

M-VLM2Vec: Structure-Preserving Matryoshka Representation Learning for Vision-Language Embeddings

Minh Khoi Nguyen Do^{1*}, Nguyen Hong Dang^{2*}, Vinh Le^{2*},
Minh Quoc Nhat Nguyen^{2*}, Linh Ngo Van², Duc Anh Nguyen^{2†}, Trung Le³

¹University of Rochester, ²Hanoi University of Science and Technology,
³Monash University

Abstract

Deploying high-dimensional Vision-Language Model (VLM) embeddings is computationally prohibitive. While Matryoshka Representation Learning (MRL) allows elastic truncation, standard MRL optimizes dimensional hierarchies semi-independently. This lack of cross-scale structural constraints creates a severe topological disconnect causing representational collapse at lower dimensions and introduces optimization trade-offs that compromise high-dimensional capacity. To solve this, we propose **M-VLM2Vec**, a framework that shifts MRL from a compromised hierarchy to a synergistic geometric structure. Without additional supervision, we integrate two novel constraints into the multimodal contrastive objective: (1) Laplacian spectral alignment to preserve the global manifold topology across scales, and (2) Stiefel-constrained orthogonal projection to bidirectionally align local contrastive landscapes. By allowing nested spaces to mutually regularize, M-VLM2Vec eliminates competing gradients, preventing ultra-low-dimensional collapse while yielding striking, consistent performance gains across all embedding scales on the Multimodal Embedding Benchmark (MMEB).

1 Introduction

Vision-Language Models (VLMs) like CLIP, ALIGN, and SigLIP (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023) yield highly expressive multimodal representations with remarkable zero-shot transferability. To maintain fidelity across complex representation spaces, recent studies enforce structural preservation in dense encoders through intra-model relational modeling (Truong et al., 2025), hierarchical relational distillation (Le et al., 2026b), and adaptive query selection mechanisms (Minh et al., 2026). However, because these systems fundamentally rely on high-dimensional embedding

spaces (e.g., $\mathbb{R}^{1024}$ or $\mathbb{R}^{4096}$), they impose severe storage overhead and latency bottlenecks in large-scale retrieval and edge deployments. Matryoshka Representation Learning (MRL) (Kusupati et al., 2022) effectively bridges this gap by forcing a single encoder to learn nested, multi-granular sub-embeddings, enabling downstream systems to dynamically truncate the vector to a lower dimension d_s without retraining.

Despite its practical utility, standard MRL suffers from a fundamental design flaw: it treats adjacent dimensional hierarchies as semi-independent optimization sub-problems. This independence creates two critical points of failure: (1) a severe **topological disconnect** where the truncated low-dimensional manifold fails to reflect high-dimensional geometry, causing structural collapse at extreme low dimensions, and (2) **competing gradients** across scales, leading to zero-sum optimization trade-offs that compromise high-dimensional capacity to accommodate lower dimensions.

We are therefore driven by a dual motive: (1) to solve the problem of *geometric structure preservation* by encouraging low-dimensional embeddings to preserve important relational structure of high-dimensional representations, and (2) to establish a rigorous *cross-dimensional consistency* that acts as a mutual regularizer. We hypothesize that tightly coupling these nested spaces geometrically will resolve optimization conflicts and synergistically boost performance across **all** dimensions. Building upon VLM2Vec (Jiang et al., 2025), we propose **Matryoshka VLM2Vec (M-VLM2Vec)**, a structure-preserving multimodal framework grounded in differential geometry and spectral graph theory. Rather than relying solely on cascaded contrastive losses, we explicitly enforce manifold consistency and spectral invariance through two rigorous regularization mechanisms. First, **Cross-Dimensional Symmetric Contrastive Consistency** utilizes orthogonal projection matri-

*Equal contribution

†Corresponding author: anhnd@soict.hust.edu.vn

ces bounded by the Stiefel manifold $\text{St}(d_s, d_l)$ to map dominant embeddings into the subordinate space. By minimizing the variation between bidirectional cross-dimensional mutual information estimators, we preserve bi-Lipschitz continuity. Crucially, this ensures the low dimensions benefit from high-dimensional richness while the high dimensions are regularized by the robust compressed core. Second, **Laplacian Spectral Alignment** provides a theoretical foundation with bounded spectral distortion, encouraging the preservation of important low-frequency structural properties—such as data clustering and intrinsic geometry—across embedding scales. By constructing the normalized discrete Laplace-Beltrami operator for both dimensional spaces, we strictly align their eigen-spectra via a symmetric Kullback-Leibler divergence.

By deeply integrating these geometric priors, M-VLM2Vec successfully turns dimensional nesting from an optimization compromise into a synergistic advantage. Our main contributions are summarized as follows:

- We identify the structural collapse and optimization trade-offs inherent in standard multimodal Matryoshka learning and propose **M-VLM2Vec**, a framework driven by cross-scale geometric preservation.
- We formulate two novel geometric consistency objectives: an orthogonal projection mechanism to enforce symmetric cross-modal mutual information, and a spectral alignment loss to preserve the Laplace-Beltrami graph topology of the latent space.
- We comprehensively evaluate on the Multimodal Embedding Benchmark (Jiang et al., 2025). Validating our core motivation, M-VLM2Vec demonstrates consistent, synergistic performance gains across **all dimensions** (rather than just mitigating low-dimension degradation), with striking improvements in ultra-low-dimensional regimes for Classification and Visual Question Answering.

2 Related Work and Background

2.1 Related Work

Evolution of Sentence Representations Foundational sentence embeddings were primarily constructed from static word vectors (Pennington

et al., 2014; Mikolov et al., 2013) and early contextualized encoders (Peters et al., 2018; Devlin et al., 2019), which often lacked specific tuning for sentence-level semantic alignment. Sentence-BERT (Reimers and Gurevych, 2019) catalyzed a shift by introducing a siamese fine-tuning objective, drastically improving efficiency for semantic similarity and dense retrieval tasks. This was followed by advancements in contrastive learning, which further refined representational discriminability (Gao et al., 2021; Nishikawa et al., 2022). Currently, the state-of-the-art has shifted toward leveraging Large Language Models, which act as high-capacity backbones or distilled teachers to generate increasingly sophisticated sentence-level manifolds (He et al., 2025; BehnamGhader et al., 2024; Tao et al., 2025).

Matryoshka Representation Learning Methods

Although sentence representation learning has advanced significantly, fixed embedding dimensionality still limits large-scale deployment due to computational and storage costs. Matryoshka Representation Learning (Kusupati et al., 2022) addresses this by learning nested embeddings in a coarse-to-fine manner, where larger prefixes yield more expressive representations. Low-dimensional prefixes can serve as standalone embeddings while requiring supervision at only $O(\log d)$ checkpoints and no extra forward passes. This enables strong compression with competitive performance and faster inference. Building on this idea, Espresso Sentence Embeddings (ESE) introduce learn-to-express and learn-to-compress mechanisms to scale representations across both dimensionality and network depth (LI et al., 2025). The Matryoshka principle has also been extended beyond single-modal setups into image generation (Gu et al., 2024), multimodal architectures (Cai et al., 2025), and LLMs (Hu et al., 2024). To mitigate cross-scale degradation inherent in dimensional truncation, recent advances employ self-distilled intra-relational chaining (Huy et al., 2026), span-aware Matryoshka distillation across heterogeneous tokenizers (Tran et al., 2026), and isotropic subspace alignment to maximize informational capacity (Hong et al., 2026). While these methods target text or single-modality representations, maintaining structural and geometric consistency across nested dimensions in vision-language embedding spaces remains under-explored.

Structural-Topological Knowledge Alignment Knowledge distillation (KD) serves as a criti-

cal paradigm for compressing high-capacity backbones into lightweight student models while preserving latent geometry. Beyond standard logit matching, modern cross-tokenizer KD addresses vocabulary and sequence alignment through dual-space weighting (Vu et al., 2026a), preference distillation (Nguyen et al., 2026), cross-chain-of-thought with optimal transport (Le Hai et al., 2026), and mixture-of-layers schemes (Vu et al., 2026b). To preserve structural manifolds across complex representations, recent approaches leverage multi-cost Wasserstein metrics (Vuong et al., 2026), token-level self-play (Le et al., 2026a), multi-granular trajectory alignment (Chi et al., 2026), teacher-anchored layer alignment (Dao et al., 2026b), and span-based representation matching (Dao et al., 2026a). Distinct from cross-model distillation, **M-VLM2Vec** re-purposes geometric alignment principles internally, enforcing cross-dimensional spectral and isometric consistency across nested sub-embeddings within a single multimodal encoder.

2.2 Background

Matryoshka Representation Learning (MRL). Let an encoder f_θ map an input x to a dense embedding $z = f_\theta(x) \in \mathbb{R}^D$. MRL (Kusupati et al., 2022) defines a set of increasing prefix dimensions $\mathcal{D} = \{d_1, \dots, d_n = D\}$. Each truncated prefix $z^{(d_i)} = z_{1:d_i}$ is evaluated via a task-specific head g_{ϕ_i} to compute a loss $\mathcal{L}_{\text{task}}^{(d_i)}$. MRL jointly optimizes all prefixes:

$$\mathcal{L}_{\text{MRL}}(\theta, \{\phi_i\}) = \sum_{i=1}^n \mathbb{E}_{(x,y)} [\mathcal{L}_{\text{task}}^{(d_i)}(x, y)]$$

This nested supervision encourages semantic richness across all scales, enabling *adaptive retrieval* using compressed prefixes for fast, memory-efficient coarse searches, followed by full-embedding re-ranking. Consequently, MRL drastically reduces storage and latency, allowing a single trained model to elastically serve diverse hardware budgets.

Contrastive Representation Learning. Following VLM2Vec (Jiang et al., 2025), our vision-language model $\mathcal{M}_\theta$ embeds a multimodal input using the representation of its final eos token, $\mathbf{z} \in \mathbb{R}^d$. For a mini-batch of size B , given queries $\mathbf{z}_i^q$ and their positive targets $\mathbf{z}_i^{t+}$, we optimize the embed-

ding space using the InfoNCE loss:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\text{sim}(\mathbf{z}_i^q, \mathbf{z}_i^{t+})/\tau)}{\sum_{j=1}^B \exp(\text{sim}(\mathbf{z}_i^q, \mathbf{z}_j^t)/\tau)}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is the temperature hyperparameter.

3 Methodology

The overall framework of our proposed method is illustrated in Figure 1, highlighting the seamless integration of global and local structural constraints.

3.1 Global Topology Preservation via Laplacian Spectral Consistency

Standard MRL isolates dimensional hierarchies, enforcing structure solely through independent point-wise contrastive losses. This neglect of the *global geometric relationship* between adjacent subspaces causes **topological drift**, where low-dimensional spaces structurally collapse under restricted capacity. To solve this degradation and actively synergize dimensions, we shift from rigid coordinate matching to *global topological matching*. By aligning the discrete Graph Laplacians of the underlying data manifolds across scales, we encourage the approximate preservation of global relational geometry, which provides bounded spectral distortion for structural consistency without overly constraining the local coordinates.

