Report

Miguel Benalcazar
migangben@gmail.com

October, 2025

Contents

1 Experiment: Multimodal Cross Retrieval using Laplacian Sheaf Theoretic

1.1 Overview

In this experiment, a multimodal cross-retrieval task was conducted by using a sheaf-theoretic Laplacian constraints over vision and language embeddings, as recommended in [1], Section C (Real-World Applications of Sheaf-FMTL). The sheaf framework was applied to the COCO dataset, where DINOv2 was used to extract visual embeddings from images, and BERT to generate textual embeddings from captions. Following the motivation in Sections C.1 and C.2 of [1], sheaf-theoretic modeling was used to capture feature vector similarities in multimodal tasks, enforcing consistency between image and text embeddings through the sheaf Laplacian operator.

This setup models local and global consistency constraints across modalities, with restriction maps defining the shared subspaces between vision and language. By incorporating these sheaf-theoretic constraints, the approach improves representation alignment for downstream cross-modal retrieval tasks (image-to-text and text-to-image), highlighting the potential of sheaf-based frameworks in multimodal learning.

1.2 Dataset

The **MS COCO (Microsoft Common Objects in Context)** dataset is a large-scale benchmark for image recognition, detection, segmentation, and captioning.

For the image–text retrieval and captioning tasks, each image in the dataset is paired with five different human-written captions. These captions are written by different annotators to describe the same image in slightly different ways, capturing various details or perspectives.

1.3 Framework

The experiment uses `facebook/dinov2-small` for visual embeddings and `distilbert-base-uncased` for textual embeddings. Each model provides a *local downstream*:

- DINOv2 last hidden state image features with dimension 384.
- DistilBERT last hidden state text features with dimension 768.

To merge the two modalities, a sheaf-theoretic Laplacian framework is applied with linear restriction maps, denoted as P_{12} and P_{21} (see Fig. 1). These maps project the visual and textual embeddings into a shared 128-dimensional subspace, where *local consistency* is enforced. This process effectively merges the modality-specific embeddings into a unified *global representation*:

$$P_{12}\,\theta_{\text{img}} \;\approx\; P_{21}\,\theta_{\text{txt}}$$

Finally, this unified representation is used for the downstream *multimodal cross-retrieval task* (image-to-text and text-to-image), where the visual embeddings and captions are compared within the common space.

1.4 Efficiency in Semantic Communication

Semantic communication (System 2 SC) focuses on transmitting meaning rather than raw signals. Smaller embeddings require less bandwidth, are faster to send, compare, and fuse. Agents can exchange compact, aligned representations instead of large modality-specific features.

- Without reduction: Vision agent sends 384-d, Text agent sends 768-d $\rightarrow$ difficult to fuse.
- With reduction: Both send 128-d $\rightarrow$ communication is efficient and aligned in the same "language of meaning".

