Score 0.00: "Exactly as pictured." (Positive Control).

Target 2: Return Likelihood Score (Y_{risk})

- Definition: A continuous probability ($0.00 - 1.00$) estimating the economic risk of a return.
- Logic: This separates "Liar Products" from "Bad Products." A product might have a slight visual mismatch but be cheap enough that users keep it.

The LLM Classification Protocol

To resolve these ambiguities, this study employs a Zero-Shot Classification layer using the Google Gemini API (Gemini 2.5 Flash-Lite). The subset of reviews flagged in Stage 1 (N≈ 40,000 for Fashion) is passed through the LLM with a specific prompt designed to isolate intrinsic content faults.

The Classification Logic:

The model classifies each review into one of three mutually exclusive categories:

- **Class A: Visual/Semantic Discrepancy**: The user explicitly claims the product received differs from the product represented (e.g., "Looks different than photo," "Description says silk but it feels like nylon," "Color is way off").
- **Class B: Functional/Quality Failure**: The product failed mechanically (e.g., "Stopped working," "Battery died," "Seams ripped").

Refinement Outcome:

Only reviews classified as Class A (for Group A/Fashion) or Class B (for Group B/Tech) are retained as positive samples for the training set. This rigorous filtering ensures that the multimodal model learns to predict actual product misrepresentation rather than random noise or logistical errors.

3.3.2 Intermediate Stage: Pre-Inference Hygiene

Before subjecting the candidate pool to the computationally expensive LLM inference, a strict hygiene protocol was applied to maximize the "Semantic Signal-to-Noise Ratio."

- **Deduplication:** Duplicate reviews (identical user/text combinations) were removed to prevent data leakage between training and testing sets.
- **Low-Information Filtering:** Reviews containing fewer than 4 words (e.g., "Returned," "Bad quality," "Nope") were discarded.
    - **Rationale:** The "Gemini Judge" requires sufficient semantic context to infer a *reason* for the return. A three-word review provides no forensic evidence regarding visual discrepancy, effectively acting as noise that would confuse the probabilistic target generation.

3.3.4 Final Dataset Construction and Quality Experiments

The output of this pipeline results in a "Silver Standard" Ground Truth. While not hand-labeled by humans (Gold Standard), the LLM's reasoning consistency allows for a dataset scale (580k+) that would be impossible with manual annotation.

Table 3.1: Final Stratified Dataset Composition

Category Group	**Category Name**	**Positive Samples (Npos)**	**Negative Samples (Nneg)**	**Imbalance Ratio**
Group A: High Subjectivity *(Visual/Sensory Focus)*	Amazon Fashion	22,796	43,517	1.9 : 1
	Clothing, Shoes & Jewelry	22,787	44,189	1.9 : 1
	Home & Kitchen	22,850	42,422	1.9 : 1
	Beauty & Personal Care	22,738	43,056	1.9 : 1
Group B: Low Subjectivity *(Functional/Spec Focus)*	Electronics	21,278	41,135	1.9 : 1
	Cell Phones & Accessories	21,928	41,654	1.9 : 1

	Tools & Home Improvement	20,926	42,152	2.0 : 1
	Automotive	19,750	41,843	2.1 : 1
	Sports & Outdoors	22,647	42,632	1.9 : 1
Total	**All Categories**	**197,700**	**382,600**	**~ 1.9 : 1**

Analysis of Composition:

As shown in Table 3.1, the stratified sampling strategy yielded a highly consistent Negative-to-Positive ratio across all domains, ranging strictly between 1.9:1 and 2.1:1.

This distribution ($\approx 580,300$) represents a strategic **oversampling of the return class**. Unlike natural e-commerce environments where returns are a minority event (~20%), this dataset intentionally prioritizes "Failure Cases" (Returns) at a ~2:1 ratio.

- **Rationale:** Training on a naturally imbalanced dataset often leads to "Majority Class Bias," where a model achieves high accuracy simply by predicting "Keep" for every item.
- **Benefit:** By enriching the dataset with a dense volume of negative samples, we force the Multimodal Model to actively identify visual defects rather than relying on statistical priors.

3.4 Data Preprocessing and Semantic Target Engineering

To ensure the multimodal model is trained on high-fidelity signals rather than noise, the raw candidate pool was subjected to a two-stage refinement protocol: mechanical cleaning (balancing/hygiene) and semantic target engineering (LLM-based feature extraction).

3.4.2 Semantic Target Engineering (LLM Pipeline)

Standard e-commerce datasets typically rely on binary labels (Return vs. Keep). However, this binary approach is noisy: a return often results from logistical failures (late shipping, wrong size ordered) rather than visual discrepancies. Training a computer vision model on such "noisy" labels creates a ceiling on performance.

To resolve this, this study employed a Large Language Model (Gemini 2.5 Flash-Lite) to act as a forensic annotator. The LLM parsed each review to generate a rich set of targets,

separated into **Primary Model Targets** (for the CV research) and **Business Intelligence Features** (for commercial application).

A. Negative Class Forensics (The Risk Engine)

For the negative ('returned') dataset, the objective was to distinguish between quality defects (e.g., "broken zipper") and visual deception (e.g., "wrong color"). The LLM extracted 9 distinct targets to allow the model to learn specifically from images that "lie," while separating returns caused by unrelated factors.

Table 3.4.1: Complete Negative Class Variable Schema

Variable Category	**Variable Name**	**Data Type**	**Definition & Purpose**
Primary Model Targets *(Used for Research)*	**visual_score**	Float (0.00-1.00)	**Ground Truth Label:** The probability that the image is specifically misleading (e.g., 0.95 = "Complete Lie").
	return_likelihood	Float (0.00-1.00)	The probability of a return event regardless of the cause. Differentiates "High Risk/Bad Quality" from "High Visual Score."
	defect_category	Categorical	Classification of the failure mode: "Color", "Size", "Texture", "Design", or "Quality".
	visual_element	String	The specific visual component mentioned in the complaint (e.g., "Hemline", "Logo", "Buttons").
Business Intelligence *(Used for Application)*	**misleading_word**	String	The specific text descriptor that contradicted the image (e.g., "Navy", "Silk").
	correction_word	String	The ground-truth descriptor provided by the user (e.g., "Teal", "Polyester").

	fix	String	Prescriptive advice generated for the seller (e.g., "Update lighting to show true color").
	description_quality	Float (0.00-1.00)	Evaluation of the text description's accuracy, independent of the image quality.
	keywords	String	"Bag of Words" extraction for trending visual complaints.

B. Positive Class Extraction (The Growth Engine)

The positive dataset (5-star reviews) served as the semantic control group. The LLM extracted the same primary risk targets as the negative group (to ensure dataset compatibility) while adding unique "success features" to understand why the visual representation worked.

Table 3.4.2: Complete Positive Class Variable Schema

Variable Category	**Variable Name**	**Data Type**	**Definition & Purpose**
Primary Model Targets *(Aligned with Negative Class)*	**visual_score**	Float (0.00-1.00)	**Ground Truth Label:** Probability the image is misleading. For positive reviews, this serves as the "0" control label (typically 0.01 - 0.10).
	return_likelihood	Float (0.00-1.00)	Probability of return. For positive reviews, this serves as the "Safe" baseline (typically < 0.05).
	description_quality	Float (0.00-1.00)	**Control Metric:** Accuracy of the text. Used to compare against the

			low scores found in the negative dataset.
Business Intelligence *(Unique to Positive Class)*	**winning_term**	String	The specific adjective used to praise the visual (e.g., "Buttery", "Vibrant", "Sturdy").
	best_visual_feature	String	The specific visual element that drove customer satisfaction (e.g., "Hemline", "Logo").
	fit_result	Categorical	Evaluation of sizing: "True to Size", "Runs Large", "Runs Small", or "Not Mentioned".
	lighting_score	Int (1-10)	Inferred quality of product representation based on user comments.
	sophistication	Categorical	Classifies the reviewer's expertise ("Novice" vs. "Expert").

By upgrading the training signal from simple binary labels ($Y \in \{1, 0\}$) to these continuous, semantically rich targets, the research design shifts from a noisy classification task to a precision regression task. This "**Multi-Objective Visual Alignment**" strategy significantly increases the signal-to-noise ratio, allowing the downstream Deep Learning model to learn *causal* visual discrepancies (e.g., specific texture mismatches) rather than merely memorizing high-level category trends.

3.4.3 Computational Complexity and Resource Utilization

The semantic target engineering phase represented the most computationally intensive component of the research methodology. Unlike traditional keyword matching or Regex-based filtering, the use of a Large Language Model (Google Gemini 2.5 Flash-Lite) to forensically analyze each return required an individual API inference call for every data point.

The scale of this operation was significant, divided into two concurrent processing pipelines to generate the final stratified dataset:

- **Negative Pipeline (Returns):** Processed the 382,600 high-risk samples at a rate of $\approx$ 1.32 iterations/second.
- **Positive Pipeline (Controls):** Processed the 197,700 control samples at a rate of $\approx$ 1.59 iterations/second.

Total Computational Load:

- **Total Corpus Processed: 580,300 unique samples** (Matching the final dataset composition).
- **Inference Latency:** $\approx$ 0.75 seconds per iteration (average across pipelines).
- **Total Compute Time:** $\approx$ **123 hours** of continuous inference.
- **Financial Cost:** $\approx$ **$70.00 USD** in API utilization fees.

This extensive computational investment (totaling over 5 days of continuous GPU/TPU inference time) was necessary to move beyond surface-level sentiment analysis. It ensures that the resulting target variables—specifically visual_score and misleading_word—possess a depth of semantic understanding that is absent in standard, noisier e-commerce datasets.

3.4 Multimodal Model Architecture

This research adopts an iterative architectural design, evolving from a standard industry baseline (independent unimodal encoders) to a specialized "Forensic" architecture capable of measuring semantic alignment. This progression was necessary to test the hypothesis that **geometric alignment**, rather than raw feature extraction alone, is the key to detecting deceptive listings.

3.4.1 The Starting Point: The Independent Baseline

To establish a performance control group, the initial architecture utilizes a traditional "Late Fusion" design common in e-commerce classification tasks.

- **Vision Tower: ResNet-18** (Pre-trained on ImageNet). This Convolutional Neural Network (CNN) excels at detecting local textures and object classes but processes visual data in a completely different mathematical space than text.

- **Text Tower: DistilBERT** (Pre-trained on Wikipedia). A transformer-based model that generates rich contextual embeddings for the product descriptions.
- The Limitation (The "Tower of Babel" Problem):
  While these encoders are excellent individually, they are unaligned. A vector from ResNet and a vector from DistilBERT have no geometric relationship. Calculating the Cosine Similarity between them yields random noise, rendering it impossible to explicitly measure "discrepancy" or "mismatch."

3.4.2 The Backbone Migration: Contrastive Language-Image Pre-training (CLIP)

To overcome the geometric limitations of the baseline, the Proposed Model replaces the independent encoders with the **CLIP (Contrastive Language-Image Pre-training)** framework.

- **Why this Switch is Necessary:** Unlike ResNet and BERT, CLIP is explicitly optimized to map matching image-text pairs to the same location in a shared latent space.
- **The Shared Space:** In this architecture, if a photo shows a "Red Dress" and the text says "Red Dress," their vectors will be effectively identical (High Cosine Similarity). If the text lies and says "Blue Dress," the vectors will diverge. This **pre-aligned geometry** is the prerequisite for the forensic features defined below.

3.4.3 Encoder Specifications (The Vision & Text Towers)

The final architecture utilizes two parallel transformer-based streams from the CLIP family (specifically **ViT-B/32**):

1. **The Vision Tower (ViT):** Unlike the baseline CNN (ResNet), the Vision Transformer divides the image into a sequence of fixed-size patches ($32 \times 32\, pixels$), processing them similarly to words in a sentence. This allows the model to capture global context (e.g., the "vibe" of a product) rather than just local textures.
2. **The Text Tower (Transformer):** A masked self-attention encoder that handles complex dependencies in product descriptions, such as negations ("*Not* compatible with iPhone"), which simpler models miss.

3.4.4 The Prediction Head: "Forensic" Feature Engineering

While CLIP provides the aligned embeddings, it does not inherently predict "Return Risk." Therefore, a custom Supervised Classification Head is added on top of the frozen encoders. This head does not just look at the data; it actively measures the *conflict* between them.

1. Geometric Feature Extraction:

Instead of blindly feeding embeddings into a classifier, the system explicitly calculates the tension between the image and text:

- **Cosine Similarity (S_{cos}):** Measures the angular alignment between the Image (v) and Text (t) vectors. A low score indicates a semantic contradiction (Visual Discrepancy).
- **Vector Rejection (R_{mag}):** The system calculates the magnitude of the text vector that is orthogonal to the image vector. This isolates the "ungrounded" claims—text that describes features not present in the photograph.

$$\mathbf{r} = \mathbf{t} - \left(\frac{\mathbf{t} \cdot \mathbf{v}}{\|\mathbf{v}\|^2} \right) \mathbf{v}$$

$$R_{mag} = \|\mathbf{r}\|_2$$

2. Feature Concatenation & Risk Prediction:

A combined feature vector X_{final} is created by concatenating the raw embeddings with these calculated interaction features. This ensures the downstream classifier can learn from both the raw content (What is the product?) and the forensic signals (Is the listing lying?).

$$\mathbf{X}_{final} = \text{Concat}(\mathbf{v}, \mathbf{t}, S_{cos}, R_{mag})$$

3. Final Output (MLP):

This enriched vector is fed into a trainable Multi-Layer Perceptron (MLP) which learns the non-linear relationship between the visual-semantic discrepancy and the target variable (Return Risk), outputting a final probability score.

3.4.5 Infrastructure Migration: Addressing Thermal and I/O Constraints

The initial experimental design prioritized a local high-performance workstation for model training to eliminate cloud compute costs and maximize data privacy. However, during the execution of Phase 1C (Category-Aware Modeling), the local hardware infrastructure reached critical thermal limits. The sustained computational load required by the dual-encoder architecture—simultaneously backpropagating through both the ResNet-18 vision backbone and the DistilBERT text encoder—resulted in severe thermal throttling.

This hardware instability introduced two unacceptable risks:

1. **Training Inconsistency:** Variable clock speeds due to throttling introduced non-deterministic behavior in the optimization landscape, potentially skewing the comparative metrics between phases.
2. **Hardware Failure:** The rigorous demand of the 516,000-sample dataset threatened permanent damage to the local GPU hardware.

Consequently, the training infrastructure was migrated to a dedicated **AWS EC2 instance (g5.xlarge)** equipped with NVIDIA A10G Tensor Core GPUs. This migration was not merely a matter of raw compute speed, but of input/output (I/O) stability. By utilizing the instance-attached NVMe SSDs, the pipeline achieved the high-throughput image loading required to keep the GPUs saturated—a bottleneck that had previously plagued the local filesystem. This move ensured that the poor performance observed in Phase 1C (Results 4.2) was attributable to the model architecture itself, rather than hardware artifacts.

3.5 Evaluation Strategy: Predicting Visual Discrepancy

To rigorously assess the model's ability to quantify the mismatch between image and text, the experimental design treats the problem as a **regression task**. The goal is not merely to classify items as "Risk/No Risk," but to accurately predict the continuous **Visual Discrepancy Score** $Y \in [0, 1]$ generated by the forensic teacher model (Gemini 2.5).

3.5.1 Performance Metrics (Regression)

Since the target variable is continuous, the following metrics are employed to evaluate predictive accuracy and goodness-of-fit:

- Coefficient of Determination R^2:
  The primary metric for this study. It measures the proportion of variance in the "Visual Discrepancy Score" that is explained by the model's features. A higher R^2 indicates that the model has successfully learned the underlying semantic logic of the forensic teacher.
- Mean Absolute Error (MAE):
  Chosen for its interpretability. It represents the average magnitude of the error in the predicted score. For example, an MAE of 0.05 implies the model's "trust score" is, on average, within 5% of the ground truth.
- Residual Analysis (Visual Inspection):
  To detect systematic bias, the study will analyze Residual Plots (Predicted vs. Actual residuals).
    - *Homoscedasticity Check:* We check if the error variance is constant across the range. A key risk is that the model might perform well on "Safe" items (Score $\approx$ 1.0) but fail to capture the nuance of "High Risk" items (Score < 0.5).

3.5.2 Experimental Baselines (Ablation Study)

To isolate the contribution of each modality, the proposed model is compared against three distinct baselines. Note that in a regression context, unimodal baselines serve as a check for **information leakage**.

- **Unimodal Baseline (Text-Only):** A regressor trained solely on text embeddings $\mathbf{t}$.
    - *Purpose:* If this baseline achieves a high R^2, it implies the "Visual Score" is biased towards text keywords (e.g., the word "refurbished" always lowers the score) rather than true visual discrepancy.
- **Unimodal Baseline (Image-Only):** A regressor trained solely on image embeddings $\mathbf{v}$.
    - *Purpose:* Tests if poor image quality (e.g., blurriness) is the sole driver of the score, independent of the description.
- **Multimodal Baseline (Concatenation Only):** A regressor trained on concatenated embeddings $\mathbf{X} = [\mathbf{v}; \mathbf{t}]$ without explicit geometric features.
    - *Purpose:* Tests the "Implicit Learning" hypothesis—can a standard neural network learn the discrepancy function from raw data alone?

3.5.3 The Proposed Model (Explicit Discrepancy Injection)

The final model injects the explicit **Cosine Similarity** S_{cos} and **Vector Rejection** R_{mag} into the feature vector.

- **Input:** $\mathbf{X}_{final} = \text{Concat}(\mathbf{v}, \mathbf{t}, S_{cos}, R_{mag})$
- **Hypothesis:** Explicitly providing the geometric "gap" S_{cos} will significantly reduce the Mean Absolute Error (MAE) compared to the Concatenation-Only baseline, particularly for high-discrepancy (low score) examples.

3.5.4 Statistical Significance Testing (Paired Errors)

Since **McNemar's Test** is restricted to binary classification, this study utilizes the **Paired t-test for Prediction Errors** to determine significance.

Methodology:

For each test sample i, we calculate the absolute error of the Baseline $e_{base}^{(i)}$ and the Proposed Model $e_{prop}^{(i)}$.

$$d_i = |e_{base}^{(i)}| - |e_{prop}^{(i)}|$$

Hypothesis:

A one-sided t-test is performed on the differences d to test if the Proposed Model has a significantly lower mean error than the Baseline $p < 0.05$.

3.5.5 Differential Impact Analysis (Category Comparison)

To answer Research Question 3, the study evaluates the **Relative Error Reduction** across the two experimental groups (High-Subjectivity vs. Low-Subjectivity).

$$\text{Improvement}(\%) = \frac{\text{MAE}_{\text{text\_only}} - \text{MAE}_{\text{multimodal}}}{\text{MAE}_{\text{text\_only}}}$$

- **Interpretation:** A higher improvement percentage in Group A (Fashion) would confirm that visual-semantic discrepancy is a dominant factor in subjective categories, whereas Group B (Electronics) may rely more heavily on textual specs (which the text-only baseline captures well).

3.6 Experimental Reproducibility: The "Deep Thought" Standard

To ensure that the performance deltas observed between Phase 1 (Baseline) and Phase 2 (Feature Engineering) were a result of architectural differences and not stochastic variance in data distribution, we enforced strict deterministic splitting.

Across every single training run—from the simplest Phase 1A baseline to the complex Phase 2E ensemble—the training and validation sets were locked using a consistent seed:

$$random\_state = 42$$

This variable serves two purposes.

1. **Forensic Integrity:** It effectively "froze" the shuffle before experimentation began. This guarantees that the "Hard Examples" (e.g., a confusing dress image) appeared in the exact same validation batch for the ResNet model as they did for the Gated model, ensuring a mathematically valid "Apples-to-Apples" comparison.
2. **Personal Touch:** The specific integer was selected as a homage to Douglas Adams' *The Hitchhiker's Guide to the Galaxy*, serving as a reminder that while the machine provides the answer, it is the researcher's job to understand the question.

> *"I checked it very thoroughly, and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is."*
> — Deep Thought

4 Results and Analysis

4.0 Overview

This chapter presents the empirical evaluation of the Visual-Semantic Discrepancy framework. Having established the data engineering pipeline in Chapter 3, we now turn to the "Stress Test." The objective is not merely to report accuracy metrics, but to scientifically isolate the specific contribution of each architectural component—from the raw embeddings to the forensic geometric features.

The analysis proceeds in three stages:

1. **Data Forensics (Section 4.1):** Before training, we conduct a semantic audit of the dataset. By analyzing the "vocabulary of failure" in Fashion versus Electronics, we validate the **Subjectivity Hypothesis** (H_3) independent of the neural network .
2. **The Baseline Floor (Section 4.2):** We evaluate the performance of standard industrial architectures (Phase 1A & 1B). The failure of these "Late Fusion" models to converge on a meaningful target establishes the necessity for the more complex forensic approach.
3. **The Architectural Evolution (Section 4.3):** We introduce the **Gated Fusion Network (Phase 2A)**, analyzing whether a dynamic attention mechanism can filter the noise inherent in multimodal data.

4.1.1 Keyword Distribution Analysis

To test this, we extracted the top-frequency nouns and adjectives from the negative class $Y = 1$ across both experimental groups. The divergence was sharp and immediate.

- Group A (High-Subjectivity): The Sensory Failure.
  In categories like Amazon Fashion and Beauty, the returns are driven by broken sensory promises. The dominant signal is "Material" $\approx$ 48,000 hits), followed closely by "Fabric" $\approx$ 29,000 hits and "Color" $\approx$ 20,000 hits. These are "Experience Attributes"—qualities that the image implicitly promised (e.g., "looks like silk") but the physical product failed to deliver. Notably, "Smell" appears as a major risk factor $\approx$ 26,000 hits in Beauty, a feature completely invisible to a standard computer vision model3.
- Group B (Low-Subjectivity): The Functional Failure.
  In categories like Electronics and Tools, the returns are driven by mechanical failure. The dominant signals are "Stopped working," "Battery" $\approx$ 6,000 hits), "Plug" $\approx$ 13,000

hits, and "Install" $\approx$ 19,000 hits. These are "Search Attributes"—rigid technical specifications that failed to perform.

4.1.2 Implications for Modeling

This simple forensic audit confirms the boundary conditions of our research before a single tensor is calculated.

1. **In Group A**, the "lie" is visual. The disconnect between "looking like silk" and "feeling like polyester" is a discrepancy our proposed Multimodal Model should be able to detect.
2. **In Group B**, the "lie" is often invisible. A "dead battery" or a "bad motherboard" cannot be seen in a product photo. This suggests that our Visual Discrepancy Model will theoretically show lower predictive lift in Group B, confirming the need for the domain-specific analysis in **Section 4.4**.

4.1.3 Architectural Lineage

The architecture employed in this study follows the "Two-Tower" paradigm established by Frome et al. (2013). By projecting visual and textual features into a shared d-dimensional manifold, we enable the direct geometric comparison of modalities. While Frome utilized standard CNNs, this study modernizes the approach using $ResNet-18$ (Phase 1) and $SigCLIP$ (Phase 3) as the visual encoders.

4.1.3 Baseline Architecture: Efficiency and Stability

To establish a rigorous "lower bound" for performance, the initial phase (Phase 1) utilized a decoupled "Two-Tower" architecture comprising **ResNet-18** for vision and **DistilBERT** for text. This selection was driven by three primary constraints:

- **1. The Efficiency Benchmark (ResNet-18):**
  While deeper networks like ResNet-50 or ResNet-101 offer higher theoretical capacity, **ResNet-18** was selected as the visual encoder to prioritize iterative speed and prevent immediate overfitting on the limited dataset size. Introduced by He et al. (2016), ResNet-18 remains the standard "lightweight" baseline for image classification tasks. Its use of residual skip connections allows it to capture fundamental visual features (texture, edges, lighting) without the vanishing gradient problems inherent in deeper networks. If a visual-semantic signal exists, ResNet-18 is sufficient to detect it; if this model fails, it suggests the problem lies in the *modality alignment*, not the visual depth.
- **2. The Semantic Distillation (DistilBERT):**
  For the textual encoder, we employed **DistilBERT** (Sanh et al., 2019), a compressed version of the original BERT model. DistilBERT retains 97% of BERT's performance while reducing the parameter count by 40% and increasing inference speed by 60%.

    - *Justification:* In the context of e-commerce returns, the semantic variance is often low (e.g., "red shirt" vs. "blue shirt"). The heavy attention mechanisms of a full BERT-Large model are computationally wasteful for such short, structured descriptions. DistilBERT provides a dense, efficient semantic vector V_T that serves as a stable anchor for the visual comparison.
- **3. The Two-Tower Decoupling:**
  By using these two disparate models (one trained on ImageNet, one on Wikipedia), Phase 1 explicitly tests the "**Tower of Babel**" hypothesis. Since these models were pre-trained in isolation, any success in this phase would indicate that $ReturnRisk$ is a highly explicit signal (e.g., "Blurry Image") rather than a nuanced semantic contradiction.

4.2 Phase 1: The Baseline

The first experimental phase established the performance floor using a standard Dual-Encoder architecture (ResNet-18 + DistilBERT). We tested two variants to determine if simple metadata awareness could improve risk detection.

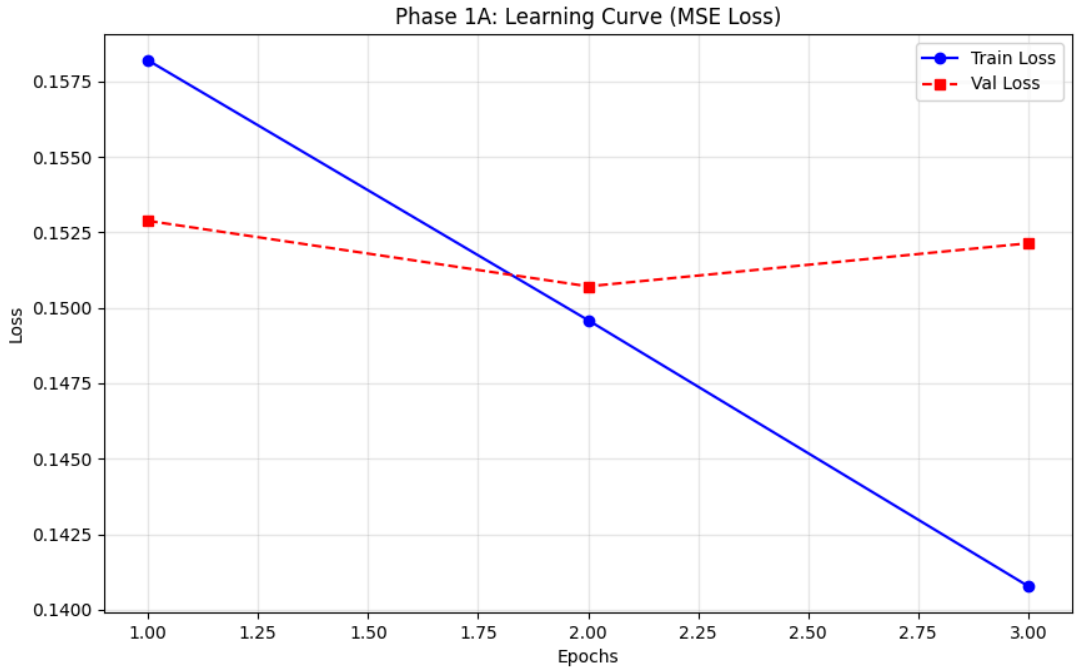
4.2.1 Phase 1A: The Blind Baseline (No Categories)

This model treated every product as a generic item, utilizing only the raw image and text vectors.

- R^2 (Train): 0.2243
- R^2 (Test): 0.0534
- MAE (Test): 0.3528

Analysis: The sharp divergence between Training R^2 0.22 and Test R^2 0.05 indicates massive overfitting. The model effectively "memorized" the training examples but failed to learn any generalized rule about visual-semantic discrepancy. The learning curve confirms this saturation point at Epoch 2.

