

# STM

## AML Challenge Report 2025/26

Luca Moresca  
Matricola: 2192853

Valerio Santini  
Matricola: 2176222

Nicholas Suozzi  
Matricola: 2203359

## 1 Proposed Method

### 1.1 Architecture

The proposed model extends the classic VSE++ scheme for joint text-image embedding [1], integrating a compositional branch based on *slot* vectors that introduce semantic granularity into the supervision signal. The goal is to allow the model to capture not only the global similarity between caption and image, but also local alignments between descriptive concepts (objects, attributes, relations) and visual content. The main component is the text adapter **TextToVis**, a two layer MLP that maps each text embedding  $t \in \mathbb{R}^{d_{text}}$  (derived from RoBERTa [2], with  $d_{text} = 1024$ ) into a visual space  $\mathbb{R}^{d_{vis}}$  [3] corresponding to the precomputed image feature embeddings (latent space VAE,  $d_{vis} = 1536$ , obtained from a DINO-ViT backbone [4]) [5]. The result is the global vector  $Z = \text{TextToVis}(t)$ , normalised L2, which acts as an aggregate representation of the caption and is the only vector used in the retrieval phase. In parallel, the auxiliary head **SlotAuxHead** is a small feed-forward network (two layer MLP with GELU) which, starting from  $t$ , generates a set of  $K = 4$  distinct vectors, called *slots*  $\mathbf{S}_T = \{\mathbf{s}_1, \dots, \mathbf{s}_K\}$ , each in  $\mathbb{R}^{d_{vis}}$  and normalised [6]. The testing phase uses only  $Z$ . However, to increase robustness and reduce the stochastic variance typical of training, the final system adopts an ensemble of models trained with different seeds but identical architecture.

### 1.2 Loss Function

During training, we optimise a total loss composed of multiple terms that combine global and compositional supervision. Specifically, the components are:

**Global triplet loss:** applied to vector  $Z$ , with mining of the most difficult negative in the batch [1]. The distance between  $Z$  and the correct image feature is minimised, while  $Z$  is penalised with respect to the most similar negative image.

**Triplet loss (max-over-slot):** calculated on the vectors  $\mathbf{S}_T$ . For each caption, its slots are compared with the images in the batch, using maximum similarity to distinguish the correct image from the negative ones [7].

**Per-slot InfoNCE:** a contrastive loss such as InfoNCE (cross-entropy) calculated separately for each slot  $\mathbf{s}_k$ . The

associated image acts as a positive for  $\mathbf{s}_k$ , while all other images in the batch are considered negative.

**Intra-Slot Diversity Loss:** a regularisation term that penalises excessive similarity between slots of the same caption [8], [7].

**Condensation Loss:** a term that pushes the global embedding  $Z$  to “condense” the information present in the slots. For each correct (caption, image) pair, we identify the slot  $\mathbf{s}_{k^*}$  most similar to the target image  $v$  and encourage  $Z$  to move closer to  $\mathbf{s}_{k^*}$  in the feature space.

### 1.3 Training Details

Optimisation is performed using Adam (initial learning rate with cosine scheduler), batch size 128, for 24 epochs. The triplet loss uses a margin of  $\alpha = 0.20$ . The weights of the various loss components are set to  $\lambda_{tri}^{(glob)} = 0.30$ ,  $\lambda_{slot} = 0.15$ ,  $\lambda_{ISDL} = 0.02$ ,  $\lambda_{cond} = 0.05$  (components not explicitly weighted have a weight of 1). These hyperparameters were chosen by validating them on the challenge’s development set.

## 2 Results and Discussion

The ensemble of models achieved a Mean Reciprocal Rank (MRR) value of 0.88141 on the challenge’s public leaderboard. Our architecture manages to exploit the richness of a multi-slot representation during training, while respecting the challenge’s deployment constraints by using only one vector for each caption in the test: compared to the official baseline, which uses a quasi-linear MLP trained with simple MSE between textual and visual embeddings, we adopt a normalised VSE++ adapter with a multi-slot auxiliary head and a combination of triplets, InfoNCE, ISDL and condensation loss.

