
Beyond Contrastive Learning — Multimodal Extension with RoBERTa and DINOv2

Alessio Pannozzo¹ Federico Raponi¹ Lorenzo Spataro¹

1. Introduction

While contrastive paradigms like CLIP (Radford et al., 2021) dominate cross-modal retrieval, they operate essentially as “black boxes,” prioritizing discriminative alignment over representational structure. To address this, we adapt the **Variational Multi-Modal Semantic-Style Translation (VMSST)** architecture (Wieting et al., 2022) to the vision-language domain.

Our primary contribution is a dual-stream VAE framework designed to explicitly disentangle shared semantic content (μ_{sem}) from modality-specific noise (z_{noise}). Unlike standard approaches, we introduce **Deep Projection Heads** and a hybrid inference strategy, leveraging deterministic centroids for alignment and stochastic sampling for reconstruction, to stabilize the joint distribution learning. We rigorously quantify the *generative cost*: our experiments on MS-COCO demonstrate that VMSST achieves competitive retrieval performance (within $\sim 6.5\%$ of a contrastive baseline) while unlocking bidirectional generation and interpretability absent in purely discriminative models.

The source code for the project is available at github.com/AP71/Multimodal-VMSST

2. Related Work

State-of-the-art retrieval utilizes contrastive dual-encoders (e.g., CLIP (Radford et al., 2021), ALBEF (Li et al., 2021)) optimized via InfoNCE loss to maximize mutual information. Despite their high retrieval accuracy, these models produce entangled embeddings incapable of data generation. In the generative landscape, Variational Autoencoders (VAEs) offer structured latent spaces. Our work builds directly on the VMSST framework (Wieting et al., 2022), originally proposed for multilingual retrieval. We extend this methodology to the heterogeneous image-text domain,

diverging from recent hybrid approaches (e.g., BLIP (Li et al., 2022)) to focus on a pure variational translation objective. This allows us to isolate and evaluate the trade-off between retrieval precision and semantic-style disentanglement.

3. Method

We adopted an offline preprocessing strategy for both the VMSST model and the CLIP baseline in order to ensure computational efficiency and focus the training budget on the alignment modules. Instead of performing end-to-end fine-tuning, we pre-computed all visual and textual embeddings using the frozen **DINOv2** (Oquab et al., 2023) and **RoBERTa Large** (Liu et al., 2019) backbones, respectively. By caching these high-dimensional representations, we bypass the computationally expensive forward passes of the backbone networks during training, significantly reducing GPU memory requirements and epoch duration.

3.1. Baselines

We rigorously evaluated the trade-off between discriminative precision and generative capabilities implementing a simplified **CLIP** (Radford et al., 2021) baseline. Leveraging the pre-computed features described above, the model projects both modalities onto a shared 512-dimensional hypersphere. It is optimized using the **Symmetric InfoNCE Loss** (Oord et al., 2018) with a learnable temperature parameter τ (initialized at $\log(\frac{1}{0.07})$) to maximize the cosine similarity of positive pairs against the $N^2 - N$ negatives:

$$\mathcal{L}_{CLIP} = \frac{1}{2}(\mathcal{L}_{I \rightarrow T} + \mathcal{L}_{T \rightarrow I}) \quad (1)$$

This baseline serves as the theoretical upper bound for retrieval accuracy in our experiments.

3.2. Contribution: The VMSST Architecture

Our primary contribution is the adaptation of the Variational Multi-Modal Semantic-Style Translation (VMSST) framework to the vision-language domain. While originally designed for multilingual text retrieval, applying this architecture to image-text matching introduces significant challenges due to the structural heterogeneity of the data.

¹Sapienza University of Rome Email: Alessio Pannozzo <pannozzo.1960374@studenti.uniroma1.it>, Federico Raponi <raponi.1963339@studenti.uniroma1.it>, Lorenzo Spataro <spataro.1946590@studenti.uniroma1.it>.

Unlike contrastive models that map inputs to a single point, VMSST maps each input x to a probability distribution in a disentangled latent space.