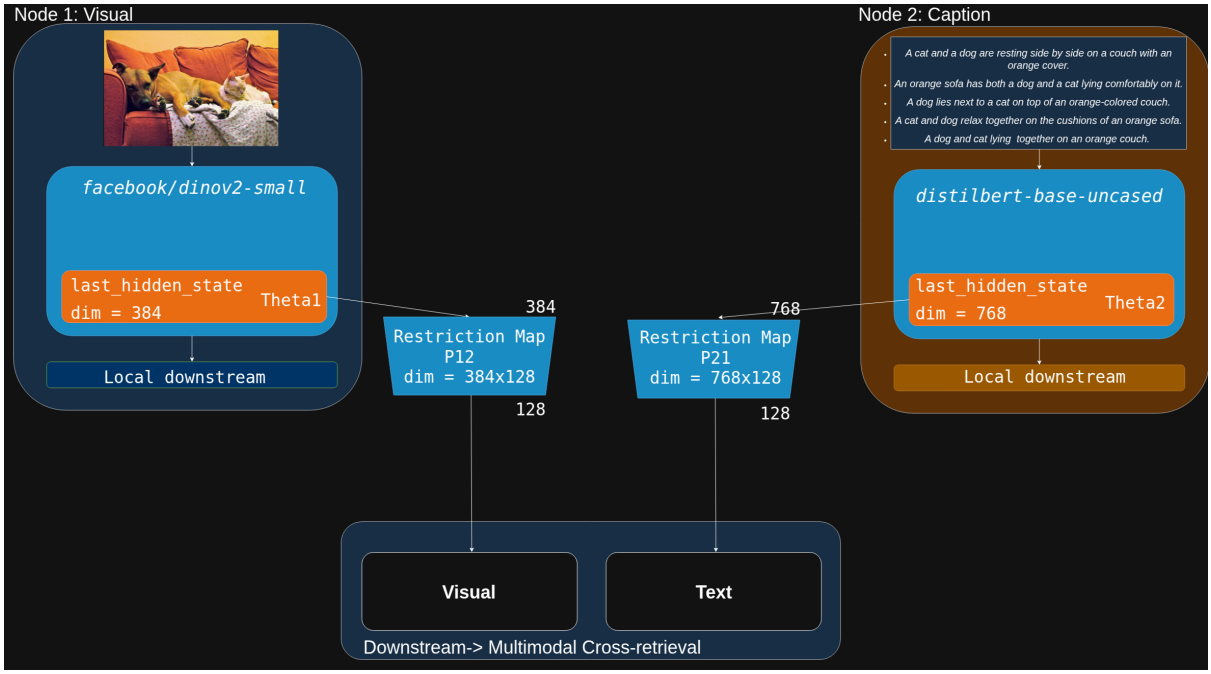

Figure 1: Framework

1.5 Experiment Setup

The visual embeddings from DINOv2 and textual embeddings from BERT are kept frozen during training. The optimization focuses on learning the restriction maps P_{12} and P_{21}, with the objective function defined as the sum of the *sheaf Laplacian loss* and the *variance regularization* of P_{12} and P_{21}.

Embeddings are extracted from the last hidden states of each model (DINOv2 for images, BERT for captions) and projected via the learned restriction maps into a shared 128-dimensional space. Cross-modal retrieval is then evaluated using standard metrics like Recall@K.

1.6 Results

The sheaf-theoretic training metrics (Table 1) show a low discrepancy norm and relatively stable variance of the restriction maps, indicating that the learned linear maps P_{12} and P_{21} successfully enforce local consistency and reduce noise in the embedding space. The cosine similarity around 0.52 suggests moderate alignment between the projected vision and text embeddings.

Table 1: Sheaf-theoretic training metrics (local consistency and embedding stability).

Epoch	Sheaf Loss	Discrepancy Norm	Cosine Similarity	Avg Var θ_1 / θ_2
6	0.0143	0.000468	0.528	0.00515 / 0.0175

The downstream retrieval results (Table 2) demonstrate that the unified 128-dimensional embeddings allow effective cross-modal retrieval. For "mage-to-text", Recall@10 reaches 0.631, showing that the correct captions are frequently retrieved in the top results. As for "Text-to-image" performance is lower, with R@10 = 0.232, likely reflecting the higher complexity of mapping a single caption to multiple possible images. Overall, incorporating the sheaf Laplacian constraints improves embedding coherence, which contributes to meaningful retrieval performance across modalities.

Table 2: Downstream multimodal cross-retrieval results on COCO (image $\leftrightarrow$ text).

Task	R@1	R@5	R@10
Image $\rightarrow$ Text	0.202	0.497	0.631
Text $\rightarrow$ Image	0.148	0.148	0.232

1.7 Conclusion

The sheaf-theoretic Laplacian constraints play a crucial role in merging the visual and textual embeddings into a shared subspace. By learning the linear restriction maps P_{12} and P_{21}, the framework enforces *local consistency* across modalities while reducing the original high-dimensional embeddings (384-d for DINOv2, 768-d for DistilBERT) into a compact 128-dimensional space. This reduction not only removes modality-specific noise but also ensures that the embeddings capture the most informative semantic features, which is particularly important for cross-modal retrieval.

On the COCO dataset, the sheaf framework effectively merges multiple captions per image into a unified representation, improving alignment with the corresponding visual embedding. This enhances embedding coherence, supports efficient computation, and handles multi-caption scenarios. While text-to-image retrieval is somewhat lower due to caption ambiguity, the sheaf Laplacian framework overall provides a principled approach for merging multimodal data and enforcing consistency for cross-modal retrieval.

References

[1] C. Ben Issaid, P. Vepakomma, and M. Bennis, "Tackling feature and sample heterogeneity in decentralized multi-task learning: A sheaf-theoretic approach," arXiv preprint arXiv:2502.01145, 2025.

[2] Benalcazar, M. (2025). *Sheaf Theory Multimodal Repository*. Available at: `https://github.com/MiguelBenalcazar/Multimodal_sheaf_theoretic`