

Rofola Mountain Club

AML Challenge Report 2025/26

Lorenzo Spataro
Matricola: 1946590

Federico Raponi
Matricola: 1963339

Marco Realacci
Matricola: 1938880

Alessio Pannozzo
Matricola: 1960374

1 Proposed Method

We employ a VAE-based architecture inspired by [1] that translates text to image embeddings through a probabilistic latent space.

1.1 Architecture

The encoder maps 1024D text embeddings through a residual MLP to 896D latent parameters μ and $\log(\sigma^2)$. During training, z is sampled via reparameterization trick [2]; at inference, $z = \mu$. The decoder projects z to 1536D through residual blocks and produces L2-normalized image embeddings.

1.2 Loss Function

The total training objective combines a contrastive alignment term with KL regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{contrastive}} + \beta \cdot \mathcal{L}_{\text{KL}} \quad (1)$$

1.2.1 Contrastive Loss

We adopt a symmetric multi-label contrastive loss inspired by supervised contrastive learning [3] and CLIP [4, 5], which handles multiple captions per image [6]. Given a batch of B text-image pairs $\{(x_i^{\text{text}}, z_i^{\text{image}}, \text{idx}_i)\}_{i=1}^B$, we compute the similarity matrix $S_{ij} = \tau \cdot (z_i^{\text{image}})^\top z_j^{\text{image}}$, where $\tau = \exp(\theta_{\text{scale}})$ is a learnable temperature (initialized to $\exp(2.6592)$, as from [7]).

With multi-label positive mask $M_{ij} = \mathbb{K}[\text{idx}_i = \text{idx}_j]$, each directional loss uses a multi-positive formulation [3]:

$$\mathcal{L}_{\text{direction}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\sum_{j=1}^B M_{ij} \exp(S_{ij})}{\sum_{j=1}^B \exp(S_{ij})} \quad (2)$$

The symmetric contrastive loss is computed as $\mathcal{L}_{\text{contrastive}} = \frac{1}{2}(\mathcal{L}_{\text{t} \rightarrow \text{i}} + \mathcal{L}_{\text{i} \rightarrow \text{t}})$, where $\mathcal{L}_{\text{t} \rightarrow \text{i}}$ evaluates the directional loss with S and $\mathcal{L}_{\text{i} \rightarrow \text{t}}$ with $S^\top$.

1.2.2 KL Divergence Loss

The VAE regularization term is $\mathcal{L}_{\text{KL}} = -\frac{1}{2B} \sum_{i=1}^B \sum_{d=1}^D (1 + \log(\sigma_{id}^2) - \mu_{id}^2 - \sigma_{id}^2)$, where μ_{id}

and σ_{id}^2 are the latent mean and variance. We employ linear warmup [8] for β :

$$\beta(\text{epoch}) = \begin{cases} \beta_{\text{max}} \cdot \frac{\text{epoch}}{E_{\text{warmup}}} & \text{if } \text{epoch} < E_{\text{warmup}} \\ \beta_{\text{max}} & \text{otherwise} \end{cases} \quad (3)$$

with $E_{\text{warmup}} = 5$ and $\beta_{\text{max}} = 0.2$ to prevent posterior collapse.

1.3 Training Details

The model is implemented within the **PyTorch Lightning** framework and trained using a supervised contrastive learning objective. For parameter optimization, we employ the **AdamW** optimizer. The learning rate is dynamically managed by a **ReduceLROnPlateau scheduler**, which monitors the **Mean Reciprocal Rank (MRR)** on the validation set. If the MRR metric fails to improve for a patience of 5 epochs, the learning rate is consequently reduced by a factor of 0.3.

2 Results and Discussion

Our model achieved an MRR of **0.86807** on the public leaderboard (baseline: 0.46205). The approach succeeds by combining three elements: (1) multi-layer residual blocks capture the complex non-linear transformation between heterogeneous embedding spaces; (2) multi-label contrastive loss directly optimizes for retrieval by attracting matching pairs while handling multiple captions per image; and (3) VAE regularization prevents overfitting and encourages a generalizable latent representation.

3 Conclusion

Combining VAE-based probabilistic modeling with contrastive learning provides an effective framework for model stitching tasks, translating between heterogeneous embedding spaces from different pre-trained models. This approach shows promise for broader applications in connecting incompatible model architectures.

[LINK TO THE REPO](#)

What We Tried

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

Method 1: Relative representations

We explored the "relative representations" framework [9] by projecting absolute embeddings into a relative coordinate system defined by their cosine similarity to a fixed set of shared anchors, this method theoretically enables zero-shot model stitching without additional training. However, we observed that this alignment yielded performance comparable to the provided baseline without significant improvement, leading us to discard this approach.