3.2.1. LATENT DISENTANGLEMENT STRATEGY

The core innovation of our architecture is the explicit disentanglement of the latent representation into two independent subspaces. We define a **Semantic Latent** (μ_{sem}, σ_{sem}) designed to capture modality-invariant information, such as the high-level concept of a scene, which serves as the sole component for cross-modal alignment and retrieval. Complementing this, we introduce a **Noise/Style Latent** ($\mu_{noise}, \sigma_{noise}$), a private subspace responsible for encoding modality-specific attributes irrelevant to alignment, such as pixel-level texture or syntactic phrasing. This separation effectively addresses the *modality gap* by offloading structural discrepancies into the noise variable, thereby allowing the semantic component to focus purely on geometric alignment.

3.2.2. DEEP PROJECTION HEADS

Standard linear mappings proved insufficient for parameterizing the complex manifolds of DINOv2 and RoBERTa. Consequently, we implemented **Deep Projection Heads** designed as non-linear MLPs equipped with residual connections. Each head consists of an input projection followed by a stack of **Residual Blocks** (He et al., 2015) where each block applies a sequence of linear transformations, LayerNorm regularization, dropout and GELU activations before the residual addition and a final output mapping. Crucially, we employ an architectural bias by using deeper heads (2 blocks) for the semantic branch to capture high-level abstractions, and shallower heads (1 block) for the noise branch, implicitly encouraging the model to store the most significant information in the shared semantic subspace.

3.2.3. SYMMETRICAL DECODERS

We employed two structurally identical decoders that allowed us to reconstruct the original backbone features. Crucially, each decoder takes as input the **concatenation** of the semantic and noise latent variables ($z_{in} = z_{sem} \oplus z_{noise}$). The architecture mirrors the deep projection heads, utilizing an initial expansion layer followed by a stack of two **Residual Blocks** and a final linear projection to map the combined latent representation back to the 1024-dimensional feature space of DINOv2 and RoBERTa, respectively.

3.2.4. LOSS FUNCTION

To handle the manifold disparity, we utilize a composite objective optimized via **Cosine Distance** ($\mathcal{D}$) rather than

MSE. Based on extensive tuning, we set the weighting factors as $\lambda_{pure} \approx 7.02$, $\lambda_{align} \approx 1.18$, while keeping regularization terms low ($\lambda_{elbo} \approx 0.004$, $\beta \approx 0.014$). The complete objective and its components are defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{cross} + \lambda_{pure}\mathcal{L}_{pure} + \lambda_{align}\mathcal{L}_{align} + \lambda_{elbo}\mathcal{L}_{self} + \beta\mathcal{L}_{KL} \quad (2)$$

$$\mathcal{L}_{cross} = \sum_{\substack{A, B \in \{i, t\} \\ A \neq B}} \mathcal{D}(D_A(\mu_{sem}^B \oplus z_{noise}^A), x_A) \quad (3)$$

$$\mathcal{L}_{pure} = \sum_{m \in \{i, t\}} \mathcal{D}(D_m(z_{sem}^m \oplus \mathbf{0}), x_m) \quad (4)$$

$$\mathcal{L}_{align} = \mathcal{D}(\mu_{sem}^i, \mu_{sem}^t) \quad (5)$$

$$\mathcal{L}_{self} = \sum_{m \in \{i, t\}} \mathcal{D}(D_m(z_{sem}^m \oplus z_{noise}^m), x_m) \quad (6)$$

$$\mathcal{L}_{KL} = \sum_{m \in \{i, t\}} \sum_{k \in \{s, n\}} D_{KL}(q_k^m(z|x) \parallel \mathcal{N}(0, I)) \quad (7)$$

Note: $i=img, t=txt, s=sem, n=noise$. In $\mathcal{L}_{self}$ and $\mathcal{L}_{pure}$, latents are sampled ($z = \mu + \sigma \odot \epsilon$) (Kingma & Welling, 2013).

1. Cross-Modal Translation (Eq. 3): This is the primary objective. We reconstruct modality A using the *deterministic semantic mean* of B combined with the *noise* of A . Crucially, we apply Noise Dropout ($p \approx 0.37$) to z_{noise} to force reliance on the semantic variable.

