Colloquium

Brandyn, Ewanek

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

Study on Pre-Purchase risk detection, moving from logistics to content governance.

Place and date

The Multi-Billion Dollar "Tax"

E-commerce Returns: ~17.6% (NRF, 2023) to 30% in Fashion

And we accept this huge waste.

Almost all predictive models focus on the **Who**. They ask: 'Is this a serial returner?' or 'Did they buy the wrong size?'

We aren't looking at the product itself. Ignoring the **What**

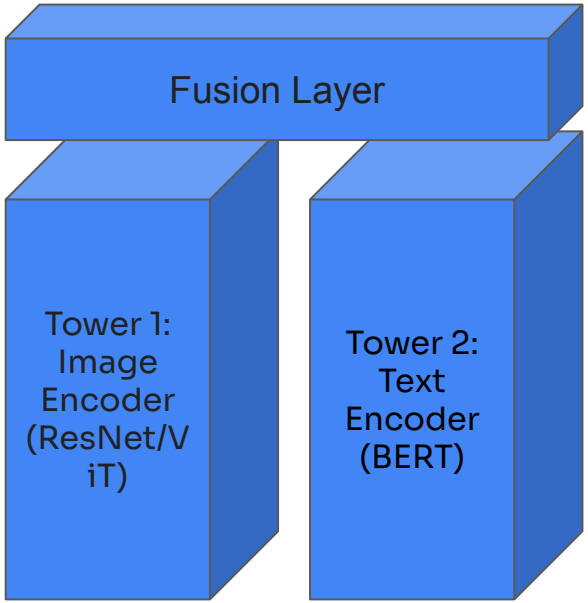
The Hypothesis ("Visual-Semantic Discrepancy")

- This is the '**lie**' hidden in the data.
- Promises 'High-Quality Silk,' but the image features—the texture, the drape, the reflection—clearly indicate 'Cheap Polyester.'

The Objective
If we can quantify —the 'lie'—we can build a pre-sale check that flags deceptive listings before the sale happens.

Source: National Retail Federation & Appriss Retail, "2023 Consumer Returns in the Retail Industry"

Methodology: The "Two-Tower" Architecture



The "**Tower of Babel**" Analogy:
From Book of Genesis much like the story of Babel, the system fails because the **components cannot speak a unified language**.

Pixels speak Pixel and **Text speaks Text**

The core assumption of my thesis was simple geometry.