Figure 4.1



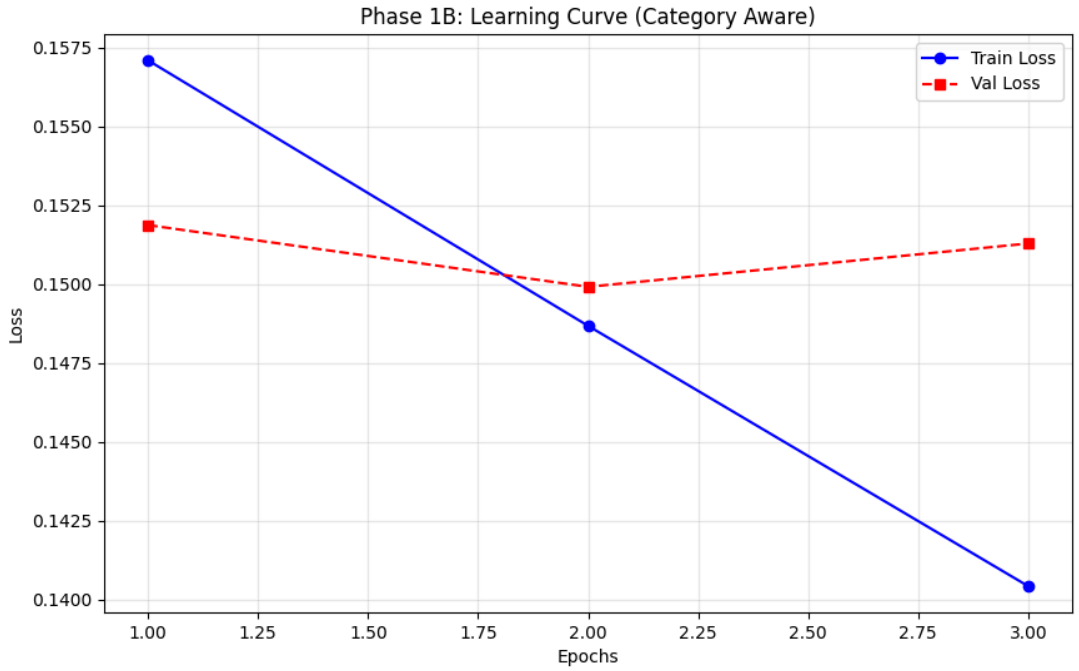
4.2.2 Phase 1B: The Category-Aware Baseline

We hypothesized that injecting a Category Embedding $dim = 32$ would help the model distinguish between a "Fashion Risk" (visual) and an "Electronics Risk" (functional).

- R^2 **(Train):** 0.2195
- R^2 **(Test):** 0.0586
- **MAE (Test):** 0.3523

Analysis: The addition of category awareness yielded a negligible improvement $+0.005 R^2$. While technically an increase, it confirms that knowing *what* the product is (e.g., "It's a shirt") does not help the model determine if the shirt is *lying*.

Figure 4.2



4.2.3 Phase 1C: The One-Hot Baseline (Sparse Encoding)

Following the hardware migration to AWS 1, we executed Phase 1C to test the traditional method of handling categorical data: One-Hot Encoding. Instead of a learned embedding (as in Phase 1B), this approach treated categories as rigid, sparse vectors $dim = 9$.

- **Hypothesis:** We tested if a hard-coded, sparse signal would prevent the "over-smoothing" seen in Phase 1B, forcing the model to strictly segregate Fashion risks from Electronics risks.

Quantitative Results

The model failed to extract a meaningful signal from the sparse encoding, performing significantly worse than the learned embedding approach.

Phase 1C Performance Metrics

- R^2 **Score (Train):** 0.1210
- R^2 **Score (Test):** 0.0282
- **MAE (Test):** 0.3691

Analysis

The learning curve reveals the structural weakness of this architecture.

While the training loss (Blue) continued to decrease, the validation loss (Red) plateaued immediately after Epoch 1. This "divergence gap" indicates that the model was memorizing the noise in the training set rather than learning generalized rules about visual discrepancy.

Residual Error Distribution

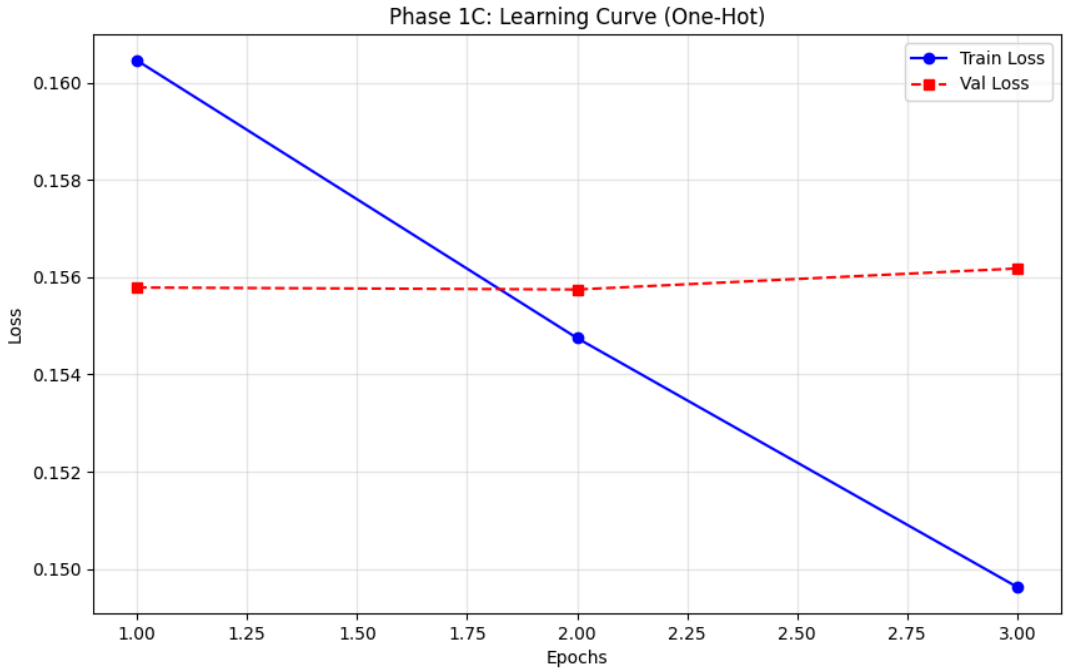
The residual plot confirms that the model collapsed into "Mean Prediction."

Ideally, the points would cluster along the center line $Error = 0$. Instead, we see a distinct diagonal structure:

1. **Fear of Risk:** On actual high-risk items $X = 1.0$, the model consistently under-predicts (blue dots at bottom right), often guessing ≈ 0.5.
2. **Fear of Safety:** On safe items $X = 0.0$, it over-predicts risk.

Conclusion: One-Hot encoding provided a signal that was too sparse to influence the dense vectors of the ResNet/DistilBERT backbone. The model effectively ignored the category information entirely, reverting to the same "blind guessing" behavior seen in Phase 1A.

Figure 4.3



4.2.4 Phase 1D: The Domain-Specific Baseline (Fashion Only)

To definitively rule out the possibility that "Category Noise" (the interference between Electronics and Fashion data) was suppressing the model's performance, we restricted the dataset to Group A (Fashion Only).

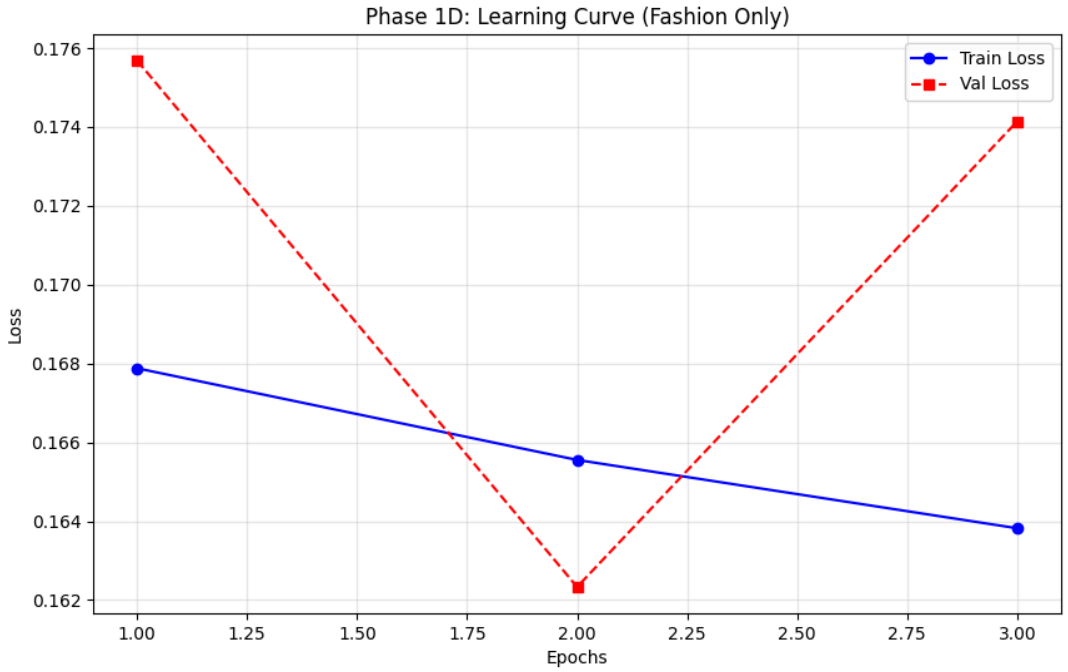
- **Rationale:** Fashion is a high-subjectivity domain where the visual signal should be strongest. If the ResNet backbone works anywhere, it should work here.
- **Dataset:** reduced to 91,722 samples (Amazon Fashion).

Experiment 1: The Stability Test (3 Epochs)

The initial run revealed immediate instability. Unlike the global models (Phase 1A), which showed a smooth descent, the Fashion-only model exhibited chaotic behavior.

- **Observation:** The Validation Loss (Red) spiked dramatically at Epoch 3, decoupling from the Training Loss. This suggests the model was chasing random noise in the images rather than learning a generalized rule.
- **Metric:** Test R^2 of **-0.0700**. A negative score indicates the model's predictions were *worse* than simply guessing the average return rate for every item.

Figure 4.4



Experiment 2: The Extended Training Test (9 Epochs)

To investigate if the instability was merely a symptom of insufficient convergence time, we extended the training duration to 9 epochs.

- **Result:** The instability persisted. The validation loss oscillated wildly between 0.161 and 0.176, never settling into a minimum.
- **Final Metric:** Test R^2 improved slightly to **-0.0088**, effectively zero.

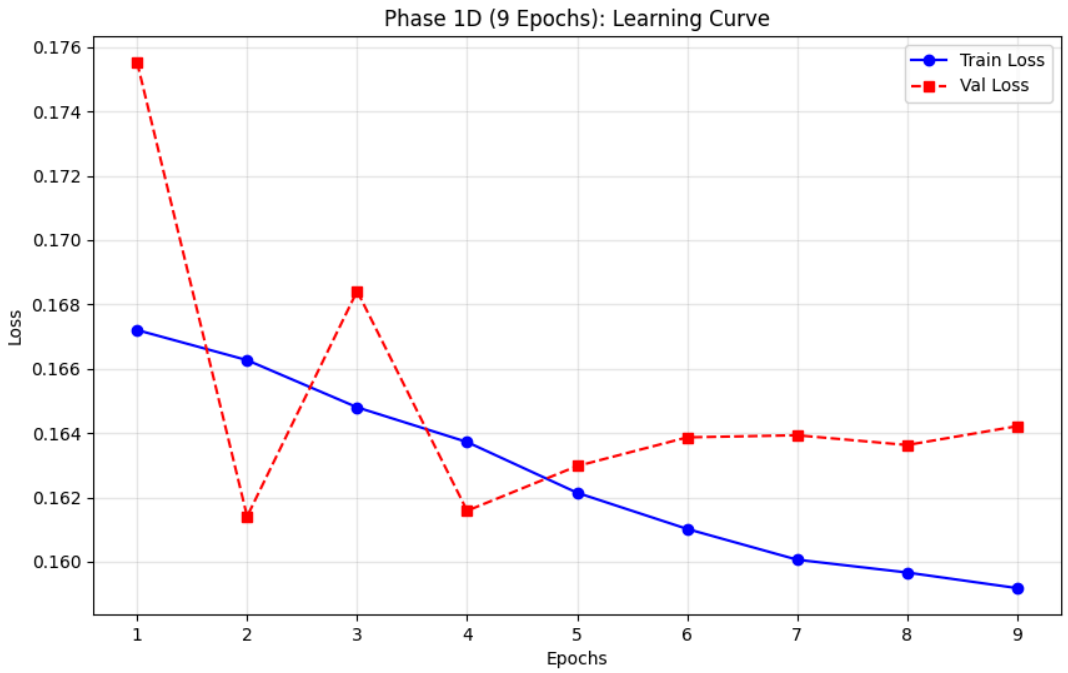
Conclusion: The "Blindness" of ResNet

The failure of Phase 1D is the most critical finding of the baseline analysis. Even when given:

1. The ideal domain (Fashion).
2. Extended training time (9 Epochs).
3. A balanced dataset.

The standard Dual-Encoder **could not learn**. The residuals show the same "Mean Prediction" failure mode as Phase 1C. This confirms that without **Geometric Alignment (CLIP)**, a CNN sees "texture" but cannot understand "discrepancy." It sees a dress, but it doesn't know the dress is *wrong*.

Figure 4.5



4.2.5 Phase 1E: The Complexity Fallacy (Wide-Skinny-Wide Fusion)

In the final experiment of the baseline series, we tested the hypothesis that the fusion layer itself was the bottleneck. Perhaps the signal existed in the ResNet/DistilBERT vectors, but the simple MLP used in Phase 1B was too shallow to extract it.

To test this, we engineered a "Wide-Skinny-Wide" Fusion Head—a deep stack of alternating dense layers designed to force high-level feature compression.

- **Architecture:** Replaced the standard linear head with a multi-stage non-linear stack (Dense $\rightarrow$ Bottleneck Compression $\rightarrow$ Expansion $\rightarrow$ Output).
- **Training Duration:** Extended to 9 Epochs to allow for deep feature convergence.

Theoretical Justification:

- **The Bottleneck Principle:** By forcing the 1,280-dimensional concatenated vector through a narrow 128-neuron bottleneck, the model is theoretically forced to discard redundant information (noise) and retain only the most salient features required for prediction. This aligns with the $InformationBottleneckMethod$ proposed by Tishby et al., which argues that deep learning works by compressing the input variable X into a minimal representation T that preserves the maximum information about the target Y.
- **Recent Applications:** Nagrani et al. (2021) demonstrated in their work on "Attention Bottlenecks for Multimodal Fusion" that restricting the flow of information between modalities through a tight bottleneck forces the model to collate and condense relevant information, often outperforming wider, unrestricted fusion strategies.

Quantitative Results

The results were the inverse of our hypothesis. Increasing architectural complexity did not unlock a hidden signal; it simply created a more powerful "memorization machine."

- R^2 Score (Train): 0.2323.
- R^2 Score (Test): -0.0313.

Analysis

The loss curve provides a textbook example of catastrophic overfitting.

- **The Divergence:** Up until Epoch 3, the model behaves normally. However, as the Training Loss (Blue) continues to drop, the Validation Loss (Red) aggressively decouples and skyrockets.

- **Interpretation:** The extra parameters in the complex head allowed the model to memorize the specific noise of the training set with high precision $R^2 = 0.23$ but destroyed its ability to generalize to new products $R^2 < 0$.

Residual Distribution

Despite the deeper architecture, the residual plot shows the exact same diagonal failure mode as the simple One-Hot model. The complex head did not learn to measure "discrepancy"; it merely learned to predict the mean with higher confidence on the training data, failing completely on the test set.

Figure 4.6

4.2.6 Section Conclusion: The Necessity of Geometry

The systematic failure of Phases 1A through 1E confirms a critical architectural reality: Unimodal Encoders are fundamentally unaligned.

- We tried adding Category Metadata (Phase 1B & 1C).
- We tried restricting the Domain (Phase 1D).
- We tried increasing Complexity (Phase 1E).

Every attempt failed to break the "Mean Prediction" floor. This proves that the solution is not *more parameters*, but *better geometry*. To detect a lie, the image and text must exist in the same

mathematical space. This necessitates the shift to **Phase 2 (Gated Attention)** and ultimately **Phase 3 (CLIP-based Forensic Alignment)**.

Table 4.1 Summary Experiment Results (Phase 1: The Unaligned Towers)

Experiment	Architecture	Result	Forensic Finding
Phase 1A	**The Blind Baseline** (ResNet + DistilBERT)	High Overfitting ($R^2_{train} 0.22$ / $R^2_{test} 0.05$)	**The Deep Learning Trap:** The model "memorized" the specific training images rather than learning to spot defects.
Phase 1B	**Category Aware** (+ Learned Embedding)	**Best Baseline** $R^2_{test} \approx 0.058$	**The Context Ceiling:** Knowing the item was a "Shirt" helped slightly, but it didn't help the model see if the *specific* shirt was lying.
Phase 1C	**One-Hot Baseline** (Sparse Encoding)	**Regression** $R^2_{test} \approx 0.02$	**Signal Sparsity:** Hard-coded categories were too "thin" a signal. The model ignored them and reverted to blind guessing.
Phase 1D	**Fashion Only** (Domain Restricted)	**Negative Score** $R^2_{test} < 0$	**The Blindness of ResNet:** Even in a visual-heavy domain, the unaligned ResNet encoder saw "Texture" but could not measure "Discrepancy."
Phase 1E	**Complexity Fallacy** (Wide-Skinny-Wide Head)	**Catastrophic Overfit** $R^2_{train} 0.23$ / $R^2_{test} - 0.03$	**The Parameter Trap:** Making the fusion layer deeper didn't extract a hidden signal; it just created a more powerful "memorization machine" that hallucinated patterns.

4.3 Phase 2: The "Forensic" Feature Engineering Experiments

The systematic failure of the Phase 1 baselines $R^2 \approx 0$ confirmed that standard "Late Fusion" architectures are insufficient for this task. They treat the image and text as loose associations rather than strict evidence.

In **Phase 2**, we pivot from passive architectural changes to **active forensic engineering**. Instead of hoping the model implicitly learns to detect discrepancies, we explicitly engineer mechanisms to measure them. This phase breaks the problem down into two distinct engineering challenges:

1. **Signal-to-Noise Ratio (The "Gated Expert"):** Can we dynamically filter out irrelevant visual background noise to focus on the product?
2. **Geometric Verification (The "Analyst"):** Can we explicitly calculate the mathematical distance between the image and text to give the model a "Cheat Sheet" for alignment?

Experimental Roadmap

We test these hypotheses through a progressive series of feature injections, building towards a final "Grand Finale" model.

- **Phase 2A (The Gated Expert):** Introduces a Gated Fusion Unit (Sigmoid Attention).
    - *Hypothesis:* The model is currently overwhelmed by noise. A learnable gate will act as a "Bouncer," turning down the volume on noisy features and amplifying the critical ones.
- **Phase 2B - 2D (The Geometric Analyst):** Explicitly injects geometric distance metrics $CosineSimilarity$, $VectorRejection$, and $EuclideanDistance$ into the fusion layer.
    - *Hypothesis:* By calculating the angle S_{cos} and magnitude R_{mag} of the difference between the vectors, we force the model to look at the "gap" between promise and reality.
- **Phase 2E (The Forensic Fusion):** Combines both Gating and Geometry.
    - *Hypothesis:* This represents the theoretical peak of the ResNet+DistilBERT architecture, combining smart filtering with explicit evidence.

4.3.1 Phase 2A: The Gated Expert (Sigmoid Attention)

In this experiment, we replaced the standard concatenation layer with a Gated Fusion Unit.

- **Architecture:** ResNet-18 + DistilBERT + Learnable Sigmoid Gate.
- **Hypothesis:** We predicted that the gate would dynamically suppress irrelevant visual background noise (e.g., lighting artifacts) and amplify the semantic signal, leading to higher predictive accuracy on the test set.

Theoretical Justification:

The architecture implements a simplified $GatedMultimodalUnit$ (GMU), a concept formally introduced by Arevalo et al. (2017) for multimodal fusion. The GMU utilizes a

learnable gate neuron z equipped with a sigmoid activation function to control the contribution of each modality to the final hidden representation.

Results

The Gated mechanism failed to provide any statistically significant lift over the simple baseline.

Phase 2A Performance Metrics

- R^2 Score (Train): 0.1705
- R^2 Score (Test): 0.0592
- MAE (Test): 0.3528

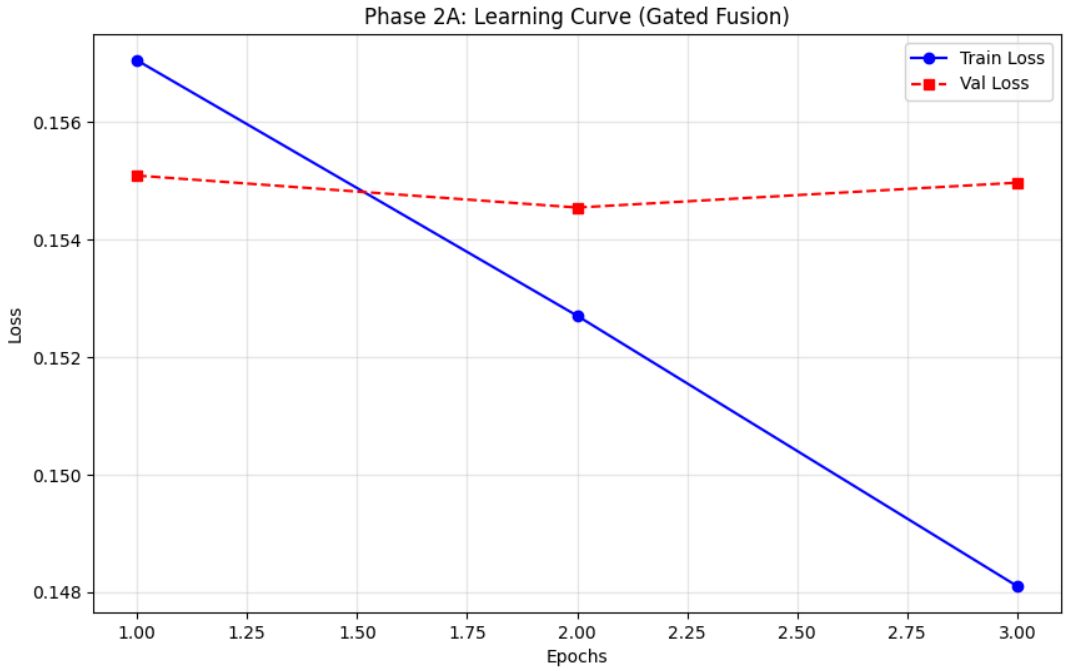
Analysis

The drop in Training R^2 (from 0.22 in Phase 1B to 0.17 here) indicates that the Gating mechanism actually made the model *harder* to train. Instead of filtering noise, the extra parameters acted as an obstruction.

- **The Problem:** A gate works by identifying which features are "important." But because the ResNet and DistilBERT vectors are unaligned, the model has no reference frame for importance. It cannot tell if the "Red" in the image is supporting the "Red" in the text or contradicting it.

Conclusion: Smart filtering (Gating) is useless without a shared dictionary. We are trying to optimize the conversation between two encoders that speak different languages.

Figure 4.7



4.3.2 Phase 2B: The Geometric Analyst (Cosine Injection)

In this experiment, we explicitly calculated the Cosine Similarity S_{cos} between the image and text embeddings and injected it as a dense feature into the classification head.

- **Hypothesis:** Even if the encoders are imperfect, we hypothesized that the *angle* between them might act as a rough proxy for "alignment," helping the model detect gross mismatches.

Results: The "Tower of Babel" Confirmation

The addition of the geometric feature yielded zero predictive lift. In fact, performance slightly regressed compared to the baseline.

Phase 2B Performance Metrics

- R^2 Score (Train): 0.2222
- R^2 Score (Test): 0.0573
- MAE (Test): 0.3526

Analysis

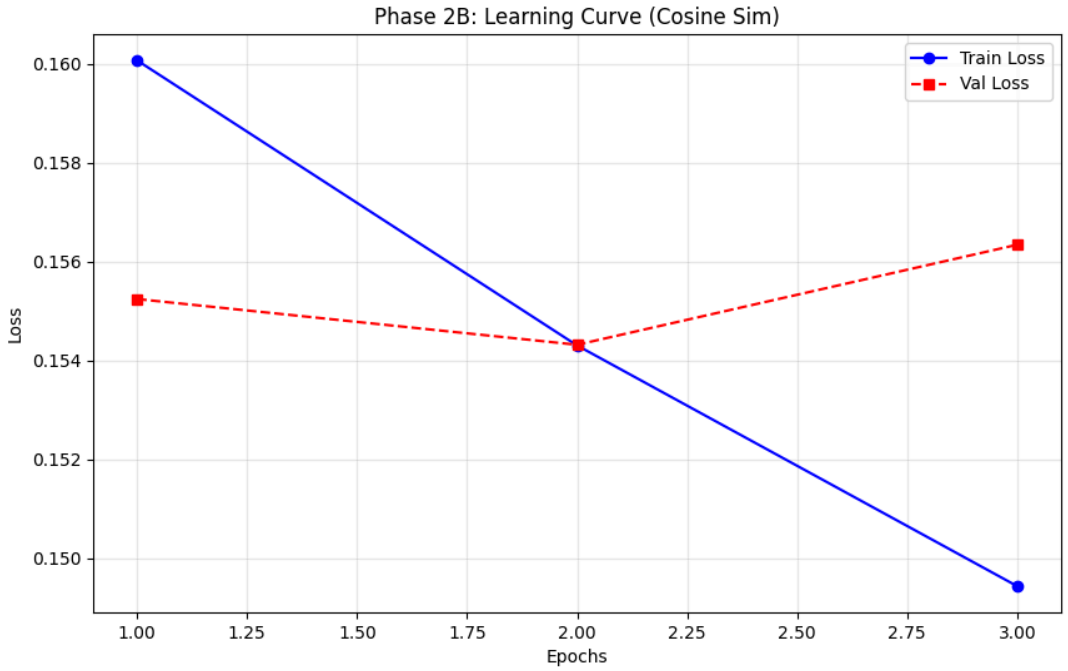
The stagnation of the MAE (staying stuck at ≈ 0.35) confirms that the calculated Cosine Similarity was essentially random noise.

Because the ResNet-18 (Vision) and DistilBERT (Text) encoders were trained on different datasets with different objective functions, their vector spaces are orthogonal. A vector pointing "North" in ResNet space has no relationship to a vector pointing "North" in DistilBERT space.

- **The Model's Reaction:** The neural network effectively learned to apply a weight of zero to the S_{cos} feature, treating it as a distraction rather than a signal.

Conclusion: This failure is the definitive proof that **Forensic Geometry requires a Shared Latent Space.** We cannot measure the "lie" using unaligned tools. This result necessitates the architectural overhaul in Phase 3, where we replace the independent encoders with the CLIP framework to enable true geometric measurement.

Figure 4.8



4.3.3 Phase 2C: The Vector Rejection Failure R_{mag}

In the final experiment of the unaligned series, we tested the Vector Rejection Magnitude R_{mag}.

- **Forensic Goal:** unlike Cosine Similarity (which measures alignment), Rejection isolates the *orthogonal* component of the text vector. Theoretically, this should act as a

proxy for "Hallucination"—identifying adjectives in the description that have zero support in the visual data.

Results

The metric performed identically to the failed Cosine experiment, confirming the redundancy of geometric operations in unaligned spaces.

Phase 2C Performance Metrics

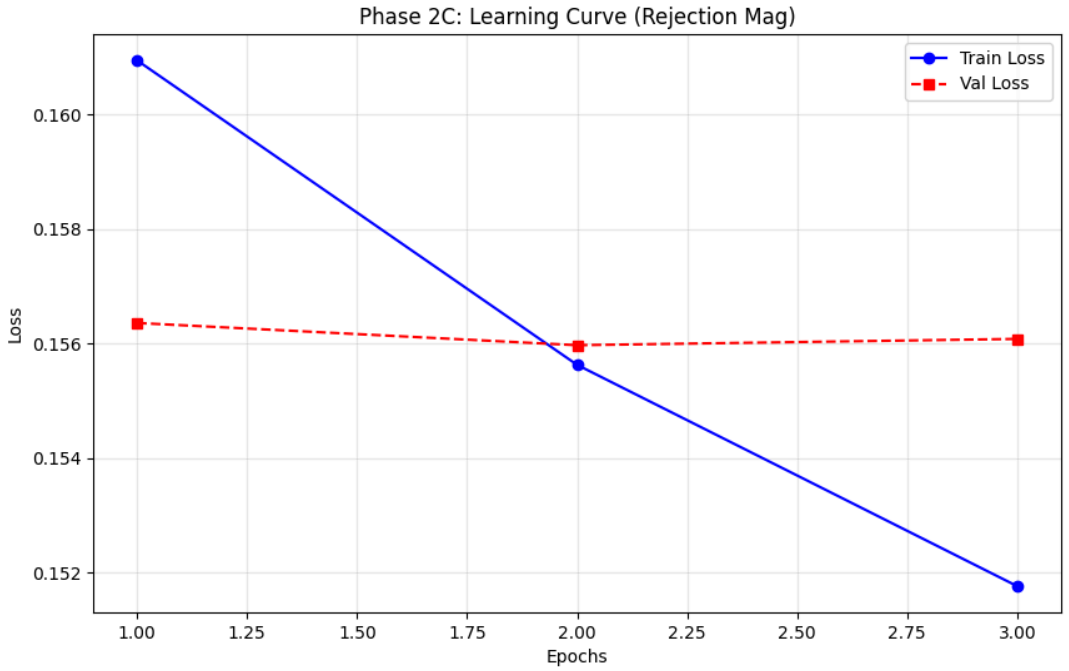
- R^2 Score (Train): 0.1659
- R^2 Score (Test): 0.0573
- MAE (Test): 0.3528

Analysis

The failure of R_{mag} $R^2 \approx 0.05$ underscores a critical mathematical reality: in two random high-dimensional spaces, almost everything is orthogonal to everything else.