2. Pure Reconstruction (Eq. 4): To prevent information leakage, this heavily weighted term forces reconstruction **solely from the semantic variable** by explicitly setting the noise vector to zero ($\mathbf{0}$). This guarantees that z_{sem} captures all essential content for zero-shot tasks.

3. Alignment (Eq. 5): We explicitly minimize the cosine distance between the semantic centroids of paired inputs. This geometric constraint creates the dense clustering required for retrieval.

4. VAE Regularization (Eq. 6, 7): We include standard ELBO terms to ensure manifold smoothness. $\mathcal{L}_{self}$ enforces reconstruction using *sampled* latents (unlike the deterministic mean used in translation), while $\mathcal{L}_{KL}$ regularizes all four latent distributions (semantic and noise for both modalities) towards $\mathcal{N}(0, I)$ using a β -warming schedule.

4. Results

Experiments were conducted on both the MS-COCO (Lin et al., 2015) and Flickr30k (Plummer et al., 2015) datasets utilizing the standard Karpathy splits:

Train → Eval	Model	Text-to-Image (T2I)			Image-to-Text (I2T)			Geometric Metrics		
		R@1	R@5	R@10	R@1	R@5	R@10	Align	Sim ⁺	Sim [−]
Training Source: MS-COCO										
COCO → COCO (Test)	CLIP	58.57	87.82	94.26	72.20	93.40	97.10	0.55	0.60	0.05
	VMSST	52.02	83.57	92.01	63.72	88.48	92.01	0.76	0.77	0.01
COCO → Flickr (ZS)	CLIP	45.52	74.50	83.54	60.80	84.00	90.10	0.48	0.59	0.11
	VMSST	41.34	70.62	79.94	54.10	81.20	88.20	0.55	0.67	0.12
Training Source: Flickr30k										
Flickr → Flickr (Test)	CLIP	55.14	82.86	89.62	67.90	91.40	95.20	0.66	0.66	0.00
	VMSST	54.82	82.76	89.04	66.10	89.70	95.00	0.60	0.66	0.06
Flickr → COCO (ZS)	CLIP	28.80	56.95	70.22	36.26	62.70	73.24	0.56	0.62	0.06
	VMSST	30.80	59.73	73.23	38.80	65.94	76.72	0.53	0.63	0.10

Table 1. **Quantitative Results & Cross-Dataset Generalization.** Comparison between CLIP (Baseline) and VMSST (Ours). Best results in each column are marked in **bold**. The table is divided by training source to highlight Zero-Shot (ZS) transfer capabilities. Note: "Train → Eval" indicates the training dataset and the evaluation target. ZS denotes Zero-Shot evaluation.

- **MS-COCO:** 5,000 images are reserved for validation and 5,000 for testing. The remaining images from the 2014 train and val sets are combined to form a training set of 113,287 images. To strictly adhere to standard evaluation protocols, we report retrieval metrics by averaging results over **5 folds of 1,000 images** each extracted from the test set.
- **Flickr30k:** 1,000 images used for validation, 1,000 for testing, and the remaining 29,783 images for training.

To rigorously assess generalization capabilities, we implemented a cross-evaluation protocol: models trained on MS-COCO were evaluated in-domain and zero-shot on Flickr30k, while models trained on Flickr30k underwent the inverse procedure. We report standard Recall@K metrics ($K \in \{1, 5, 10\}$) for bidirectional retrieval tasks. Additionally, to probe the manifold structure, we monitor the Alignment Score and the average cosine similarities for positive (Sim_{pos}) and negative (Sim_{neg}) pairs.

4.1. Main Results

Table 1 summarizes the performance across diverse training scales. On the large-scale **COCO** → **COCO** benchmark, VMSST achieves competitive results (T2I R@1: 52.02%) compared to the CLIP baseline (58.57%), with a $\approx 6.5\%$ gap representing the trade-off for enforcing a generative manifold. Interestingly, on the smaller **Flickr** → **Flickr** split, this gap virtually vanishes (54.82% vs 55.14% in T2I R@1), suggesting that VMSST’s variational constraints prevent overfitting on smaller datasets. Critically,