Similarity Graph and Affinity Kernel. Rather than operating on the true underlying data manifold—which is computationally intractable—we construct a mini-batch similarity graph as a stochastic approximation of the local manifold geometry. This construction is analogous to Monte-Carlo approximations commonly used in spectral clustering and manifold learning. We formalize the graph construction through the following details:

- **Cosine similarity computation:** For ℓ_2 normalized batch embeddings $Z \in \mathbb{R}^{B \times d}$, we compute pairwise similarities $S = ZZ^\top$, where $S_{ij} = \langle z_i, z_j \rangle$.
- **Exponential affinity kernel:** We apply a temperature-scaled exponential kernel $A_{ij} = \exp(S_{ij}/\tau)$, where $\tau > 0$ controls the localized heat diffusion bandwidth.
- **Diagonal masking:** Crucially, to satisfy the conditions of our theoretical guarantees (The-

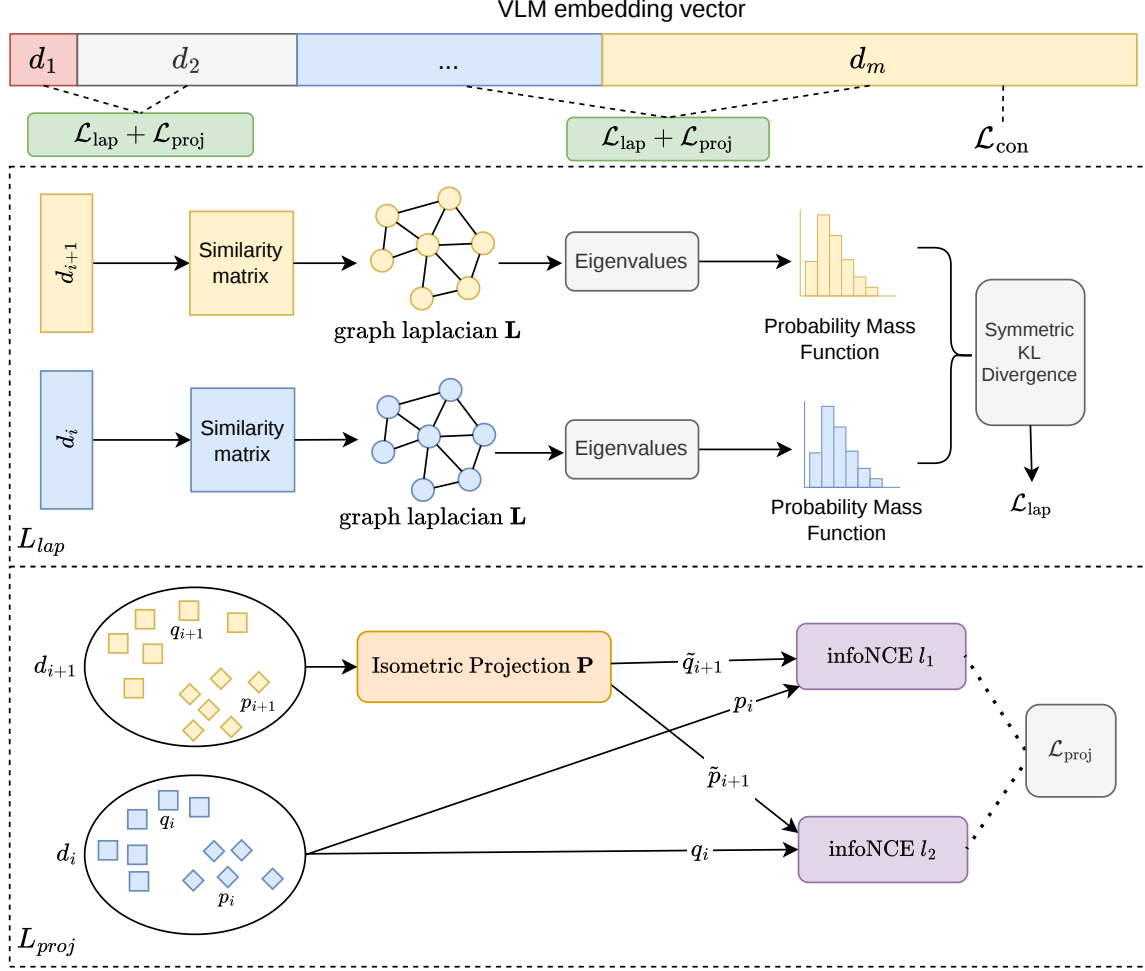


Figure 1: Overview of our proposed M-VLM2Vec framework. The architecture combines Laplacian spectral alignment ($\mathcal{L}_{\text{lap}}$) for structural topology consistency across nested dimensions and isometric projection ($\mathcal{L}_{\text{proj}}$) with InfoNCE loss to enforce cross-dimensional invariance.

orem 1) and prevent self-loops from dominating the graph structure, self-similarities are explicitly removed before graph construction by setting $A_{ii} = 0$.

Symmetric Normalized Graph Laplacian. We capture the graph topology using the symmetric normalized Graph Laplacian $\mathbf{L} \in \mathbb{R}^{B \times B}$, which corrects for irregular sampling densities:

$$\mathbf{D}_{ii} = \sum_{j=1}^B \mathbf{A}_{ij}$$

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$$

Eigenspectrum as a Topological Fingerprint. The manifold’s structural complexity is compactly summarized by its real eigenvalues, representing the “resonant frequencies” of the graph:

$$\sigma(\mathbf{L}) = \{\lambda^{(1)}, \lambda^{(2)}, \dots, \lambda^{(B)}\}.$$

Because we utilize the normalized Laplacian, these eigenvalues are strictly bounded and sorted such that:

$$0 = \lambda^{(1)} \leq \lambda^{(2)} \leq \dots \leq \lambda^{(B)} \leq 2.$$

By analyzing the distribution of these frequencies, we capture the complete hierarchical geometry of the representation space, from distinct global categories down to highly localized sample variations:

- **Low Frequencies (Macro-Geometry):** The smallest non-zero eigenvalues, particularly the Fiedler value (Fiedler, 1973), $\lambda^{(2)}$, quantify manifold bottlenecks (the minimum Cheeger cut (Cheeger, 2015)), dictating the macro-level separation of major semantic categories.
- **High Frequencies (Micro-Geometry):** The largest eigenvalues capture localized intra-cluster variance, fine-grained spacing, and

point-wise noise within isolated semantic neighborhoods.

While the Laplacian spectrum is not a complete graph invariant, it serves as a robust descriptor regularizing: (1) Graph conductance via the Fiedler value; (2) Connectivity structure; (3) Diffusion behavior; (4) Low-frequency global organization; and (5) Bottleneck structures.

Spectral Loss Formulation. To rigorously quantify and align these topological summaries across dimensions d_s and d_l , we normalize the spectra into 1D Probability Mass Functions (PMFs):

$$p_i = \frac{\lambda^{(i)}}{\sum_{k=1}^B \lambda^{(k)}}$$

We eliminate structural collapse and enforce spectral consistency by minimizing the symmetric Kullback-Leibler (KL) divergence between adjacent spectral distributions $\mathcal{P}_{d_s}$ and $\mathcal{P}_{d_l}$:

$$\mathcal{L}_{\text{lap}}^{(d_s, d_l)} = D_{\text{KL}}(\mathcal{P}_{d_s} \| \mathcal{P}_{d_l}) + D_{\text{KL}}(\mathcal{P}_{d_l} \| \mathcal{P}_{d_s}) \quad (1)$$

3.2 Local Alignment via Orthogonal Projection Consistency

While spectral consistency preserves global topology, optimizing adjacent spaces semi-independently still invites competing gradients at the local level. To eliminate these optimization trade-offs and drive mutual cross-scale enhancement, we explicitly align the local contrastive landscapes of queries and positives across scales.

Isometric Projection Operator. To map high-dimensional representations to lower dimensions without collapsing the feature space, we constrain the linear projection operator $\mathbf{P}$ to the Stiefel manifold (Edelman et al., 1998):

$$\mathbf{P} \in \text{St}(d_s, d_l) = \left\{ \mathbf{M} \in \mathbb{R}^{d_s \times d_l} \mid \mathbf{M}\mathbf{M}^\top = \mathbf{I}_{d_s} \right\}$$

This enforces an isometric mapping that preserves vector geometry. We project high-dimensional queries and positives as:

$$\tilde{q}_l = \mathbf{P}q_l, \quad \tilde{p}_l = \mathbf{P}p_l$$

Cross-Dimensional Contrastive Loss & Consistency. We evaluate the alignment of projected features with native lower-dimensional counterparts through asymmetric, cross-dimensional InfoNCE losses:

$$\ell_1 = -\log \frac{\exp(\langle \tilde{q}_l, p_s \rangle / \tau)}{\sum_{k=1}^B \exp(\langle \tilde{q}_l, p_s^{(k)} \rangle / \tau)}$$

$$\ell_2 = -\log \frac{\exp(\langle q_s, \tilde{p}_l \rangle / \tau)}{\sum_{k=1}^B \exp(\langle q_s^{(k)}, \tilde{p}_l \rangle / \tau)}$$

Crucially, to ensure our embedding manifold is perfectly nested and symmetric—allowing the robust low-dimensional core to regularize the high-dimensional space, and vice-versa—we minimize the discrepancy between these bidirectional landscapes:

$$\mathcal{L}_{\text{proj}}^{(d_s, d_l)} = \|\ell_1 - \ell_2\|_1 \quad (2)$$

3.3 Unified Multi-Scale Objective

Our framework fundamentally transitions Matryoshka learning from a compromised hierarchy to a synergistic geometric structure. For a predefined set of nesting dimensions:

$$\mathcal{D} = \{d_1 < d_2 < \dots < d_n\}$$

We optimize standard in-dimension InfoNCE contrastive losses $\mathcal{L}_{\text{con}}^{(d_i)}$ alongside our cross-dimensional regularizers for all adjacent pairs (d_i, d_{i+1}) . The final joint objective optimizes encoder parameters θ and Stiefel operators $\mathbf{P}$ to elevate representational capacity holistically:

$$\begin{aligned} \mathcal{J}(\theta, \mathbf{P}) = & \sum_{i=1}^n \mathcal{L}_{\text{con}}^{(d_i)} + \lambda_1 \sum_{i=1}^{n-1} \mathcal{L}_{\text{lap}}^{(d_i, d_{i+1})} \\ & + \lambda_2 \sum_{i=1}^{n-1} \mathcal{L}_{\text{proj}}^{(d_i, d_{i+1})} \end{aligned} \quad (3)$$

where λ_1 and λ_2 govern the global spectral consistency and local projection alignment, respectively. Crucially, the projection matrices $\mathbf{P}$ act as learnable parameters that are driven by the local consistency objective $\mathcal{L}_{\text{proj}}$. The Stiefel constraint ($\mathbf{P}\mathbf{P}^\top = \mathbf{I}_{d_s}$) is essential because an unconstrained projection matrix could minimize the alignment objective through degenerate solutions, such as shrinking feature magnitudes or discarding informative directions. By enforcing this orthogonality during training, the optimization preserves the geometry of the retained subspace and prevents dimensional collapse.

4 Theoretical Analysis

A fundamental question arising from our formulation is whether minimizing the Kullback–Leibler divergence between normalized spectral distributions effectively regularizes the manifold structure. We answer this affirmatively by showing that our

spectral consistency loss bounds the distortion of the graph’s isoperimetric structure, as characterized by the Fiedler value (Fiedler, 1973). Let h_G denote the Cheeger constant (Cheeger, 2015) (graph conductance) of the similarity graph, which measures the difficulty of partitioning the graph into two disjoint but internally dense semantic clusters. By the discrete Cheeger inequality (Chung, 1997), the Cheeger constant (Cheeger, 2015) is bounded by the Fiedler value (Fiedler, 1973) $\lambda^{(2)}$ of the normalized graph Laplacian $\mathbf{L}$:

$$\frac{\lambda^{(2)}}{2} \leq h_G \leq \sqrt{2\lambda^{(2)}}.$$

Consequently, preserving the Fiedler value (Fiedler, 1973) across embedding dimensions is equivalent to preserving the semantic clusterability of the representation space. The following theorem establishes that our spectral consistency loss $\mathcal{L}_{\text{lap}}$ directly controls this distortion.

Theorem 1 (Bounded Distortion of the Spectral Gap). *Assume that the affinity matrix $\mathbf{A}$ has a zero diagonal, i.e.,*

$$\mathbf{A}_{ii} = 0.$$

Let $\mathbf{L}_{d_s}$ and $\mathbf{L}_{d_l}$ denote the normalized graph Laplacians at the small and large embedding dimensions, respectively, and let $\lambda_{d_s}^{(i)}, \lambda_{d_l}^{(i)}$ denote their corresponding sorted eigenvalues. If the symmetric spectral consistency loss satisfies

$$\mathcal{L}_{\text{lap}}^{(d_s, d_l)} \leq \epsilon,$$

then the distortion of the Fiedler value is bounded by

$$\left| \lambda_{d_s}^{(2)} - \lambda_{d_l}^{(2)} \right| \leq B\sqrt{2\epsilon},$$

where B denotes the batch size.