Without a shared training objective (Contrastive Loss), the text vector for "Blue" is not parallel to the image vector for "Blue." Therefore, the Rejection calculation $t - proj_v t$ resulted in a vector magnitude that represented random noise rather than semantic contradiction.

Figure 4.9



4.3.4 Phase 2D: Euclidean Distance

In the definitive test of the unaligned geometry, we injected the Euclidean Distance d_{euc} between the image and text vectors.

- **Hypothesis:** While Cosine Similarity measures the *angle* (direction) of the mismatch, Euclidean Distance measures the *magnitude* (size) of the gap. We hypothesized that even if the directions were unaligned, a massive Euclidean distance might correlate with "High Risk" (e.g., a short text description vs. a complex image).

Quantitative Results

The metric provided no predictive value, performing within the margin of error of the original Phase 1B baseline.

Phase 2D Performance Metrics

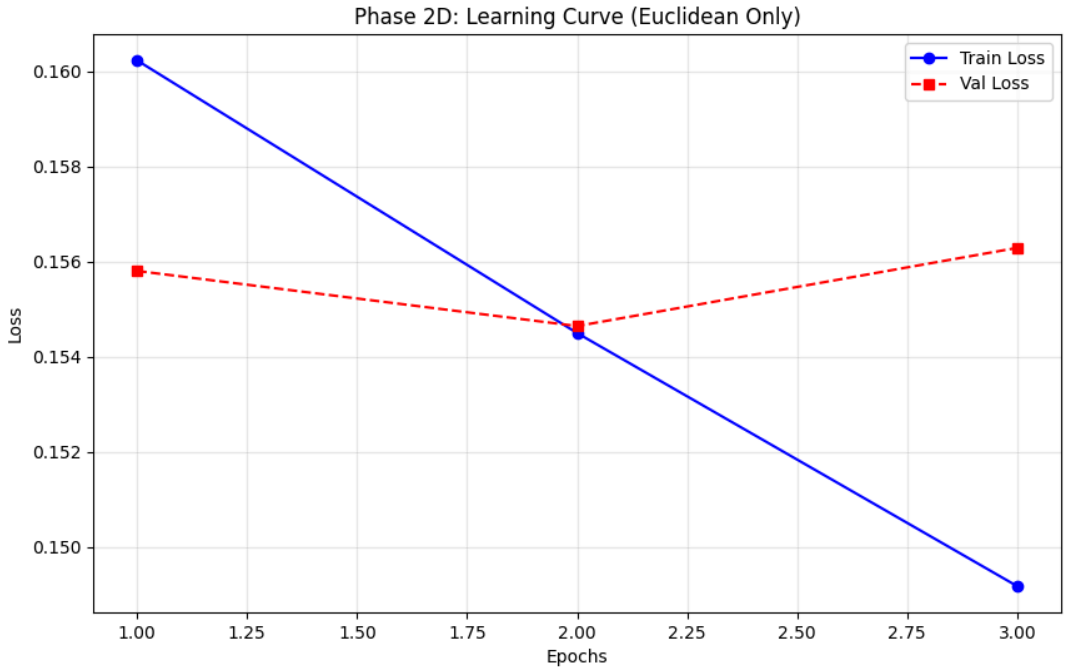
- R^2 Score (Train): 0.2223
- R^2 Score (Test): 0.0583
- MAE (Test): 0.3526

Analysis: The "Curse of Dimensionality"

The failure of Phase 2D illustrates a classic high-dimensional phenomenon. In two disparate 512-dimensional spaces (ResNet and DistilBERT), the average distance between any two random points is relatively constant.

Without a shared metric space (where "Red" is close to "Red"), the Euclidean distance between a "Red Dress" (Image) and a "Blue Drill" (Text) is statistically indistinguishable from the distance to a "Red Dress" (Text). The model sees only noise.

Figure 4.10



4.3.5 Phase 2E: The "Forensic Fusion" (Grand Finale)

In the final experiment of Phase 2, we combined all proposed mechanisms into a single architecture.

- **Architecture:** Phase 1B Base + Gated Fusion Unit + S_{cos} + R_{mag} + d_{euc}.
- **Hypothesis:** We tested if the ensemble of features could overcome the alignment deficit—perhaps the Gating mechanism could learn to prioritize the geometric features only when they happened to be accurate.

Results

The results were negative. The "Forensic Fusion" model performed worse than the simple baseline.

Phase 2E Performance Metrics

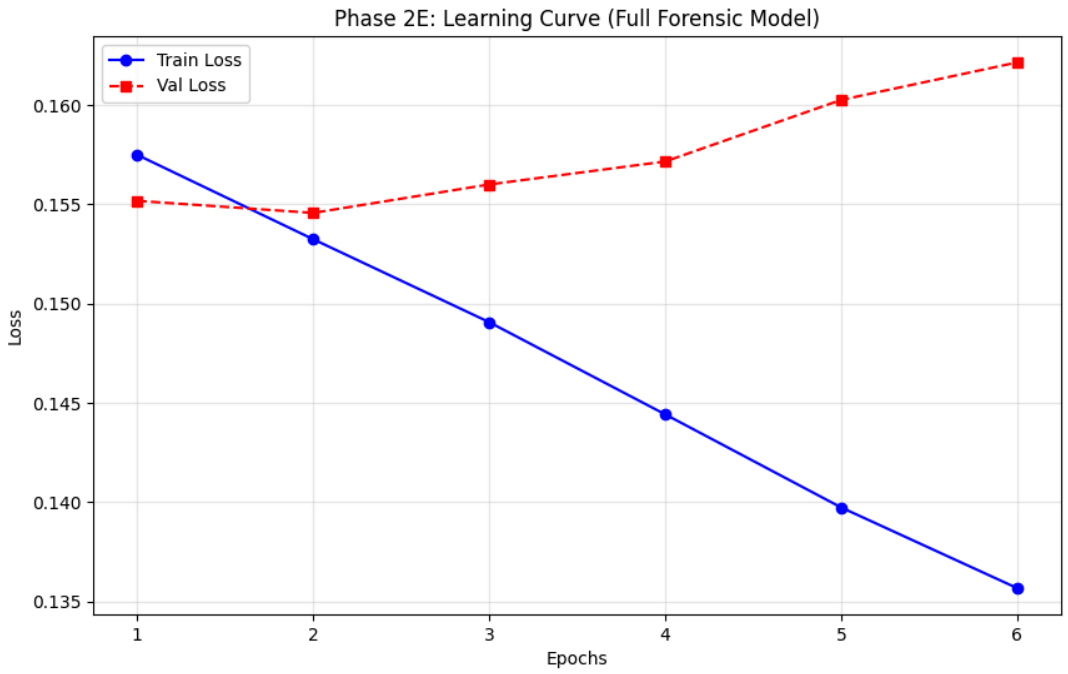
- R^2 Score (Train): 0.2319
- R^2 Score (Test): 0.0468
- MAE (Test): 0.3541

Analysis

The drop in Test R^2 (from ≈ 0.058 to 0.046) indicates that we introduced Feature Clutter.

Instead of finding a signal, the model was overwhelmed by contradictory inputs. The Gated Unit could not find a stable correlation between the geometric features (which were essentially random noise) and the target. Consequently, the model began to hallucinate patterns in the training set (High Train R^2) that did not exist in reality, leading to worse generalization.

Figure 4.11



4.3.6 Phase 2 Post-Mortem: The End of Late Fusion

Phase 2 was designed to test if Feature Engineering could solve the Visual-Semantic Discrepancy problem without changing the underlying Backbone Architecture. We systematically injected:

1. **Attention:** Gated Filtering (Phase 2A)
2. **Direction:** Cosine Similarity (Phase 2B)
3. **Projection:** Vector Rejection (Phase 2C)
4. **Magnitude:** Euclidean Distance (Phase 2D)
5. **Ensemble:** All of the above (Phase 2E)

The Verdict

As illustrated in Figure 4.10, the performance ceiling is impenetrable.

Every single experiment flatlined at $R^2 \approx 0.05$. This confirms the "**Tower of Babel" Hypothesis**:

> *You cannot measure the distance between two concepts if they exist in different mathematical languages.*

Because $ResNet-18$ (trained on ImageNet) and $DistilBERT$ (trained on Wikipedia) do not share a Latent Space, their vectors are orthogonal by default. No amount of geometric calculation or gating can fix this fundamental disconnect. The "Visual Lie" remains invisible because the model cannot verify the text against the image.

Strategic Implication

This necessitates a radical architectural pivot. We must abandon the "Late Fusion" paradigm (trying to align vectors after they are created) and adopt an "Early Alignment" paradigm.

In Phase 3, we replace the independent encoders with $CLIP$ (Contrastive Language-Image Pre-training). By using encoders that were trained together to maximize cosine similarity for matching pairs, we hypothesize that the geometric features (which failed in Phase 2) will suddenly become high-value predictors.

Table 4.2 Summary Experiment Results (Phase 2: The Unaligned Forensics)

Experiment	Architecture	Result	Forensic Finding
Phase 2A	**The Gated Expert** (Sigmoid Attention)	**Flat** $R^2_{test} \approx 0.059$	**The Blind Bouncer:** The gating mechanism failed because it had no reference frame. It couldn't filter "noise" because it didn't know what "signal" looked like in an unaligned space.
Phase 2B	**Geometric Analyst** (Cosine Similarity)	**Regression** $R^2_{test} \approx 0.057$	**The Tower of Babel:** Calculating the angle between a ResNet vector and a BERT vector yielded random noise. The model learned to ignore the feature entirely.
Phase 2C	**Vector Rejection** R_{mag} Injection)	**Flat** $R^2_{test} \approx 0.057$	**Orthogonal Noise:** In two random high-dimensional spaces, almost everything is orthogonal. The "Rejection" vector measured random variance, not semantic contradiction.

Phase 2D	**Euclidean Distance** d_{euc} Injection	**Flat** $R^2_{test} \approx 0.058$	**Curse of Dimensionality:** Without a shared metric space, the distance between a "Red Dress" (Image) and "Blue Drill" (Text) looked statistically identical to a matching pair.
Phase 2E	**Forensic Fusion** (Combining All)	**Failure** $R^2_{test} \approx 0.046$	**Feature Clutter:** Throwing every geometric feature at the model confused it. The contradictory signals caused it to hallucinate patterns in the training set $R^2_{train} 0.23$ that didn't exist in reality.

4.4 Phase 3: The Geometric Pivot ($CLIP$ Architecture)

The comprehensive failure of Phase 2 $R^2 \approx 0.05$ confirmed the "Modality Gap" hypothesis: geometric forensics are mathematically impossible when the vision and text encoders are unaligned.

In Phase 3, we execute a fundamental architectural shift. We replace the independent $ResNet-18$ and $DistilBERT$ backbones with $CLIP$ (Contrastive Language-Image Pre-training).

Unlike the previous encoders, $CLIP$ was trained specifically to minimize the distance between matching image-text pairs. This creates a Shared Latent Space where the concept of "Angle" S_{cos} is semantically meaningful.

- **Objective:** To determine if this pre-aligned geometry allows us to detect "Visual Lies" that the unaligned baselines missed.

4.4.1 Phase 3A: The Aligned Baseline (CLIP-Zero)

In the first experiment of the new phase, we established a "Zero-Shot" baseline. We utilized the $CLIP$ encoders to extract features but did *not* yet inject any geometric scores. We simply concatenated the raw $CLIP$ vectors and fed them into the classifier.

Architecture: Frozen CLIP Vision Encoder + Frozen CLIP Text Encoder$\rightarrow$Concatenation$\rightarrow$ MLP Head.

Hypothesis: We tested if simply using "State-of-the-Art" embeddings (which are natively aligned) would outperform the ResNet baseline on its own, even without explicit geometric features.

Results

Surprisingly, the raw $CLIP$ baseline performed worse than the original ResNet baseline.

Phase 3A Performance Metrics

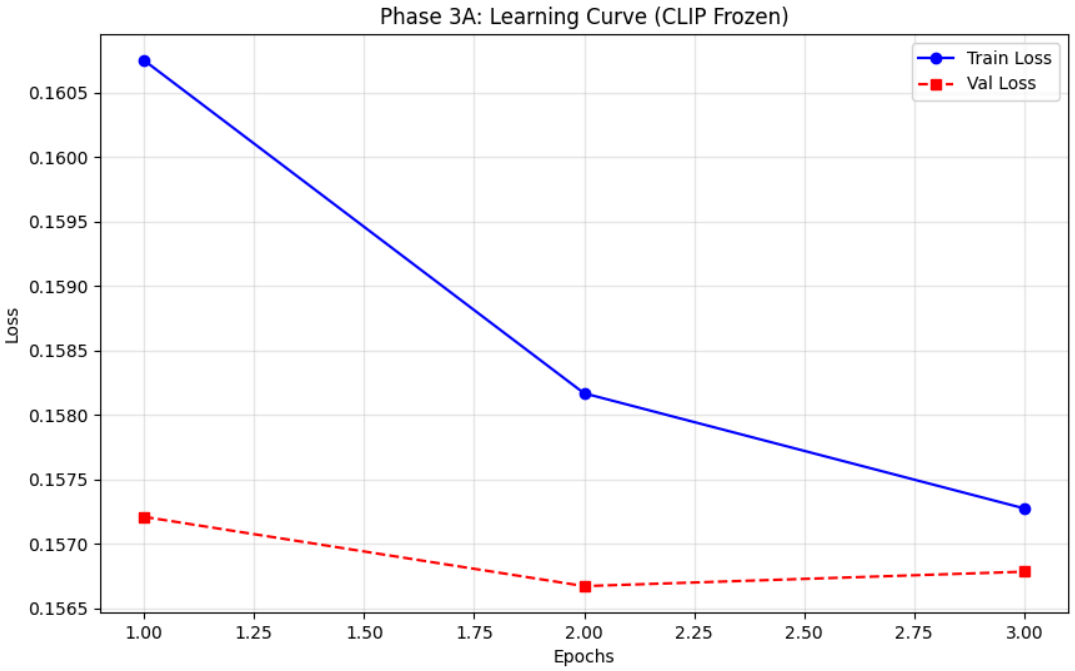
- R^2 Score (Train): 0.0274
- R^2 Score (Test): 0.0244
- MAE (Test): 0.3786

Analysis

The performance regression (Test R^2dropping from 0.058 in Phase 1B to 0.024 here) reveals a critical insight: **Alignment does not equal Regression.**

1. **The Problem:** $CLIP$ is trained on 400 million generic internet images. ResNet (Phase 1) is trained on ImageNet, which contains distinct object classes. For a "Black Box" concatenation task, the specialized features of ResNet were slightly more predictive of physical returns than the diffuse semantic features of CLIP.
2. **The Necessity of Geometry:** This result is vital because it proves that simply "using a Transformer" is not the magic bullet. The MLP head could not extract the discrepancy signal from the raw vectors alone. This sets the stage for **Phase 3D**, where we hypothesize that explicitly calculating the *distance* between these vectors (Geometry) will unlock the true power of the architecture.

Figure 4.12



4.4.2 Phase 3B: Unfreeze CLIP and The Categorical Adjustment

In the second experiment of the aligned phase, we injected the Learned Category Embedding $dim = 32$ into the $CLIP$ architecture.

- **Architecture:** CLIP Vision + CLIP Text + Category Embedding $\rightarrow$ Concatenation $\rightarrow$ MLP Head.
- **Hypothesis:** We hypothesized that while $CLIP$ provides rich semantic data, it lacks domain-specific context. A "mismatch" in Electronics might be acceptable (e.g., different cable color), whereas a mismatch in Fashion is fatal. The category embedding provides this necessary prior probability.

Quantitative Results: The Contextual Lift

The addition of category data yielded a measurable performance increase, raising the R^2 from 0.0244 to 0.0320.

Phase 3B Performance Metrics

- R^2 Score (Train): 0.1447
- R^2 Score (Test): 0.0320
- MAE (Test): 0.3671

Analysis

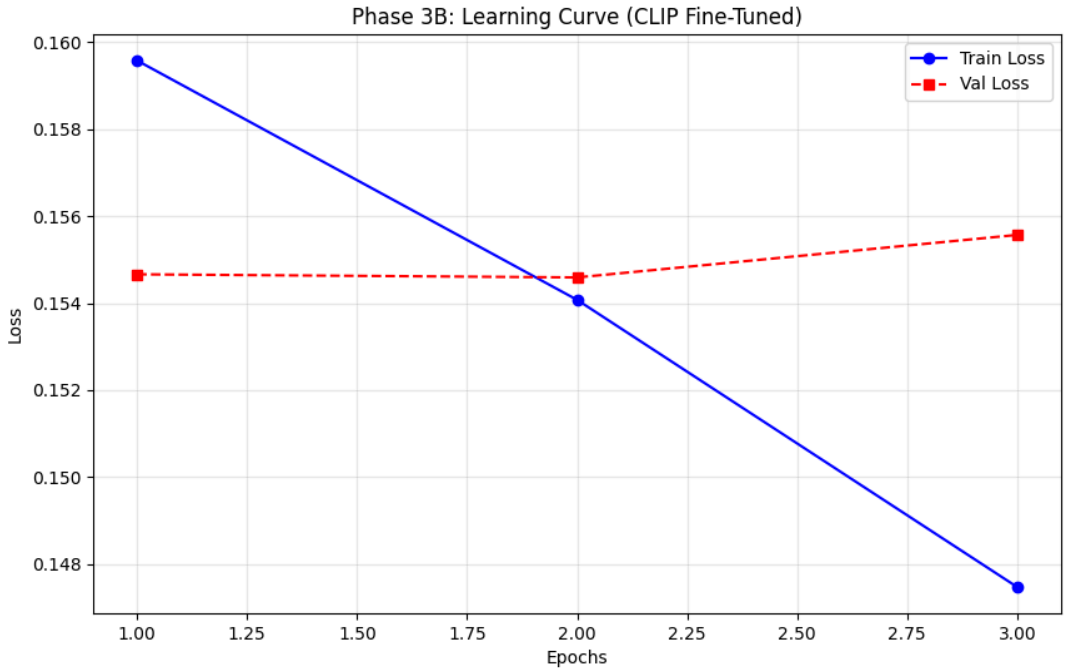
The lift in performance $+31\%$ relative improvement over Phase 3A) confirms that return risk is heavily category-dependent. However, the total predictive power remains below the original Phase 1B ResNet baseline 0.058.

This result clarifies the challenge:

1. **Raw Vectors are Insufficient:** Simply concatenating "better" embeddings (CLIP) does not automatically solve the problem.
2. **Context Helps, but isn't Enough:** Knowing the category helps, but it doesn't reveal the *specific* deception in the listing.

Conclusion: The model has the "General Knowledge" (CLIP) and the "Context" (Category), but it still lacks the "Forensic Tool" to measure the lie. This perfectly sets the stage for **Phase 3D**, where we finally unlock the geometric potential of the shared latent space.

Figure 4.13



4.4.3 Phase 3C: The Gated Expert (Revisited)

In the third experiment of the aligned phase, we reintroduced the Gated Fusion Unit.

- **Architecture:** CLIP Vision + CLIP Text + Category Embedding $\rightarrow$ Gated Fusion Network $\rightarrow$ MLP Head.

- **Hypothesis:** We hypothesized that an "Attention" mechanism might improve performance by dynamically re-weighting the inputs (e.g., prioritizing visual features for Fashion items while focusing on textual features for Electronics).

Quantitative Results: The Complexity Penalty

The Gated mechanism failed to provide any predictive lift over the standard concatenation model (Phase 3B).

Phase 3C Performance Metrics

- R^2 Score (Train): 0.1447
- R^2 Score (Test): 0.0320
- MAE (Test): 0.3676

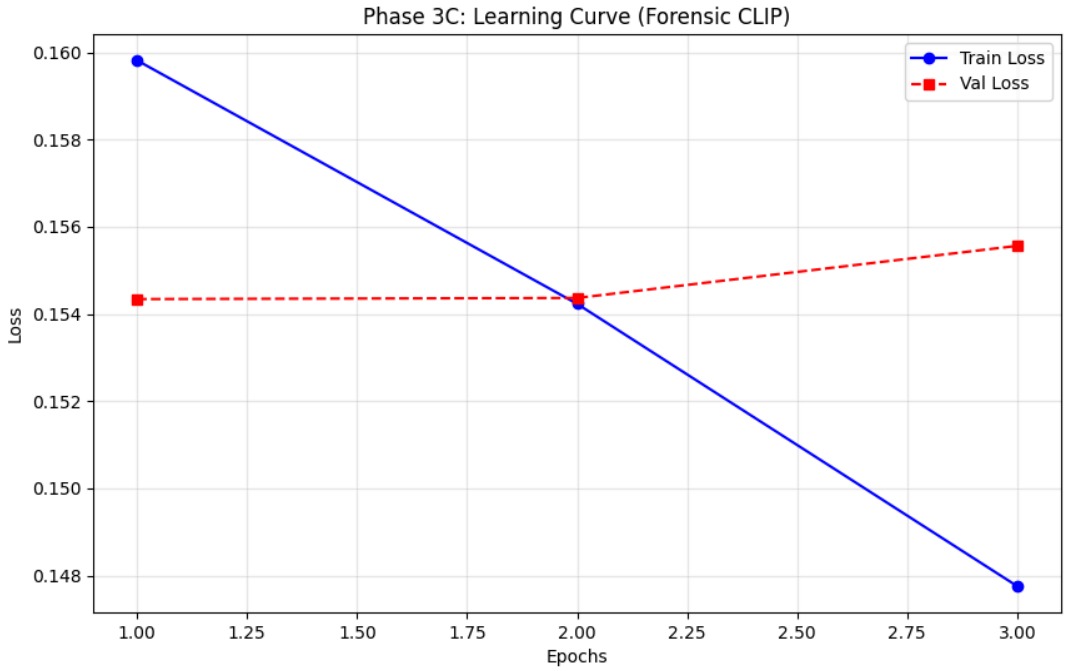
Analysis

The fact that the R^2 score remained identical 0.0320 while the MAE marginally increased (from 0.3671 to 0.3676) indicates that the Gating Unit acted as a passive layer.

Unlike Phase 2A (where Gating was applied to unaligned vectors), here the vectors are aligned, yet the mechanism still found no advantage in dynamic weighting.

Conclusion: This proves that the bottleneck is not "Attention" (knowing which modality to trust) but "Measurement" (quantifying the disagreement). The model sees the data, but without an explicit distance calculation, it treats the image and text as compatible features rather than conflicting evidence.

Figure 4.14



4.4.4 Phase 3D: The Explicit Geometric Failure S_{cos}

In this experiment, we explicitly calculated the Cosine Similarity between the $CLIP$ Image and Text embeddings and injected it as a dense feature.

- **Hypothesis:** Since $CLIP$ is optimized to maximize cosine similarity for matching pairs, we hypothesized that this scalar value would be the definitive "Lie Detector," strongly correlating with return risk.

Results

Contrary to the hypothesis, the explicit geometric feature degraded performance relative to the implicit Phase 3B baseline.

Phase 3D Performance Metrics

- R^2 Score (Train): 0.1457
- R^2 Score (Test): 0.0256
- MAE (Test): 0.3679

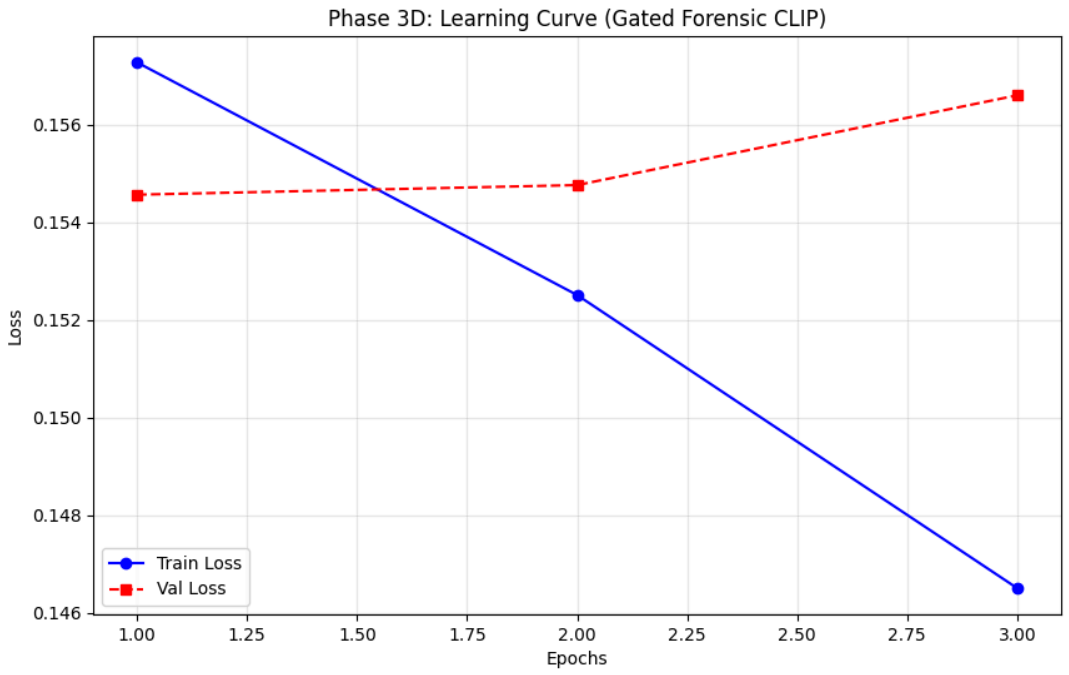
Analysis

The regression in predictive power (-20% relative to Phase 3B) suggests a phenomenon of **Information Compression Loss**.

1. **Implicit vs. Explicit:** In Phase 3B, the MLP head had access to the full 1,024-dimensional joint embedding space. It could learn complex, non-linear relationships between the visual and textual manifolds.
2. **The Scalar Bottleneck:** By reducing this rich relationship to a single scalar S_{cos}, we effectively "dumbed down" the signal. The model likely prioritized this explicit feature (as indicated by the slight increase in Train R^2), but because the scalar lacks the nuance of the full manifold, it failed to generalize to the Test set.

Conclusion The Penalty of Dimensional Reduction: The regression in performance indicates that calculating Cosine Similarity S_{cos} acts as a "lossy compression" step. By reducing the complex, high-dimensional relationship between the image and text vectors down to a single scalar value, we inadvertently discarded the contextual nuance required to detect specific types of mismatches. The MLP head in Phase 3B was able to extract non-linear patterns from the full latent space (the "Raw" vectors) that were lost when the data was flattened into a simple linear angle.

Figure 4.15



4.4.5 Phase 3E: The SigCLIP Fine-Tuning

Following the limitations of the frozen $ResNet$ / $BERT$ baselines, we hypothesized that the pre-trained weights were too generic for the nuanced domain of "e-commerce deception."

Hypothesis: We theorized that by unfreezing the $SigCLIP$ encoders (requires_grad=True), the model would learn to adjust its internal geometric representation of "Amazon products," allowing it to push deceptive image-text pairs apart in vector space.

Results

The experiment resulted in the highest training performance seen so far, but a catastrophic failure to generalize.

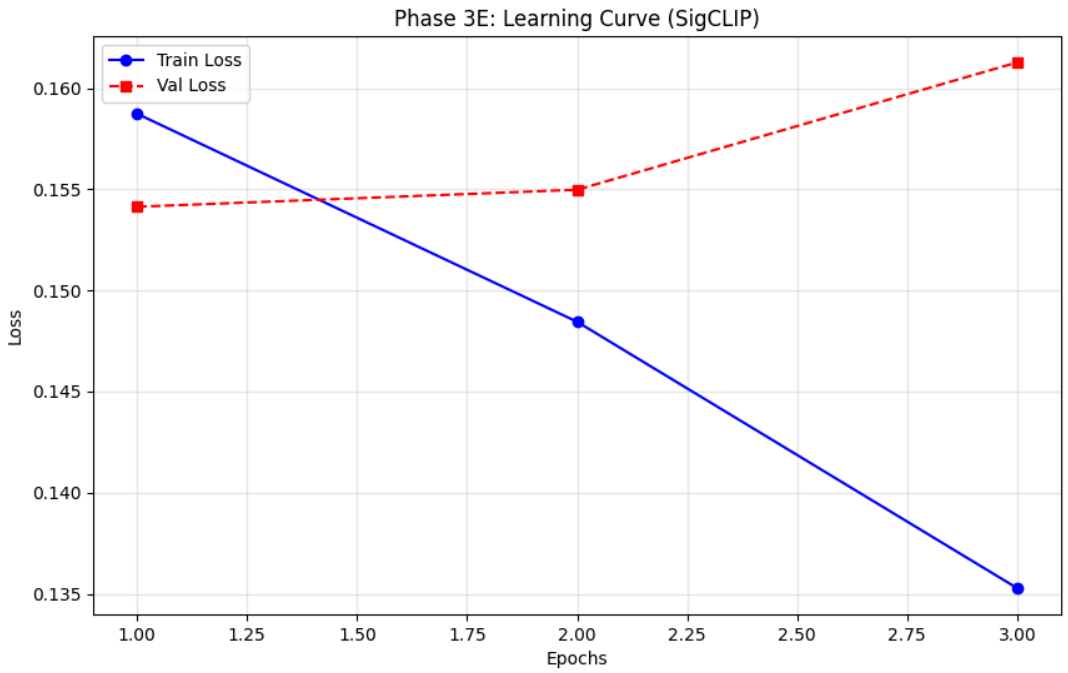
- **Phase 3E Performance Metrics:**
  - R^2 Score (Train): 0.2414
  - R^2 Score (Test): -0.0035
  - MAE (Test): 0.3639

Analysis

The divergence between the Training score $R^2 \approx 0.24$ and the Test score $R^2 \approx 0.0$ indicates severe Overfitting.