in the challenging **Flickr** → **COCO** Zero-Shot setting (transferring from small to large domain), VMSST demonstrates superior robustness, **outperforming** the CLIP baseline across all retrieval metrics (T2I R@1: 30.80% vs 28.80%; R@10: 73.23% vs 70.22%). This indicates that while CLIP relies on large-scale data to refine its boundaries, VMSST learns more transferable semantic abstractions. Geometrically, the proper normalization of CLIP reveals a structural convergence: both models now exhibit orthogonal negative pairs ($\text{Sim}_{neg} \approx 0$). However, VMSST maintains superior latent cohesion on the source domain ($\text{Sim}_{pos} \approx 0.77$ vs 0.60 on COCO), confirming that it prioritizes semantic density—a prerequisite for generation.

5. Conclusion

Our investigation quantifies the structural trade-off between discriminative precision and generative density. The results demonstrate that while the "generative constraint" incurs a moderate precision cost on large-scale benchmarks (COCO), it offers distinct advantages in data efficiency and robustness. Most notably, in the **Flickr** → **COCO** Zero-Shot transfer, VMSST surpasses the contrastive baseline, suggesting that the variational regularization effectively mitigates overfitting when training data is scarce. By forcing the model to reconstruct the input via a disentangled semantic bottleneck, VMSST learns robust conceptual representations rather than exploiting superficial discriminative features. Future integration of this architecture into Large Multimodal Models (LMMs) could provide a path to unify retrieval and generation within a single, geometrically consistent embedding space.

References

- Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. *CoRR*, abs/1907.10902, 2019. URL <http://arxiv.org/abs/1907.10902>.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition, 2015. URL <http://arxiv.org/abs/1512.03385>.
- Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes, 2013. URL <http://arxiv.org/abs/1312.6114>.
- Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S., Xiong, C., and Hoi, S. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation, 2021. URL <http://arxiv.org/abs/2107.07651>.
- Li, J., Li, D., Xiong, C., and Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, 2022. URL <https://arxiv.org/abs/2201.12086>.
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. Microsoft COCO: Common Objects in Context, 2015. URL <http://arxiv.org/abs/1405.0312>.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. URL <http://arxiv.org/abs/1907.11692>.
- Oord, A. v. d., Li, Y., and Vinyals, O. Representation Learning with Contrastive Predictive Coding, 2018. URL <http://arxiv.org/abs/1807.03748>.
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labetut, P., Joulin, A., and Bojanowski, P. DINOv2: Learning Robust Visual Features without Supervision, 2023. URL <http://arxiv.org/abs/2304.07193>.
- Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. 2015. doi: 10.1109/ICCV.2015.303. URL <http://ieeexplore.ieee.org/document/7410660/>.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision, 2021. URL <http://arxiv.org/abs/2103.00020>.
- Wieting, J., Clark, J. H., Cohen, W. W., Neubig, G., and Berg-Kirkpatrick, T. Beyond contrastive learning: A variational generative model for multilingual retrieval. 2022. doi: 10.48550/arXiv.2212.10726. URL <http://arxiv.org/abs/2212.10726>.

6. Appendix

6.1. Qualitative Analysis

In this section, we provide additional qualitative analysis of the latent space representations learned by our proposed model compared to the standard CLIP architecture. Figures 1 through 4 present t-SNE visualizations of the image-text embeddings on both the MS COCO and Flickr30k datasets.

The visualizations are organized to show a side-by-side comparison between **our proposed method** (Left) and the **CLIP reference model** (Right). We analyze the embedding topology in both training (*supervised*) and zero-shot (*unseen data*) scenarios to demonstrate the semantic clustering capabilities and generalization potential of our approach.