Theorem 1 provides a theoretical justification for our formulation. Since the variation in the Fiedler value (Fiedler, 1973) is explicitly bounded by $\mathcal{O}(\sqrt{\epsilon})$, the Cheeger constant (Cheeger, 2015) bounds of the higher-dimensional subspace d_l are inherited by the lower-dimensional subspace d_s with controlled distortion. Consequently, if the higher-dimensional representation learns well-separated semantic clusters, the lower-dimensional representation is provably protected against topological collapse. It should be noted that because our affinity matrix is constructed over a mini-batch, this theoretical guarantee bounds the structural distortion of the stochastic Monte-Carlo approximation of the local manifold, rather than the complete

global data manifold itself. The proof is provided in Appendix B.

Theorem 2 (Optimality and Consistency of Stiefel Projection (Edelman et al., 1998)). *Let $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_l})$. Let $\mathbf{P} \in \mathbb{R}^{d_s \times d_l}$ with $d_s < d_l$ be a linear projection from a large representation space to a smaller one. Given query-positive pairs with representations (q_l, q_s) and (p_l, p_s) , define the cross-dimensional projections*

$$\tilde{q}_l = \mathbf{P}q_l, \quad \tilde{p}_l = \mathbf{P}p_l.$$

(i) Subject to the constraint $\|\mathbf{A}\|_F^2 = d_s$, the Stiefel condition (Edelman et al., 1998) $\mathbf{P}\mathbf{P}^\top = \mathbf{I}_{d_s}$ is the unique global maximizer of the functional

$$\mathcal{F}(\mathbf{A}) = H(\mathbf{A}\mathbf{z}) - \text{KL}\left(\mathcal{N}(\mathbf{0}, \mathbf{A}\mathbf{A}^\top) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_s})\right)$$

and simultaneously maximizes the InfoNCE discriminability bound $\hat{I}_B(\tilde{q}_l; p_s)$.

(ii) If the cross-dimensional projection loss is bounded by ϵ , i.e., $\mathcal{L}_{\text{proj}} \leq \epsilon$, then the mutual information gap satisfies

$$|I(\tilde{q}_l; p_s) - I(q_s; \tilde{p}_l)| \leq \epsilon + \mathcal{O}(B^{-1}),$$

where B denotes the batch size. Furthermore, $\mathcal{L}_{\text{proj}} = 0$ almost surely if and only if the cross-covariances commute:

$$\mathbf{P} \mathbb{E}[q_l p_s^\top] = \mathbb{E}[q_s p_l^\top] \mathbf{P}^\top.$$

Theorem 2 establishes the geometric and information-theoretic foundations underlying our method. Geometrically, Part (i) shows that restricting the projection to the Stiefel manifold (Edelman et al., 1998) provides an optimal covariance structure that encourages information preservation under projection while preventing dimensional collapse and maintaining the isotropic geometry of the source space. From an information-theoretic perspective, Part (ii) shows that minimizing the cross-dimensional projection loss induces a bounded, symmetric, and bidirectional information channel between the asymmetric embedding spaces. Together, these theoretical motivations suggest that the resulting multi-scale representations remain structurally aligned across dimensions, promoting the approximate preservation of discriminative information, enabling dynamic adjustment of embedding dimensionality at inference time with minimal degradation in retrieval performance. The proof is provided in Appendix B.

5 Experiments

5.1 Experimental Setup

Datasets. To comprehensively evaluate our approach, we report results on datasets sourced from the Massive Multimodal Embedding Benchmark (MMEB) (Jiang et al., 2025), focusing on two representative task families that rely critically on embedding quality. First, for **Classification**, which requires models to capture the overall semantics of multimodal inputs, we evaluate on ImageNet-1K, SUN397, N24News, VOC2007, and HatefulMemes. Second, regarding **Visual Question Answering (VQA)**, which focuses on reasoning and modeling semantic relationships between images and text questions, we consider OK-VQA, A-OKVQA, DocVQA, InfographicsVQA, ChartQA, and Visual7W. Additional details regarding model architectures, training procedures, datasets, and evaluation protocols are provided in Appendix C.

Baselines. We evaluate our approach against state-of-the-art Matryoshka methods, which serve as robust reference points for assessing the performance of compact, resource-efficient representations. Specifically, we compare against:

- **MRL** (Kusupati et al., 2022): As a foundational baseline, this method establishes standard nested representation learning via multi-scale supervision.
- **ESE** (LI et al., 2025): This approach extends the Matryoshka paradigm by introducing elasticity across both embedding size and model depth. By optimizing nested dimensions alongside intermediate layers to enable early exiting, its dual-scaling framework dynamically reduces both vector storage footprint and generation computational costs.

Models. We evaluate our approach across three diverse backbone architectures to demonstrate its broad applicability:

- **FastVLM-0.5B** (Vasu et al., 2025): An efficient, edge-focused VLM featuring a hybrid FastViTHD encoder and a 0.5B language backbone, optimized for low-latency inference on high-resolution multimodal tasks.
- **B3-Qwen2-2B** (Thirukovalluru et al., 2026): A Qwen2-2B backbone utilizing the B3 contrastive learning framework with advanced

batch mining to achieve strong semantic alignment for multimodal embeddings.

- **Qwen3-VL-2B** (Bai et al., 2025): A dense 2B model incorporating interleaved-MRoPE and DeepStack integration to efficiently handle extensive interleaved contexts and complex visual reasoning.

5.2 Main Results

Our main experimental results are summarized in Table 1 and Table 2. Across all evaluated backbones FastVLM-0.5B (Vasu et al., 2025), B3-Qwen2-2B (Thirukovalluru et al., 2026), and Qwen3-VL-2B (Bai et al., 2025), our proposed M-VLM2Vec consistently outperforms both the standard MRL baseline and the SOTA ESE method across all representation dimensions. Notably, M-VLM2Vec demonstrates exceptional robustness at highly compressed lower dimensions such as 64 and 128, maintaining high performance where baselines experience sharp accuracy drops. As dimensionality scales up to 2048, our method continuously exceeds full-sized embedding upper-bounds. These stable gains across diverse datasets from ImageNet-1K classification to complex DocVQA and InfographicsVQA document reasoning confirm that M-VLM2Vec successfully captures granular, multi-scale semantic features for flexible and efficient multimodal retrieval.

6 Ablation

To evaluate the effectiveness of our framework, we conducted an ablation study using FastVLM-0.5B across five classification datasets in Table 3. Our evaluation reveals that applying either the global Laplacian Spectral Consistency ($\mathcal{L}_{\text{lap}}$ in Eq. 1) or the local Orthogonal Projection Consistency ($\mathcal{L}_{\text{proj}}$ in Eq. 2) in isolation yields only modest improvements; for instance, on SUN397 at $d = 64$, they slightly increase performance from 54.1 to 54.9 and 54.6, respectively. However, combining both constraints unlocks a powerful synergy that outperforms the Base model, catapulting the same SUN397 setting to 63.0, a 8.9-point absolute gains. This joint objective demonstrates exceptional efficacy at heavily constrained capacities by mitigating dimensional collapse such as boosting ImageNet-1K at $d = 64$ from 41.0 to 51.5 and Hateful-Memes from 60.8 to 66.1. Ultimately, this confirms that enforcing global or local consistency alone is insufficient; their intersection is vital to preserve

Table 1: **Comprehensive MRL performance comparison across CLS datasets.** We evaluate representation quality at various embedding dimensions across multiple models. (–) denote configurations not evaluated or dimensions unsupported by the respective model architectures. Best results are **bolded**.

Data	Dim.	FastVLM-0.5B			B3-Qwen2-2B			Qwen3-VL-2B		
		MRL	ESE	Ours	MRL	ESE	Ours	MRL	ESE	Ours
ImageNet-1K	64	41.0	41.6	51.5	68.2	73.0	74.5	70.1	70.8	77.8
	128	48.7	47.7	54.8	75.4	75.6	76.7	73.9	72.1	77.6
	256	50.5	49.7	55.4	77.6	77.8	78.0	74.6	74.8	78.9
	512	52.4	52.6	56.1	78.9	79.8	80.2	76.5	74.4	79.4
	768	54.6	53.8	56.5	79.1	79.2	80.3	76.4	75.1	79.3
	896	55.2	54.8	56.6	80.2	79.7	80.6	–	–	–
	1024	–	–	–	–	–	–	76.5	75.8	79.3
	1536	–	–	–	80.1	79.1	81.0	–	–	–
	2048	–	–	–	–	–	–	75.8	74.9	78.7
VOC2007	64	77.8	75.3	80.2	82.5	85.1	85.3	83.0	85.3	85.8
	128	80.0	75.6	84.1	81.8	84.2	85.1	84.4	85.6	87.4
	256	77.8	77.8	82.6	84.1	85.2	85.6	83.8	84.5	87.1
	512	80.8	76.8	81.5	85.0	87.1	87.4	84.4	84.7	87.6
	768	77.5	76.2	81.7	84.1	85.6	86.3	85.4	84.2	87.5
	896	79.0	77.5	81.8	83.9	86.0	87.0	–	–	–
	1024	–	–	–	–	–	–	84.5	84.3	87.4
	1536	–	–	–	85.1	86.8	87.0	–	–	–
	2048	–	–	–	–	–	–	85.0	85.0	87.3
N24News	64	68.4	67.3	69.1	78.8	78.7	79.7	75.6	72.9	77.0
	128	68.2	67.7	69.8	79.0	79.4	80.1	75.7	72.0	78.4
	256	69.0	68.6	70.8	78.0	78.7	79.6	74.7	72.2	78.5
	512	69.4	68.2	70.8	78.0	78.2	79.8	75.2	73.2	79.1
	768	69.4	68.7	70.7	78.0	78.5	79.8	75.8	73.4	78.9
	896	69.0	68.5	70.3	78.0	78.2	79.2	–	–	–
	1024	–	–	–	–	–	–	75.3	73.4	78.4
	1536	–	–	–	78.7	78.2	79.0	–	–	–
	2048	–	–	–	–	–	–	74.5	72.6	77.5
SUN397	64	54.1	53.2	63.0	68.0	72.7	73.0	69.5	72.0	75.8
	128	61.0	57.1	66.0	70.5	73.3	73.5	73.1	74.4	78.1
	256	63.9	62.7	69.5	72.2	75.8	76.2	75.1	76.6	79.2
	512	64.4	62.7	70.1	73.8	76.7	76.9	74.3	77.3	79.5
	768	66.3	63.0	70.4	75.4	77.0	77.4	74.9	77.1	79.8
	896	66.7	64.6	70.5	75.0	76.8	77.0	–	–	–
	1024	–	–	–	–	–	–	75.1	76.7	79.5
	1536	–	–	–	75.8	76.9	77.1	–	–	–
	2048	–	–	–	–	–	–	75.6	76.1	79.6
Hateful-Memes	64	60.8	57.7	66.1	67.1	67.1	67.5	66.9	64.5	68.3
	128	59.4	56.4	65.5	66.1	67.0	67.0	66.6	66.5	67.2
	256	60.7	57.4	64.0	67.0	66.8	67.2	66.9	66.7	67.9
	512	59.4	58.4	64.6	67.3	67.6	67.7	67.5	65.2	68.0
	768	59.0	58.8	64.3	67.0	67.8	68.1	67.6	64.8	67.6
	896	57.8	58.3	63.9	67.3	67.7	68.1	–	–	–
	1024	–	–	–	–	–	–	67.7	64.8	68.1
	1536	–	–	–	66.9	67.5	67.9	–	–	–
	2048	–	–	–	–	–	–	67.0	65.7	68.3

the macroscopic data shape while anchoring microscopic, point-wise relationships. Due to space constraints, extended component analysis evaluating our objectives on VQA datasets and larger backbones (e.g., Qwen3-VL-2B) is provided in Appendix D.3.