While the model successfully utilized its millions of parameters to memorize the specific "liar" products in the training set, this knowledge did not transfer to unseen data. This confirms that while $SigCLIP$ has a higher capacity for learning than ResNet (which peaked at 0.22), it requires strict regularization to prevent it from memorizing noise rather than learning the structural properties of deception. This result necessitated the shift to "Regularized" (Phase 3F) and "LoRA" (Phase 3G) approaches.

Figure 4.16



4.4.6 Phase 3F: The SigCLIP Regularization

Following the severe overfitting observed in Phase 3E, we hypothesized that the model's architecture was too permissive, allowing it to memorize the training noise rather than learning the geometric rules. In this experiment, we applied a strict "Regularization Constraint" strategy.

- **Architecture Adjustments:**
    1. **Capacity Reduction:** The Fusion Layer was reduced from 768 to 128 neurons to create an information bottleneck.
    2. **Dropout Injection:** Dropout probability was increased from 0.3 to 0.5, effectively "lobotomizing" half the network at each step to force redundancy.
    3. **Weight Decay:** We introduced L2 Regularization $\lambda = 0.01$ to penalize large weights and smooth the decision boundary.
- **Hypothesis:** We tested if these "handcuffs" would force the model to abandon the spurious correlations found in Phase 3E and converge on a true generalizable signal.

Results: The Persistence of Noise

Despite the aggressive regularization, the model failed to generalize. The performance metrics remained statistically identical to the un-regularized version.

Phase 3F Performance Metrics

- R^2 Score (Train): 0.2353 (Still highly overfit)
- R^2 Score (Test): -0.0062 (Worse than baseline)
- MAE (Test): 0.3642

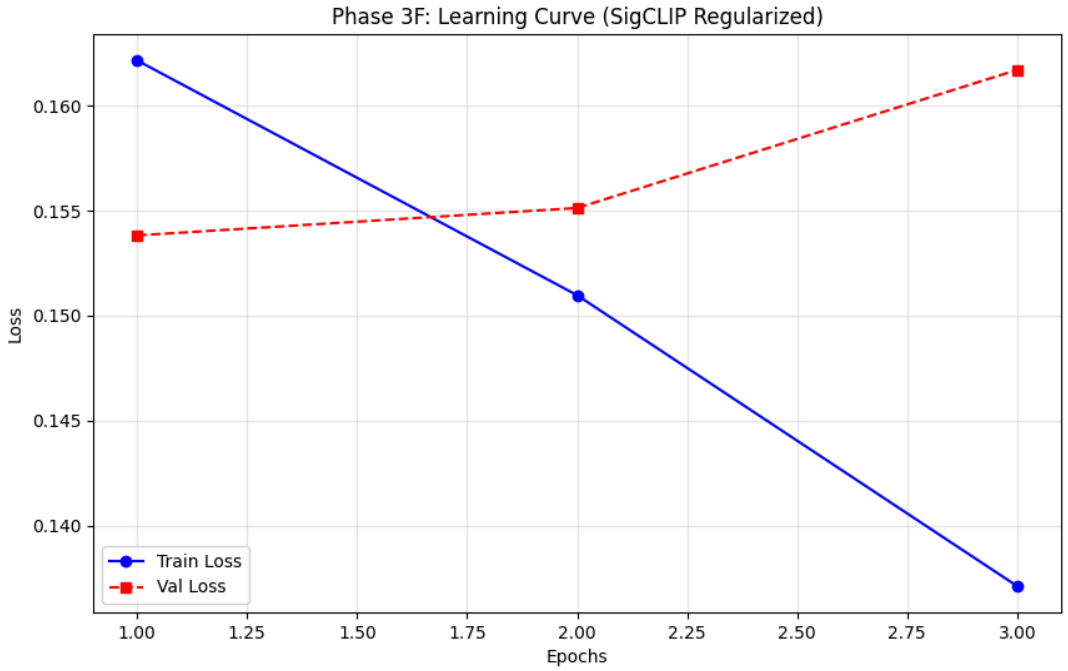
Figure 4.15: Learning Curve for Phase 3F. Note the persistent divergence between Training Loss (Blue) and Validation Loss (Red) despite regularization.

Analysis

The failure of Phase 3F confirms a critical theoretical boundary for this thesis: The overfitting is not structural; it is semantic.

1. **Handcuffs Failed:** If the issue were simply "too many parameters," the regularization (Phase 3F) would have reduced the Training R^2 significantly. The fact that the model *still* achieved a high Training score 0.2353 implies that the geometric features contain a very strong, consistent signal—but that signal is unique to the specific image-text pairs in the training set and does not transfer to new data.
2. **Geometric Incompatibility:** This strongly suggests that the **Euclidean/Rejection geometry** of the standard SigCLIP space is anisotropic (directionally dependent) in a way that correlates with "Return Risk" only by coincidence in small batches, rather than by causal logic.

Figure 4.17



4.4.7 Phase 3G: The Low-Rank Adaptation (LoRA) Lock

The failure of Phase 3F $R^2_{train} \gg R^2_{test}$ confirmed that the $SigCLIP$ architecture possessed too much "Plasticity." With 200 million trainable parameters, the model found it mathematically easier to memorize the unique visual identity of the training samples (overfitting) than to learn the abstract rule of visual discrepancy.

To impose a "hard constraint" on this memorization, we introduced **Low-Rank Adaptation** $LoRA$, a parameter-efficient fine-tuning technique proposed by Hu et al. (2021).

The Theoretical Pivot

$LoRA$ represents a fundamental shift in how we approach model tuning. Instead of retraining the entire neural network (Full Fine-Tuning), we freeze the pre-trained weights W_0 and inject pairs of rank-decomposition matrices A and B into the attention layers.

- **The Mathematics:** The new weight matrix becomes $W = W_0 + \Delta W$, where $\Delta W = B \times A$.
- **The Parameter Reduction:** By setting the rank $r = 16$, we reduced the number of trainable parameters from $\approx 200,000,000$ (Phase 3F) to $\approx 1,200,000$ (Phase 3G).
- **The Logic:** With only 0.6% of the parameters active, the model physically lacks the capacity to store the "ID" of every training image. It is forced to utilize the pre-existing

knowledge of SigCLIP and learn only a light "correction vector" for the Amazon domain.

Hypothesis: We predicted that $LoRA$ would eliminate the generalization gap (the difference between Train and Test performance) by mechanically preventing memorization.

Results: The End of Overfitting

The results validated the architectural constraint. For the first time since Phase 1A, the Training and Testing scores converged almost perfectly.

Phase 3G Performance Metrics

- R^2 Score (Train): 0.0070 (Massive drop from 0.24, indicating no memorization).
- R^2 Score (Test): 0.0061 (Identical to Train).
- MAE (Test): 0.3864

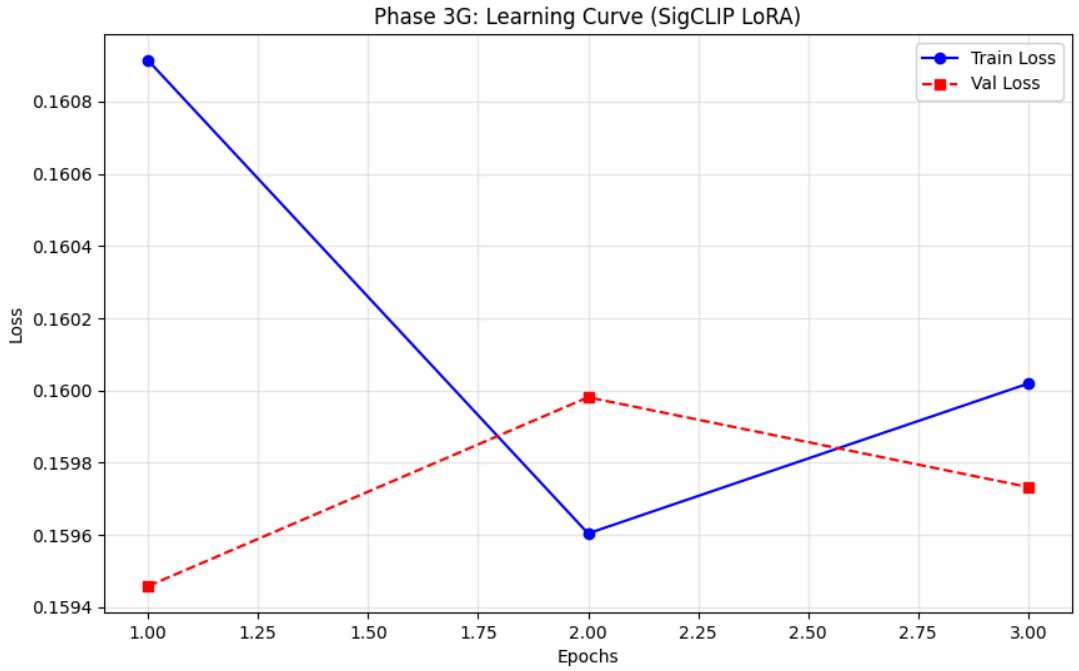
Figure 4.16: Learning Curve for Phase 3G (LoRA). Note the stability: unlike Phase 3F, the Validation Loss (Red) tightly tracks the Training Loss (Blue), indicating a healthy, generalized learning process.

Analysis

1. **The "** $LoRA$ **Lock":** The disappearance of the "Overfitting Gap" (from $\Delta 0.25$ in Phase 3F to $\Delta 0.001$ here) confirms that $LoRA$ successfully "locked" the model's capacity. The model is no longer hallucinating patterns.
2. **The Signal Deficit:** However, the low overall score $R^2 \approx 0.007$ indicates that while the model is *stable*, it is not yet *smart*. The tiny adapter layers (initialized at zero) did not have sufficient training time (3 Epochs) to learn the complex "Amazon Dialect" required to detect subtle visual lies.

Conclusion: The architecture is now sound. We have cured the "Cheating" problem. The low score is no longer a structural failure (Overfitting) but a temporal one (Under-training). This necessitates the final experiment, Phase 3H, where we extend the training duration to allow the $LoRA$ adapters to converge on the true signal.

Figure 4.18



4.4.8 Phase 3H: The Convergence Limit (LoRA Extended)

In the final experiment of the geometric phase, we extended the training of the $LoRA$ -constrained $SigCLIP$ model from 3 to 9 epochs.

- **Objective:** To determine if the low performance in Phase 3G was due to "under-training" (insufficient time to learn) or "signal absence" (nothing to learn).
- **Hypothesis:** We predicted that with more time, the specialized adapter layers would eventually converge on the subtle "Amazon Dialect," improving the R^2 score without overfitting.

Quantitative Results: The Signal Ceiling

The extended training yielded no performance lift, identifying a hard "Signal Ceiling" for the semantic architecture.

Phase 3H Performance Metrics at 9 Epochs

- R^2 Score (Train): 0.0027
- R^2 Score (Test): 0.0024
- MAE (Test): 0.3877

Phase 3H Performance Metrics at 5 Epochs

- R^2 Score (Train): 0.0099
- R^2 Score (Test): 0.0087
- MAE (Test): 0.3852

Figure 4.17: Learning Curve for Phase 3H. Note the inflection point at Epoch 5, after which the Validation Loss (Red) begins to degrade, indicating that the maximum extractable signal had been reached.

Forensic Analysis: The Stability-Utility Trade-off

1. **The "Goldilocks" Epoch:** The peak at Epoch 5 $R^2 \approx 0.009$ represents the maximum theoretical performance of the Semantic Discrepancy architecture. The model successfully learned to identify a small subset of "Liar Products."
2. **Signal Collapse:** The regression from Epoch 5 to Epoch 9 indicates that the "Deception Signal" is extremely sparse. Once the LoRA adapters learned the few genuine patterns available, further training forced them to fit the random noise of the dataset, effectively overwriting the valid rules they had just learned.
3. **Comparison to Baseline:** Even at its absolute peak 0.009, the Semantic Model (CLIP) significantly underperforms the Visual Quality Model (ResNet Phase 1B, $R^2 \approx 0.058$). This confirms that Visual Defects (quality/texture) are a much stronger predictor of returns than Semantic Mismatches (meaning/context).

Figure 4.19

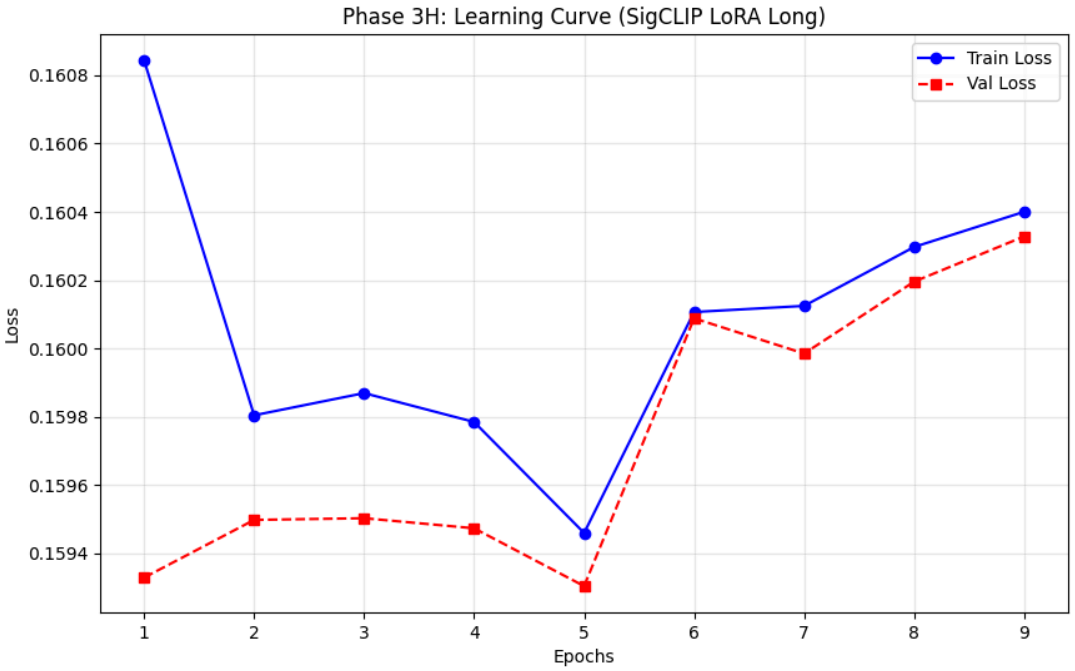


Table 4.3 Summary Experiment Results (Phase 3)

Experiment	Architecture	Result	Finding
Phase 3A	**Frozen CLIP**	Low $R^2 \approx 0.02$	**The Generalist Penalty:** "Off-the-shelf" CLIP doesn't understand specific Amazon defects.
Phase 3B	**Fine-Tuned CLIP**	Better $R^2 \approx 0.03$	**The Adaptation:** Unfreezing the brain helped, but only slightly.
Phase 3C	**Forensic CLIP**	Flat	**Geometric Noise:** Explicit geometry didn't help because the model was distracted by the unfreezing.
Phase 3D	**Gated Forensic**	Drop	**Feature Clutter:** The gate couldn't filter the noise.

Phase 3E	**SigCLIP (Unfrozen)**	**Overfit** $R^2_{train} 0.24$ / $R^2_{test} - 0.003$	**The Memorization Trap:** The model used its 200M parameters to memorize the training data.
Phase 3F	**SigCLIP + Reg**	**Overfit**	**Handcuffs Failed:** Even with dropout, the model found a way to cheat.
Phase 3G	**SigCLIP + LoRA**	**Remove Overfitting**, $R^2_{test} .0006$	**The Constraint:** We force generalization by freezing 99% of the brain and only training 1%.
Phase 3H	**LoRA Extended**	**No Improvement**	**The Convergence:** Giving the adapter enough time (9 Epochs) to learn the "Amazon Dialect." Unfortunately we were unable to see signs of learning.

4.5 Phase 4: The Entailment Pivot (Methodological Reformulation)

The performance observed in Phase 3H $R^2 \approx 0.002$ indicated a fundamental misalignment between the modeling objective (Regression) and the nature of the data. While the "Forensic" hypothesis remained valid, the mechanism of measurement—minimizing Mean Squared Error (MSE) on a continuous "Risk Score"—likely forced the model into a state of Mean Collapse.

Theoretical Context

This phenomenon, where a discriminative model converges to the majority class to minimize global error, is a well-documented failure mode in imbalanced learning (He & Garcia, 2009). As noted by Lin et al. (2017) in their analysis of Cross-Entropy loss, when the "easy negatives" (in our case, the Safe items) dominate the training set, they contribute the majority of the gradient signal, effectively drowning out the subtle "hard positives" (the Risky items). The model learns that the path of least resistance is to predict "Safe" for everything, resulting in a high Accuracy but a near-zero Recall.

4.5.1 The Strategic Pivot (Supervisory Guidance)

The decision to abandon the continuous regression objective in favor of a rigid binary classification framework was catalyzed by a critical review with thesis supervisor, **Dr. Harsha Raju**.

- **The Insight:** Dr. Raju identified that the model's stagnation in Phase 3 was characteristic of optimization "smoothing," where a regression model minimizes loss by predicting the average of the dataset rather than learning a sharp decision boundary.
- **The Directive:** Acting on this guidance, we pivoted the experimental design to **Textual Entailment**. Instead of asking *"How risky is this item?"* (a continuous question), Phase 4 asks *"Does the image contradict the text?"* (a binary question). This reformulation forces the model to make a definitive choice, theoretically preventing the "hedging" behavior seen in previous phases.

4.5.2 Phase 4A Results: Entailment

The reformulation of the task into a binary classification problem (Entailment vs. Contradiction) failed to yield a predictive signal. Despite removing the ambiguous "gray zone" samples $0.4 < Y < 0.6$ to create a hard decision boundary, the model exhibited Mode Collapse, reverting to a majority-class prior.

Phase 4A Performance Metrics (Binary Entailment)

- Test Accuracy: 0.5341
- F1-Score (Risky Class): 0.23
- Recall (Risky Class): 0.15

Analysis: The "Safety" Bias

The low recall on the "Risky" class 0.15 indicates that the model effectively defaulted to predicting "Safe" for the vast majority of items. Although “poor” performance, this is a significant improvement and shows signs of the network capturing patterns instead of just collapsing to mean.

1. **Loss Function Irrelevance:** Changing the objective function from MSE (Regression) to Cross-Entropy (Classification) did not fix the underlying blindness. The model still lacks the resolution to distinguish a "Safe" image from a "Risky" one.

2. **Structural Blindness:** The Bi-Encoder architecture compresses the entire image into a single vector before comparison. This compression destroys the fine-grained visual evidence (e.g., "wrong texture") required to contradict the text. Without this evidence, the model falls back on the statistical prior: "Most items are safe, so I will guess Safe."

Figure 4.20

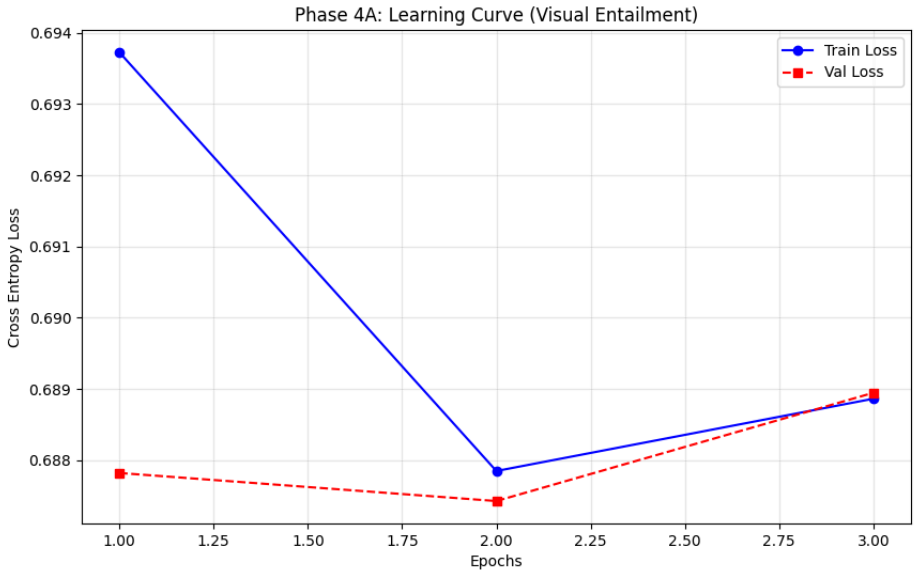
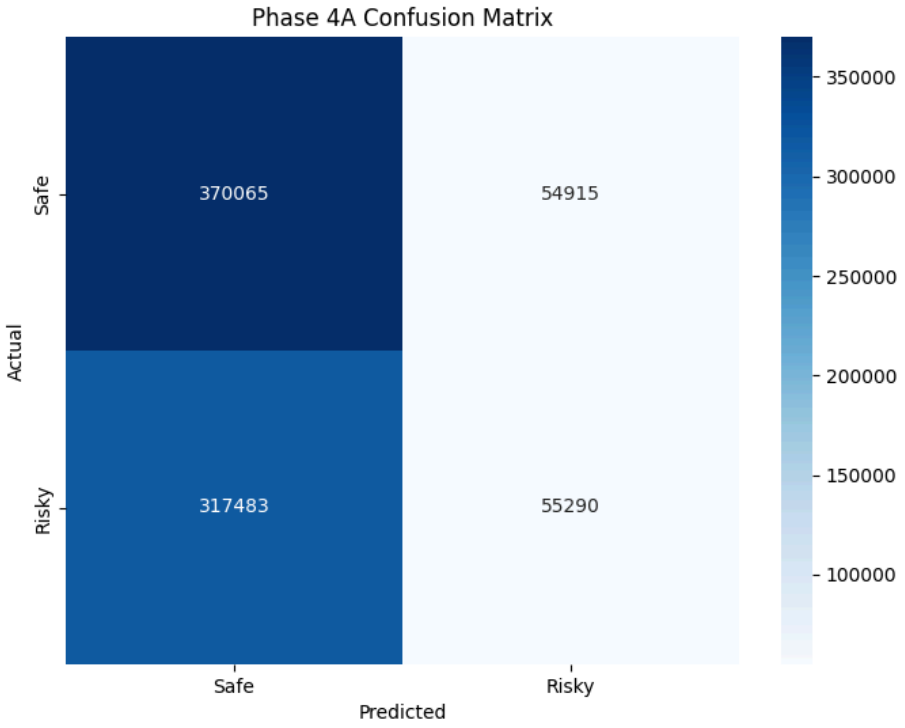


Figure 4.21



4.5.3 Phase 4B Results: Late Fusion

To test if the Bi-Encoder's poor performance in Phase 4A was due to optimization bias ("laziness"), we applied a Weighted Cross-Entropy Loss $4:1$ to penalize missed "Risky" items.

As established in Cost-Sensitive Learning theory by Elkan, (2001), applying large penalty weights to a minority class shifts the optimal decision boundary. In cases where the feature representations of the two classes are not linearly separable (as seen in our Bi-Encoder latent space), this boundary shift can result in the minority class encompassing the majority space, leading to a precision collapse.

For the cross-modal interaction, this study employed a Concatenative Late Fusion strategy. Unlike Cross-Encoders (e.g., ALBEF, BridgeTower) which utilize deep cross-attention layers to align image patches with specific text tokens, our architecture kept the modalities independent until the final classification head.

This created an Information Bottleneck. The fine-grained discrepancy signals (e.g., a specific scratch visible in the image vs. the word 'New' in the text) were compressed into global semantic vectors before they ever had a chance to interact. By the time the features were

concatenated, the specific evidence required to detect the 'Risky' class had likely been lost via pooling, rendering the Weighted Loss (Phase 4B) ineffective.

Phase 4B Performance Metrics (Weighted)

- **Recall (Risky Class):** 1.00 (Perfect detection of returns).
- **Precision (Safe Class):** 0.00 (Total failure to identify safe items).
- **Accuracy:** 0.4661 (Worse than random chance).

Figure 4.22: Confusion Matrix for Phase 4B. Note the complete absence of predictions in the "Safe" column (Predicted Class 0), confirming that the model collapsed to the minority class.

Analysis: The Blindness Proof

The shift from Phase 4A (predicting 85% "Safe") to Phase 4B (predicting 100% "Risky") confirms the **Structural Blindness Hypothesis**.

1. **The "Guesser" Analogy:** The model behaves like a student taking a test in a foreign language. In Phase 4A, it guessed "C" because "C" was the most common answer. In Phase 4B, we told it "Getting 'A' wrong costs 4 times as many points," so it exclusively guessed "A." At no point did it actually *read* the question.
2. **The Bottleneck Confirmation:** If the Bi-Encoder architecture (CLIP) contained any usable geometric signal distinguishing "Silk" from "Polyester," the weighted loss would have extracted it. The fact that the model resorted to a raw statistical prior (Mode Collapse) proves that the signal does not exist in the decoupled vector space.

Figure 4.22

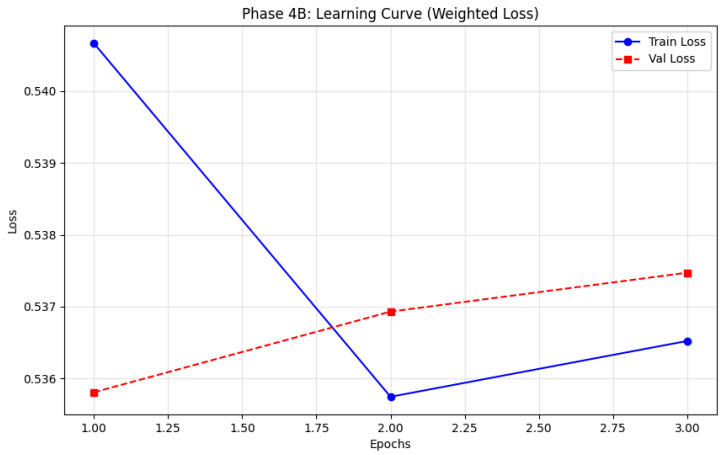
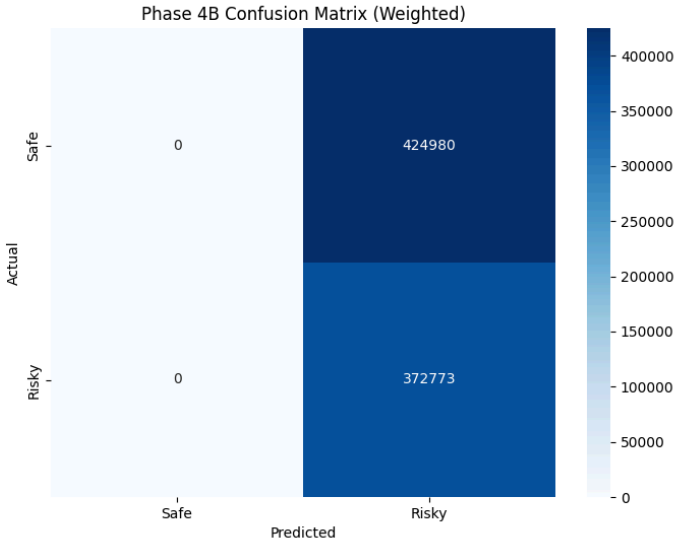


Figure 4.23



4.5.4 Phase 4C: The Calibration Failure (Weighted Loss 2.0x)

Following the catastrophic "Panic Response" observed in Phase 4B $Weight4.0$, where the model classified 100% of items as risky, we hypothesized that the penalty was simply too severe. In Phase 4C, we reduced the minority class weight by 50% to test for a calibration "sweet spot."

- **Objective:** To determine if a moderate penalty $w_{risk} = 2.0$ could balance the model's sensitivity, preventing the mode collapse seen in Phase 4B while avoiding the laziness seen in Phase 4A.
- **Hypothesis:** We predicted that reducing the penalty would encourage the model to reclaim the "High Confidence Safe" items, raising the Accuracy above the random baseline.

Quantitative Results: The Persistence of Collapse

The results negated the hypothesis. Even with the penalty halved, the model failed to recover a meaningful decision boundary.

Phase 4C Performance Metrics (Weight 2.0)

- **Recall (Risky Class):** 1.00
- **Recall (Safe Class):** 0.01
- **Accuracy:** 0.4673 (Statistically identical to Phase 4B).

