The Gradient Descenders: AML Challenge Report 2025/26

Thomas Rames
Matricola: 2245966

Adam Maciejak
Matricola: 2117730

Stella Lin
Matricola: 2239170

Rafal Ciolek
Matricola: 2248504

Maria Rita Nogueira Lopes
Matricola: 2247799

1 Proposed Method

Our final best-performing submission on the public leaderboard is a weighted ensemble of predictions from two approaches: a zero-shot transformation with added neural refinement and a VAE-based projection network.

- **Architecture:**

  L-ortho: The adapter consists of two stages. First, Procrustes analysis with orthogonal transformation (L-ortho, Maiorca et al. [1]) computes a closed-form transformation matrix between RoBERTa text embeddings ($d = 1024$) and DINOv2 image embeddings ($d = 1536$). L2 normalization is applied to inputs during fitting. Second, a small MLP is initialized with this matrix: Linear($1536 \rightarrow 2048$) $\rightarrow$ LayerNorm $\rightarrow$ GELU $\rightarrow$ Dropout(0.1) $\rightarrow$ Linear($2048 \rightarrow 1024$) $\rightarrow$ LayerNorm $\rightarrow$ GELU $\rightarrow$ Dropout(0.1) $\rightarrow$ Linear($1024 \rightarrow 1536$).

  VAE: Inspired by Liu et al. [2], this adapter projects normalized text embeddings ($d = 1024$) using an encoder (Linear($1024 \rightarrow 2048$) $\rightarrow$ GELU $\rightarrow$ LayerNorm $\rightarrow$ Linear($2048 \rightarrow 3072$)) which produces the mean and log-variance of a latent Gaussian distribution. A latent vector is obtained via the reparameterization trick and passed through a decoder (Linear($1536 \rightarrow 2048$) $\rightarrow$ GELU $\rightarrow$ LayerNorm $\rightarrow$ Linear($2048 \rightarrow 1536$)) to generate the image embedding.

- **Loss Function:**

  L-ortho: We optimize using SigLipLoss [3], a sigmoid-based contrastive loss operating on pairwise similarities with learnable logit scale and bias parameters.

  VAE: We use a composite objective combining a CLIP-style contrastive reconstruction loss and a KL-divergence regularization term weighted by $\lambda = 0.01$. The learnable logit scale is initialized to $\log(1/0.07)$ and clamped to the range $[1, 100]$ for stability.

- **Training Details:**

  L-ortho: We train for 40 epochs using AdamW (weight decay 0.01) with a batch size of 256 and gradient clipping (max norm 1.0). The learning rate is 1×10^{-4} with a cosine annealing schedule and a 5-epoch warmup. Early stopping is set to patience 8.

  VAE: We train for 20 epochs using Adam (weight decay 1×10^{-4}) with a batch size of 256. The learning rate is set to 1×10^{-3} with a cosine annealing schedule. Validation is performed using the CLIP reconstruction loss to select the best model.

2 Results and Discussion

The final weighted ensemble obtained a public leaderboard MRR of **0.85801** using exclusively competition-provided data. The performance improvement over individual models is attributed to the combination of distinct architectural biases: the Procrustes-initialized MLP prioritizes global linear alignment, while the VAE incorporates probabilistic regularization to encourage latent space smoothness. Averaging these predictions reduces the variance often associated with training on limited datasets. Additionally, the L-Ortho baseline achieved a standalone score of 0.78349. This result indicates that while the cross-modal relationship is predominantly linear, non-linear neural refinement is necessary to optimize retrieval accuracy.

3 Conclusion

We progressed from preliminary embedding distribution analysis and simple linear baselines to advanced generative methods, including Flow Matching and Variational Autoencoders. Through hyperparameter optimization and architectural experimentation, we determined that while linear methods like L-Ortho provide a strong geometric foundation, they are insufficient on their own. Our final approach demonstrates that ensembling the structural alignment of the Modified L-Ortho transformation with the regularized latent space of a VAE yields the optimal strategy for cross-modal translation.

[Github Repository]

Supplementary: What We Tried

This section details the various approaches explored during the competition that were not part of our final submission.

Method 1: Simple Linear Mapping

We established a baseline using a single linear layer to map the text dimension to the image dimension, optimized with MSE loss. This approach achieved a poor MRR, indicating the cross-modal relationship is non-linear.

Method 2: Residual MLP

We attempted to improve the baseline by increasing depth and adding residual connections inspired by ResNet [4]. While training stability improved, retrieval performance remained suboptimal compared to generative methods.

Method 3: Contrastive Projection Learning

We trained an MLP with residual connections using CLIP contrastive loss to align embeddings via symmetric cross-entropy optimization. This achieved an MRR of 0.24 on our validation set. It significantly underperformed compared to the VAE, so we did not keep it.