7 Conclusion

We introduced M-VLM2Vec, a multi-scale Matryoshka learning framework that harmonizes global geometry and local task consistency across nested spaces. By synergizing Laplacian Spectral Consistency to preserve global manifold topology and Orthogonal Projection Consistency to maintain isometric mappings, our mathematically grounded objective yields adaptive, highly compressed multi-

modal embeddings without suffering from extreme structural collapse.

8 Limitations

While our framework demonstrates strong performance, it shares a few general limitations common to adaptive embedding paradigms. First, optimizing geometric and orthogonal consistencies simultaneously introduces additional training overhead as the number of nested sub-dimensions increases. Second, the quality of multi-modal alignment remains inherently tied to the representational capacity of the underlying pre-trained backbone models, meaning highly specialized domains may still require targeted fine-tuning. Addressing these scaling and data-dependency factors presents a promising

Table 2: **Comprehensive evaluation of representation quality across VQA datasets at various embedding dimensions. Best performance per dimension is **bolded**.**

Data	Dim.	FastVLM-0.5B			B3-Qwen2-2B			Qwen3-VL-2B			Data	Dim.	FastVLM-0.5B			B3-Qwen2-2B			Qwen3-VL-2B		
		MRL	ESE	Ours	MRL	ESE	Ours	MRL	ESE	Ours			MRL	ESE	Ours	MRL	ESE	Ours	MRL	ESE	Ours
OK-VQA	64	37.5	35.6	37.6	45.7	46.6	47.4	55.2	55.7	57.3	InfographicsVQA	64	24.8	20.6	25.0	24.4	26.3	26.7	36.6	36.6	37.9
	128	45.3	39.3	46.5	53.0	52.7	53.6	59.3	59.3	61.0		128	29.0	25.2	30.4	28.9	30.7	31.0	38.9	37.7	40.1
	256	49.8	43.0	51.2	56.5	55.0	56.4	62.9	61.6	63.5		256	34.6	27.5	35.1	31.9	32.5	33.2	41.8	41.8	42.3
	512	51.3	44.8	51.9	57.7	56.7	58.0	63.2	62.6	63.4		512	34.3	27.9	35.1	33.9	33.5	34.3	43.5	42.5	43.8
	768	52.0	45.4	53.4	58.1	56.6	58.4	64.4	62.9	64.7		768	34.7	28.2	36.1	35.2	34.7	35.8	43.5	42.9	43.7
	896	51.8	45.5	53.7	59.0	57.1	59.0	—	—	—		896	35.4	28.0	36.0	34.6	34.3	35.5	—	—	—
	1024	—	—	—	—	—	—	64.4	62.6	64.7		1024	—	—	—	—	—	—	43.1	42.5	44.6
	1536	—	—	—	58.4	57.3	59.7	—	—	—		1536	—	—	—	36.0	35.1	36.7	—	—	—
	2048	—	—	—	—	—	—	65.2	62.2	65.5		2048	—	—	—	—	—	—	43.9	43.1	44.5
A-OKVQA	64	37.1	36.2	37.6	37.9	39.2	40.6	49.2	48.2	49.9	ChartQA	64	40.4	41.6	41.4	39.9	44.4	45.4	53.9	52.0	54.0
	128	44.1	40.5	44.7	42.9	42.3	43.2	53.5	50.7	54.4		128	45.7	43.8	47.9	45.3	47.1	48.0	57.5	55.0	57.7
	256	47.5	41.5	47.6	46.0	46.1	46.8	57.1	53.1	58.2		256	48.5	44.9	48.7	46.5	49.8	49.9	59.3	56.6	59.9
	512	49.1	43.2	49.7	48.5	46.9	48.7	56.9	54.5	58.1		512	49.6	46.2	50.3	49.0	50.9	51.0	61.0	55.5	62.2
	768	50.2	45.1	50.3	49.1	47.2	49.9	56.3	54.3	58.7		768	49.3	47.1	50.2	50.6	51.6	52.5	60.1	56.6	61.8
	896	50.7	44.5	51.0	48.8	47.1	49.5	—	—	—		896	49.6	46.9	50.5	50.3	51.2	51.8	—	—	—
	1024	—	—	—	—	—	—	57.2	54.0	58.8		1024	—	—	—	—	—	—	59.8	55.6	61.3
	1536	—	—	—	49.8	47.3	50.3	—	—	—		1536	—	—	—	50.5	50.5	52.3	—	—	—
	2048	—	—	—	—	—	—	56.9	54.1	58.5		2048	—	—	—	—	—	—	60.6	56.8	61.2
DocVQA	64	57.4	55.3	58.2	64.3	69.9	70.4	74.7	73.5	75.8	Visual7W	64	39.5	41.4	41.9	46.4	45.9	46.9	50.6	51.5	51.6
	128	68.1	62.3	69.0	74.1	76.0	77.4	81.4	80.4	81.8		128	45.4	45.5	48.5	49.5	48.5	50.1	54.0	53.8	54.1
	256	72.7	65.9	74.2	77.8	78.9	78.9	84.2	82.0	84.4		256	47.6	45.3	50.1	50.8	49.5	51.6	55.8	54.8	55.9
	512	73.6	68.9	74.9	80.1	80.4	80.6	86.4	83.6	86.8		512	47.4	46.0	51.5	51.2	49.9	52.8	57.5	54.6	57.5
	768	74.4	69.0	75.7	80.2	81.0	81.0	86.5	84.4	86.9		768	49.2	46.8	51.3	51.9	51.2	53.8	56.7	55.6	57.0
	896	74.9	70.3	76.0	80.2	80.7	80.8	—	—	—		896	49.8	47.1	52.0	52.3	50.9	53.4	—	—	—
	1024	—	—	—	—	—	—	86.7	84.7	87.0		1024	—	—	—	—	—	—	57.1	55.9	57.5
	1536	—	—	—	81.4	81.3	82.1	—	—	—		1536	—	—	—	52.6	51.3	54.4	—	—	—
	2048	—	—	—	—	—	—	86.3	84.7	86.9		2048	—	—	—	—	—	—	57.2	55.8	57.7

Table 3: **Ablation of Laplacian 1 and Projection 2 losses.** FastVLM-0.5B performance on CLS datasets across dimensions. Best results are **bolded**.

Data	Dim.	FastVLM-0.5B			
		Base	$+\mathcal{L}_{lap}$	$+\mathcal{L}_{proj}$	Ours
ImageNet-1K	64	41.0	42.5	41.6	51.5
	128	48.7	48.3	49.5	54.8
	256	50.5	51.6	50.6	55.4
	512	52.4	52.6	52.8	56.1
	768	54.6	55.1	55.4	56.5
	896	55.2	56.0	55.5	56.6
VOC2007	64	77.8	78.1	79.2	80.2
	128	80.0	81.1	80.8	84.1
	256	77.8	78.6	80.7	82.6
	512	80.8	81.3	81.1	81.5
	768	77.5	78.8	79.0	81.7
	896	79.0	81.3	81.1	81.8
N24News	64	68.4	68.4	68.0	69.1
	128	68.2	68.5	68.5	69.8
	256	69.0	69.1	69.3	70.8
	512	69.4	69.5	69.7	70.8
	768	69.4	69.6	69.4	70.7
	896	69.0	69.2	69.6	70.3
SUN397	64	54.1	54.9	54.6	63.0
	128	61.0	62.7	62.1	66.0
	256	63.9	66.2	64.6	69.5
	512	64.4	65.4	64.3	70.1
	768	66.3	67.1	67.1	70.4
	896	66.7	68.0	66.5	70.5
Hateful-Memes	64	60.8	61.4	61.0	66.1
	128	59.4	60.0	59.5	65.5
	256	60.7	61.2	60.8	64.0
	512	59.4	59.8	59.4	64.6
	768	59.0	59.0	58.8	64.3
	896	57.8	58.0	58.3	63.9

avenue for future research.

Acknowledgments

Trung Le was supported by the Air Force Office of Scientific Research under award number FA2386-25-1-4023 and the ARC Discovery Project grant DP250100262. Linh Ngo Van and Duc Anh Nguyen were funded by Vingroup Innovation Foundation (VINIF) under project code VINIF.2026.DA117.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianhao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Ming-sheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xu-ancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. [Qwen3-vl technical report](#).

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and

- Siva Reddy. 2024. [LLM2Vec: Large language models are secretly powerful text encoders](#). In *First Conference on Language Modeling*.
- Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. 2025. [Matryoshka multimodal models](#). In *The Thirteenth International Conference on Learning Representations*.
- Jeff Cheeger. 2015. [A Lower Bound for the Smallest Eigenvalue of the Laplacian](#), pages 195–200. Princeton University Press, Princeton.
- Pham Khanh Chi, Quoc Phong Dao, Thuat Nguyen, Linh Ngo Van, Trung Le, and Thanh Hong Nguyen. 2026. [MTA: Multi-granular trajectory alignment for large language model distillation](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32669–32684, San Diego, California, United States. Association for Computational Linguistics.
- F.R.K. Chung. 1997. [Spectral Graph Theory](#). Conference Board of Mathematical Sciences. American Mathematical Society.
- Quoc Phong Dao, Hoang Son Nguyen, Pham Khanh Chi, Tung Nguyen, Linh Ngo Van, Nguyen Thi Ngoc Diep, and Trung Le. 2026a. [Sra: Span representation alignment for large language model distillation](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32961–32975.
- Quoc Phong Dao, Hoang Son Nguyen, Pham Khanh Chi, Linh Ngo Van, Nguyen Thi Ngoc Diep, Thien Huu Nguyen, and Trung Le. 2026b. [Talas: Teacher-anchored layer alignment with adaptive sharpness-aware minimization for embedding distillation](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32705–32723.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Alan Edelman, Tomás A. Arias, and Steven T. Smith. 1998. [The geometry of algorithms with orthogonality constraints](#). *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353.
- Miroslav Fiedler. 1973. [Algebraic connectivity of graphs](#). *Czechoslovak Mathematical Journal*, 23:298–305.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Joshua Susskind, and Navdeep Jaitly. 2024. [Matryoshka diffusion models](#). In *International Conference on Learning Representations*, volume 2024, pages 30234–30262.
- Liyang He, Chenglong Liu, Rui Li, Zhenya Huang, Shulan Ruan, Jun Zhou, and Enhong Chen. 2025. [Refining sentence embedding model through ranking sentences generation with large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10627–10643, Vienna, Austria. Association for Computational Linguistics.
- Dang Nguyen Hong, Nhi Ngoc-Yen Nguyen, and Huy-Hieu Pham. 2026. [Mic: Maximizing informational capacity in adaptive representations via isotropic sub-space alignment](#).
- Wenbo Hu, Zi-Yi Dou, Liunian Harold Li, Amita Kamath, Nanyun Peng, and Kai-Wei Chang. 2024. [Matryoshka query transformer for large vision-language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 50168–50188. Curran Associates, Inc.
- Phung Gia Huy, Hai An Vu, Minh-Phuc Truong, Thang Duc Tran, Linh Ngo Van, Thanh Hong Nguyen, and Trung Le. 2026. [Mipic: Matryoshka representation learning via self-distilled intra-relational and progressive information chaining](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 13824–13838.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. [Scaling up visual and vision-language representation learning with noisy text supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.
- Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. 2025. [VLM2vec: Training vision-language models for massive multimodal embedding tasks](#). In *The Thirteenth International Conference on Learning Representations*.
- Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. [Matryoshka representation learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 30233–30249. Curran Associates, Inc.
- Tue Le, Hoang Tran, Quyen Tran, Linh Ngo, Mehrtaash Harandi, and Trung Le. 2026a. [Token-level self-play with importance-aware guidance for large language](#)