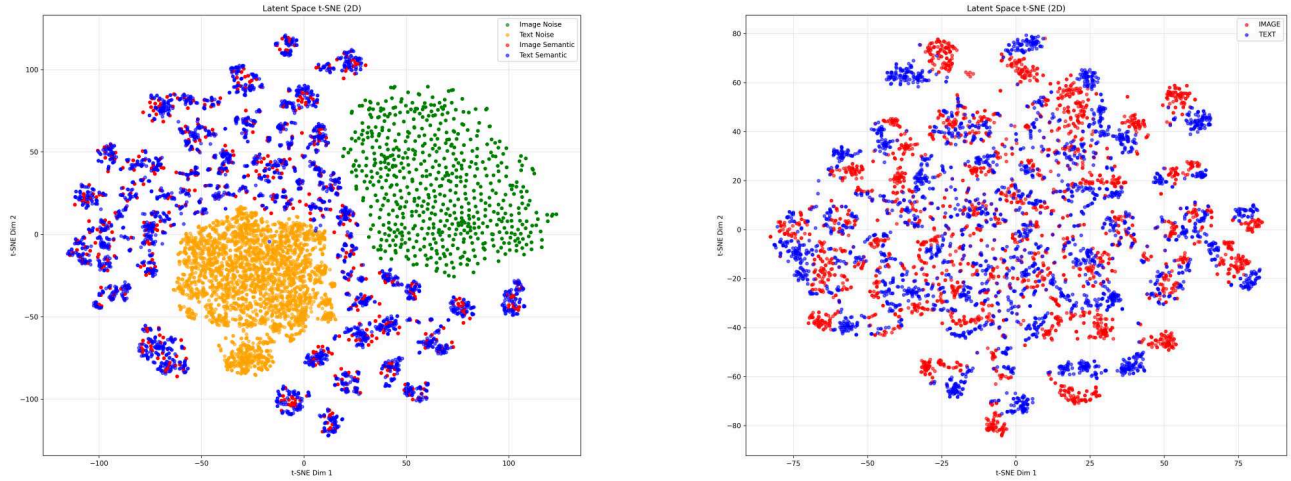


Figure 1. **Training Set Analysis (MS COCO).** t-SNE visualization of the latent embedding space.

Left: Our proposed model showing the learned semantic distribution.

Right: The CLIP reference model.

Comparing the two distributions highlights how our method structures the latent space relative to the pre-trained CLIP standard.

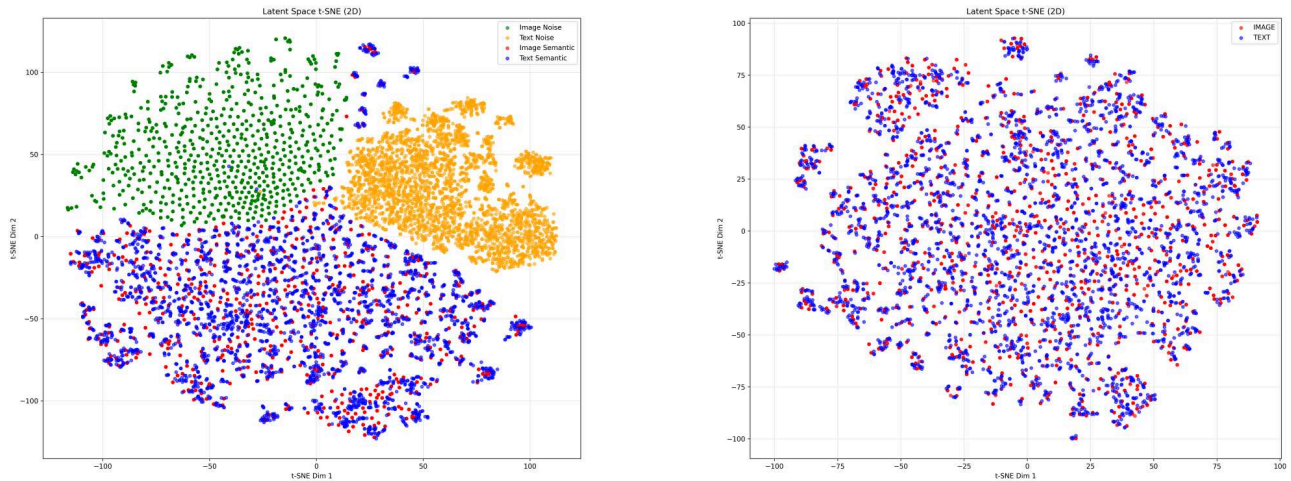


Figure 2. **Training Set Analysis (Flickr30k).**

Comparison of embedding densities.

Left: Our proposed model.

Right: CLIP reference.

