

Untrained Retriever

AML Challenge Report 2025/26 • [GitHub Repository](#)

Alan Mitouamona
2249767

Emanuele Iaccarino
2192710

Zuzana Miciakova
1919155

Ghulam Mujtaba
2184696

Abstract

We present a slightly chaotic, but ultimately fruitful, exploration of cross-modal retrieval via embedding translation: from RoBERTa text vectors (1024-D) to DINOv2-giant image latents (1536-D). Guided by recent literature, we tried over 25 approaches—some reasonable, some questionable at 3 AM. Our best result came from a VAE ensemble with minimal KL regularization, reaching MRR = 0.86830. This report details what worked, what didn't, and how a bit of geometric humility outperformed a lot of clever tricks.

1 Proposed Method

1.1 Motivation

Each image in the dataset comes with up to five captions—semantically related but not identical. Trying to squeeze all of them into a single point in embedding space using deterministic models led to underfitting and conflicting gradients. Instead, we treat the problem as inherently probabilistic: the same image can be described in multiple valid ways, so our model should reflect that. Following the direction of [1] and [3], we hypothesize that a VAE can better accommodate semantic dispersion while still aligning meaningfully with DINOv2 geometry.

1.2 Model Architecture

Our architecture comprises a VAE with matched latent and output dimensions (1536-D), avoiding compression-induced distortion. The model uses:

Encoder:

- Input: RoBERTa text embedding (1024-D)
- MLP with 3 layers, 8192 hidden units, GELU activations
- Outputs: $\mu \in \mathbb{R}^{1536}$, $\log \sigma^2 \in \mathbb{R}^{1536}$
- Regularization: LayerNorm and Dropout (p=0.4)

Latent sampling:

$$z = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

Only applied during training; inference uses the deterministic mean μ .

Decoder:

- Input: $z \in \mathbb{R}^{1536}$
- MLP (same as encoder) to predict DINOv2 latent

1.3 Training Objective

We minimize:

$$\mathcal{L} = \text{MSE}(f(x), y) + \lambda_{KL} \cdot \text{KL}(q(z|x) \| \mathcal{N}(0, I)), \quad \lambda_{KL} = 10^{-5}$$

The low KL coefficient regularizes the latent space without degrading reconstruction. The effectiveness of this balance aligns with insights from [3], where small KL helps retain geometry while modeling uncertainty.

1.4 Training Protocol

- 5-fold cross-validation; splits respect image ID boundaries
- Optimizer: AdamW, LR = 3×10^{-4} , weight decay = 4×10^{-4}
- Scheduler: CosineAnnealingLR
- Batch size: 256; Epochs: 60
- Embedding-space mixup: $\alpha = 0.6$

1.5 Ensemble Strategy

At inference, predictions from all 5 VAE folds are normalized and averaged:

$$\hat{y} = \frac{1}{5} \sum_{i=1}^5 \frac{\mu_i}{\|\mu_i\|}$$

This improves stability and retrieval consistency.

2 Results and Evaluation

Model	MRR
VAE Ensemble (5-fold)	0.86830
Best Single Fold	0.86241

Table 1: Performance on validation set.

3 Explored Alternatives

3.1 Diagnostic Suite and Embedding Geometry

Before testing learning-based models, we conducted diagnostic analyses to assess the feasibility of geometric alignment between RoBERTa and DINOv2 embeddings. Following the protocol of [3, 4, 1], we measured:

(1) Procrustes Alignment: We compared orthogonal and full affine mappings between modalities. Orthogonal Procrustes (enforcing $W^T W = I$) underperformed (MRR ≈ 0.81), consistent with [2], confirming that the mapping is anisotropic and requires axis scaling. Affine mappings performed slightly better, but still lacked flexibility for high-quality retrieval.

(2) Mutual k-NN Overlap: We computed k-NN overlap (k=10) between source (text) and projected target (image) embeddings. The low overlap rate (<20%) indicated high local distortion and motivated learning richer mappings beyond simple alignment.

(3) Rank Correlation and Cosine Statistics: We observed poor rank correlation between source and target neighborhoods, and significant norm anisotropy—validating that linear methods enforcing isotropy or cosine uniformity are structurally incompatible with this task.

3.2 Geometry-Preserving Projections

Orthogonal Procrustes (Linear): A baseline applying rotation-only mapping to preserve pairwise angles. While fast and interpretable, the constraint $W^T W = I$ limits its capacity to model anisotropic spaces. Confirmed by low MRR (0.81) and supported by [3, 2].

Angle-Preserving Loss: A regularization loss penalizing deviations in cosine similarity across the mapped space:

$$\mathcal{L}_{\text{angle}} = \sum_{i,j} |\cos(Wx_i, Wx_j) - \cos(x_i, x_j)|$$

Results were consistent with theory: angle preservation preserved local geometry but restricted the necessary deformation to match DINOv2 space. Performance degraded compared to unregularized MLP.

3.3 Contrastive Learning Approaches

InfoNCE [6]: Trained a deterministic MLP with standard contrastive loss using softmax-normalized similarities. Achieved strong MRR (0.84), but heavily reliant on batch size and temperature tuning. False negatives (e.g., semantically similar captions for different images) led to instability, echoing limitations discussed in [7].

SupCon (Multi-Positive) [7]: Treated all captions of an image as positive samples in supervised contrastive

loss. The model over-clustered captions, reducing inter-class separation. Resulted in moderate MRR (0.81). Effective for classification, but suboptimal for retrieval tasks requiring fine-grained ranking.