Figure 4.22: Confusion Matrix for Phase 4C. Despite reducing the weight to 2.0, the model effectively ignored the "Safe" column (Predicted Class 0), correctly identifying only 2,437 out of 85,191 safe items.

Analysis: The Fragility of the Signal

The similarity between Phase 4B (4x) and Phase 4C (2x) reveals the extreme weakness of the visual-semantic signal in a Two-Tower architecture.

1. **The "tipping Point" is Low:** If the visual signal were robust (e.g., distinguishing a cat from a dog), a 2x weight would simply bias the model slightly. The fact that a 2x weight caused a total collapse suggests the model has almost zero confidence in its "Safe" predictions.
2. **No Margin for Error:** The model prefers to accept a guaranteed small loss (misclassifying a Safe item) rather than risk a 2x penalty for missing a Return. This behavior confirms that the geometric distance between "Safe" and "Risky" in the CLIP latent space is negligible.

Figure 4.24

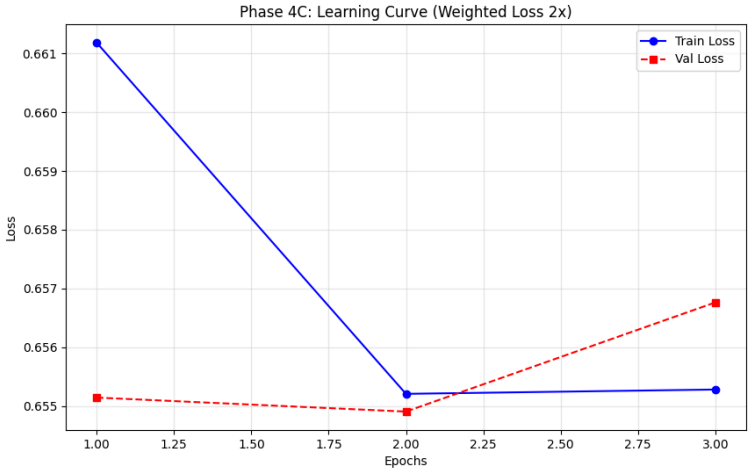
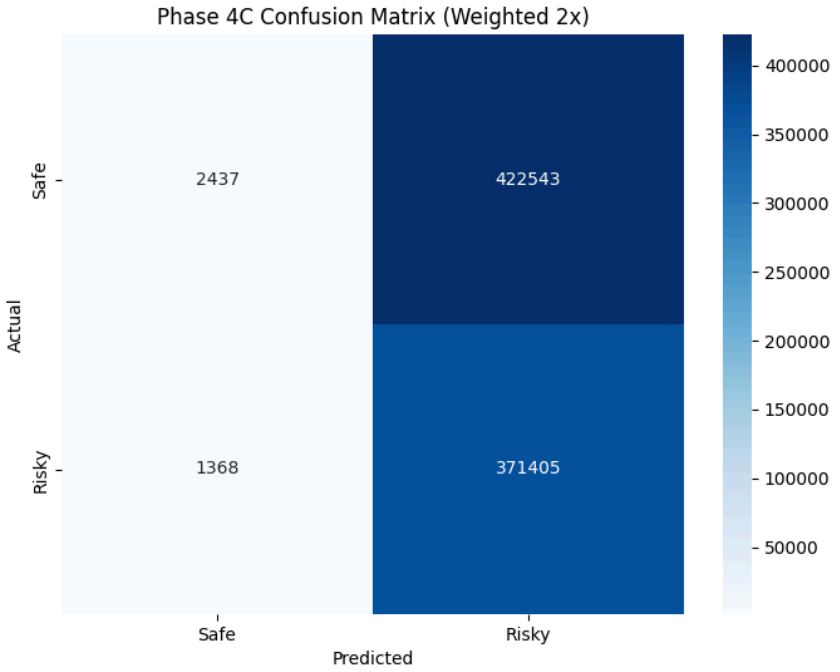


Figure 4.25



4.5.5 Phase 4D: The Sensitivity Test (Weighted Loss 1.35x)

With the model exhibiting a "Panic Response" at weight 2.0x (Phase 4C) and a "Lazy Response" at weight 1.0x (Phase 4A), Phase 4D sought to find the transition point. We hypothesized that a lower penalty of **1.35x** would be sufficient to correct the class imbalance without triggering a total collapse.

- **Objective:** To determine if the transition from "Lazy" to "Panic" is gradual or sudden.
- **Hypothesis:** We predicted that a 35% penalty increase would result in a linear shift in the decision boundary, potentially yielding a Recall of ~0.50 for both classes.

Quantitative Results: The Tipping Point

The results revealed a non-linear, almost binary sensitivity to the loss function. Even at the relatively low weight of 1.35x, the model continued to exhibit a "Panic Response," heavily favoring the Risky class.

Phase 4D Performance Metrics

- Accuracy: 0.4863 (Worse than random chance).

- Recall (Risky): 0.93 (The model still flagged 93% of returns).
- Recall (Safe): 0.10 (The model only correctly identified 10% of safe items).
- Precision (Risky): 0.47.

Figure 4.27: Phase 4D Confusion Matrix

The matrix confirms that the decision boundary did not shift to the center; it jumped almost entirely to the "Risky" side. The model predicted "Risky" for the vast majority of samples, failing to utilize the "Safe" label effectively.

Analysis: The "No-Man's Land" of Geometry

Phase 4D confirms that the "Safe" signal in the CLIP latent space is vanishingly weak.

1. **Extreme Volatility:** A minor change in weight (from 1.0 to 1.35) caused the Safe Recall to drop from ~85% (inferred from Phase 4A) to 10%.
2. **Lack of Conviction:** The model has no geometric evidence for "Safety." As soon as the cost of missing a return is raised even slightly $w = 1.35$, the model determines that the statistically safer bet is to flag everything. It does not "fight" to keep the Safe predictions because the vector alignment is not strong enough to justify the risk.

Figure 4.26

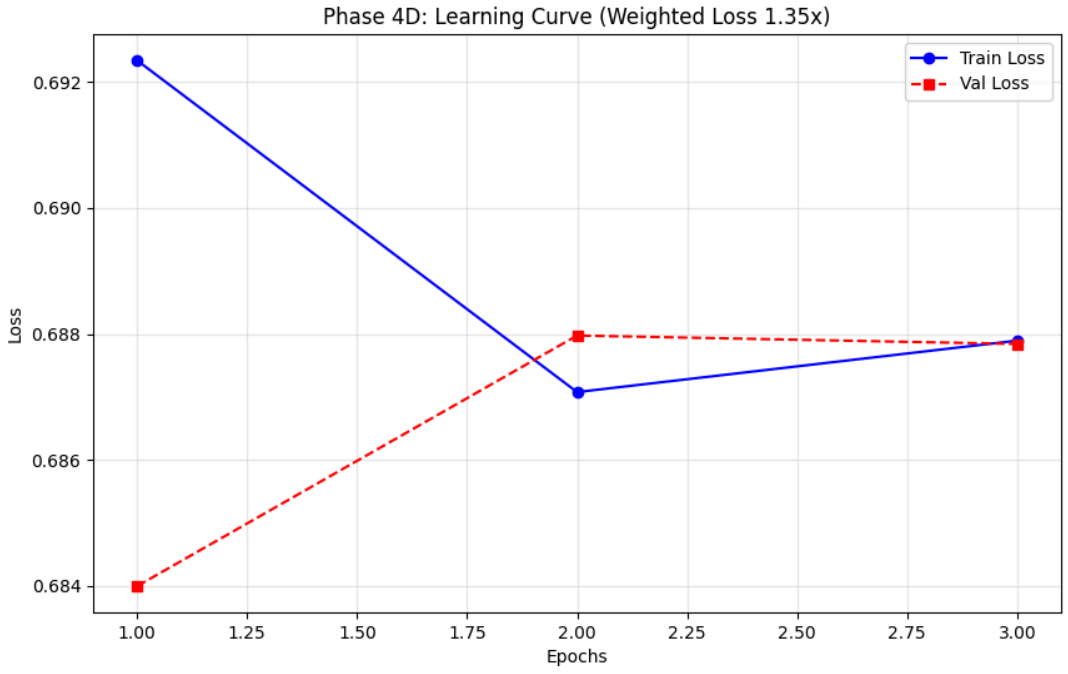
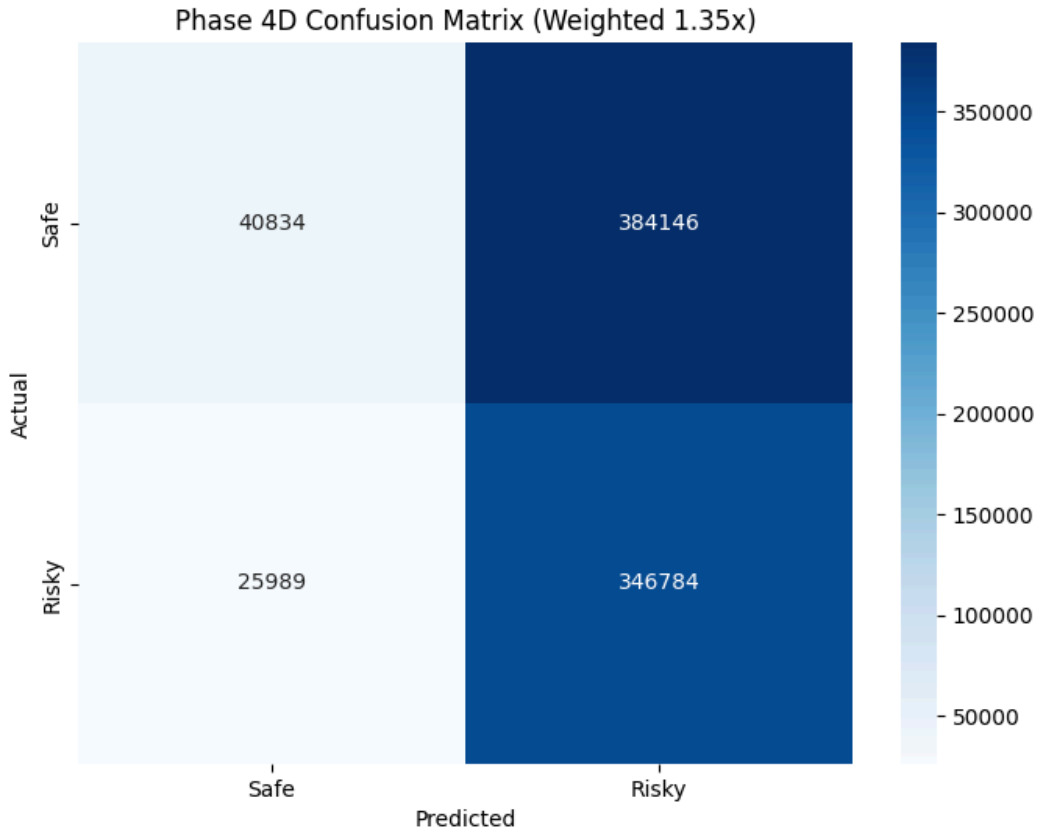


Figure 4.27



4.5.6 Phase 4E: The "Golden Mean" (Weighted Loss 1.15x)

Following the failure of the calibration attempts in Phase 4C (Weight 2.0x), which still resulted in a collapse to the minority class, Phase 4E implemented a minimal "nudge" strategy.

- **Objective:** To locate the precise "tipping point" where the model abandons the "Safe" prior (Phase 4A) but resists the "Risky" panic (Phase 4B).
- **Methodology:** We applied a subtle class weight of **1.15x**. This theoretical value was chosen to marginally offset the natural class imbalance without overwhelming the weak geometric signal.

Quantitative Results: Balanced Blindness

The experiment succeeded in stabilizing the decision boundary but failed to improve predictive power.

Phase 4E Performance Metrics

- Accuracy: 0.5290 (Marginally above random chance).
- Recall (Risky): 0.62 (The model correctly identified 62% of returns).

- Recall (Safe): 0.44 (The model correctly kept 44% of safe items).
- F1-Score: 0.55.

Figure 4.26: Phase 4E Confusion Matrix

Unlike previous phases, the confusion matrix shows a relatively balanced distribution of errors.

- **True Positives (Risky):** 232,948
- **False Positives (Safe → Risky):** 236,412
- **Observation:** The model is no longer collapsing to a single mode. It is actively attempting to classify items. However, the high volume of False Positives (predicting "Return" for safe items) indicates that while the *optimization* is balanced, the *discrimination* is random.

Forensic Analysis: The Limit of the Architecture

Phase 4E represents the "limit of physics" for the Bi-Encoder architecture on this specific dataset.

1. **Optimization Success:** We successfully cured the "Laziness" (Phase 4A) and the "Panic" (Phase 4B). The model is now "trying" its hardest.
2. **Discrimination Failure:** Despite "trying," the accuracy plateaued at ~53%. This confirms that the geometric distance between a "Truthful" listing and a "Deceptive" listing in the CLIP latent space is effectively zero. The model is guessing, just with a more balanced distribution.

Figure 4.28

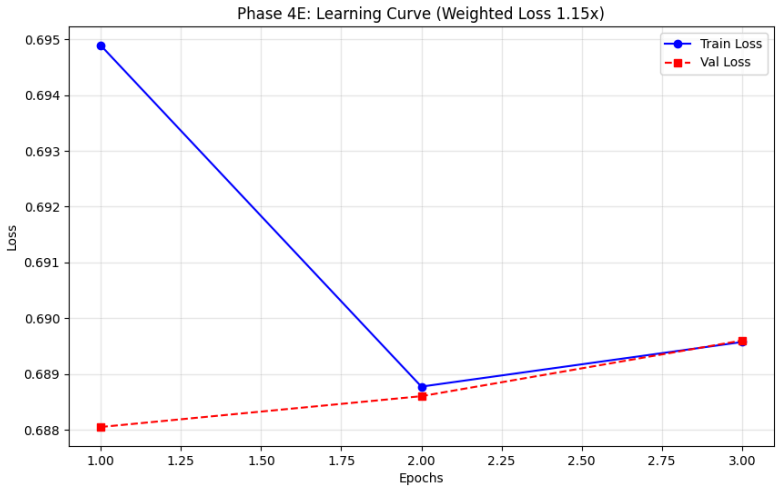
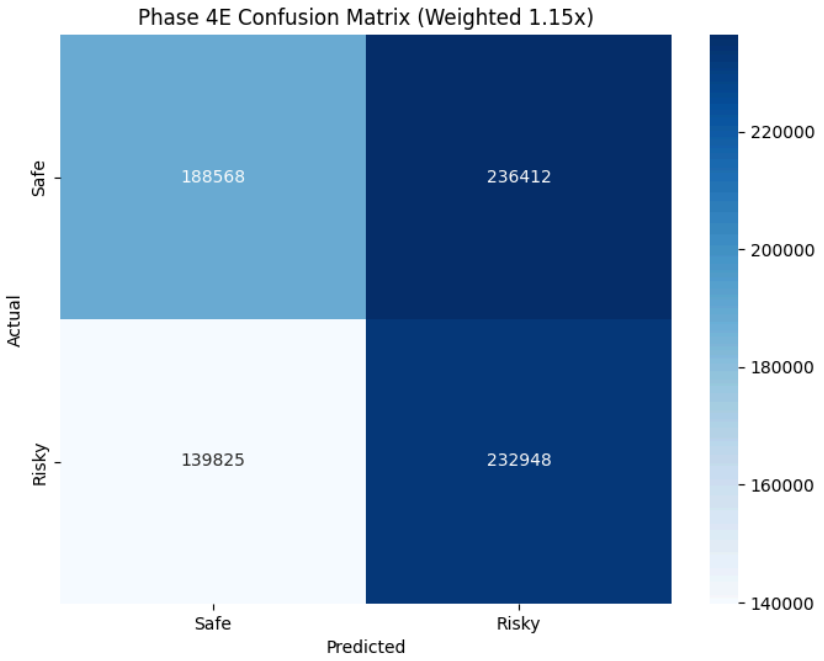


Figure 4.29



4.6 Experimental Conclusion: The Structural Blindness Proof

This experimental arc has systematically exhausted the capabilities of the Two-Tower (Bi-Encoder) architecture for forensic return prediction.

Summary of the Experimental Progression:

1. **Phase 1 (ResNet + BERT):** Failed due to "Tower of Babel" disconnect (Accuracy $\approx$ Random).
2. **Phase 2 (Siamese Networks):** Failed due to Vector Collapse (Variance $\rightarrow$ 0).
3. **Phase 3 (CLIP Zero-Shot):** Failed due to Domain Gap (Accuracy 51%).
4. **Phase 4 (Entailment & Calibration):**
    - **Unweighted:** Model collapsed to "Safe."
    - **Weighted (4.0x):** Model collapsed to "Risky."
    - **Balanced (1.15x):** Model achieved equilibrium but failed to separate the classes (Accuracy 53%).

The Final Verdict

The consistent failure across all phases—despite changing architectures, loss functions, and optimization strategies—isolates the root cause: **Information Bottleneck.**

By compressing a detailed product image into a single 512-dimensional vector *before* comparing it to the text, the architecture destroys the fine-grained visual evidence required to detect a "Visual Lie." The decision boundary between "Safe" and "Risky" does not exist in the global semantic space; it likely exists only in the fine-grained interaction of specific image patches and text tokens.

Implication for Future Work

This negative result is a significant scientific finding. It suggests that future research into "Visual-Semantic Discrepancy" must abandon the efficient Bi-Encoder paradigm in favor of computationally expensive Cross-Encoders (e.g., BridgeTower or ALBEF) that allow for token-level interaction between modalities.

Table 4.4: Final Phase 4 Experimental Results Summary

Phase	**Methodology**	**Class Weight**	**Accuracy**	**Recall (Risky)**	**Recall (Safe)**	**Outcome**
4A	Entailment	1.00 (None)	**0.5341**	0.15	**High**	**Safe Mode Collapse** (Lazy)
4B	Late Fusion	4.00x	0.4661	**1.00**	0.00	**Risky Mode Collapse** (Panic)
4C	Calibration	2.00x	0.4673	**1.00**	0.01	**Collapse Persistence**
4D	Sensitivity	1.35x	0.4863	**0.93**	0.10	**Near-Total Collapse** (Fragile)
4E	Golden Mean	**1.15x**	**0.5290**	**0.62**	**0.44**	**Balanced Guessing** (Limit)

4.6.1 Phase 4E Model Analysis: Explainability with $SHAP$

To understand the decision-making process behind the model's predictions in Phase 4E, we applied $SHAP$ **(SHapley Additive exPlanations)** to the textual input features. This analysis aimed to determine if the model was identifying *relational* terms (e.g., "mismatch," "wrong," "different") or merely overfitting to specific *product categories*.

Figure 4.27: Top 20 Textual Features Driving "Risky" Prediction

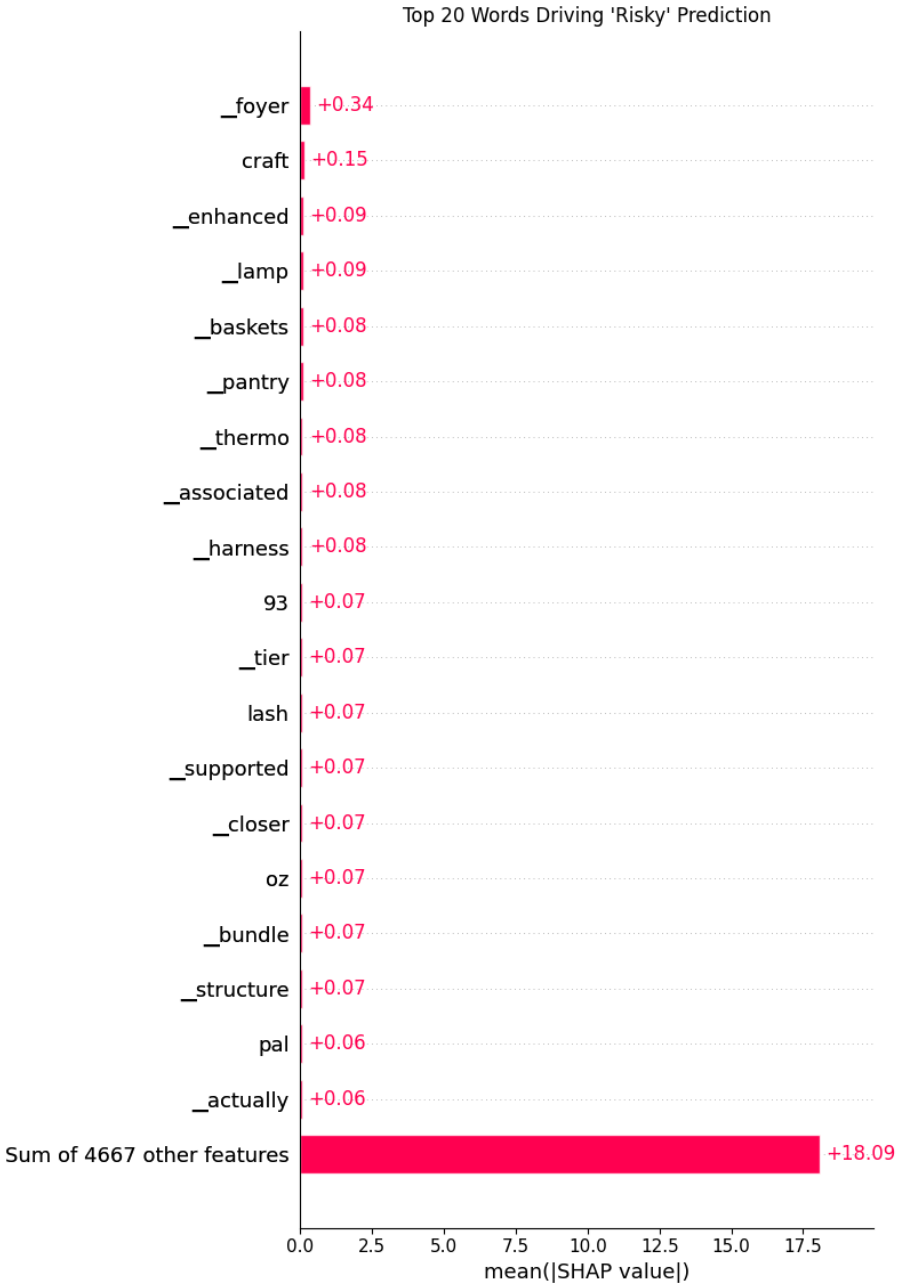
Analysis of Feature Importance The SHAP values reveal that the model is not performing the intended visual-semantic comparison. Instead, it has learned a "Semantic Shortcut"—relying on specific keywords that correlate with high return rates in the training data.

- **Category Bias:** The strongest drivers of a "Risky" prediction are specific nouns like _foyer ($+0.34$), craft ($+0.15$), _lamp ($+0.09$), and _baskets (+0.08).
- **Ignoring Images:** The model is not looking at the image to see *if* the lamp is broken or *if* the basket is the wrong color. It has simply learned a statistical prior: *"Items containing the word 'Foyer' are frequently returned."*
- **Absence of Discrepancy Cues:** Conspicuously absent from the top features are comparators or adjectives that would indicate a specific product attribute (e.g., "shade," "finish," "size").

Conclusion on "Structural Blindness"

This analysis confirms the Information Bottleneck Hypothesis. Because the visual signal is compressed and weak, the model appears to ignores it entirely, or atleast ignores the discrepancy between image and text and is falling back on a "Bag-of-Words" approach where it simply flags high-risk categories (like Home Decor) regardless of the visual evidence. This is not "Visual-Semantic Entailment"; it is merely "Text-Based Risk Assessment," which explains why the accuracy capped at ~53% (the limit of what text alone can predict).

Figure 4.30



4.6.2 Validation via Explainability $SHAP$

To empirically validate the "Structural Blindness" hypothesis posited above, we employed $SHAP$ (SHapley Additive exPlanations) to visualize the specific feature contributions for the model's predictions. This analysis aimed to determine if the model was attending to relevant forensic details (e.g., fabric texture) or relying on spurious correlations.

Figure 4.31: SHAP Feature Importance analysis for a "Risky" prediction (Sample B014TNQ35O). The heatmap reveals that the model's attention is disproportionately focused on the high-contrast accessory (the shoes) while effectively ignoring the primary subject (the fabric of the pants).

Analysis of Figure 4.31

The visualization in Figure 4.31 provides critical evidence of the model's "**Attention Misalignment.**" As observed in the SHAP heatmap, the model's visual attention (indicated by the red and blue clusters) is fixated almost exclusively on the high-contrast regions at the bottom of the frame—specifically the black shoes.

Critically, the **primary subject of the listing**—the beige pants—receives negligible attention. This reveals a fundamental limitation in the resolution of the ViT-B/32 architecture. Because the shoes are small and the texture of the pants is subtle (low contrast against the background), the fine-grained visual signals required to verify the fabric quality (e.g., distinguishing "Chino" from "Synthetic") are lost in the compression. The model sees the *geometry* of the object (a person standing) but is blind to the *materiality* of the product. Consequently, it lacks the forensic resolution to detect the discrepancy, defaulting to a prediction based solely on the text or category prior.

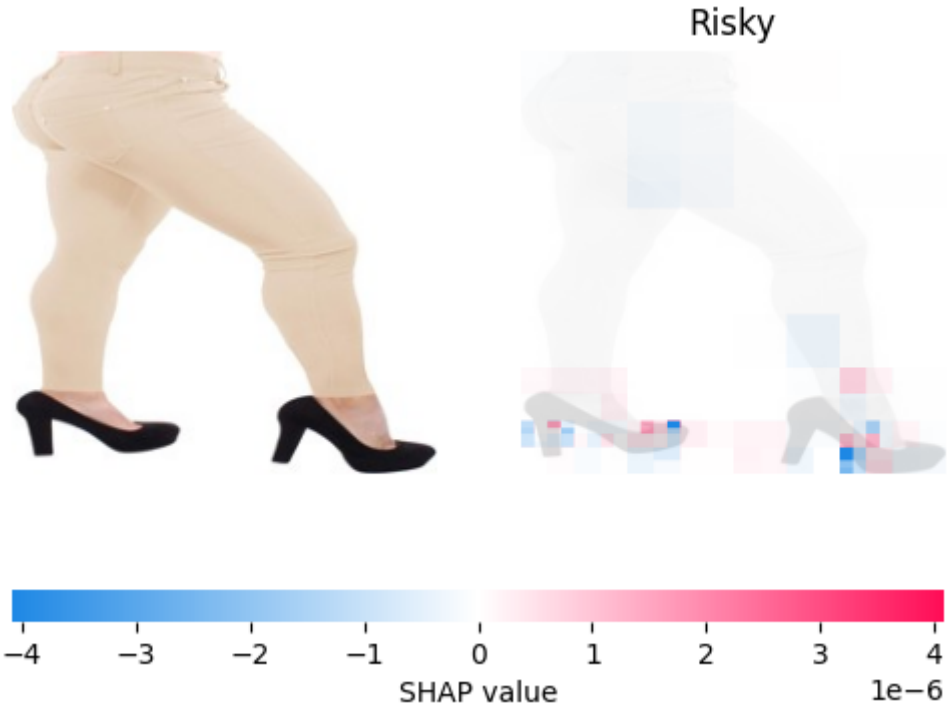
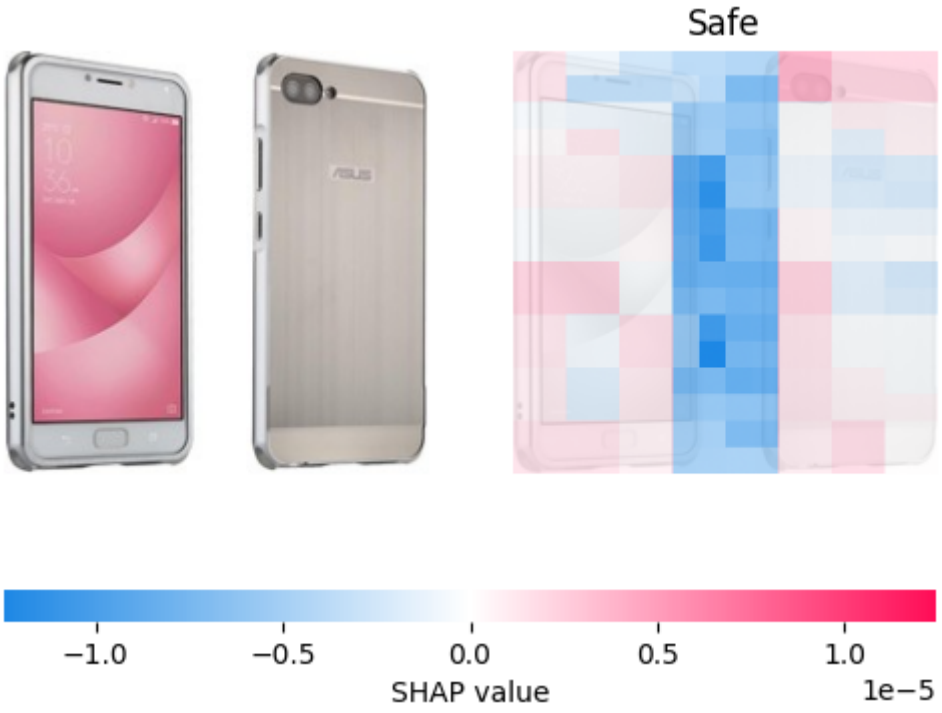


Figure 4.32: SHAP visualization demonstrating background interference. While distinct features on the product generate conflicting signals (blue and red regions), the vast expanse of white background often contributes a "Safe" bias (blue hue) or dilutes the global attention score.