The visualization demonstrates the ability of our model to form semantic clusters on the Flickr30k dataset comparable to the reference architecture.

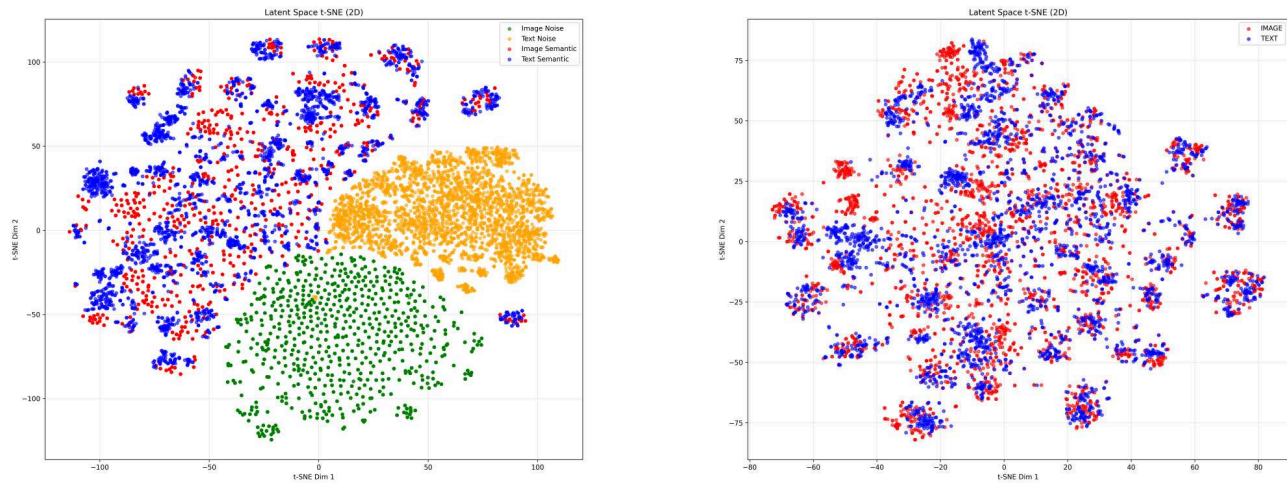


Figure 3. Zero-Shot Generalization (MS COCO).

Left: Inference results using our proposed model on unseen data.

Right: CLIP reference results.

This comparison assesses the stability of the learned features when transferring to the zero-shot task.

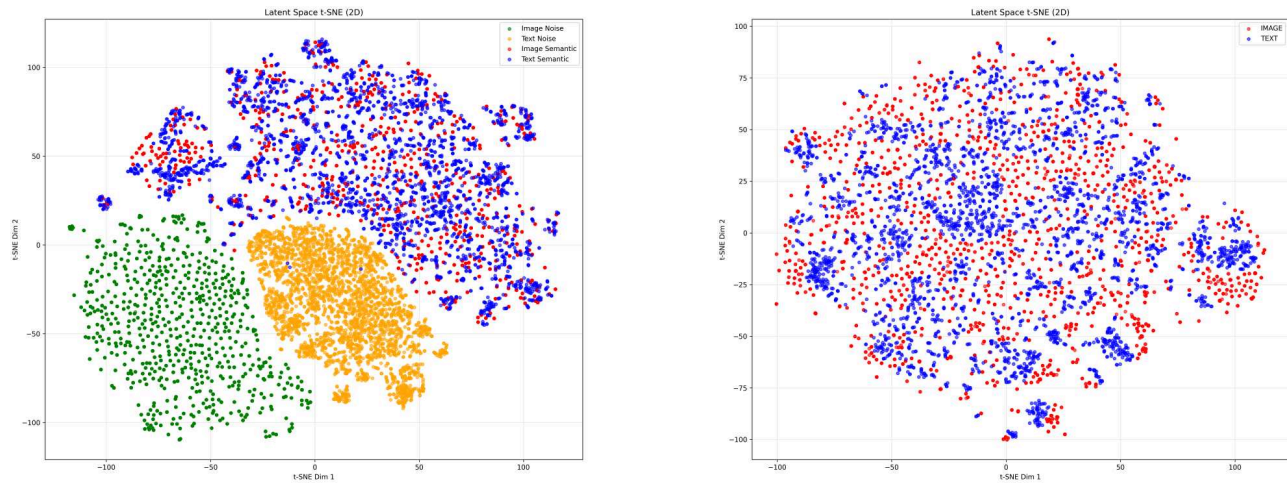


Figure 4. Zero-Shot Generalization (Flickr30k).