Distance = Discrepancy

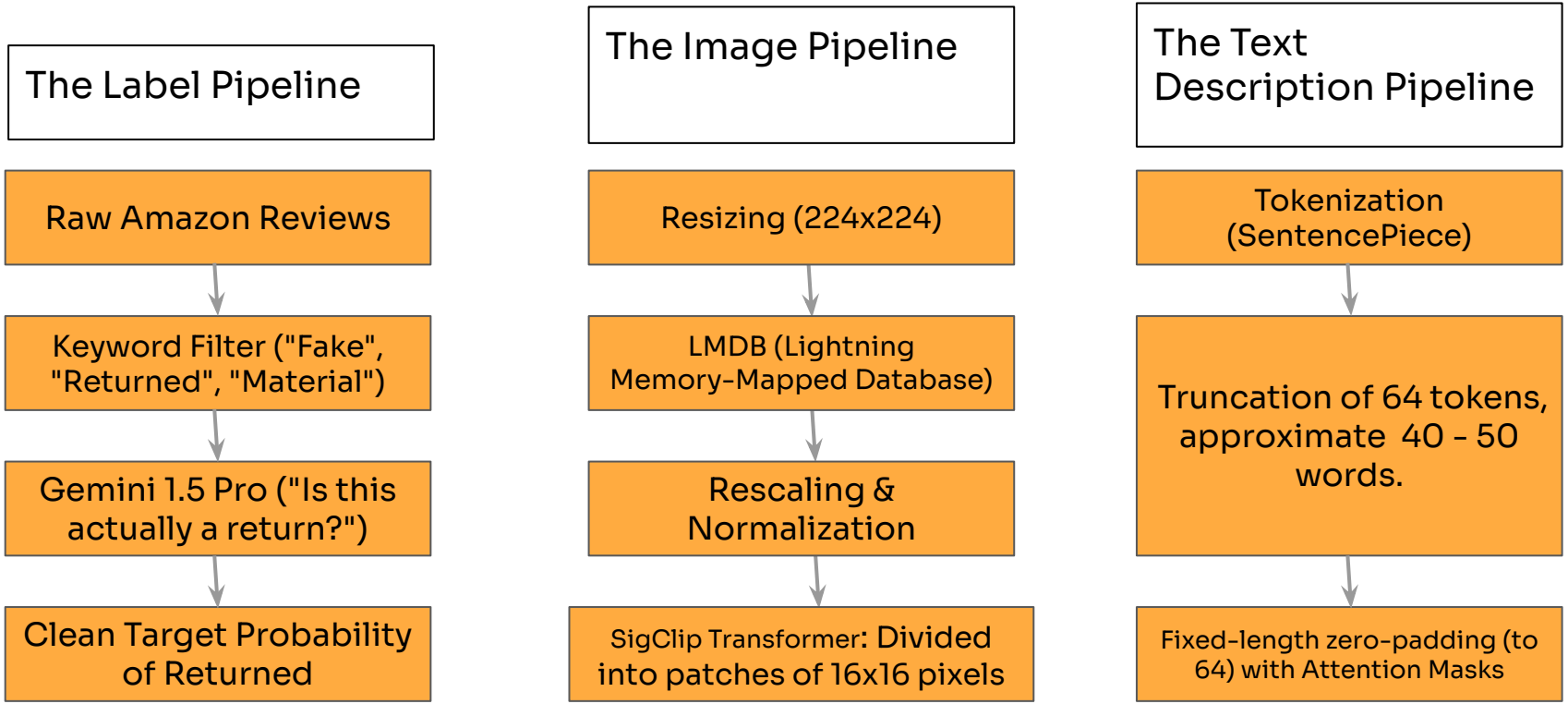
And then that,

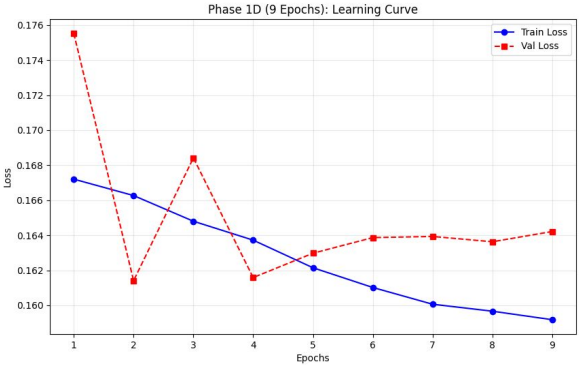
Discrepancy = Return Risk

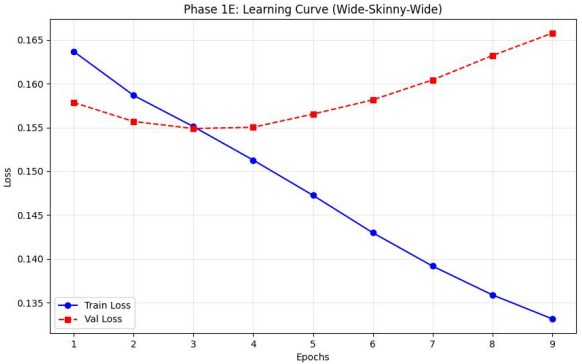
But as we'll see... just because they are in the same room, doesn't mean they are speaking the same language.

Source: Book of Genesis

Data Transformations: Labels, Images, and Text







Phase 1: The Baseline & The "Meltdown"

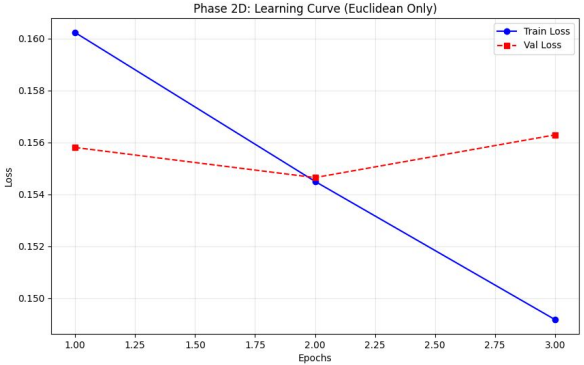
Objective: Establish a simple baseline model to detect "Visual-Semantic Discrepancy" before building the complex Two-Tower model with distance metrics.