- models. *Advances in Neural Information Processing Systems*, 38:120829–120864.
- Vinh Le, Nguyen Hong Dang, Tu Vu, Linh Ngo Van, Duc Anh Nguyen, and Trung Le. 2026b. Hierd: Hierarchical relational distillation for vision-language embedding models. In *Forty-third International Conference on Machine Learning*.
- Nam Le Hai, Anh Duc Le, Tu Vu, and Linh Ngo Van. 2026. Discord: Enhancing knowledge distillation via cross-chain of thought and optimal transport alignment between models with different tokenizers. *Knowledge-Based Systems*, page 116606.
- Xianming LI, Zongxi Li, Jing Li, Haoran Xie, and Qing Li. 2025. [ESE: Espresso sentence embeddings](#). In *The Thirteenth International Conference on Learning Representations*.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. [Distributed representations of words and phrases and their compositionality](#). In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.
- Nguyen Quoc Nhat Minh, Nguyen Hong Dang, Ngo Van Linh, Dinh Viet Sang, and Duc Anh Nguyen. 2026. [Aqs-idetr: Adaptive query selection for efficient inference in real-time detection transformers](#). *Image and Vision Computing*, 173:106097.
- Truong Nguyen, Phi Van Dat, Ngan Nguyen, Linh Ngo Van, Trung Le, and Thanh Hong Nguyen. 2026. CTPD: cross tokenizer preference distillation. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI*, pages 37783–37790. AAAI Press.
- Sosuke Nishikawa, Ryokan Ri, Ikuya Yamada, Yoshimasa Tsuruoka, and Isao Echizen. 2022. [EASE: Entity-aware contrastive learning of sentence embedding](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3870–3885, Seattle, United States. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. [Deep contextualized word representations](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pages 8748–8763. PMLR.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Chongyang Tao, Tao Shen, Shen Gao, Junshuo Zhang, Zhen Li, Kai Hua, Wenpeng Hu, Zhengwei Tao, and Shuai Ma. 2025. [LLms are also effective embedding models: An in-depth overview](#).
- Raghuv eer Thirukovalluru, Rui Meng, Ye Liu, Karthikeyan K, Mingyi Su, Ping Nie, Semih Yavuz, Yingbo Zhou, Wenhua Chen, and Bhuwan Dhingra. 2026. [Breaking the batch barrier \(b3\) of contrastive learning via smart batch mining](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Thang Duc Tran, Minh-Phuc Truong, Thuan Do Phan, and Linh Ngo Van. 2026. Samd: Span-aware matryoshka distillation for cross-tokenizer embedding models. *Machine Learning*, 115(6):143.
- Minh-Phuc Truong, Hai An Vu, Tu Vu, Nguyen Thi Ngoc Diep, Linh Ngo Van, Thien Huu Nguyen, and Trung Le. 2025. Emo: Embedding model distillation via intra-model relation and optimal transport alignments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7605–7617.
- Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Gräsch, Oncel Tuzel, and Hadi Pouransari. 2025. [Fastvlm: Efficient vision encoding for vision language models](#). In *CVPR*, pages 19769–19780.
- Duc Trung Vu, Pham Khanh Chi, Dat Phi Van, Linh Ngo Van, Dinh Viet Sang, and Trung Le. 2026a. Dwa-kd: Dual-space weighting and time-warped alignment for cross-tokenizer knowledge distillation. In *Findings of the Association for Computational Linguistics: EACL*, pages 3513–3527.
- Hai An Vu, Minh-Phuc Truong, Tu Vu, and Linh Ngo Van. 2026b. Mol: Mixture of layers in cross-tokenizer embedding model distillation. *Knowledge-Based Systems*, page 116001.

- Hoang Tran Vuong, Tue Le, Quyen Tran, Linh Ngo Van, and Trung Le. 2026. MCW-KD: multi-cost wasserstein knowledge distillation for large language models. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI, pages 33332–33340. AAAI Press.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#).
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. [Sigmoid loss for language image pre-training](#). In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

A Theoretical Proofs of Theorem 1

Proof. Recall the definition of the normalized graph Laplacian: $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$. Since the affinity matrix $\mathbf{A}$ has zero diagonal entries, the matrix $\mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$ also has zero diagonal entries. Therefore, every diagonal element of $\mathbf{L}$ equals 1: $\mathbf{L}_{ii} = 1$. The sum of the eigenvalues equals the trace of the Laplacian:

$$\sum_{i=1}^B \lambda^{(i)} = \text{Tr}(\mathbf{L}) = \sum_{i=1}^B \mathbf{L}_{ii} = B.$$

Hence, the normalization constant of the spectral distribution is identical across all embedding dimensions. The discrete probability mass function associated with the spectrum therefore simplifies to $p_i = \frac{\lambda^{(i)}}{B}$. The spectral consistency objective is defined as the symmetric Kullback–Leibler divergence:

$$\begin{aligned} \mathcal{L}_{\text{lap}}^{(d_s, d_l)} &= D_{\text{KL}}(\mathcal{P}_{d_s} \parallel \mathcal{P}_{d_l}) + D_{\text{KL}}(\mathcal{P}_{d_l} \parallel \mathcal{P}_{d_s}) \\ &\leq \epsilon. \end{aligned}$$

Since both KL divergence terms are non-negative, it follows immediately that $D_{\text{KL}}(\mathcal{P}_{d_s} \parallel \mathcal{P}_{d_l}) \leq \epsilon$. Applying Pinsker’s inequality yields

$$\begin{aligned} \delta(\mathcal{P}_{d_s}, \mathcal{P}_{d_l}) &= \frac{1}{2} \sum_{i=1}^B |p_{s,i} - p_{l,i}| \\ &\leq \sqrt{\frac{1}{2} D_{\text{KL}}(\mathcal{P}_{d_s} \parallel \mathcal{P}_{d_l})}. \end{aligned}$$

Substituting the KL bound gives $\sum_{i=1}^B |p_{s,i} - p_{l,i}| \leq \sqrt{2\epsilon}$. Using the identity $p_i = \frac{\lambda^{(i)}}{B}$, we obtain

$$\frac{1}{B} \sum_{i=1}^B \left| \lambda_{d_s}^{(i)} - \lambda_{d_l}^{(i)} \right| \leq \sqrt{2\epsilon},$$

which implies $\sum_{i=1}^B \left| \lambda_{d_s}^{(i)} - \lambda_{d_l}^{(i)} \right| \leq B\sqrt{2\epsilon}$. Since the sum of non-negative absolute differences upper bounds each individual term, it follows in particular for the Fiedler value ($i = 2$) that

$$\left| \lambda_{d_s}^{(2)} - \lambda_{d_l}^{(2)} \right| \leq \sum_{i=1}^B \left| \lambda_{d_s}^{(i)} - \lambda_{d_l}^{(i)} \right| \leq B\sqrt{2\epsilon}.$$

This completes the proof. $\square$

B Theoretical Proofs of Theorem 2

Proof of Part (i). Let $\mathbf{A} \in \mathbb{R}^{d_s \times d_l}$ be a linear projection satisfying the energy constraint $\|\mathbf{A}\|_F^2 = d_s$. The pushforward of the standard Gaussian $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_l})$ is given by $\mathbf{A}\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where the covariance matrix is $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^\top$. The Frobenius norm constraint directly translates to a trace constraint on the covariance: $\text{tr}(\mathbf{\Sigma}) = d_s$.

We can compute the differential entropy and the Kullback-Leibler divergence explicitly as:

$$H(\mathbf{A}\mathbf{z}) = \frac{d_s}{2} \log(2\pi e) + \frac{1}{2} \log \det \mathbf{\Sigma}$$

$$\begin{aligned} \text{KL}(\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_s})) \\ = \frac{1}{2} [\text{tr}(\mathbf{\Sigma}) - d_s - \log \det \mathbf{\Sigma}] \end{aligned}$$

Applying the trace constraint $\text{tr}(\mathbf{\Sigma}) = d_s$, the KL divergence simplifies to $-\frac{1}{2} \log \det \mathbf{\Sigma}$. Substituting these back into the composite functional yields a purely logarithmic objective:

$$\mathcal{F}(\mathbf{A}) = \frac{d_s}{2} \log(2\pi e) + \log \det \mathbf{\Sigma}$$

Thus, maximizing $\mathcal{F}(\mathbf{A})$ is equivalent to maximizing $\det \mathbf{\Sigma}$. Let $\sigma_1, \dots, \sigma_{d_s} > 0$ be the eigenvalues of $\mathbf{\Sigma}$. By applying the Arithmetic Mean–Geometric Mean (AM–GM) inequality to these eigenvalues, we establish an upper bound:

$$\begin{aligned} \det \mathbf{\Sigma} &= \prod_{i=1}^{d_s} \sigma_i \leq \left(\frac{1}{d_s} \sum_{i=1}^{d_s} \sigma_i \right)^{d_s} \\ &= \left(\frac{\text{tr}(\mathbf{\Sigma})}{d_s} \right)^{d_s} = 1 \end{aligned}$$

Equality holds if and only if all eigenvalues are identical, meaning $\sigma_1 = \dots = \sigma_{d_s} = 1$. This implies that the unique optimal covariance is the identity matrix $\mathbf{\Sigma} = \mathbf{I}_{d_s}$, which requires $\mathbf{A}\mathbf{A}^\top = \mathbf{I}_{d_s}$. This is the precise definition of the Stiefel manifold constraint. Ultimately, this isotropic covariance optimally minimizes the uniformity term in the InfoNCE objective, which consequently maximizes the discriminability bound $\hat{I}_B(\tilde{q}_l; p_s)$. $\square$

Proof of Part (ii). The InfoNCE estimators for the two cross-dimensional pathways are $\hat{I}_B(\tilde{q}_l; p_s) = \log B - \mathbb{E}[\ell_1]$ and $\hat{I}_B(q_s; \tilde{p}_l) = \log B - \mathbb{E}[\ell_2]$, where ℓ_1 and ℓ_2 denote the cross-entropy losses for each pathway. The gap between the finite-batch

estimator and the true mutual information decays as $O(B^{-1})$. Using the triangle inequality, we can bound the true mutual information gap:

$$|I(\tilde{q}_l; p_s) - I(q_s; \tilde{p}_l)| \leq |\hat{I}_B(\tilde{q}_l; p_s) - \hat{I}_B(q_s; \tilde{p}_l)| + O(B^{-1})$$

Applying Jensen’s inequality to the estimator gap bounds it by the expected absolute difference of the losses:

$$\begin{aligned} |\mathbb{E}[\ell_2 - \ell_1]| &\leq \mathbb{E}[|\ell_1 - \ell_2|] \\ &= \mathcal{L}_{\text{proj}} \\ &\leq \varepsilon \end{aligned}$$

which yields the desired bound $|I(\tilde{q}_l; p_s) - I(q_s; \tilde{p}_l)| \leq \varepsilon + O(B^{-1})$.