Method 2: Flow Matching (Random Noise)

We implemented a generative flow model based on *Flow matching for generative modeling* [10] to transform a random noise vector into a target image embedding. This method failed because the task was too complex. The model was unable to learn the mapping from uninformative random noise to a specific, correct point in the high-dimensional image embedding space.

Method 3: Flow Matching (Text Embeddings)

We adapted the flow model to start from a simple projection of the text embedding, rather than random noise. The model's task was simplified to *refine* this initial *educated guess* into the final target image embedding. This approach also failed, likely because the Flow Matching loss optimizes the *velocity* of the path, not the final destination.

Method 4: Latent Space Mapping via MLP (Contrastive Learning)

Building on the *Prior* concept [5], we implemented a 3-layer MLP with LayerNorm and GELU activations to directly map text embeddings into the image embedding space. To align the representations, we optimized a multi-label symmetric contrastive loss with a learnable temperature parameter. This approach started overfitting the training data after few epochs: the validation loss started rising. The best score that we achieved on the public leaderboard with this model was 85,789% after a 30 epochs training.

Method 5:

Our final model, described in the first page.

References

- [1] Thomas Theodoridis et al. "Cross-modal Variational Alignment of Latent Spaces". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, June 2020, pp. 4127–4136. ISBN: 978-1-7281-9360-1. DOI: [10.1109/CVPRW50498.2020.00488](https://doi.org/10.1109/CVPRW50498.2020.00488). URL: <https://ieeexplore.ieee.org/document/9150585/> (visited on 11/17/2025).
- [2] Yankun Chen et al. "Auto-Encoding Variational Bayes". In: *Cambridge Explorations in Arts and Sciences* 2.1 (Feb. 2024). ISSN: 2977-1595. DOI: [10.61603/ceas.v2i1.33](https://doi.org/10.61603/ceas.v2i1.33). URL: <https://ceasjournal.com/index.php/CEAS/article/view/33> (visited on 11/17/2025).
- [3] Prannay Khosla et al. *Supervised Contrastive Learning*. Version Number: 5. 2020. DOI: [10.48550/ARXIV.2004.11362](https://doi.org/10.48550/ARXIV.2004.11362). URL: <https://arxiv.org/abs/2004.11362> (visited on 11/17/2025).
- [4] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. Version Number: 1. 2021. DOI: [10.48550/ARXIV.2103.00020](https://doi.org/10.48550/ARXIV.2103.00020). URL: <https://arxiv.org/abs/2103.00020> (visited on 11/17/2025).
- [5] Aditya Ramesh et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. Version Number: 1. 2022. DOI: [10.48550/ARXIV.2204.06125](https://doi.org/10.48550/ARXIV.2204.06125). URL: <https://arxiv.org/abs/2204.06125> (visited on 11/17/2025).
- [6] Zifeng Wang et al. *MedCLIP: Contrastive Learning from Unpaired Medical Images and Text*. Version Number: 1. 2022. DOI: [10.48550/ARXIV.2210.10163](https://doi.org/10.48550/ARXIV.2210.10163). URL: <https://arxiv.org/abs/2210.10163> (visited on 11/17/2025).
- [7] Zhirong Wu et al. *Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination*. arXiv:1805.01978 [cs]. May 2018. DOI: [10.48550/ARXIV.1805.01978](https://doi.org/10.48550/ARXIV.1805.01978). URL: <http://arxiv.org/abs/1805.01978> (visited on 11/17/2025).
- [8] Samuel R. Bowman et al. "Generating Sentences from a Continuous Space". In: (2015). Publisher: arXiv Version Number: 4. DOI: [10.48550/ARXIV.1511.06349](https://doi.org/10.48550/ARXIV.1511.06349). URL: <https://arxiv.org/abs/1511.06349> (visited on 11/17/2025).
- [9] Luca Moschella et al. *Relative representations enable zero-shot latent space communication*. Version Number: 2. 2022. DOI: [10.48550/ARXIV.2209.15430](https://doi.org/10.48550/ARXIV.2209.15430). URL: <https://arxiv.org/abs/2209.15430> (visited on 11/17/2025).
- [10] Yaron Lipman et al. *Flow Matching for Generative Modeling*. Version Number: 2. 2022. DOI: [10.48550/ARXIV.2210.02747](https://doi.org/10.48550/ARXIV.2210.02747). URL: <https://arxiv.org/abs/2210.02747> (visited on 11/17/2025).