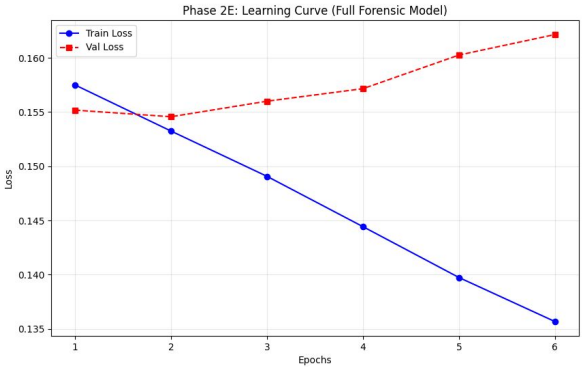
Architecture: A standard **Dual-Encoder** setup using **ResNet-18** (for images) and **DistilBERT** (for text).

Phase 1C: The "One-Hot" Lesson

- **Hypothesis**: Adding explicit category data (One-Hot Encoded) would improve baseline accuracy.
- **Consequence**: Created a "Power Virus" workload. The GPU performed billions of unnecessary calculations (x * 0), locking silicon in P0 max-voltage state.
- **Result**: Extreme throttling and physical fan failure.

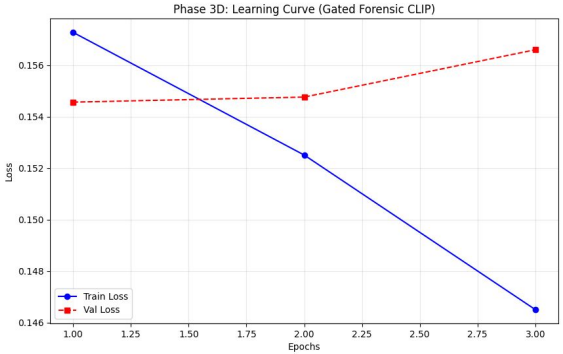
Result of Phase 1: Unaligned towers, even on just fashion items and with variations of fusion later was unable to provide meaningful results. **Best R-squared .058%**

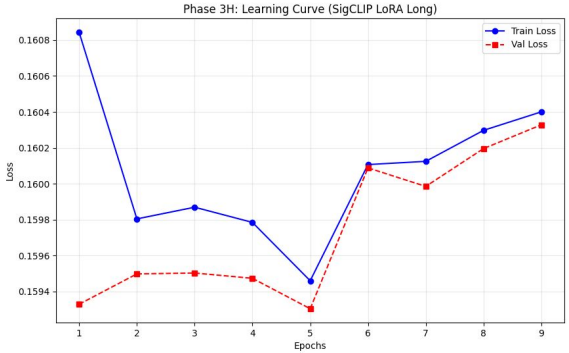




Phase 2: The "Metric Learning" Experiment

- **Hypothesis:** If the model won't learn the distance naturally, we will calculate it manually and feed it as an input.
- **Method:** A multi-modal regressor that combines ResNet, DistilBERT, and Explicit Geometric Features
    - Forensic Feature Engineering
        - Cosine, Euclidean, Rejection.
- **Result:** The model still struggled, proving that calculating the distance doesn't help if the vectors aren't aligned in the first place.
    - Performed the best with Euclidean Distance only, adding in all distance metrics together confused the model versus helping.
- **Takeaway**: You cannot measure the distance between two maps if they are in different languages.
    - ResNet speaks "Pixels."
    - DistilBERT speaks "Words."
- Even with explicit math, the Modality Gap was too wide.





Phase 3: The Foundation Model Pivot (CLIP vs. SigLIP)

Problem: Phase 2 proved we could not align Text and Images from scratch (insufficient data).

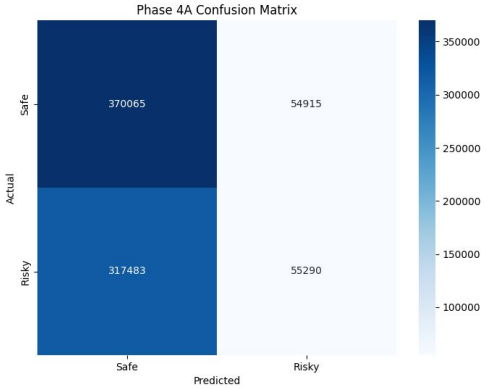
Solution: Pivot to Foundation Models (CLIP & SigLIP) pre-trained on billions of image-text pairs.