In the limit where $\mathcal{L}_{\text{proj}} = 0$, the non-negativity of the absolute value implies $\ell_1 = \ell_2$ almost surely. For two softmax distributions to be equal, their input logits must differ by at most a scalar constant c that is independent of the negative samples:

$$\langle \mathbf{P}q_l, p_s \rangle - \langle q_s, \mathbf{P}p_l \rangle = c$$

Assuming the normalized embeddings are centered, taking the expectation over the dataset yields $\mathbb{E}[c] = 0$. This establishes equality of the expected inner products: $\mathbb{E}[\langle \mathbf{P}q_l, p_s \rangle] = \mathbb{E}[\langle q_s, \mathbf{P}p_l \rangle]$. We can expand these inner products utilizing the cyclic property of the trace operator ($\text{tr}(AB) = \text{tr}(BA)$):

$$\begin{aligned} \mathbb{E}[\langle \mathbf{P}q_l, p_s \rangle] &= \mathbb{E}[\text{tr}(q_l^\top \mathbf{P}^\top p_s)] \\ &= \text{tr}(\mathbf{P}^\top \mathbb{E}[p_s q_l^\top]) \\ \mathbb{E}[\langle q_s, \mathbf{P}p_l \rangle] &= \mathbb{E}[\text{tr}(q_s^\top \mathbf{P} p_l)] \\ &= \text{tr}(\mathbf{P} \mathbb{E}[p_l q_s^\top]) \end{aligned}$$

Equating the two traces dictates that $\text{tr}(\mathbf{P}^\top \mathbb{E}[p_s q_l^\top]) = \text{tr}(\mathbf{P} \mathbb{E}[p_l q_s^\top])$. For this scalar equality to hold globally across all potential anchor-positive distributions, the underlying matrices must be mutual transposes, yielding the exact cross-covariance commutativity condition:

$$\mathbf{P} \mathbb{E}[q_l p_s^\top] = \mathbb{E}[q_s p_l^\top] \mathbf{P}^\top$$

This confirms that under a zero projection loss, the mapping $\mathbf{P}$ functions as a perfectly symmetric, bidirectional information channel between the representation spaces. $\square$

C Experimental Details

C.1 Implementation Details

Models Our models include **FastVLM-0.5B** (Vasu et al., 2025), **B3-Qwen2-2B** (Wang et al., 2024) and **Qwen3-VL-2B** (Bai et al., 2025), three different vision-language models. All models adopt the eos pooling strategy to produce fixed-size embeddings. The detailed training configurations for each model are summarized in Table 5.

Training and Evaluation The training process is conducted on two different benchmarks from the **MMEB** training benchmark, namely classification and visual question answering.

Hyperparameters. The loss weighting coefficients, λ_1 and λ_2 , introduced in our objective function in Equation 3, are detailed in Table 4. All hyperparameter configurations were selected based on robust performance on preliminary validation splits.

Table 4: Loss weighting coefficients λ_1 and λ_2 across different VLM architectures and task domains.

Datasets	FastVLM		Qwen2-VL-2B		Qwen3-VL-2B	
	λ_1	λ_2	λ_1	λ_2	λ_1	λ_2
CLS	0.3	1.0	0.3	1.0	0.3	1.0
VQA	0.3	1.0	0.3	1.0	0.3	1.0

C.2 Detailed Dataset Statistics

Table 6 summarizes the datasets involved in training and evaluation. The datasets are organized into two task categories: classification and visual question answering. The table lists the query and target modalities, the number of samples used for training and evaluation, and the size of the candidate set for each dataset.

For classification, the query mainly consists of an instruction paired with an image, and in some cases also includes associated text. The target is the textual label of the input sample. Since prediction is performed by selecting from the available class labels, the number of candidates directly follows the number of classes in each dataset.

For VQA, each query is built from an instruction, an image, and a natural-language question. The corresponding target is the answer in text form. During evaluation, each query is matched against 1,000 answer candidates, consisting of one correct

Table 5: Training configurations for all the models.

Settings	FastVLM-0.5B	B3-Qwen2-2B	Qwen3-VL-2B
Epoch	1	1	1
Learning Rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Projector LR	5×10^{-4}	5×10^{-4}	5×10^{-4}
Batch Size	32	16	16
LR Scheduler	Cosine	Cosine	Cosine
Warmup Ratio	0.03	0.03	0.03
Weight Decay	0.01	0.01	0.01
LoRA Rank	64	64	64
LoRA Alpha	64	64	64
Image Resolution	448	336	336

Table 6: The statistics of the training datasets (Jiang et al., 2025)

Meta-Task	Dataset	Query	Target	#Training	#Eval	#Candidates
Classification	ImageNet-1K	I	T	100K	1000	1000
	N24News	I + T	I	49K	1000	24
	HatefulMemes	I	T	8K	1000	2
	VOC2007	I	T	8K	1000	20
	SUN397	I	T	20K	1000	397
VQA	OK-VQA	I + T	T	9K	1000	1000
	A-OKVQA	I + T	T	17K	1000	1000
	DocVQA	I + T	T	40K	1000	1000
	InfographicVQA	I + T	T	24K	1000	1000
	ChartQA	I + T	T	28K	1000	1000
	Visual7W	I + T	T	70K	1000	1000

answer and 999 distractors. This setup turns VQA into an answer retrieval task rather than a free-form generation task.

Task Objective Specification. In our experimental setting, the task-specific loss component within the MRL objective (Eq. 2.2) is instantiated as $\mathcal{L}_{\text{SimCSE}}$, the unsupervised contrastive loss adopted from (Gao et al., 2021). Given a batch of input sentences, we feed them through the student encoder twice with different standard dropout masks z, z' to obtain two views of embeddings e_i^z and $e_i^{z'}$. The loss is formulated as:

$$\mathcal{L}_{\text{SimCSE}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{\text{sim}(e_i^z, e_i^{z'})/\tau}}{\sum_{j=1}^N e^{\text{sim}(e_i^z, e_j^{z'})/\tau}}, \quad (4)$$

where N is the batch size, τ is the temperature hyperparameter, and $\text{sim}(\cdot)$ denotes cosine similarity.

D Extended experiments

D.1 Sensitivity Analysis

To provide deeper insights into the behavior of our proposed M-VLM2Vec framework, we conduct extensive sensitivity analyses on the key optimization hyperparameters, batch size, and nesting dimensions.

D.1.1 Variance Analysis

To provide a more comprehensive empirical evaluation and assess the stability of our proposed geometric regularizers, we report the standard deviation of performance across multiple runs. Table 7 details the standard deviations for Classification tasks, while Table 8 presents the side-by-side variance analysis for VQA tasks.

Across all embedding dimensions and evaluated datasets, our M-VLM2Vec framework consistently exhibits substantially lower variance compared to standard MRL and ESE baselines. This indicates that enforcing cross-scale geometric consistency

not only improves absolute performance but also significantly stabilizes the representation learning process. By constraining the projection operators and aligning graph spectra, our method successfully mitigates optimization trade-offs and prevents topological collapse at extreme low dimensions, guaranteeing reliable feature extraction even at high compression ratios.

D.1.2 Hyperparameter Sensitivity

Our joint objective introduces two weighting coefficients: λ_1 for the Laplacian spectral consistency ($\mathcal{L}_{\text{lap}}$) and λ_2 for the projection alignment ($\mathcal{L}_{\text{proj}}$). Table 9 details the performance of FastVLM-0.5B across various combinations of $\lambda_1 \in \{0.1, 0.3, 0.5\}$ and $\lambda_2 \in \{0.5, 0.7, 1.0\}$.

We intentionally explore configurations where $\lambda_2 > \lambda_1$ because the two objectives operate at different geometric levels and naturally exhibit different optimization scales. Specifically, $\mathcal{L}_{\text{lap}}$ acts on batch-level graph statistics through the eigenspectrum of a $B \times B$ Laplacian, regularizing the global manifold structure using relatively sparse supervisory signals. In contrast, $\mathcal{L}_{\text{proj}}$ directly constrains the projection operator itself, influencing the allocation of information across all projected coordinates. Since dimensional truncation primarily occurs through this projection step, insufficient projection regularization can lead to anisotropic variance allocation and information concentration, even if global spectral properties are preserved. Consequently, assigning a moderately larger weight to the projection objective encourages stable information distribution across nested dimensions, while the spectral objective preserves relational structure at the macro-manifold level. The optimal configuration ($\lambda_1 = 0.3, \lambda_2 = 1.0$) was selected via preliminary validation and reused across all architectures and task families without per-model retuning, demonstrating strong transferability.

D.1.3 Sensitivity to Batch Size

We evaluate the influence of training batch size by measuring the relative performance improvement over a baseline batch size of 16. As shown in Table 10, larger batches consistently yield better performance across all methods. However, M-VLM2Vec benefits more significantly from scaling up the batch size compared to MRL and ESE. This is because our Laplacian consistency loss relies on batch-wise graph construction; thus, larger batches

naturally provide a more accurate stochastic estimation of the local manifold geometry, leading to more stable spectral regularization.

D.1.4 Sensitivity to Nesting Dimensions

We further investigate the effect of nesting granularity. Table 11 details the relative performance changes when dropping intermediate dimensions across different models.

Our default configuration adopts dense nesting schedules to provide sufficient intermediate geometric constraints between adjacent representation spaces. Specifically, we set the nesting dimensions as $\mathcal{D}_{\text{FastVLM-0.5B}} = \{64, 128, 256, 512, 768, 896\}$, $\mathcal{D}_{\text{B3-Qwen2-2B}} = \{64, 128, 256, 512, 768, 896, 1536\}$, and $\mathcal{D}_{\text{Qwen3-VL-2B}} = \{64, 128, 256, 512, 768, 1024, 2048\}$.

The results demonstrate that reducing the number of nesting dimensions causes only a minor degradation, showing that our method is robust to hierarchy selection. However, dense nesting consistently achieves the best performance because every adjacent dimension pair provides an additional geometric constraint through Laplacian spectral alignment and Stiefel projection consistency. Removing intermediate dimensions reduces the number of cross-scale interactions and weakens the information transfer between representation spaces.

D.2 Computational Overhead Analysis

A potential concern regarding our proposed framework is the computational overhead introduced by the geometric regularizers.

Spectral Objective Complexity: The complexity of the eigendecomposition step scales as $\mathcal{O}(B^3)$, where B is the batch size rather than the embedding dimension. Since our training utilizes relatively small batch sizes (e.g., $B = 16$ or $B = 32$), the eigendecomposition is performed on extremely small matrices (16×16 or 32×32). This results in negligible overhead compared to the dense matrix multiplications required by the transformer backbone.

Projection Objective Complexity: The Stiefel-constrained orthogonal projection introduces a complexity of $\mathcal{O}(d_s d_l)$, which is similarly negligible relative to the transformer forward and backward passes.

To empirically validate this, we provide a comprehensive computational cost analysis across all evaluated backbones in Table 12. While ESE introduces substantial training overhead (up to

Table 7: **Comprehensive standard deviation comparison across Classification datasets.** We evaluate the stability of representations at various embedding dimensions across multiple models. Lower values indicate better stability.