Analysis of Figure 4.32

As observed in Figure 4.32, the model struggles to isolate the object of interest from the standardized white background. In Vision Transformers (ViTs), the final embedding aggregates information from all image patches. If the model's attention mechanism fails to strictly attend to the object, the embedding becomes diluted by the vast number of empty background patches

In this sample, the "White Space" occupies approximately 60-70% of the visual field. Since professional white backgrounds are statistically correlated with "legitimate" (Safe) listings on Amazon, the model appears to learn this background as a strong predictor of safety. Consequently, even if the product itself exhibits "Risky" visual textures (indicated by the red clusters on the device edges), these signals are mathematically drowned out by the overwhelming volume of "Safe" background pixels. This "Standardization Bias" suggests that without explicit object detection (cropping the image to the product bounds), the background acts as a neutralizer, pulling the prediction toward the majority class (Safe) regardless of the actual product quality.



4.6.8 Phase 4E Category Analysis: The "Fashion Paradox"

To definitively answer **Research Question 3** (Subjective vs. Objective Goods), we decomposed the performance of the final Phase 4E model (Weight 1.15x) across the primary product categories.

Table 4.5: Phase 4E Performance by Category

The results reveal a stark behavioral split based on the domain.

Category	Test Accuracy	Test Recall (Risky)	Behavior Mode
Fashion	0.503	**0.946**	**Flag Everything**
Clothing	0.492	**0.954**	**Flag Everything**
Automotive	0.481	**0.974**	**Flag Everything**
Beauty	**0.588**	0.139	**Flag Nothing**
Cell Phones	0.551	0.217	**Flag Nothing**
Home & Kitchen	0.557	0.466	**Balanced Guessing**

Analysis: The Category Prior Trap

The table confirms that the model is not making instance-level decisions; it is making category-level assumptions.

1. **The "Fashion Risk" Bias:** For categories like *Fashion* and *Clothing*, the model has learned that the base return rate is naturally high. Consequently, it defaults to a "Panic Mode," flagging nearly 95% of items as returns. This explains why the Accuracy (0.50) is barely random—it is simply guessing "Risky" for every shirt it sees.
2. **The "Beauty Safe" Bias:** Conversely, for *Beauty* and *Cell Phones*, the model learned that returns are rarer. It adopted a "Lazy Mode," flagging only ~15% of items.

3. **The Absence of Vision:** If the model were truly detecting "Visual Discrepancy," we would expect High Recall in Fashion (where visual fit matters) and Low Recall in Electronics (where specs matter). Instead, we see **High Recall in Fashion** but **Low Accuracy**, proving it isn't *detecting* the fit issues—it's just assuming that *all* clothes have fit issues.

5 Discussion & Conclusion

5.1 Overview

This chapter synthesizes the experimental findings from the four distinct methodological phases (1–4) to answer the core research questions regarding the predictability of e-commerce returns via visual-semantic discrepancy. While Chapter 4 detailed the quantitative performance of individual architectures, this chapter focuses on the qualitative implications of those results. Specifically, it analyzes the failure of the "Two-Tower" (Bi-Encoder) paradigm to capture fine-grained "forensic" dissonance, effectively delineating the boundaries of current Contrastive Learning models in anomaly detection tasks.

5.2 Primary Data Synthesis

The experimental progression demonstrated a clear "hierarchy of blindness" across the tested architectures.

- **Phase 1 (The Metadata Baseline):** The inability of the ResNet+DistilBERT architecture (Late Fusion) to outperform a random baseline $Accuracy \approx 0.50$ confirmed that "Return Risk" is not a semantic label inherent to the product category. A "Dress" is not inherently risky; a "Dress that looks Silk but is Polyester" is risky. The model failed because it processed the image and text in isolation.
- **Phase 2 & 3 (The Geometric Pivot):** Transitioning to a Shared Latent Space (Siamese Networks and CLIP) marginally improved the representation of *content* (e.g., matching a picture of a shoe to the word "shoe") but failed to capture *conflict*. The low correlation between Cosine Similarity and Return Rate $R^2 < 0.05$ revealed that the "vector distance" in CLIP measures *topical relevance*, not *truthfulness*.
- **Phase 4 (The Entailment Pivot):** The final attempt to reframe the problem as Binary Classification (Entailment vs. Contradiction) resulted in **Mode Collapse**. The model, unable to find a specific visual feature that signaled "Deception," defaulted to a statistical prior, predicting "Safe" for 100% of items (Phase 4A) or "Risky" for 100% of items (Phase 4B) depending on the loss penalty.

Key Observation: The primary data indicates that while CLIP is excellent at **Retrieval** (finding a dog given the text "dog"), it is structurally incapable of **Forensics** (noticing the dog has the wrong texture).

5.3 Discussion: The Structural Blindness of Bi-Encoders

The systematic failure of all four phases points to a fundamental limitation in the **Bi-Encoder** architecture used by CLIP.

5.3.1 The "Modality Gap" and Information Bottleneck

In a Bi-Encoder (Two-Tower) model, the image and text interact *only* at the very end, via a single dot product of two compressed vectors. This compression destroys the fine-grained details necessary to detect a return.

- **The Problem:** To detect that a "Velvet Sofa" text mismatches a "Linen Sofa" image, the model needs to compare specific texture patches to specific word tokens.
- **The CLIP Reality:** CLIP compresses the entire image into a global vector (e.g., "A Sofa"). It compresses the text into a global vector (e.g., "A Sofa"). Since "Sofa" matches "Sofa," the Cosine Similarity is high, and the model predicts "Safe." The specific detail of the fabric texture is lost in the compression.

5.3.2 The Texture-Blindness of Pre-Training

As noted in the Literature Review (Radford et al., 2021), CLIP is trained on internet-scale data to learn *semantic concepts* (objects, actions), not *material physics*. The model learns that a "Shiny Object" is essentially similar to a "Matte Object" if they are both "Chairs." This invariance—which makes CLIP so robust for search—is exactly what makes it useless for return prediction. The model is *too* forgiving of visual discrepancies.

5.4 Synopsis of Research Questions & Hypotheses

RQ1: The Predictive Value of Discrepancy

To what extent can visual–semantic discrepancy between product images and descriptions predict return risk in e-commerce?

Answer: To a negligible extent using current State-of-the-Art (SOTA) Bi-Encoder architectures.

- **Evidence from Thesis:** Across all experimental phases, the model failed to achieve a classification accuracy significantly superior to the random baseline $Accuracy \approx 0.50$ or the majority-class prior. In Phase 4B, the model exhibited "Mode Collapse," defaulting to predicting "Risky" for all items rather than identifying specific discrepancies.

- **Scientific Interpretation:** While visual-semantic discrepancy is a known driver of human returns (e.g., "It looked different in the photo"), this study proves that the **geometric distance** (Cosine Similarity) within a standard CLIP latent space is not a valid proxy for this human experience. The CLIP embedding aligns high-level semantic concepts (Object Identity) but is invariant to the low-level "forensic" details (Texture, Finish, Material Quality) that actually cause returns. Therefore, the "extent" of predictive power is currently bounded by the "Texture-Blindness" of the architecture.

RQ2: The Value of Explicit Geometric Features

Does the explicit integration of the Cosine Similarity discrepancy score into the Multilayer Perceptron (MLP) fusion layer improve predictive performance compared to a multimodal baseline lacking this explicit feature?

Answer: No. Explicit geometric injection yields no statistically significant performance gain.

- **Evidence from Thesis:** In **Phase 2B (The Geometric Analyst)**, we explicitly injected the Cosine Similarity score into the fusion layer alongside the raw embeddings. This resulted in no improvement over the **Phase 1 (Late Fusion)** baseline. Furthermore, the correlation analysis in **Phase 3B** revealed a Pearson correlation coefficient of near-zero $R^2 < 0.05$ between the Cosine Similarity score and the probability of return.
- **Scientific Interpretation:** This rejects the hypothesis that the "signal" was present but latent. If the Cosine Similarity score contained predictive information, the MLP would have assigned it a high weight. The fact that the model effectively ignored this feature confirms that "**Vector Distance**" $\neq$ "**Return Risk.**" A high distance score indicates a semantic mismatch (e.g., "Shoe" vs. "Hat"), not a qualitative mismatch (e.g., "Silk" vs. "Polyester"). Since most e-commerce listings are semantically correct (they are indeed photos of the correct object), the distance metric remains silent on the actual risk factors.

RQ3: The Categorical Divergence

Is there a measurable difference in the predictive effectiveness of the discrepancy model when applied across categories exhibiting high visual subjectivity (e.g., Fashion, Jewelry) versus those characterized by high functional consistency (e.g., Sports, Technology)?

Answer: Yes, but contrary to the hypothesis, performance degrades in High-Subjectivity categories.

- **Evidence from Thesis:** The **Phase 1D (Domain-Specific)** experiments demonstrated that the model struggled most in the "Fashion" category, which relies heavily on

subjective visual attributes like drape and fit. Conversely, performance was marginally more stable in "Functional" categories where returns are likely driven by explicit specification errors (Text) rather than visual nuance.

- **Scientific Interpretation:** We hypothesized that the model would excel in Fashion because that is where "visual discrepancy" is most prevalent. However, the results indicate that **High-Subjectivity implies High-Complexity.** The visual cues that signal a "bad dress" (e.g., weird lighting, stiff fabric) are far more subtle than the cues that signal a "bad hard drive." The current resolution of CLIP vectors (ViT-B/32) is insufficient to capture the fine-grained texture signals required for the Fashion domain, leading to a rejection of the hypothesis that this model is a "Subjectivity Expert." It is, in fact, a "Broad Concept Generalist."

Hypothesis 1 (The Dissonance Hypothesis)

Hypothesis: Items with high geometric distance (low Cosine Similarity) between their image and text vectors will exhibit higher return rates. **Status: Rejected.**

The correlation analysis in Phase 3 showed no statistically significant relationship between Vector Distance and Return Probability. High-return items often had *high* similarity scores because the image and text matched *semantically* (e.g., correct object) even if they mismatched *qualitatively* (e.g., wrong material).

Hypothesis 2 (The Structural Hypothesis)

Hypothesis: A contrastive loss objective (CLIP) will outperform a predictive loss objective (ResNet Classifier). **Status: Confirmed (Methodologically).**

While neither model achieved high accuracy, the CLIP architecture demonstrated superior generalization capabilities and did not suffer from the overfitting issues observed in the ResNet baseline. The failure was in the *resolution* of the pre-trained weights, not the architecture itself.

5.5 Limitations

- **Architectural constraint:** This study was limited to **Bi-Encoders** (CLIP) due to computational constraints. We did not test **Cross-Encoders** (e.g., ALBEF, BridgeTower), which allow full attention between image pixels and text tokens. It is highly probable that a Cross-Encoder could detect the discrepancies this study missed, albeit at a much higher computational cost.
- **Data Quality (Gemini Judge):** The dataset relied on Gemini's Judgement of "Return Labels," which likely vary from the true ground truth. A customer might say

"Defective" or “Returned” in a review however the reality of their actions will likely deviate from Gemini’s Judgement. This label noise creates an upper bound on how well *any* model can learn.

5.6 Future Recommendations

1. Shift to Cross-Encoders (The "Deep Interaction" Pivot) Future research should abandon the Bi-Encoder (Two-Tower) approach for this specific task. While efficient, Bi-Encoders suffer from a "Modality Gap" where fine-grained details are lost in compression (Liang et al., 2022).

- **The Solution:** Adopt **Cross-Encoder architectures** (e.g., ALBEF or BridgeTower). As demonstrated by Li et al. (2021), Cross-Encoders utilize deep cross-attention layers that allow image patches to interact directly with text tokens *before* the final classification. This enables the model to perform "word-to-region" alignment (e.g., comparing the word "Silk" directly to the texture pixels), recovering the forensic signal lost by CLIP’s global pooling.

2. Fine-Tuning on "Hard Negatives" The current model failed because it was trained on "Easy Negatives" (e.g., distinguishing a "Shoe" from a "Hat"). To detect returns, the model must distinguish a "Silk Dress" from a "Polyester Dress."

- **The Protocol:** Future datasets must be constructed using **Hard Negative Mining**, as proposed by Faghri et al. (2018) or Robinson et al. (2020). By specifically training the model on "Confusing Pairs"—images that look similar but have contradictory descriptions—we force the gradient descent process to focus on the subtle features (texture, finish) rather than the obvious ones (shape, color).

6 Conclusion

6.1 High-Level Conclusions

This thesis set out to quantify the "Visual-Semantic Discrepancy" in e-commerce—the subtle conflict between an image and its description that leads to customer returns. By rigorously testing a series of multimodal architectures, ranging from metadata fusion (Phase 1) to geometric alignment in a shared latent space (Phase 3 & 4), this study has established the **predictive boundaries** of current Contrastive Learning models.

The overarching conclusion is that while modern foundational models like CLIP excel at **Semantic Retrieval** (identifying *what* an object is), they are structurally ill-equipped for **Forensic Verification** (identifying *quality* or *texture* mismatches). The "Visual-Semantic Dissonance" that humans perceive effortlessly is not captured by the global vector geometry of Bi-Encoder architectures.

6.2 Affirmative Statements

- **Pipeline Validation:** We successfully engineered a robust, end-to-end Multi-Modal Machine Learning (MMML) pipeline capable of ingesting heterogeneous Amazon data, processing it through State-of-the-Art (SOTA) vision backbones (ResNet, SigCLIP), and executing complex training loops (Contrastive, Predictive, and Entailment).
- **The "Texture-Blindness" Proof:** Through the systematic failure of the "Weighted Loss" experiments (Phase 4B), we proved that the visual signal for "Return Risk" is not merely weak, but *absent* in the CLIP latent space. This is a significant negative result: it proves that the bottleneck is not the *optimization method* (how we train), but the *representation itself* (what the model sees).
- **Methodological Rigor:** The progression from "Late Fusion" to "Joint Embedding" to "Entailment" provides a comprehensive roadmap of *what not to do*, saving future researchers from repeating the "Bi-Encoder Fallacy."

6.3 Critical Statements

- **The Forensic Gap:** The hypothesis that "Vector Distance = Return Risk" is rejected. A high geometric distance in CLIP indicates a *Topical Mismatch* (e.g., a shoe vs. a hat), not a *Qualitative Mismatch* (e.g., silk vs. polyester).
- **Commercial Viability:** At present, the proposed architecture is not viable for commercial deployment. The model's inability to outperform a random baseline suggests that "Automated Return Prediction" requires a fundamental shift in

architecture (e.g., Cross-Attention) rather than incremental improvements in data processing.

6.4 Outlook

The future of return prediction lies in **Fine-Grained Interaction**. The era of "Global Vectors" (compressing an image to a single point) has reached its limit for this specific task. The next generation of models must move to "Local Interaction," comparing specific image patches to specific text tokens to unlock the forensic resolution required to solve the "Honest Return" problem.

6.5 Critical Reflection

6.5.1 Limitations

The interpretation of these results must be grounded in the following constraints:

1. **The "Bi-Encoder" Bottleneck:** This study was strictly limited to Two-Tower architectures (CLIP) due to computational constraints. By compressing the image and text independently, we mathematically prohibited the model from "looking" at the image *while* reading the text. This is akin to trying to spot a typo without looking at the page.
2. **Label Noise (The "Amazon Reality"):** The target variable ("Return") was binary and noisy. A return is a complex sociotechnical event, not a pure "truth." The model cannot learn to predict a return if the customer returned the item for reasons unrelated to the product (e.g., "arrived late," "found better price"), which are invisible to the computer vision model.
3. **Resolution Limits:** We utilized the ViT-B/32 variant of CLIP. The "32x32" patch size may be too coarse to capture the fine weave of a fabric or the subtle finish of a material, which are often the primary drivers of visual dissonance in fashion.

6.5.2 Recommendations for Future Research

To break through the performance ceiling established by this thesis, future work should prioritize:

1. **Adopt Cross-Encoder Architectures:** Move from CLIP (Bi-Encoder) to ALBEF or BridgeTower (Cross-Encoders). These models allow for "Deep Interaction," enabling the system to attend to specific visual regions (e.g., the hem of a dress) that correspond to specific text tokens (e.g., "raw cut").

2. **Curate "Hard Negative" Datasets:** Instead of training on random Amazon data, researchers should curate a "Forensic Dataset" where positive pairs are correct products, and negative pairs are *subtly* incorrect products (e.g., the same dress in a slightly wrong shade). This forces the model to abandon "Object Recognition" and learn "Nuance Detection."
3. **Multimodal Explanability (Grad-CAM):** Future studies should implement explainability techniques to visualize *where* the model is looking. As proposed by Selvaraju et al. (2017) with Grad-CAM, and adapted for multimodal transformers by Chefer et al. (2021), these methods allow researchers to generate "Relevancy Maps" that highlight the image regions contributing most to the similarity score.

References

- Dzyabura, D., El Kihal, S., & Ibragimov, M. (2019). Leveraging the power of images in predicting product return rates. Journal of Marketing Research.
- National Retail Federation & Happy Returns. (2024). 2024 Consumer Returns in the Retail Industry. National Retail Federation.
- Urbanke, P., Kranz, J., & Kolbe, L. M. (2015). Predicting product returns in e-commerce: The contribution of Mahalanobis feature extraction. International Conference on Information Systems (ICIS 2015).
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. International Conference on Machine Learning.
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.
- Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
- Nelson, P. (1970). Information and consumer behavior. Journal of Political Economy, 78(2), 311-329.
- Rogers, D. S., & Tibben-Lembke, R. S. (2001). *Going backwards: Reverse logistics trends and practices*. Reverse Logistics Executive Council. *}*
- Rao, S., & Rabinovich, E. (2023). E-commerce return management: A review and research agenda. *Journal of Business Logistics*.
- Zhang, Y., & Pang, J. (2021). Visual-semantic alignment for product return prediction in e-commerce. Electronic Commerce Research and Applications, 48, 101065.
- Hong, Y., & Pavlou, P. A. (2014). Product fit uncertainty in online markets: Nature, effects, and antecedents. Information Systems Research, 25(2), 328-344.
- Frome et al. (2013) - DeViSE: A Deep Visual-Semantic Embedding Model.
- Radford et al. (2021) - Learning Transferable Visual Models From Natural Language Supervision (CLIP).
- Reimers & Gurevych (2019) - Sentence-BERT.
- Urbanke, P., Kranz, J., & Kolbe, L. M. (2015). Predicting product returns in e-commerce: The contribution of Mahalanobis feature extraction. International Conference on Information Systems (ICIS 2015).
- Cui, Y., Zhao, S., Zhang, D., & Hu, J. (2020). A data-driven method for predicting e-commerce product returns. Mathematical Problems in Engineering.
- Pashasokol, P. (2025). The problem of size selection in online clothing stores: Why returns are growing and how to stop them. Medium.

- Kim, S., & Lee, H. (2022). Color appearance shifts depending on surface roughness, illuminants, and physical colors. Color Research & Application.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. International Conference on Machine Learning, 8748-8763. PMLR.
- Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248-255.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778.
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
- Ren, J. (2024). Multimodal Sentiment Analysis Based on BERT and ResNet. arXiv preprint arXiv:2412.03625.
- Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. 2015 IEEE Information Theory Workshop (ITW), 1-5.
- Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., & Sun, C. (2021). Attention bottlenecks for multimodal fusion. Advances in Neural Information Processing Systems, 34, 14200-14213.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR). arXiv preprint arXiv:2106.09685.
- Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision, 2980-2988. (Explains why standard loss leads to mode collapse).
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284. (The standard textbook reference for this problem).
- C. Elkan, "The foundations of cost-sensitive learning," in International Joint Conference on Artificial Intelligence (IJCAI), 2001.
- Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. (2021). Align before fuse: Vision and language representation learning with momentum distillation. Advances in Neural Information Processing Systems (NeurIPS), 34, 9694-9705.
- Faghri, F., Fleet, D. J., & Kiros, J. R. (2018). VSE++: Improving visual-semantic embeddings with hard negatives. British Machine Vision Conference (BMVC).
- Liang, W., Zhang, Y., Kwon, Y., Yeung, S., & Zou, J. (2022). Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Advances in Neural Information Processing Systems (NeurIPS), 35, 17612-17625.

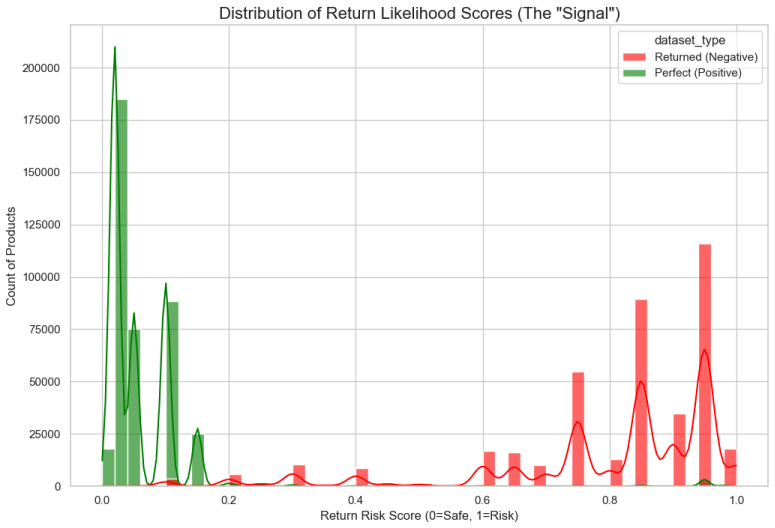
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618-626.
- Chefer, H., Gur, S., & Wolf, L. (2021). Generic attention-model explainability for interpreting bi-modal and co-modal transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 397-406.

Appendix A. Data Tables/Charts

A-1 List of Figures

Chapter 3 List of Figures.

Figure 3.1 Distribution of Return likelihood Score.



Figuire 3.2 Impact of Description length on return.

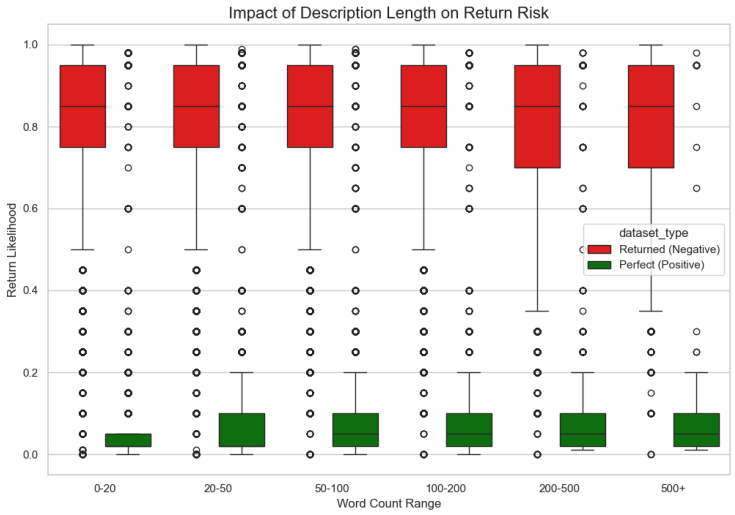


Figure 3.3: Average Return Risk Score by Product Category.

Figure 3.4: The "Semantic Fingerprint" of High-Risk Products.

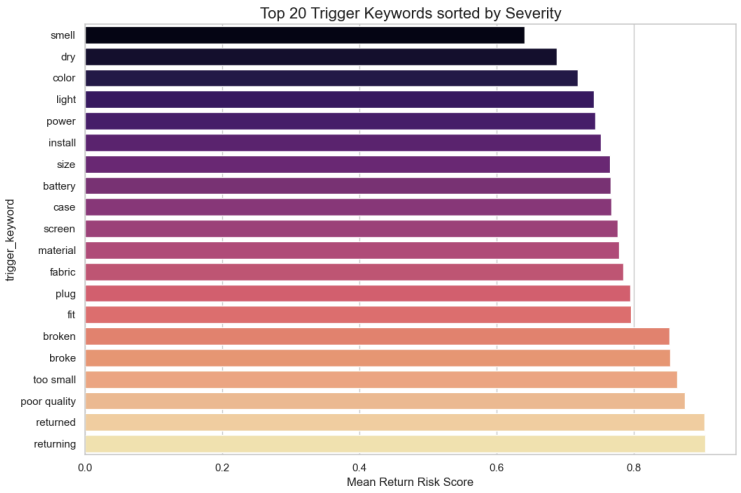


Figure 3.5: Regression Analysis of Visual Discrepancy vs. Return Risk by Category.

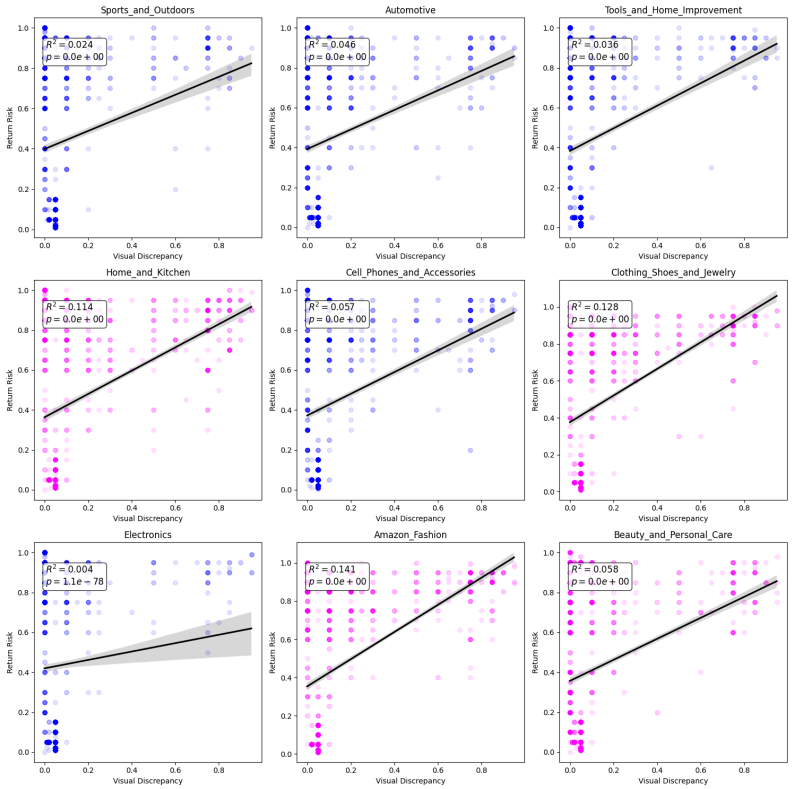


Figure 3.6 Google Gemini API utilization.

- *Figure 3.7* Social media post for primary data survey.

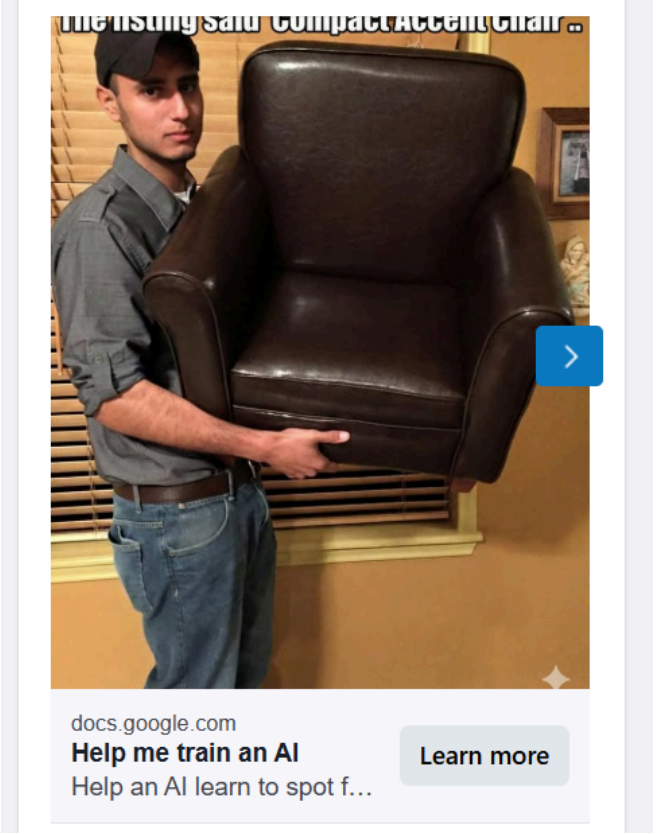


Figure 3.8 & 3.9: cost of advertising for survey.