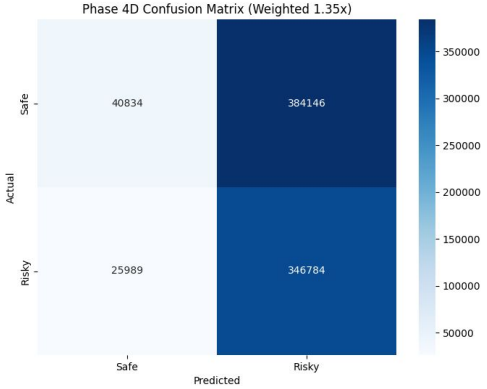
Hypothesis: These models already "know" what a shoe looks like; we just need to teach them what a return looks like.

Architectures:

- OpenAI CLIP: Uses Contrastive Loss. Good **general** knowledge, but struggled with fine-grained product details (Phase 3A-3D).
- Google SigCLIP: Uses Sigmoid Loss for Image-Text Pre-training. Proved superior at handling dense, noisy e-commerce data (Phase 3E-3H).

Results:
Takeaway: Complex "Forensic" gating (Phase 3D) was unnecessary. The raw semantic signal from SigLIP (optimized via **LoRA**) provided the strongest correlation to return risk.





Phase 4: The Entailment Pivot

From Regression to Entailment: Phase 3 (Regression) failed due to "Mean Collapse" (predicting the average).

New Hypothesis: We reformulated the task from "How risky is this?" (**Continuous**) to "Does this image contradict the text?" (**Binary Classification**).

Entailment:

- Safe Zone (Class 0): Return Likelihood ≤ 0.4
- Danger Zone (Class 1): Return Likelihood ≥ 0.6
- Force the model to make a hard decision preventing mean collapse.

Phase 4A (Unweighted): The "Lazy" Response.

- High Accuracy (53%), but 0.15 Recall on Risk.
- It ignored the "Visual Lie."

Phase 4B-D (Weighted 1.35x - 4.0x): The "Panic" Response

- We penalized missing a return.
- Result: 100% Recall, but 0% Precision.

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

- Result: Balanced guessing (Recall 0.62), but accuracy remained stuck at ~53%.

Phase 4: The Proof of Structural Blindness

The Scientific Conclusion: **Structural Blindness**

The Findings: We cured the optimization issues, but the model still couldn't distinguish Safe from Risky.

Hypothesis of inability to classify: **Information Bottleneck.**

- The Bi-Encoder (Two-Tower) architecture **compresses** the image into a single vector before comparing it to text.
- Visual-Semantic Discrepancies (e.g., a scratch, a wrong texture) are **fine-grained** details that are destroyed during this compression.

Verdict: A Two-Tower architecture is geometrically incapable of forensic return detection. Partly because of information loss and partly because there are likely many other factors to this problem.

Phase	Meth.	Class Weight	Accuracy	Recall (Risky)	Recall (Safe)
4A	Entailment	1.00 (None)	**0.5341**	0.15	**High**
4B	Late Fusion	4.00x	0.4661	**1.00**	0.00
4C	Calibration	2.00x	0.4673	**1.00**	0.01
4D	Sensitivity	1.35x	0.4863	**0.93**	0.10
4E	Golden Mean	**1.15x**	**0.5290**	**0.62**	**0.44**

RQ1: Can Visual-Semantic Discrepancy Predict Return Risk?

The Question: To what extent can the discrepancy between images and descriptions predict return risk in e-commerce using current Bi-Encoders?

The Verdict: Negligible. (With current SOTA architectures).

Evidence:

- Across all phases (1–4), accuracy never significantly exceeded the random baseline.
- Phase 4B Result: The model exhibited "Mode Collapse," defaulting to 100% Risky rather than identifying specific flaws.

Scientific Interpretation:

"**Texture-Blindness**": CLIP/SigLIP are "Object Identity" experts (e.g., "Is this a shoe?"), not "Forensic" experts (e.g., "Is the leather cheap?").

Geometric Distance Failure: The cosine distance in latent space measures semantic difference, not qualitative difference. The vector for a "Good Shirt" and a "Bad Shirt" are too similar.

RQ2: Do Explicit Geometric Features Improve Prediction?