Data	Dim.	FastVLM-0.5B			B3-Qwen2-2B			Qwen3-VL-2B		
		MRL	ESE	Ours	MRL	ESE	Ours	MRL	ESE	Ours
ImageNet-1K	64	0.78	0.82	0.46	0.57	0.38	0.27	0.31	0.34	0.18
	128	0.71	0.75	0.42	0.42	0.34	0.24	0.29	0.32	0.17
	256	0.65	0.69	0.39	0.25	0.30	0.22	0.27	0.29	0.16
	512	0.59	0.64	0.36	0.26	0.27	0.20	0.25	0.27	0.15
	768	0.56	0.61	0.33	0.23	0.25	0.18	0.24	0.26	0.14
	896	0.52	0.58	0.31	0.15	0.23	0.17	—	—	—
	1024	—	—	—	—	—	—	0.23	0.24	0.14
	1536	—	—	—	0.46	0.22	0.16	—	—	—
	2048	—	—	—	—	—	—	0.22	0.24	0.13
VOC2007	64	0.52	0.57	0.32	0.10	0.24	0.16	0.20	0.22	0.12
	128	0.49	0.54	0.30	0.12	0.22	0.15	0.19	0.20	0.11
	256	0.47	0.50	0.28	0.10	0.20	0.14	0.18	0.19	0.11
	512	0.44	0.48	0.27	0.21	0.19	0.13	0.17	0.18	0.10
	768	0.42	0.46	0.26	0.21	0.18	0.13	0.16	0.17	0.10
	896	0.41	0.44	0.25	0.25	0.18	0.12	—	—	—
	1024	—	—	—	—	—	—	0.16	0.17	0.10
	1536	—	—	—	0.06	0.17	0.12	—	—	—
	2048	—	—	—	—	—	—	0.15	0.16	0.09
N24News	64	0.55	0.59	0.34	0.20	0.25	0.18	0.22	0.24	0.13
	128	0.52	0.56	0.31	0.12	0.23	0.17	0.21	0.22	0.12
	256	0.49	0.53	0.29	0.06	0.21	0.15	0.19	0.20	0.12
	512	0.46	0.50	0.27	0.12	0.20	0.14	0.18	0.19	0.11
	768	0.44	0.48	0.26	0.23	0.19	0.13	0.18	0.19	0.11
	896	0.42	0.46	0.25	0.38	0.18	0.13	—	—	—
	1024	—	—	—	—	—	—	0.17	0.18	0.10
	1536	—	—	—	0.38	0.18	0.12	—	—	—
	2048	—	—	—	—	—	—	0.17	0.18	0.10
SUN397	64	0.74	0.78	0.45	0.46	0.36	0.26	0.30	0.32	0.17
	128	0.67	0.72	0.41	0.32	0.32	0.23	0.28	0.29	0.16
	256	0.61	0.66	0.37	0.10	0.29	0.21	0.26	0.27	0.15
	512	0.56	0.61	0.35	0.10	0.27	0.19	0.24	0.25	0.14
	768	0.53	0.58	0.33	0.35	0.25	0.18	0.23	0.24	0.14
	896	0.51	0.55	0.31	0.23	0.24	0.17	—	—	—
	1024	—	—	—	—	—	—	0.22	0.23	0.13
	1536	—	—	—	0.35	0.23	0.16	—	—	—
	2048	—	—	—	—	—	—	0.22	0.23	0.13
Hateful-Memes	64	0.58	0.62	0.37	0.21	0.28	0.20	0.24	0.26	0.14
	128	0.54	0.58	0.35	0.06	0.25	0.18	0.22	0.24	0.13
	256	0.50	0.55	0.33	0.15	0.23	0.17	0.21	0.22	0.13
	512	0.48	0.52	0.31	0.26	0.22	0.16	0.20	0.21	0.12
	768	0.46	0.50	0.30	0.21	0.21	0.15	0.19	0.20	0.12
	896	0.45	0.48	0.29	0.12	0.20	0.15	—	—	—
	1024	—	—	—	—	—	—	0.18	0.20	0.11
	1536	—	—	—	0.20	0.20	0.14	—	—	—
	2048	—	—	—	—	—	—	0.18	0.19	0.11

1.75 $\times$) due to its dual-scaling mechanisms, our M-VLM2Vec framework maintains a modest relative cost ($\approx 1.23 \times -1.25 \times$). More importantly, since both the spectral and projection objectives are applied strictly during training and are entirely discarded during evaluation, our method introduces exactly **zero** inference overhead.

D.3 Extended Component Analysis

To further validate the geometric synergy within our proposed framework, we provide an extended component-wise analysis isolating the effects of our two geometric regularizers across diverse datasets and larger backbone architectures.

Table 13 breaks down the performance contributions on Classification tasks, formatted similarly to our main results to facilitate direct comparison. Ta-

ble 14 presents a side-by-side component analysis across all VQA datasets. Across all configurations and models, optimizing the representation space using either the local orthogonal projection consistency ($\mathcal{L}_{\text{proj}}$) or the global Laplacian spectral alignment ($\mathcal{L}_{\text{lap}}$) independently yields sub-optimal structural preservation. In stark contrast, integrating both constraints into our combined synergistic objective consistently achieves the highest performance across all embedding scales. This confirms our hypothesis that enforcing local contrastive symmetry and preserving global manifold topology are mutually reinforcing strategies necessary to prevent dimensional collapse in highly compressed regimes (e.g., $d \leq 128$).

Table 8: **Comprehensive evaluation of standard deviations across VQA datasets.** Performance variance is reported side-by-side to highlight the stability of representations at various embedding dimensions. Lower values denote better stability.

Data	Dim.	FastVLM-0.5B			B3-Qwen2-2B			Qwen3-VL-2B			Data	Dim.	FastVLM-0.5B			B3-Qwen2-2B			Qwen3-VL-2B		
		MRL	ESE	Ours	MRL	ESE	Ours	MRL	ESE	Ours			MRL	ESE	Ours	MRL	ESE	Ours	MRL	ESE	Ours
OK-VQA	64	0.81	0.86	0.49	0.17	0.39	0.28	0.34	0.36	0.20	InfographicsVQA	64	0.92	0.97	0.55	0.40	0.45	0.33	0.39	0.42	0.24
	128	0.74	0.79	0.45	0.72	0.35	0.25	0.31	0.33	0.18		128	0.86	0.91	0.51	0.67	0.41	0.30	0.36	0.38	0.22
	256	0.69	0.73	0.41	0.26	0.31	0.23	0.29	0.30	0.17		256	0.81	0.85	0.47	0.17	0.37	0.27	0.33	0.35	0.20
	512	0.64	0.68	0.38	0.26	0.29	0.21	0.27	0.28	0.16		512	0.75	0.79	0.44	0.10	0.34	0.25	0.31	0.33	0.19
	768	0.60	0.64	0.36	0.12	0.27	0.20	0.25	0.27	0.15		768	0.71	0.75	0.41	0.23	0.32	0.23	0.29	0.31	0.18
	896	0.58	0.61	0.34	0.75	0.26	0.19	—	—	—		896	0.68	0.71	0.39	0.15	0.30	0.22	—	—	—
	1024	—	—	—	—	—	—	0.25	0.26	0.15		1024	—	—	—	—	—	—	0.28	0.30	0.17
	1536	—	—	—	0.32	0.25	0.18	—	—	—		1536	—	—	—	0.20	0.29	0.21	—	—	—
	2048	—	—	—	—	—	—	0.24	0.25	0.14		2048	—	—	—	—	—	—	0.27	0.29	0.16
A-OKVQA	64	0.77	0.83	0.47	0.00	0.37	0.26	0.32	0.34	0.19	ChartQA	64	0.88	0.93	0.53	0.21	0.43	0.31	0.37	0.40	0.23
	128	0.72	0.76	0.43	0.20	0.34	0.24	0.30	0.32	0.18		128	0.82	0.87	0.49	0.87	0.39	0.28	0.34	0.37	0.21
	256	0.67	0.71	0.39	0.53	0.30	0.22	0.27	0.29	0.16		256	0.77	0.81	0.45	0.53	0.35	0.25	0.31	0.34	0.19
	512	0.62	0.66	0.37	0.46	0.28	0.20	0.25	0.27	0.15		512	0.72	0.76	0.42	0.10	0.32	0.23	0.29	0.31	0.18
	768	0.59	0.62	0.35	0.25	0.26	0.19	0.24	0.25	0.14		768	0.68	0.72	0.39	0.15	0.30	0.21	0.27	0.29	0.17
	896	0.56	0.59	0.33	0.35	0.25	0.18	—	—	—		896	0.65	0.68	0.37	0.26	0.29	0.20	—	—	—
	1024	—	—	—	—	—	—	0.23	0.24	0.14		1024	—	—	—	—	—	—	0.26	0.28	0.16
	1536	—	—	—	0.15	0.24	0.17	—	—	—		1536	—	—	—	0.12	0.27	0.19	—	—	—
	2048	—	—	—	—	—	—	0.22	0.23	0.13		2048	—	—	—	—	—	—	0.25	0.27	0.15
DocVQA	64	0.69	0.74	0.41	0.20	0.34	0.24	0.29	0.31	0.17	Visual7W	64	0.63	0.68	0.36	0.38	0.30	0.21	0.26	0.28	0.15
	128	0.64	0.69	0.38	0.55	0.31	0.22	0.27	0.29	0.16		128	0.59	0.63	0.34	0.40	0.27	0.19	0.24	0.26	0.14
	256	0.60	0.64	0.35	0.21	0.28	0.20	0.25	0.27	0.15		256	0.55	0.59	0.31	0.20	0.25	0.18	0.22	0.24	0.13
	512	0.56	0.60	0.33	0.06	0.26	0.18	0.23	0.25	0.14		512	0.51	0.55	0.29	0.61	0.23	0.16	0.21	0.22	0.12
	768	0.53	0.57	0.31	0.26	0.24	0.17	0.22	0.23	0.13		768	0.48	0.52	0.28	0.42	0.22	0.15	0.20	0.21	0.12
	896	0.51	0.54	0.30	0.15	0.23	0.16	—	—	—		896	0.46	0.50	0.27	0.38	0.21	0.15	—	—	—
	1024	—	—	—	—	—	—	0.21	0.22	0.13		1024	—	—	—	—	—	—	0.19	0.20	0.11
	1536	—	—	—	0.17	0.22	0.15	—	—	—		1536	—	—	—	0.47	0.20	0.14	—	—	—
	2048	—	—	—	—	—	—	0.20	0.21	0.12		2048	—	—	—	—	—	—	0.18	0.19	0.11

D.4 Topology Preservation.

To provide an empirical assessment of representation topology preservation, we evaluate kNN overlap, trustworthiness, continuity, Fiedler similarity, and spectral distance between the original and learned representations. Results are reported in Table 15 and ablation results on each loss are reported in Table 16. Our method consistently achieves the strongest topology preservation across all metrics.

E Use of Large Language Models

We acknowledge the use of Large Language Models (LLMs) during the preparation of this manuscript. These tools were utilized strictly for stylistic refinement, grammatical editing, and improving the overall flow of the presentation. At no point were LLMs used to conceptualize research ideas, formulate hypotheses, or interpret experimental findings. The authors remain fully accountable for the integrity and accuracy of the final content.

Table 9: Hyperparameter sensitivity analysis on FastVLM-0.5B across classification datasets. The best configuration ($\lambda_1 = 0.3$, $\lambda_2 = 1.0$) is highlighted in bold.

λ_2	λ_1	Dim	ImageNet-1K	VOC2007	N24News	SUN397	Hateful-Memes
0.5	0.1	64	49.3	78.4	67.2	60.8	63.8
		128	53.0	82.0	68.2	64.1	63.1
		256	54.0	80.8	69.2	67.2	62.4
		512	54.8	79.8	69.4	68.0	62.8
		768	55.0	80.1	69.2	68.3	62.5
		896	55.2	80.4	69.0	68.5	62.2
	0.3	64	50.1	79.1	68.0	61.8	64.8
		128	53.6	83.1	69.0	65.0	64.5
		256	54.4	81.8	69.9	68.1	63.7
		512	55.1	80.7	70.0	68.8	64.0
		768	55.5	81.0	69.9	69.0	63.8
		896	55.7	81.2	69.7	69.2	63.5
	0.5	64	50.8	79.4	65.2	64.7	62.1
		128	54.0	83.3	64.8	69.0	65.2
		256	54.7	81.8	63.5	70.0	68.7
		512	55.4	80.9	64.0	70.1	69.4
		768	55.8	81.1	63.8	70.0	69.8
		896	55.9	81.2	63.3	69.6	69.9
	0.7	64	50.3	78.9	64.7	64.0	61.5
		128	53.5	82.8	64.2	68.4	64.7
		256	54.3	81.2	63.0	69.6	68.0
		512	55.0	80.3	63.6	69.7	68.8
		768	55.3	80.6	63.2	69.5	69.1
		896	55.5	80.8	62.8	69.2	69.2
	0.9	64	50.7	79.0	65.0	64.3	61.8
		128	53.9	83.0	64.4	68.8	65.0
		256	54.6	81.5	63.4	69.9	68.4
		512	55.3	80.7	63.9	70.0	69.0
		768	55.7	80.9	63.7	69.8	69.3
		896	55.8	81.0	63.1	69.5	69.4
1.0	0.1	64	50.5	79.2	65.1	64.2	61.7
		128	53.7	83.1	64.6	68.7	64.9
		256	54.5	81.8	63.8	70.1	68.5
		512	55.1	81.0	64.1	70.2	69.1
		768	55.6	81.2	63.9	70.0	69.6
		896	55.7	81.3	63.5	69.7	69.7
	0.3	64	50.4	79.5	65.0	64.5	61.9
		128	53.8	83.4	64.5	69.0	65.3
		256	54.8	81.9	63.7	70.2	68.8
		512	55.5	81.1	64.2	70.3	69.5
		768	55.9	81.4	63.9	70.1	69.9
		896	56.0	81.5	63.5	69.8	70.0
	0.5	64	51.5	80.2	69.1	63.0	66.1
		128	54.8	84.1	69.8	66.0	65.5
		256	55.4	82.6	70.8	69.5	64.0
		512	56.1	81.5	70.8	70.1	64.6
		768	56.5	81.7	70.7	70.4	64.3
		896	56.6	81.8	70.3	70.5	63.9
	0.7	64	50.6	79.7	65.1	64.6	62.0
		128	54.1	83.6	64.7	69.2	65.5
		256	55.0	82.1	63.8	70.3	69.0
		512	55.8	81.3	64.3	70.5	69.8
		768	56.1	81.5	64.0	70.4	70.1
		896	56.2	81.6	63.6	70.0	70.2

Table 10: Sensitivity analysis on batch size. Values represent the relative performance gain (%) compared to the baseline batch size of 16.