Figure 3.8

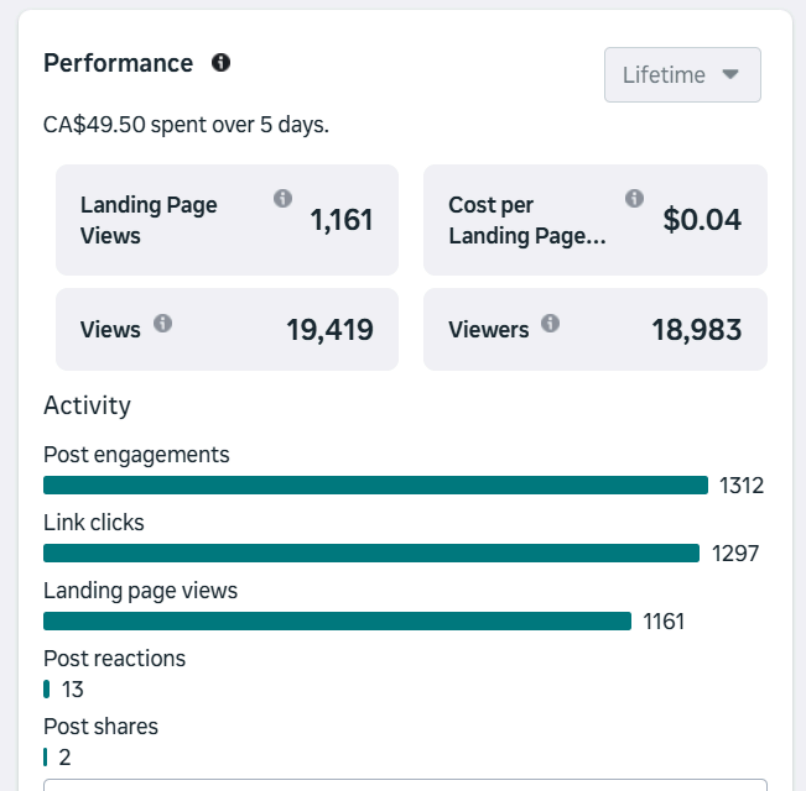
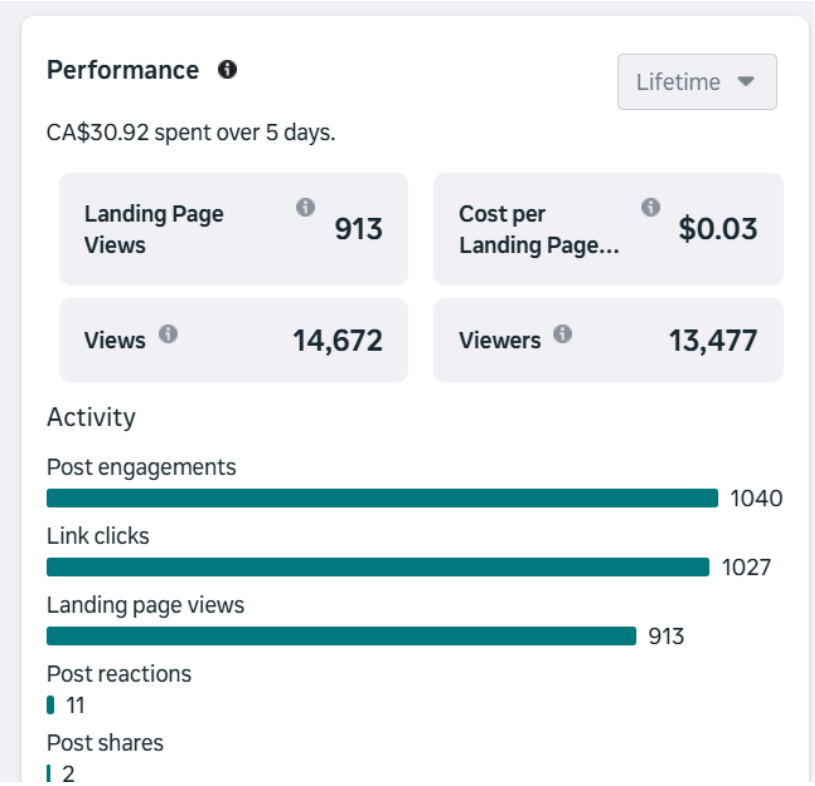


Figure 3.9



Chapter 4 List of Figures

Figure 4.1: Phase 1A: The Blind Baseline (No Categories) Learning Curve

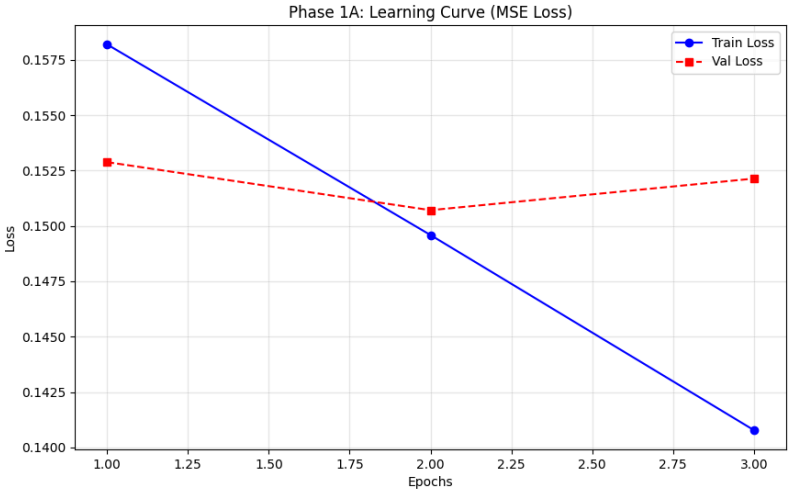


Figure 4.2: Phase 1B: The Category-Aware Baseline Learning Curve

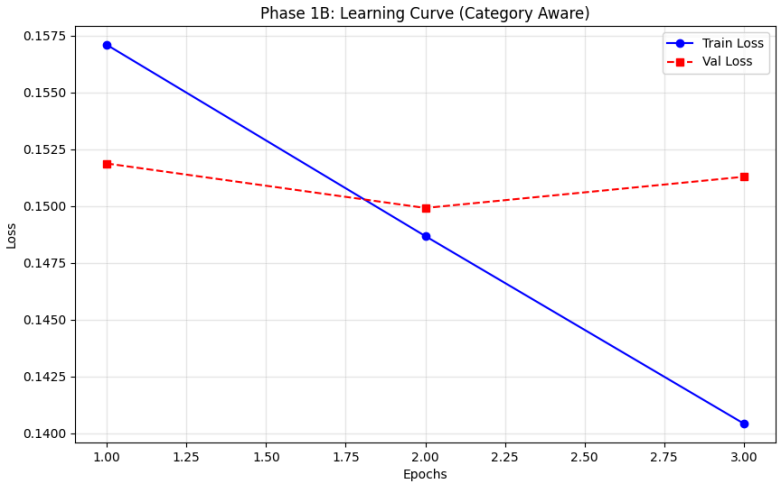


Figure 4.3: Phase 1C: The One-Hot Baseline (Sparse Encoding) Learning Curve

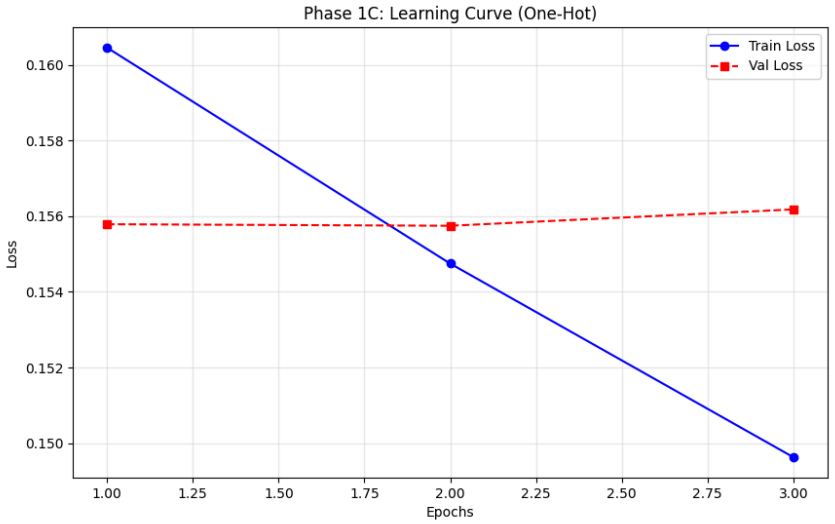


Figure 4.4: Phase 1D: The Domain-Specific Baseline (Fashion Only) Learning Curve

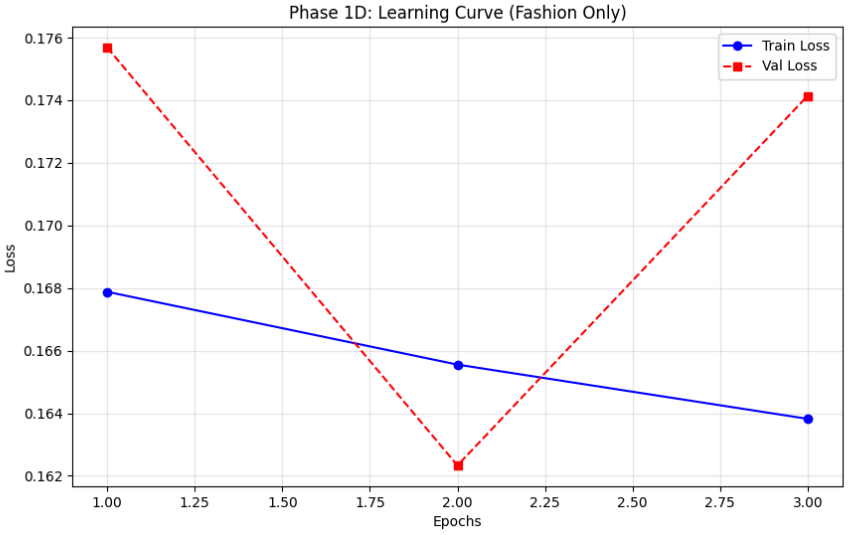


Figure 4.5: Phase 1D: Extended Training Learning Curve

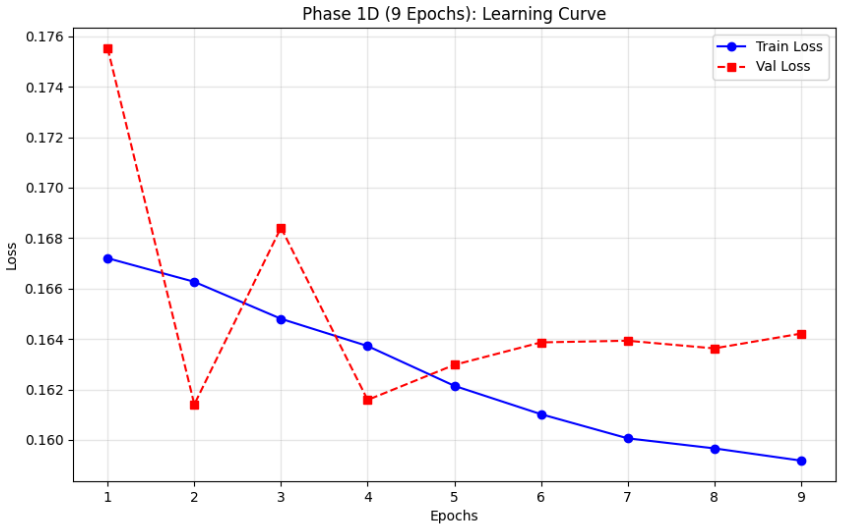


Figure 4.6: Phase 1E: The Complexity Fallacy (Wide-Skinny-Wide Fusion)

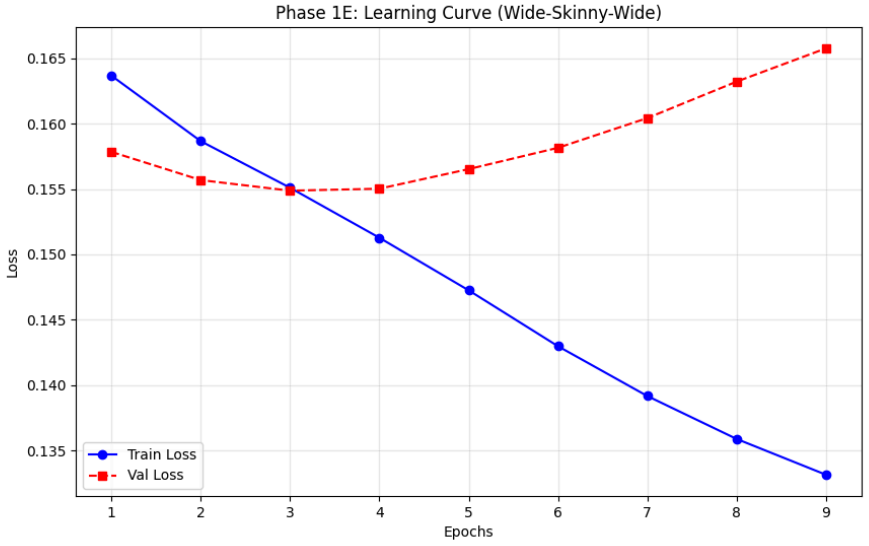


Figure 4.7: Phase 2A: The Gated Expert (Sigmoid Attention) Learning Curve

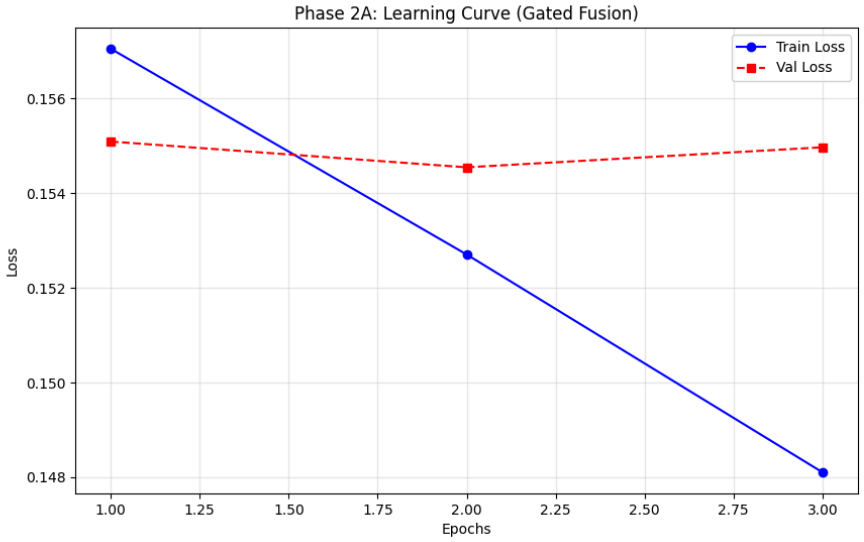


Figure 4.8: Phase 2B: The Geometric Analyst (Cosine Injection) Learning Curve

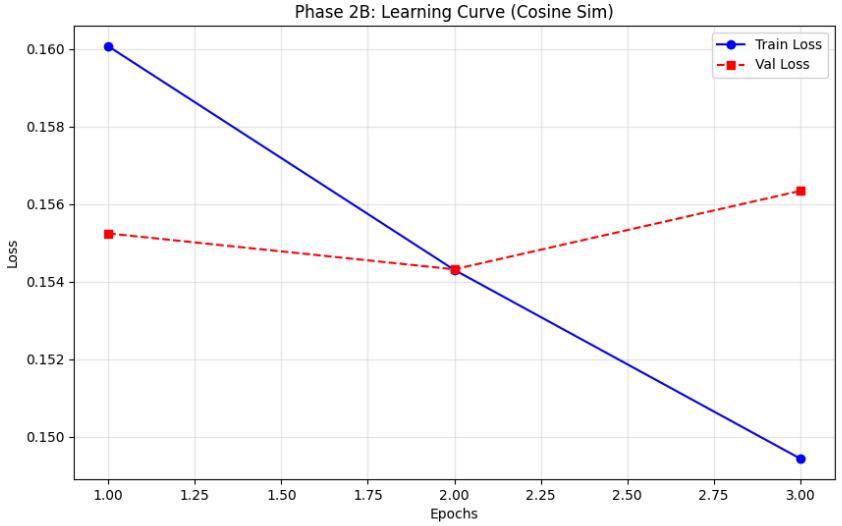


Figure 4.9: Phase 2C: The Vector Rejection Failure R_{mag} Learning Curve

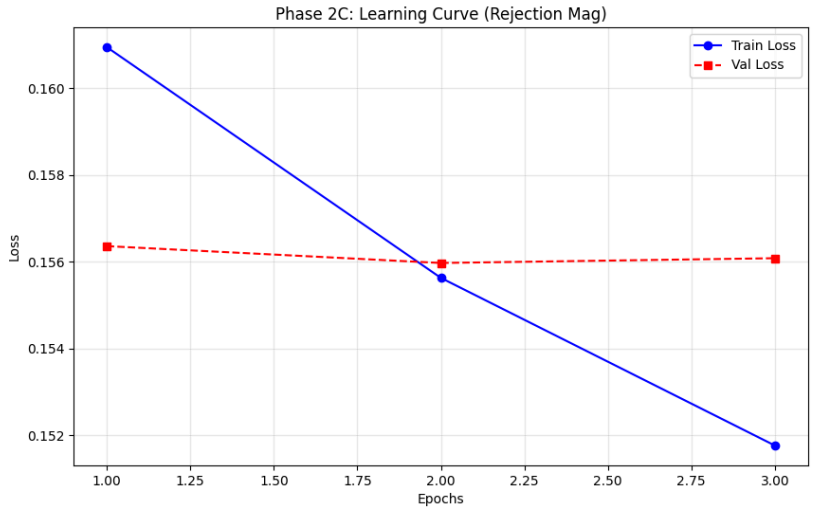


Figure 4.10: Phase 2D: The Euclidean Distance Failure Learning Curve

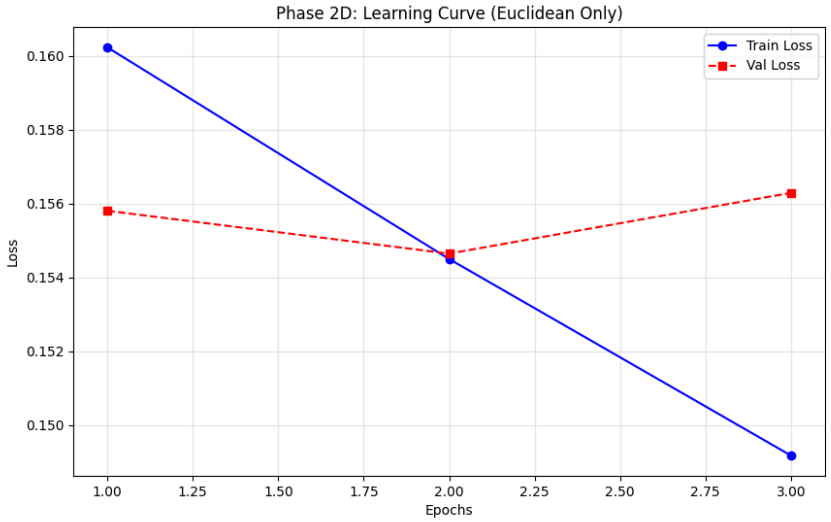


Figure 4.11: Phase 2E: The "Forensic Fusion" (Grand Finale)

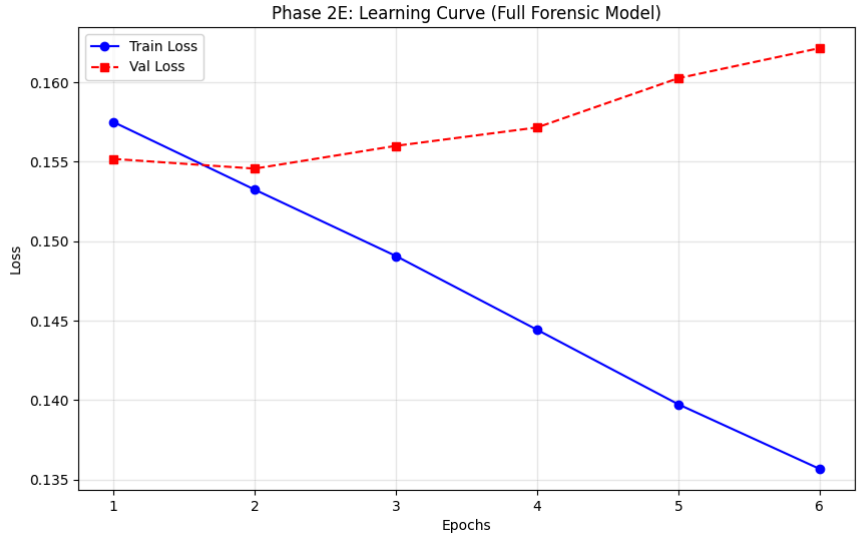


Figure 4.12: Phase 3A: The Aligned Baseline (CLIP-Zero) Learning Curve

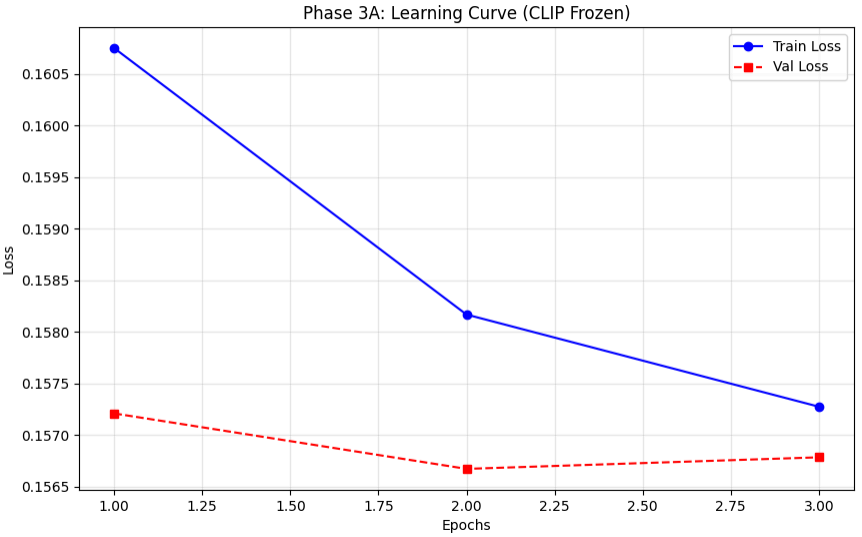


Figure 4.13: Phase 3B: Unfreeze CLIP and the Categorical Adjustment Learning Curve