Left: Our proposed model.

Right: CLIP reference.

The plots illustrate the generalization capacity of our approach, maintaining semantic coherence even without direct fine-tuning on the target data.

6.2. Ablation Study: Preventing Shortcut Learning

To validate the architectural modifications introduced for the vision-language domain, we conducted a comparative study against the original VMSST parameterization (Wieting et al., 2022).

Experimental Setup We trained the model using the standard hyperparameters proposed in the original paper, specifically removing our proposed disentanglement constraints:

- **No Pure Reconstruction:** $\lambda_{pure} = 0$. The decoder has full access to the noise variable without penalty.
- **Loss Configuration:** We utilized the original weighting scheme where alignment is prioritized, but no explicit mechanism forces the semantic variable to carry reconstruction information.

Results: The "Manifold Collapse" Phenomenon We report metrics on the **MS-COCO 1K (5-fold)** splits. The results (Table 2) reveal a critical failure mode compared to our final method. While the baseline achieves moderate performance on this easier split (Mean R@1 $\approx$ 19.6%), the geometric metrics elucidate why it fails to generalize further. The model suffers from **Partial Manifold Collapse**: both Sim_{pos} (0.99) and Sim_{neg} (0.89) are extremely high.

This indicates that without the regularization provided by $\mathcal{L}_{pure}$, the encoder minimizes the alignment loss by mapping *all* inputs—regardless of content—into a narrow cone in the latent space. Because distinct semantic concepts are not sufficiently orthogonal (Alignment Score $\approx$ 0.10 vs 0.76), the model lacks the discriminative power required for high-precision retrieval.

Configuration	T2I R@1	I2T R@1	Align	Sim ⁺	Sim [−]
Original VMSST (1K Fold)	18.60%	20.74%	0.10	0.99	0.89
Ours (Composite)	52.02%	63.72%	0.76	0.77	0.01

Table 2. **Impact of Disentanglement Constraints.** Comparison between the Original VMSST configuration (evaluated on 1K Folds) and our Final Composite Model. Note the high Sim[−] in the baseline, indicating a collapsed latent space.

6.3. Hyperparameters optimization

We conducted an automated hyperparameter search utilizing the Optuna framework (Akiba et al., 2019) so we could determine the optimal balance between generative capability and retrieval precision. The search space was navigated using a **Tree-structured Parzen Estimator (TPE)** sampler, paired with a **Median Pruner** to terminate unpromising trials early and ensure computational efficiency.

The optimization objective was to maximize the **Recall@1** metric on the validation set (val_R@1). We targeted critical hyperparameters including the learning rate, batch size, dropout rates (standard and noise dropout), and the weighting factors of the composite loss function ($\lambda_{pure}, \lambda_{align}, \lambda_{elbo}, \beta$).

Following 30 optimization trials, the best configuration achieved a val_R@1 of 28.92. The identified optimal values, including the ones not optimized, which were adopted for the final training described in Section 3, are:

Hyperparameter	Value
<i>Architecture</i>	
Input Dimension	1024
Latent Dimension	512
Encoder Hidden Dimension	1024
Decoder Hidden Dimension	2048
<i>Optimization</i>	
Batch Size	64
Max Epochs	50
Warm-up Epochs	10
Learning Rate	1.34×10^{-4}
Weight Decay	0.01
<i>Regularization & Loss Weights</i>	
Dropout	0.12
Noise Dropout (p_{noise})	0.37
Pure Reconstruction Weight (λ_{pure})	7.02
Alignment Weight (λ_{align})	1.18
ELBO Weight (λ_{elbo})	0.0044
KL Divergence Weight (β_{kl})	0.0138

Table 3. Final Hyperparameters Configuration. The values were determined through Optuna optimization or fixed based on architectural constraints.