Sigmoid Loss (SigLIP) [8]: Removed softmax normalization in favor of independent pairwise sigmoid classification. This yielded training stability and was batch-size agnostic. MRR was competitive (0.84), but remained limited by the deterministic formulation.

3.4 Soft-Label and Distributional Methods

NDGL [9]: Used all negatives (rather than hardest) and aligned predicted similarity distributions with those from a teacher model. While conceptually robust, performance was bottlenecked by the teacher’s quality. Errors in the teacher’s rankings propagated into the student, resulting in MRR 0.83.

CUSA [10]: Trained with cross-modal soft labels—using SBERT similarity for text and CLIP similarity for images. However, SBERT and CLIP differ semantically from RoBERTa and DINOv2, introducing domain shift. The result was label inconsistency and noisy supervision. MRR 0.82.

3.5 Relative Representations and Anchoring

Relative Embedding Vectors [5]: We projected each text input into a similarity vector over K=500 fixed anchor images. While this captured coarse alignment, it led to significant information loss due to projection dimensionality ($1024 \rightarrow 500$), anchor selection sensitivity, and poor reversibility. MRR 0.79. Aligns with limitations discussed in [3].

3.6 Flow-Based Mapping

CrossFlow [11]: Modeled mapping as a learned velocity field through an ODE. Although theoretically continuous and invertible, this required solving differential equations during training—computationally expensive and numerically unstable. MRR plateaued near 0.82.

FlowTok [12]: Replaced ODE integration with residual transformer steps, learning discrete flows. Improved efficiency, but performance remained capped (0.81) likely due to limited sequence modeling capacity in single-vector settings.

3.7 Cycle-Consistency and Multi-Direction Models

Vec2Vec with Cycle Loss [3]: Constructed dual encoders (text \rightarrow latent and image \rightarrow latent) with round-trip cycle constraints. The cycle loss enforced invertibility but

limited the model’s ability to deform representations appropriately for retrieval. MRR 0.80. Reverse mapping (image→text) also introduced training instability, being an ill-posed task in our setup.

3.8 Hybrid Architectures and Adversarial Variants

LABridge-Inspired OU Decoder [13]: Implemented progressive decoding using a shallow OU-inspired architecture. Instability during training and limited benefit over direct regression discouraged further exploration in time-limited settings.

Transformer Residual Correction: Used a frozen linear layer followed by a 4-layer transformer to learn residual mappings. Found marginal gains over baseline; attention was underutilized on fixed-length vector inputs. MRR 0.81.

Dual Encoder to Shared Space: Built symmetric encoders for both text and image to project into a shared 768-D latent space (with InfoNCE). R@10 0.66 on validation. Performance was promising, but final output violated dimensionality constraints (not in DINOv2 space), making it ineligible.

3.9 Incompatible Loss Functions

VAE + Triplet Loss: Attempted replacing MSE with a triplet margin loss:

$$\mathcal{L}_{texttriplet} = \max(0, d(x, x^+) - d(x, x^-) + \text{margin})$$

This conflicted with the KL regularization, which promotes isotropy and smoothness, while triplet enforces sharp local separability. The optimization collapsed—MRR dropped to 0.06. Highlights the need for compatible objectives in probabilistic models.

4 Conclusion

This project turned into a comprehensive (and slightly chaotic) dive into cross-modal embedding translation. Across 25+ modeling strategies, we confronted the mismatch between geometric elegance and practical performance, learning that many theoretically sound approaches fail under real-world constraints.

Still, the score we obtained almost certainly didn’t win the hackathon. But it taught us what matters: respecting anisotropy, avoiding over-regularization, and not forcing isotropy where it doesn’t belong. More broadly, it showed how pretrained embedding spaces resist simplistic alignment—especially when each image is described in five subtly different ways.

This post-mortem report documents both what worked and what didn’t—and why reading too many papers at 3 AM might lead you to build adversarial discriminators for a retrieval problem that only needed regression.

Future Work: We closed this sprint with more ideas than time to test—possibly too many. Looking ahead, we believe progress lies in:

- **Enhancing latent flexibility** by replacing the VAE’s Gaussian prior with normalizing flows or diffusion models;
- **Designing asymmetric contrastive losses** that better reflect the 1-to-many mapping from caption to image;
- **Refining soft supervision** through distillation or pseudo-labeling strategies that respect the native geometry of DINOv2 space.

The most promising solutions likely live at the intersection of geometry-respecting design and probabilistic modeling—with less brute force, and more signal.

References

- [1] Huh et al., *The Platonic Representation Hypothesis*, ICML 2024.
- [2] Bansal et al., *Revisiting Model Stitching*, arXiv 2023.
- [3] Jha et al., *Harnessing the Universal Geometry of Embeddings*, NeurIPS 2024.
- [4] Moschella et al., *Shared Global and Local Geometry of LM Embeddings*, arXiv 2024.
- [5] Moschella et al., *Relative Representations Enable...*, NeurIPS 2022.
- [6] Radford et al., *Learning Transferable Visual Models*, ICML 2021.
- [7] Khosla et al., *Supervised Contrastive Learning*, NeurIPS 2020.
- [8] Zhai et al., *Sigmoid Loss for Language Image Pre-Training*, ICCV 2023.
- [9] *Negative Distribution Guided Learning*, arXiv 2023.
- [10] *Cross-modal and Uni-modal Soft-label Alignment*, arXiv 2024.
- [11] Li et al., *Flowing from Words to Pixels*, arXiv 2024.
- [12] *FlowTok: Flowing Seamlessly Across Text and Images*, arXiv 2024.
- [13] Shao et al., *LABridge: Text-Image Alignment via OU Process*, arXiv 2024.