## 3 Conclusions

We have presented an image caption retrieval system based on VSE++ with an auxiliary slot branch and new loss functions aimed at embedding diversity and condensation. These extensions improve the alignment between textual descriptions and fine grained images, transferring both global information and local details to a single global embedding.

## 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: Geometric Baseline and Residual Adapter Framework

We attempted to implement a two stage text to image alignment system operating in the latent space of two pre trained encoders (text in  $\mathbb{R}^{1024}$ , VAE images in  $\mathbb{R}^{1536}$ ). In the first stage, pairs of parallel *anchors* ( $A_X, A_Y$ ) are constructed: a diverse subset of images is selected using farthest point sampling and, for each, the most similar positive caption in cosine is selected; the embeddings are standardised and padded to handle dimensional mismatch. Several zero-shot geometric translators are defined on these anchors, evaluated using correlation metrics and retrieval metrics (MRR, NDCG, Recall@k) on an image level split; the method with maximum MRR is chosen as the baseline  $\hat{y}^{\text{geo}}(x)$ .

In the second stage, a ResidualAdapter  $f_\theta(x, \hat{y}^{\text{geo}})$  refines the baseline: an MLP receives  $[x, \hat{y}^{\text{geo}}]$  and produces a residual projected orthogonally to  $\hat{y}^{\text{geo}}$ , combined via a learnable scalar coefficient and renormalised [9], [10]. Training uses AdamW and a composite loss (InfoNCE with negative bank, cosine alignment, norm regularisation, neighbourhood preservation, and consistency in relative coordinates with respect to anchors), with controlled noise on queries. In validation, global retrieval is measured and, optionally, a re-ranking that blends absolute and relative similarity is applied; the best model is then used to translate the test set and generate the submission.

### Method 1 bis: Geodesic Residual Adapter

Building on the anchor based geometric baseline of Method 1, we introduce a geodesic extension that explicitly models the non-linear structure of the visual manifold. Instead of relying solely on Euclidean distances in the anchor space, we construct a  $k$ -NN graph over the image anchors  $A_Y$ , with edge weights given by local Euclidean distances. On this graph we precompute shortest path distances (Dijkstra) between all pairs of anchors, obtaining an approximation of geodesic distances along the data manifold. At inference time, for each query  $x$  we first compute its Euclidean similarity to the text anchors  $A_X$ , select a top- $k$  subset, and then reweight the corresponding visual anchors in  $A_Y$  using a Gaussian kernel defined on a combined distance that blends local Euclidean and precomputed geodesic components. This yields a geodesic prediction  $\hat{y}^{\text{geo}}(x)$  that better preserves the curved geometry of the visual space.

## References

- [1] Fartash Faghri et al. “Vse++: Improving visual-semantic embeddings with hard negatives”. In: *arXiv preprint arXiv:1707.05612* (2017).
- [2] Yinhan Liu et al. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019).
- [3] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. “Revisiting model stitching to compare neural representations”. In: *Advances in neural information processing systems* 34 (2021), pp. 225–236.
- [4] Mathilde Caron et al. “Emerging properties in self-supervised vision transformers”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660.
- [5] Mayug Maniparambil et al. “Do vision and language encoders represent the world similarly?” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 14334–14343.
- [6] Dongwon Kim, Namyup Kim, and Suha Kwak. “Improving cross-modal retrieval with set of diverse embeddings”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 23422–23431.
- [7] Yale Song and Mohammad Soleymani. “Polysemantic visual-semantic embedding for cross-modal retrieval”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 1979–1988.
- [8] Hani Alomari et al. “Maximal Matching Matters: Preventing Representation Collapse for Robust Cross-Modal Retrieval”. In: *arXiv preprint arXiv:2506.21538* (2025).
- [9] Luca Moschella et al. “Relative representations enable zero-shot latent space communication (2023)”. In: *arXiv preprint arXiv:2209.15430* (2023).
- [10] Valentino Maiorca et al. “Latent space translation via semantic alignment”. In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 55394–55414.