Method 4: Semantic Alignment

We explored and implemented all of the methods described by Maiorca et al. [1], with the most promising being Procrustes and L-Ortho. Procrustes proved to be effective for model initialization. L-Ortho, despite being a linear solution, was effective as a standalone, achieving a score of 0.7834. We further optimized the number of anchors using Optuna, which boosted performance.

Method 5: Flow Matching

We explored following up our VAE with a Flow Matching model as described by Liu et al. [2]. We began by implementing a simplified MLP-based architecture following the methodology of Hawley [5], incorporating the Rectified Flow technique and higher-order integration schemes such as Runge-Kutta 4. However, we found this baseline approach underperformed. Consequently, we transitioned to the architecture proposed by Liu et al. [2] (CrossFlow), enhancing it with Classifier-Free Guidance and a Transformer-based backbone. Despite these architectural improvements, the model proved computationally expensive and RAM intensive during training. It achieved a maximum score of 0.51635, so due to the high resource cost and slow convergence, we opted to switch approaches.

Metric Analysis: Centered Kernel Alignment

We implemented CKA to compare the quality of some of the model representations. We attempted to use CKA as a loss function component to enforce representational similarity, but this did not improve convergence or final retrieval scores, so the idea was abandoned.

Hyperparameter Optimization: Optuna

For promising architectures, we utilized Optuna for automated hyperparameter tuning. This allowed us to extract maximum performance from simpler architectures and identify what hyperparameters were the most impactful for the model's MRR score on our validation set.

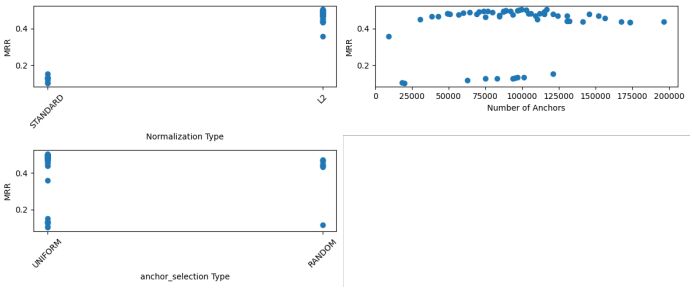

Figure 1: Hyperparameter impact on L-Ortho score.

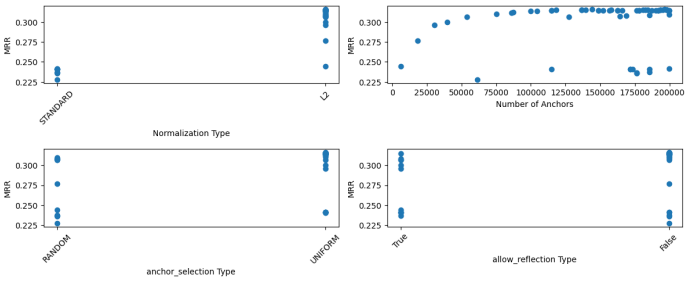

Figure 2: Hyperparameter impact on Procrustes score.

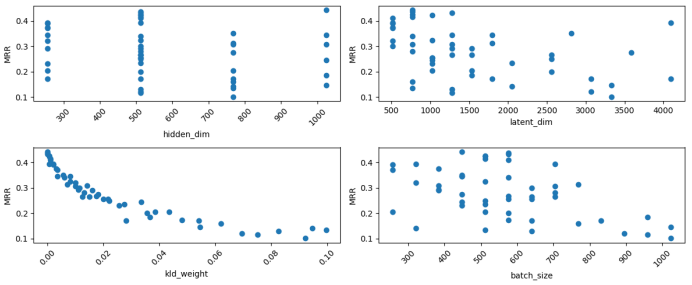

Figure 3: Hyperparameter impact on VAE score (10 epochs).

References

[1] Valentino Maiorca et al. *Latent Space Translation via Semantic Alignment*. arXiv.org, 2023. URL: https://arxiv.org/abs/2311.00664 (visited on 11/17/2025).

[2] Qihao Liu et al. "Flowing from words to pixels: A noise-free framework for cross-modality evolution". In: 2025, pp. 2755–2765.

[3] Xiaohua Zhai et al. *Sigmoid Loss for Language Image Pre-Training*. arXiv.org, 2023. URL: https://arxiv.org/abs/2303.15343.

[4] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[5] Scott H. Hawley. “Flow With What You Know”. In: *The Fourth Blogpost Track at ICLR 2025*. https://d2jud02ci9yv69.cloudfront.net/2025-04-28-flow-with-what-you-know-38/blog/flow-with-what-you-know/. 2025. URL: https://openreview.net/forum?id=3XlPoJeDu4.