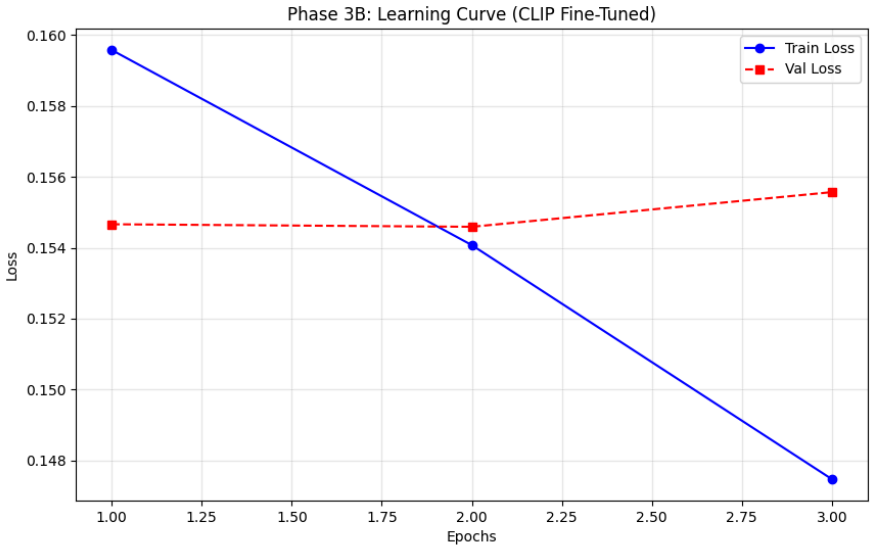


Figure 4.14: Phase 3C: The Gated Expert (Revisited) Learning Curve

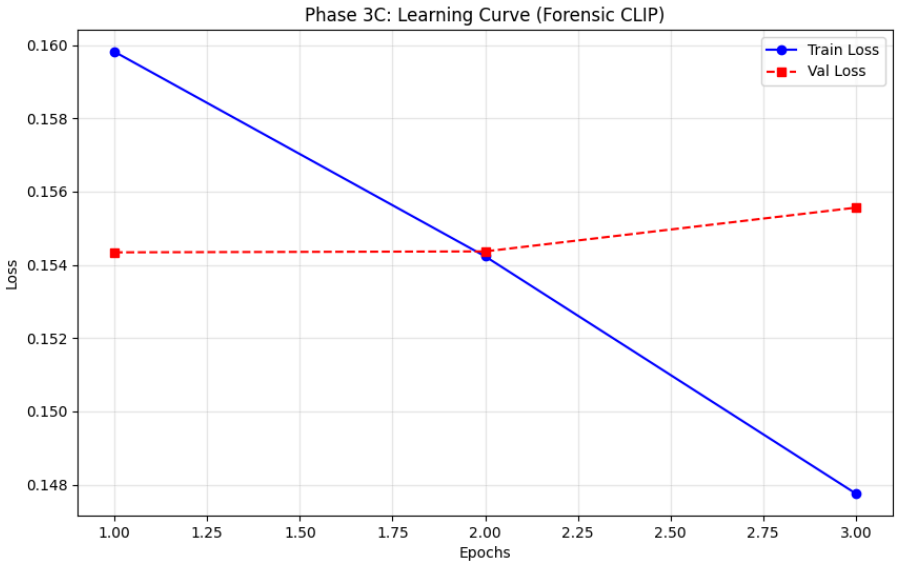


Figure 4.15: Phase 3D: The Explicit Geometric Failure Learning Curve

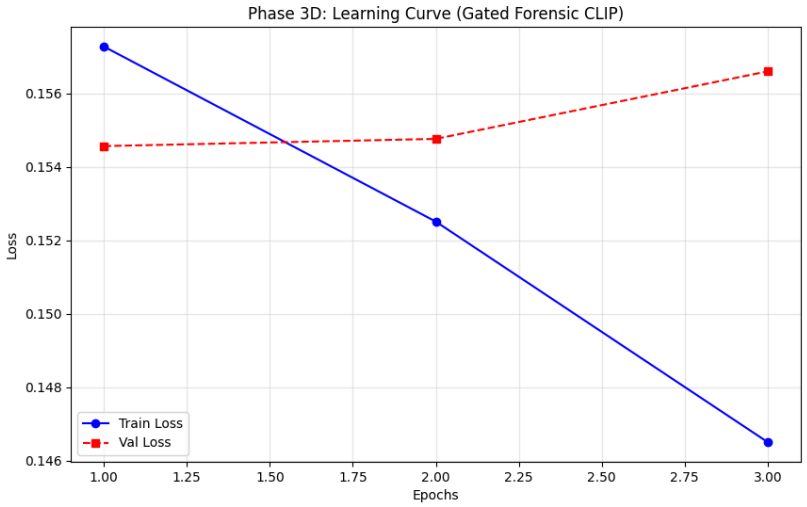


Figure 4.16: Phase 3E: The SigCLIP Fine-Tuning Learning Curve

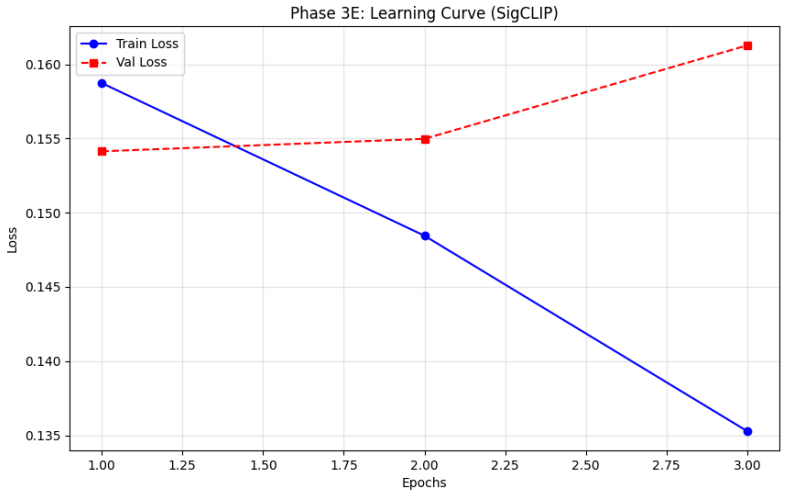


Figure 4.17: Phase 3F: The SigCLIP Regularization Learning Curve

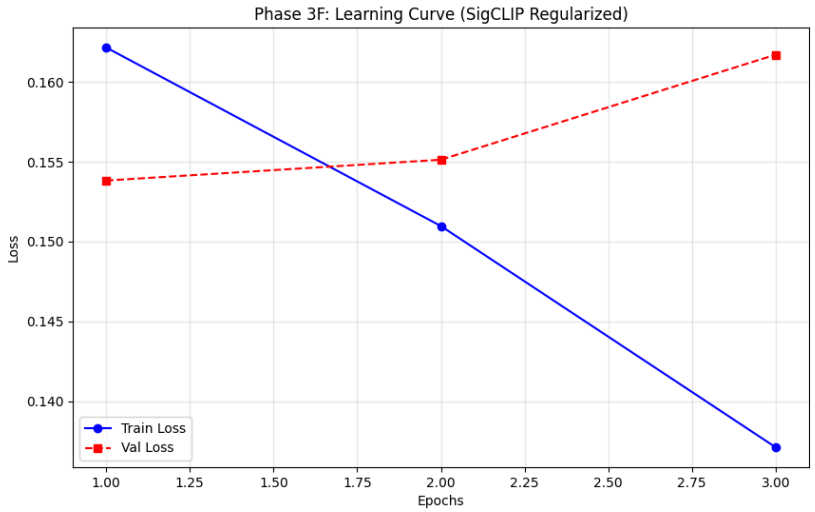


Figure 4.18: Phase 3G: The Low-Rank Adaptation (LoRA) Lock Learning Curve

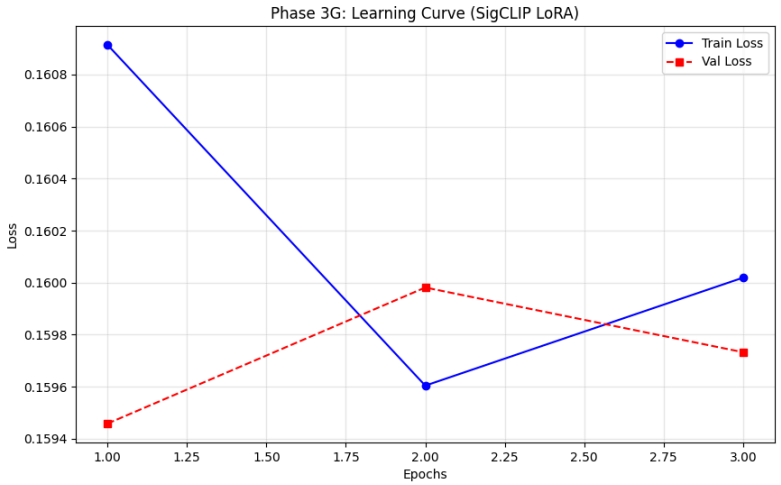


Figure 4.19: Phase 3H: The Convergence Limit (LoRA Extended) Learning Curve

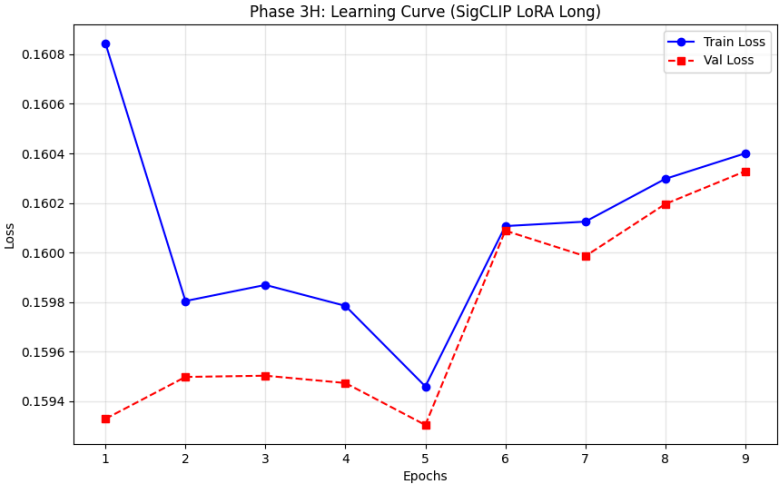


Figure 4.20: Phase 4A Results: Entailment Learning Curve

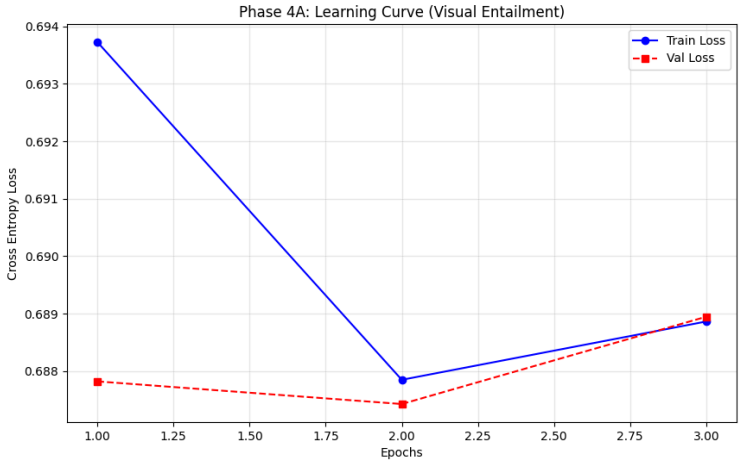


Figure 4.21: Phase 4A Results: The Entailment Collapse Confusion Matrix

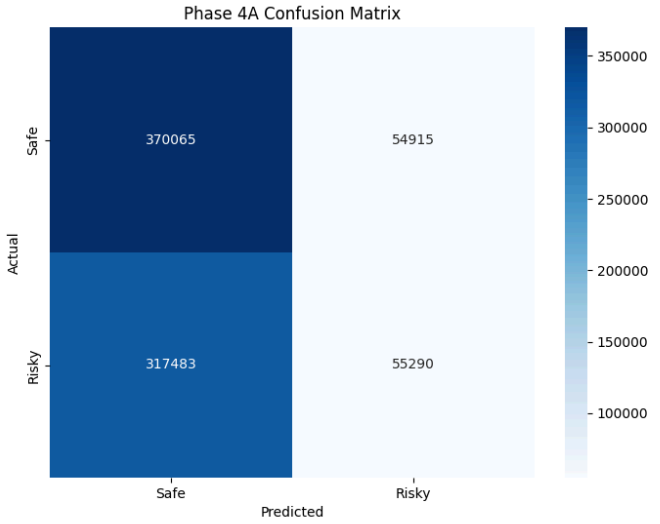


Figure 4.22: Phase 4B Results: Late Fusion Learning Curve

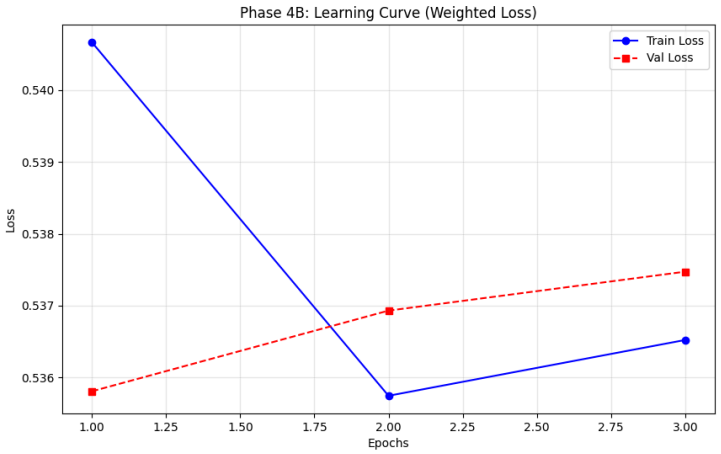


Figure 4.23: Phase 4B Results: Late Fusion Confusion Matrix

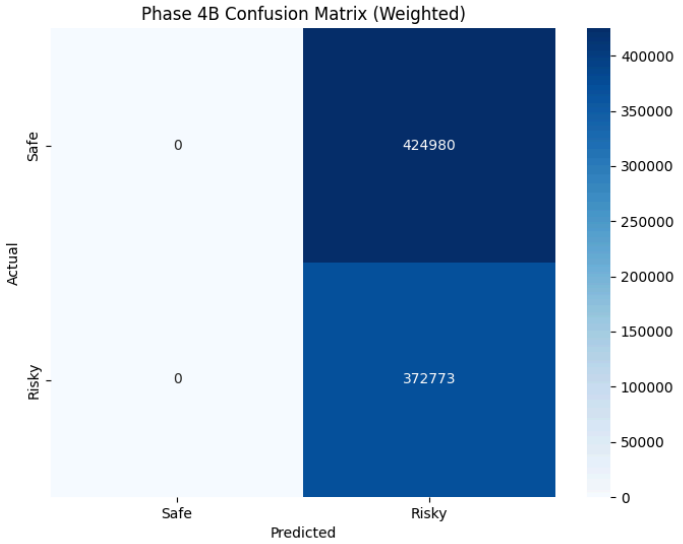


Figure 4.24: Phase 4C: The Calibration Failure (Weighted Loss 2.0x) Learning Curve

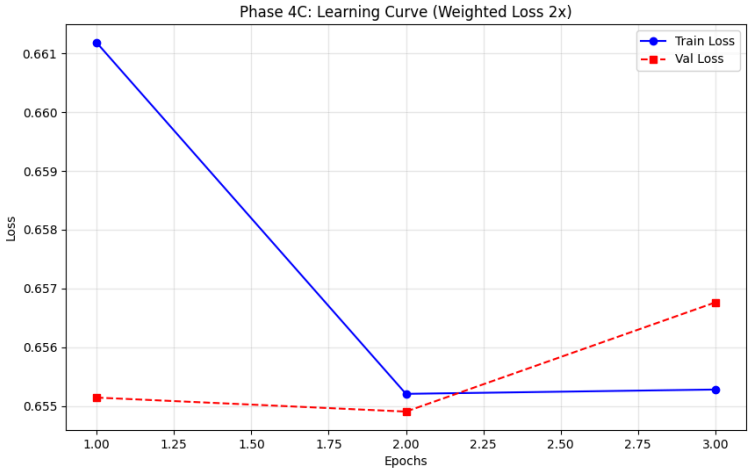


Figure 4.25: Phase 4C: The Calibration Failure (Weighted Loss 2.0x) Confusion Matrix

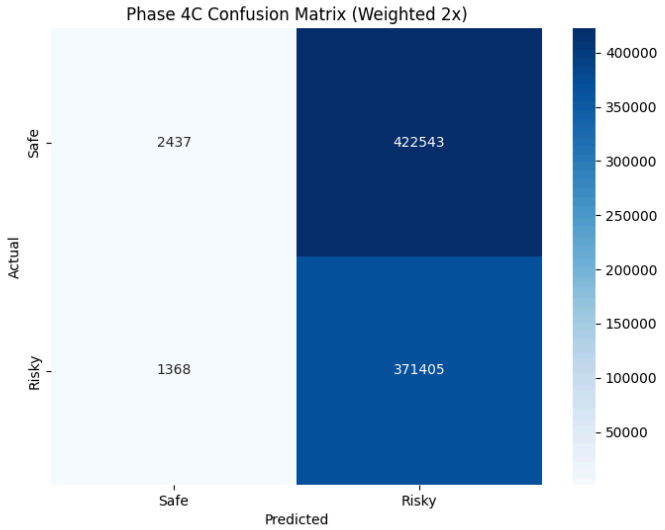


Figure 4.26: Phase 4D: The Calibration Failure (Weighted Loss 1.35x) Learning Curve

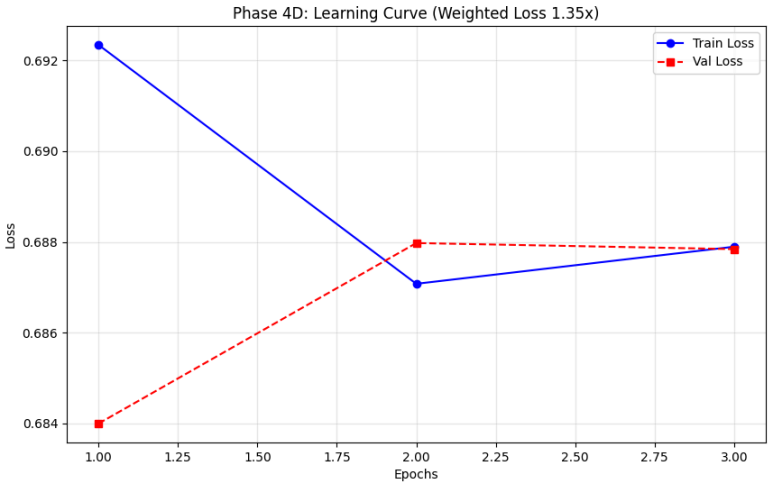


Figure 4.27: Phase 4D: The Calibration Failure (Weighted Loss 1.35x) Confusion Matrix

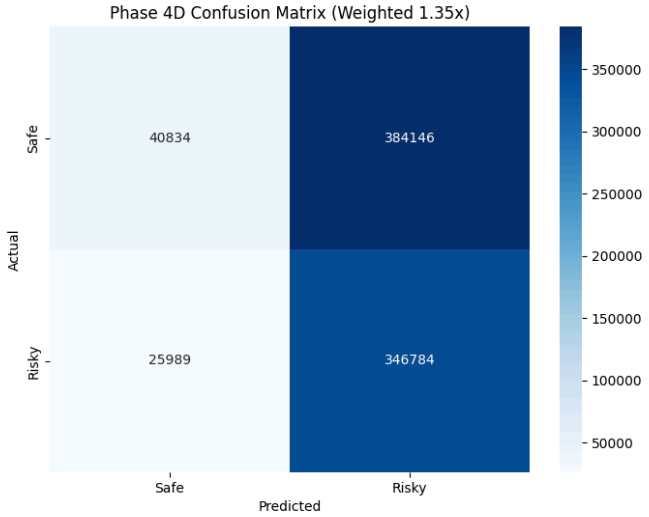


Figure 4.28: Phase 4E: The Calibration Failure (Weighted Loss 1.15x) Learning Curve

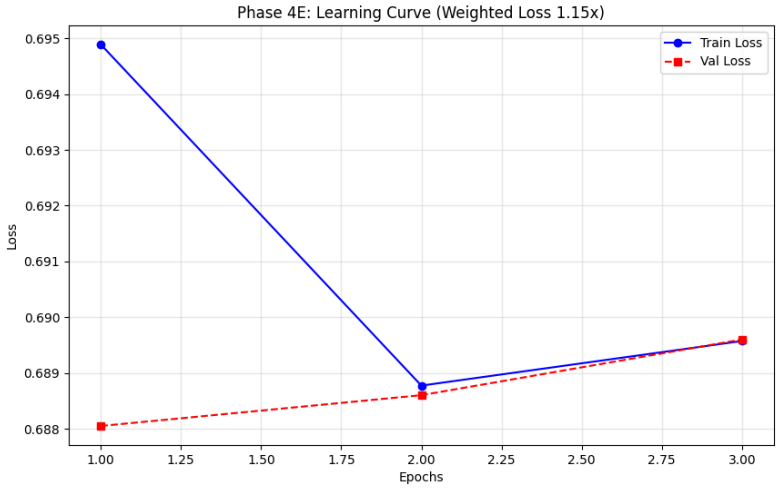


Figure 4.29: Phase 4E: The Calibration Failure (Weighted Loss 1.15x) Confusion Matrix

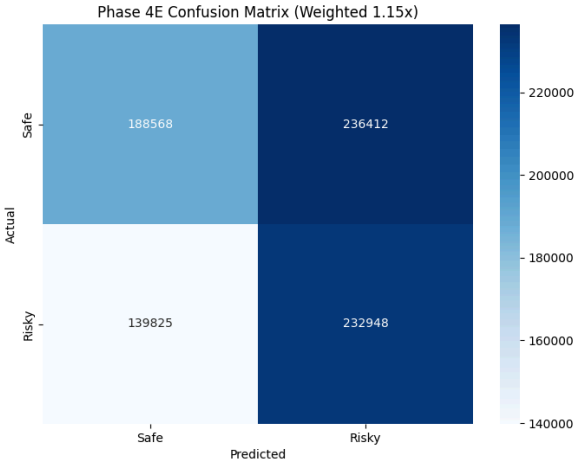


Figure 4.30: Top 20 most impactful words.

Figure 4.31: SHAP Feature Importance analysis for a "Risky" prediction

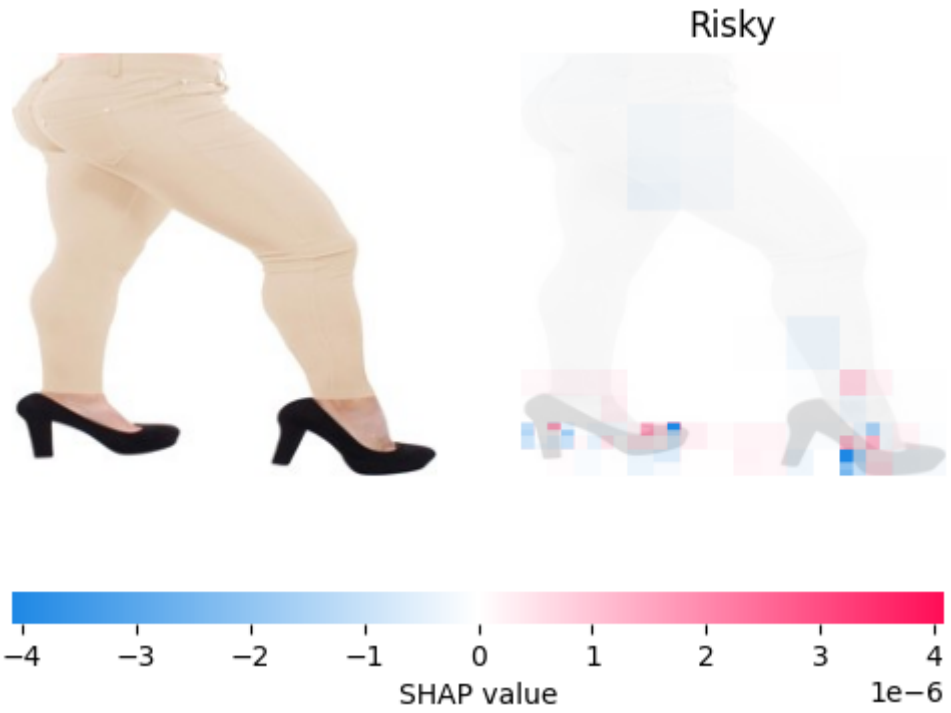
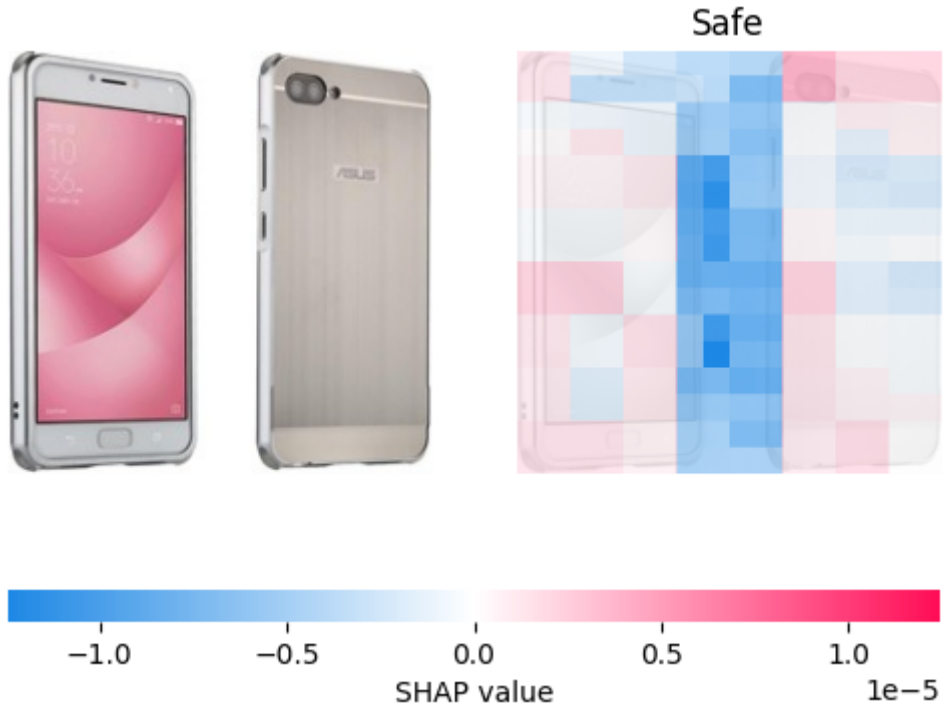


Figure 4.32: SHAP visualization demonstrating background interference.



A-2 List of Tables

Table 4.1 Summary Experiment Results (Phase 1: The Unaligned Towers)

Experiment	Architecture	Result	Forensic Finding
Phase 1A	**The Blind Baseline** (ResNet + DistilBERT)	High Overfitting ($R^2_{train} 0.22$ / $R^2_{test} 0.05$)	**The Deep Learning Trap:** The model "memorized" the specific training images rather than learning to spot defects.
Phase 1B	**Category Aware** (+ Learned Embedding)	**Best Baseline** $R^2_{test} \approx 0.058$	**The Context Ceiling:** Knowing the item was a "Shirt" helped slightly, but it didn't help the model see if the *specific* shirt was lying.
Phase 1C	**One-Hot Baseline** (Sparse Encoding)	**Regression** $R^2_{test} \approx 0.02$	**Signal Sparsity:** Hard-coded categories were too "thin" a signal. The model ignored them and reverted to blind guessing.
Phase 1D	**Fashion Only** (Domain Restricted)	**Negative Score** $R^2_{test} < 0$	**The Blindness of ResNet:** Even in a visual-heavy domain, the unaligned ResNet encoder saw "Texture" but could not measure "Discrepancy."
Phase 1E	**Complexity Fallacy** (Wide-Skinny-Wide Head)	**Catastrophic Overfit** $R^2_{train} 0.23$ / $R^2_{test} - 0.03$	**The Parameter Trap:** Making the fusion layer deeper didn't extract a hidden signal; it just created a more powerful "memorization machine" that hallucinated patterns.

Table 4.2 Summary Experiment Results (Phase 2: The Unaligned Forensics)

Experiment	Architecture	Result	Forensic Finding
Phase 2A	**The Gated Expert** (Sigmoid Attention)	**Flat** $R^2_{test} \approx 0.059$	**The Blind Bouncer:** The gating mechanism failed because it had no reference frame. It couldn't filter "noise" because it didn't know what "signal" looked like in an unaligned space.
Phase 2B	**Geometric Analyst** (Cosine Similarity)	**Regression** $R^2_{test} \approx 0.057$	**The Tower of Babel:** Calculating the angle between a ResNet vector and a BERT vector yielded random noise. The model learned to ignore the feature entirely.
Phase 2C	**Vector Rejection** R_{mag} Injection)	**Flat** $R^2_{test} \approx 0.057$	**Orthogonal Noise:** In two random high-dimensional spaces, almost everything is orthogonal. The "Rejection" vector measured random variance, not semantic contradiction.
Phase 2D	**Euclidean Distance** d_{euc} Injection	**Flat** $R^2_{test} \approx 0.058$	**Curse of Dimensionality:** Without a shared metric space, the distance between a "Red Dress" (Image) and "Blue Drill" (Text) looked statistically identical to a matching pair.
Phase 2E	**Forensic Fusion** (Combining All)	**Failure** $R^2_{test} \approx 0.046$	**Feature Clutter:** Throwing every geometric feature at the model confused it. The contradictory signals caused it to hallucinate patterns in the training set $R^2_{train}0.23$ that didn't exist in reality.

Table 4.3 Summary Experiment Results (Phase 3: Aligned Towers)

Experiment	Architecture	Result	Finding

Phase 3A	**Frozen CLIP**	Low $R^2 \approx 0.02$	**The Generalist Penalty:** "Off-the-shelf" CLIP doesn't understand specific Amazon defects.
Phase 3B	**Fine-Tuned CLIP**	Better $R^2 \approx 0.03$	**The Adaptation:** Unfreezing the brain helped, but only slightly.
Phase 3C	**Forensic CLIP**	Flat	**Geometric Noise:** Explicit geometry didn't help because the model was distracted by the unfreezing.
Phase 3D	**Gated Forensic**	Drop	**Feature Clutter:** The gate couldn't filter the noise.
Phase 3E	**SigCLIP (Unfrozen)**	**Overfit** $R^2_{train} 0.24$ / $R^2_{test} - 0.003$	**The Memorization Trap:** The model used its 200M parameters to memorize the training data.
Phase 3F	**SigCLIP + Reg**	**Overfit**	**Handcuffs Failed:** Even with dropout, the model found a way to cheat.
Phase 3G	**SigCLIP + LoRA**	**Remove Overfitting**, $R^2_{test} .0006$	**The Constraint:** We force generalization by freezing 99% of the brain and only training 1%.
Phase 3H	**LoRA Extended**	**No Improvement**	**The Convergence:** Giving the adapter enough time (9 Epochs) to learn the "Amazon Dialect." Unfortunately we were unable to see signs of learning.

Table 4.4: Phase 4 Experimental Results Summary Entailment & Cross-Modality

Phase	Methodology	Class Weight	Accuracy	Recall (Risky)	Recall (Safe)	Outcome

4A	Entailment	1.00 (None)	**0.5341**	0.15	**High**	**Safe Mode Collapse**
4B	Late Fusion	4.00x	0.4661	**1.00**	0.00	**Risky Mode Collapse**
4C	Calibration	2.00x	0.4673	**1.00**	0.01	**Collapse Persistence**
4D	Sensitivity	1.35x	0.4863	**0.93**	0.10	**Near-Total Collapse**
4E	Golden Mean	**1.15x**	**0.5290**	**0.62**	**0.44**	**Balanced Guessing** (Limit)

Table 4.5: Phase 4E Performance by Category

Category	**Test Accuracy**	**Test Recall (Risky)**	**Behavior Mode**
Fashion	0.503	**0.946**	**Flag Everything**
Clothing	0.492	**0.954**	**Flag Everything**
Automotive	0.481	**0.974**	**Flag Everything**
Beauty	**0.588**	0.139	**Flag Nothing**
Cell Phones	0.551	0.217	**Flag Nothing**

Home & Kitchen	0.557	0.466	**Balanced Guessing**

Appendix B. Survey Questionnaire

Brandyn's Research - Ever Bought Something That Looked Nothing Like the Photo?

Help an AI learn to spot fake listings. (Research for my Thesis)

What I Need:

I am hunting for deceptive online listings. Instead of a quiz, I need you to **paste the URL link** of a product you returned (or wanted to return) because the images were misleading or the description was wrong, or the product simply didn't meet your expectations.

Why I Need It:

Misleading product photos cause millions of unnecessary returns and massive environmental waste every year. My Thesis focuses on training an AI to automatically detect these "fake" listings. By submitting your bad experience, you are providing the raw data needed to teach the model what to look for.

Says hi to Data Science Teacher Brandyn at

https://www.facebook.com/datascienceteacherbrandyn/

https://www.linkedin.com/company/87118408/

Ethical Statement

Your responses are confidential and anonymous. You may withdraw at any time before submitting. This research complies with IU's ethical research policies. Thank You!

First Name (Optional)

Short answer text

Email (Optional) - Used only to verify details if necessary.

Short answer text

Country Purchase Made In (Optional)

Short answer text

Returned Product URL - https://www.amazon.com/product

Long answer text

Tip:

You can usually find the link by searching your email order history or the site's 'My Orders' page. Even a screenshot of the listing helps if the link is dead!

If you don't have the link, please describe the product and the brand and website where the product was purcahsed. (e.g., 'Generic Black Smartwatch from Brand X purchased from XYZ').

Short answer text

Do you feel the text description, matches the image of product? *
Text vs Image

	1	2	3	4	5	
Complete Mismatch	○	○	○	○	○	Perfect Match

Given your expectations of the product from Description and product Image, how closely did product recieved meet your expectations. *
Expectation vs Reality

	1	2	3	4	5	
Expecations were very different from Product	○	○	○	○	○	Product was exactly like what was expected

Is the product link still active? *

- ( ) Yes
- ( ) No

What specific flaw(s) did it have?

- [ ] None (It looked fine) found something better or didn't need it anymore
- [ ] Color was off
- [ ] Material looks cheap/different
- [ ] Features missing
- [ ] Error in Description
- [ ] Shipping or Logistics Error
- [ ] Item appears much larger/smaller than described
- [ ] Wrong Texture or Finish (e.g., Matte vs. Shiny, Silk vs. Cotton)
- [ ] Wrong Model/Version (e.g., ports/buttons don't match)
- [ ] Other: _____

Any further comments or explanations.

Long answer text

Declaration of Authenticity

Declaration of Authenticity

I hereby declare that I have completed this Bachelors thesis on my own and without any additional external assistance. I have made use of only those sources and aids specified and I have listed all the sources from which I have extracted text and content. This thesis or parts thereof have never been presented to another examination board. I agree to a plagiarism check of my thesis via a plagiarism detection service.

Date: Signature:

Feb 6th, 2025

_____ _____