The Question: Does manually calculating and injecting "Cosine Similarity" or "Euclidean Distance" into the model improve performance?

The Verdict: No. (Explicit injection yields very limited gains).

Evidence:

- **Phase 2:** Adding the explicit Cosine Score, Euclidean distance metrics to the fusion layer resulted in **zero improvement** over the baseline.

Scientific Interpretation:

- **Vector Distance *not equal* Return Risk:** A large distance usually means a "Hallucination" (e.g., Text says "Hat", Image shows "Shoe").
- **Reality of Returns:** Most returns are "Subtle Deceptions" (e.g., Polyester vs. Silk). These have a small geometric distance, so the metric remains silent on the actual risk factors.

RQ3: Is there a measurable difference in the predictive effectiveness of the discrepancy model by category?

The Question: Does the model perform better in "High-Subjectivity" categories (Fashion) or "Functional" categories (Electronics)?

The Verdict: Performance degrades in High-Subjectivity domains.

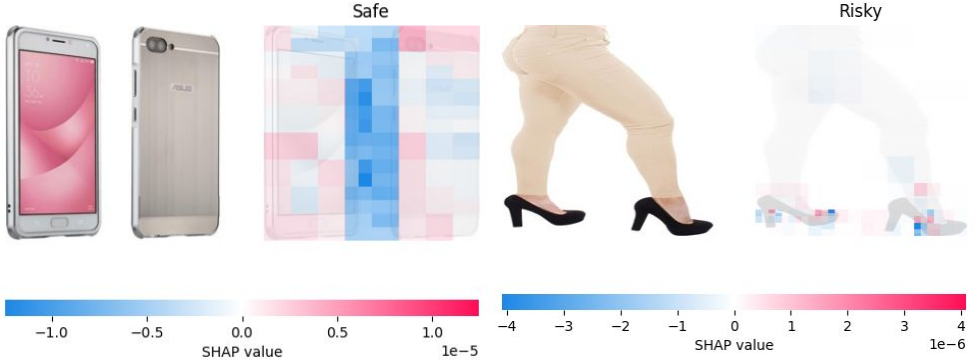
Evidence (Phase 4E Category Metrics):

- **High-Subjectivity (Fashion/Clothing):**
  - Accuracy: 0.492 (Worse than random).
  - Recall (Risky): 0.954 (It flagged 95% of clothing as risky).
- **High-Consistency (Beauty/Electronics):**
  - **B**eauty Accuracy: 0.588 (Best performing category).
  - Electronics Accuracy: 0.529 (Marginally better than Fashion).

Scientific Interpretation:

- In Fashion, return reasons are subtle (e.g., "This fabric feels cheap"). The model cannot see "feel," so it detects maximum uncertainty and defaults to predicting "Risk" for everything.
- In Beauty and Electronics, products often appear in standardized, high-quality packaging.
- **Conclusion:** The model is not a "Subjectivity Expert." It relies on proxies rather than forensic analysis of the product itself.

RQ3: Is there a measurable difference in the predictive effectiveness of the discrepancy model by category?



The Scientific Conclusion: This is Overfitting to Noise. The model found a statistical coincidence and latched onto it as a rule. It proves the model is "**Right for the wrong reasons**."

The model is not looking at a tear or a stain. It is looking at the shape of the object. The huge amount of **noise in images** and the relatively small 'target item' in image explains why the model found it difficult to use images and relied heavily on certain words, likely overfitting to the text.

Next Steps

- **Object Isolation and Pre-processing:** Implementing an object detection pipeline to tightly crop the product will force the model to look at the fabric and details rather than just the object's general shape.
- **Training on "Hard Negatives":** Future research should train models on a synthetic dataset of explicit "lies" (e.g., pairing an image of cheap polyester with a text description claiming "100% Silk") to force the network to learn visual-semantic discrepancies.
- **Iterative Explainability (XAI):** Explainability tools need to be integrated earlier in the training loop to debug the "Black Box" and ensure the model is actually looking at the product's flaws rather than text noise.

Images Generated with Gemini (Nano Banana)

THANK YOU

#2305, 5 Buttermill Ave
Vaughan, On, Canada
L4K 0J5

Brandyn Ewanek
+ 1-403-554-0069
brandyn.ewanek@IU-study.org
brandyn.ewanek@gmail.com