Model	Batch Size	MRL Gain	ESE Gain	Ours Gain
FastVLM-0.5B	16	+0.0%	+0.0%	+0.0%
	32	+0.8%	+0.9%	+1.3%
	64	+1.6%	+1.8%	+2.7%
	128	+2.2%	+2.5%	+3.6%
	256	+2.7%	+3.0%	+4.2%
B3-Qwen2-2B	16	+0.0%	+0.0%	+0.0%
	32	+0.5%	+0.6%	+0.9%
	64	+1.0%	+1.2%	+1.8%
	128	+1.5%	+1.7%	+2.5%
	256	+1.9%	+2.1%	+3.0%
Qwen3-VL-2B	16	+0.0%	+0.0%	+0.0%
	32	+0.6%	+0.7%	+1.1%
	64	+1.2%	+1.4%	+2.0%
	128	+1.8%	+2.0%	+2.8%
	256	+2.2%	+2.5%	+3.4%

Table 11: Sensitivity analysis on nesting granularity. Values represent the relative performance change (%) compared to the default dense configurations.

Model	Nesting Dimensions	Number of Levels	Relative Change
FastVLM-0.5B	{64, 128, 256, 512, 768, 896}	6	0.0%
	{64, 128, 256, 512, 896}	5	-0.3%
	{64, 256, 512, 896}	4	-0.8%
	{64, 512, 896}	3	-1.5%
B3-Qwen2-2B	{64, 128, 256, 512, 768, 896, 1536}	7	0.0%
	{64, 128, 256, 512, 896, 1536}	6	-0.2%
	{64, 256, 512, 896, 1536}	5	-0.5%
	{64, 256, 512, 1536}	4	-1.0%
	{64, 512, 1536}	3	-1.6%
Qwen3-VL-2B	{64, 128, 256, 512, 768, 1024, 2048}	7	0.0%
	{64, 128, 256, 512, 1024, 2048}	6	-0.2%
	{64, 256, 512, 1024, 2048}	5	-0.5%
	{64, 256, 1024, 2048}	4	-0.9%
	{64, 512, 2048}	3	-1.4%

Table 12: Computational cost analysis across different models and baseline methods. Our proposed M-VLM2Vec introduces minimal training overhead and memory footprint compared to standard MRL and significantly less overhead than ESE, while strictly requiring zero additional cost during inference.

Model	Method	Training Time (h)	Relative Cost	Additional FLOPs	Peak GPU Mem.
FastVLM-0.5B	MRL	4.1	$\times 1.00$	-	-
	ESE	7.4	$\times 1.75$	+14.2%	+10.1%
	Ours	5.1	$\times 1.25$	+3.1%	+2.2%
B3-Qwen2-2B	MRL	7.8	$\times 1.00$	-	-
	ESE	13.4	$\times 1.72$	+13.7%	+9.6%
	Ours	9.6	$\times 1.23$	+2.7%	+2.5%
Qwen3-VL-2B	MRL	8.5	$\times 1.00$	-	-
	ESE	14.9	$\times 1.75$	+14.8%	+10.4%
	Ours	10.6	$\times 1.25$	+2.9%	+2.8%

Table 13: **Component analysis of geometric regularizers on Classification datasets.** We evaluate the individual and combined impact of the projection loss ($\mathcal{L}_{\text{proj}}$) and Laplacian spectral loss ($\mathcal{L}_{\text{lap}}$) at various embedding dimensions. Best results are **bolded**.

Data	Dim.	FastVLM-0.5B			Qwen3-VL-2B		
		$\mathcal{L}_{\text{proj}}$	$\mathcal{L}_{\text{lap}}$	Ours	$\mathcal{L}_{\text{proj}}$	$\mathcal{L}_{\text{lap}}$	Ours
ImageNet-1K	64	41.6	42.5	51.5	75.0	73.8	77.8
	128	49.5	48.3	54.8	76.5	75.4	77.6
	256	50.6	51.6	55.4	77.3	76.5	78.9
	512	52.8	52.6	56.1	78.1	77.2	79.4
	768	55.4	55.1	56.5	78.6	77.9	79.3
	896	55.5	56.0	56.6	—	—	—
	1024	—	—	—	78.8	78.0	79.3
	2048	—	—	—	78.1	77.6	78.7
VOC2007	64	79.2	78.1	80.2	84.2	85.1	85.8
	128	80.8	81.1	84.1	85.0	86.0	87.4
	256	80.7	78.6	82.6	85.4	86.3	87.1
	512	81.1	81.3	81.5	85.9	86.7	87.6
	768	79.0	78.8	81.7	86.0	86.9	87.5
	896	81.1	81.3	81.8	—	—	—
	1024	—	—	—	86.2	87.0	87.4
	2048	—	—	—	86.1	86.8	87.3
N24News	64	68.0	68.4	69.1	76.8	75.7	77.0
	128	68.5	68.5	69.8	77.5	76.9	78.4
	256	69.3	69.1	70.8	77.9	77.4	78.5
	512	69.7	69.5	70.8	78.3	77.8	79.1
	768	69.4	69.6	70.7	78.1	77.7	78.9
	896	69.6	69.2	70.3	—	—	—
	1024	—	—	—	77.9	77.5	78.4
	2048	—	—	—	77.3	76.9	77.5
SUN397	64	54.6	54.9	63.0	73.9	75.2	75.8
	128	62.1	62.7	66.0	75.0	76.8	78.1
	256	64.6	66.2	69.5	76.0	77.8	79.2
	512	64.3	65.4	70.1	76.6	78.2	79.5
	768	67.1	67.1	70.4	76.9	78.6	79.8
	896	66.5	68.0	70.5	—	—	—
	1024	—	—	—	76.8	78.4	79.5
	2048	—	—	—	76.5	78.0	79.6
Hateful-Memes	64	61.0	61.4	66.1	66.6	67.0	68.3
	128	59.5	60.0	65.5	66.9	67.1	67.2
	256	60.8	61.2	64.0	67.2	67.4	67.9
	512	59.4	59.8	64.6	67.5	67.6	68.0
	768	58.8	59.0	64.3	67.3	67.7	67.6
	896	58.3	58.0	63.9	—	—	—
	1024	—	—	—	67.4	67.8	68.1
	2048	—	—	—	67.2	67.6	68.3

Table 14: **Component analysis of geometric regularizers on VQA datasets.** Performance is measured side-by-side to highlight the consistent gains of the combined objective. Best results are **bolded**.

Data	Dim.	FastVLM-0.5B			Qwen3-VL-2B			Data	Dim.	FastVLM-0.5B			Qwen3-VL-2B		
		$\mathcal{L}_{\text{proj}}$	$\mathcal{L}_{\text{lap}}$	Ours	$\mathcal{L}_{\text{proj}}$	$\mathcal{L}_{\text{lap}}$	Ours			$\mathcal{L}_{\text{proj}}$	$\mathcal{L}_{\text{lap}}$	Ours	$\mathcal{L}_{\text{proj}}$	$\mathcal{L}_{\text{lap}}$	Ours
OK-VQA	64	37.2	36.6	37.6	56.4	55.7	57.3	InfographicsVQA	64	23.6	24.3	25.0	36.5	37.2	37.9
	128	45.9	44.8	46.5	60.1	59.3	61.0		128	29.4	30.1	30.4	38.8	39.4	40.1
	256	50.8	49.7	51.2	62.8	62.1	63.5		256	34.1	34.6	35.1	41.2	41.7	42.3
	512	51.6	50.9	51.9	62.7	63.0	63.4		512	34.5	34.9	35.1	42.8	43.2	43.8
	768	52.9	52.6	53.4	63.9	64.1	64.7		768	35.3	35.9	36.1	43.0	43.4	43.7
	896	53.2	53.0	53.7	—	—	—		896	35.2	35.7	36.0	—	—	—
	1024	—	—	—	63.8	63.9	64.7		1024	—	—	—	42.7	43.1	44.6
	2048	—	—	—	64.4	64.0	65.5		2048	—	—	—	42.5	42.9	44.5
A-OKVQA	64	36.9	35.7	37.6	49.0	48.4	49.9	ChartQA	64	40.0	40.8	41.4	53.5	53.7	54.0
	128	44.0	42.6	44.7	53.8	52.9	54.4		128	46.5	47.2	47.9	57.1	57.3	57.7
	256	47.2	46.3	47.6	57.6	56.9	58.2		256	47.8	48.6	48.7	59.0	59.5	59.9
	512	49.0	48.0	49.7	57.5	57.0	58.1		512	49.1	50.0	50.3	60.8	61.3	62.2
	768	49.9	49.3	50.3	58.0	57.8	58.7		768	49.6	50.1	50.2	60.5	61.6	61.8
	896	50.6	49.8	51.0	—	—	—		896	49.9	50.2	50.5	—	—	—
	1024	—	—	—	58.1	57.6	58.8		1024	—	—	—	60.3	61.1	61.3
	2048	—	—	—	57.9	57.5	58.5		2048	—	—	—	60.1	60.8	61.2
DocVQA	64	56.8	57.5	58.2	74.1	74.8	75.8	Visual7W	64	40.8	40.1	41.9	50.8	50.4	51.6
	128	67.5	68.5	69.0	80.7	81.2	81.8		128	47.5	46.9	48.5	53.6	53.2	54.1
	256	72.7	73.6	74.2	83.5	84.0	84.4		256	49.6	48.7	50.1	55.4	55.1	55.9
	512	73.8	74.4	74.9	85.6	85.9	86.8		512	50.9	50.1	51.5	56.9	56.8	57.5
	768	74.5	75.1	75.7	86.0	86.4	86.9		768	50.8	50.4	51.3	56.6	56.5	57.0
	896	74.9	75.5	76.0	—	—	—		896	51.6	50.9	52.0	—	—	—
	1024	—	—	—	86.2	86.6	87.0		1024	—	—	—	57.0	57.0	57.5
	2048	—	—	—	86.1	86.3	86.9		2048	—	—	—	57.1	57.2	57.7

Table 15: Topology preservation comparison. $\uparrow$ indicates higher is better and $\downarrow$ indicates lower is better.

Method	kNN $\uparrow$	Trust. $\uparrow$	Cont. $\uparrow$	Fiedler $\uparrow$	Spectral Dist. $\downarrow$
MRL	0.844	0.912	0.904	0.869	0.193
ESE	0.856	0.920	0.915	0.886	0.168
Ours	0.892	0.944	0.937	0.949	0.088

Table 16: Ablation of topology preservation objectives.

Method	kNN $\uparrow$	Trust. $\uparrow$	Cont. $\uparrow$	Fiedler $\uparrow$	Spectral Dist. $\downarrow$
MRL	0.844	0.912	0.904	0.869	0.193
+ L_{lap}	0.861	0.923	0.918	0.927	0.109
+ L_{proj}	0.871	0.932	0.926	0.887	0.174
Full	0.892	0.944	0.937	0.949	0.088