

PGA-KD: Principal Geometry Alignment and Semantic Consistency for VLM2Vec Distillation

The-Anh Tran^{1*}, Thanh Xuan Nguyen^{1*}, Duc Anh Nguyen^{1†},
Dinh Viet Sang¹, Linh Ngo Van¹, Thien Huu Nguyen²

¹Hanoi University of Science and Technology, Vietnam,

²University of Oregon, USA

Abstract

Adapting large vision–language models for embedding tasks (VLM2Vec) is standard practice, but distilling them into compact architectures remains challenging due to capacity mismatches in tokenization, dimensionality, and multimodal representations. To address this, we propose a robust distillation framework based on two complementary mechanisms: Principal Geometry Alignment (PGA) to stabilize training by aligning principal geometric components, and a Mutual Information Maximization objective to enforce global semantic consistency across intra- and inter-modal pairs. By jointly constraining geometric structure and semantic dependency, our method effectively mitigates model capacity discrepancies. Experimental results on the Massive Multimodal Embedding Benchmark (MMEB) across all four meta-tasks—under both In-Distribution and Out-of-Distribution settings—demonstrate that our approach consistently outperforms state-of-the-art distillation methods. Our source code is available at <https://github.com/theanh-ktmt/PGA-KD>.

1 Introduction

Large vision–language models (VLMs), integrating large language models (LLMs) with visual encoders, have demonstrated strong performance in multimodal generation tasks (Wang et al., 2025; Lu et al., 2025; Bai et al., 2025). Through the VLM2Vec framework (Jiang et al., 2025), these models can be transformed into multimodal embedding engines. Their superior feature extraction allows for broad application across downstream tasks, such as image classification, image-text retrieval, and visual question answering (VQA) (Radford et al., 2021). Despite their efficacy, the intensive computational and memory requirements of VLMs present significant barriers to deployment

on resource-constrained edge devices. To overcome these limitations, knowledge distillation (KD) (Hinton et al., 2015) has emerged as a pivotal technique. By transferring complex multimodal reasoning from expansive teacher models to compact student architectures, KD facilitates edge deployment under strict performance constraints without compromising accuracy.

In the context of knowledge distillation for multimodal embedding frameworks like VLM2Vec, the primary objective is preserving the spatial characteristics of the teacher’s representation space during knowledge transfer. Current research focuses on two critical pillars: maintaining geometric structure and ensuring cross-modal consistency. First, preserving the geometric structure maintains relational distances between data samples within a batch; essentially, relative differences between inputs should remain consistent across models. To address this, RKD (Park et al., 2019) maintains distance-wise information between all pairs and angle-wise information between all triplets within a batch. Recognizing that geometric relations extend beyond simple distances and angles, CKD (Xu et al., 2025) introduced general group-level relations, utilizing an aggregate function to extract and align relative differences within sample groups.

Second, ensuring cross-modal consistency allows the student to accurately capture the sophisticated textual and visual associations learned by the teacher. For instance, Align-KD (Feng et al., 2025) proposed an Attention Mimicking mechanism at the initial transformer layer to guide how text tokens should attend to visual regions. Alternatively, EM-KD (Feng et al., 2026) constructs and aligns affinity matrices to capture cross-modal correlations inherent in each model. A comprehensive review of vision-language frameworks and classic distillation paradigms is provided in Appendix A.

While these established methods provide a foundation for relational and cross-modal distillation,

* Equal contribution.

† Corresponding author.

they suffer from two critical limitations: **Geometric Misalignment** and **Semantic Inconsistency**. First, while the teacher’s core semantic knowledge is primarily encoded within its principal eigenspaces (low-frequency components), the representations also contain a substantial *spectral tail* of high-frequency components. These high-frequency components are not meaningless: they often capture fine-grained detail such as textual or document-level cues, alongside variation tied to the teacher’s over-parameterization (Ettori, 2026). However, given the large capacity gap, a compact student cannot faithfully absorb this dense high-frequency structure. We therefore use the term *geometric noise* in a relative sense: relative to the student’s limited capacity, these components act as noise during distillation, diverting the student from the teacher’s principal structure even though they remain informative in the teacher itself. Current distillation frameworks, such as RKD (Park et al., 2019) and CKD (Xu et al., 2025), indiscriminately transfer all geometric relational information without distinguishing high-energy from low-energy structure. Consequently, the student is forced to inherit this high-frequency structure, leading to Geometric Misalignment.

Second, concentrating the entirety of cross-modal knowledge through a single transmission channel—such as the first transformer layer in Align-KD (Feng et al., 2025) or affinity matrices in EM-KD (Feng et al., 2026)—is inherently inefficient and difficult to control. This bottleneck leads to the Semantic Inconsistency problem, where the student merely mimics surface-level patterns instead of learning actual semantic relationships between text and image. This disconnect results in significant performance degradation in complex tasks like visual question answering (VQA), where the student struggles to accurately ground textual queries in corresponding visual regions.

To address these limitations, we propose **Principal Geometry Alignment (PGA-KD)**, a distillation framework designed for heterogeneous VLM2Vec models. PGA-KD consists of two core components. First, the **Principal Geometry Alignment (PGA)** module addresses Geometric Misalignment. While existing methods transfer all relational data, PGA applies spectral decomposition to filter out high-frequency geometric noise caused by architectural redundancy. By aligning only the principal geometry (low-frequency components), the student learns essential structural infor-

mation representing actual data semantics. Second, to resolve Semantic Inconsistency across different model architectures, the **Semantic Consistency Learning (SCL)** objective is introduced. SCL employs a Mutual Information (MI) Maximization (Almudévar et al., 2025) strategy through four specific pathways: intra-modal (text-to-text, image-to-image) and inter-modal (text-to-image, image-to-text). This multi-pathway approach enables the student to accurately capture fine-grained multi-modal dependencies and semantic alignments. By avoiding the limitations of single-channel methods, SCL provides a more efficient mechanism for transferring knowledge between models with different structures.

We conduct extensive experiments on the Massive Multimodal Embedding Benchmark (MMEB) (Jiang et al., 2025), covering all four meta-tasks under both ID and OOD evaluations. The results demonstrate that PGA-KD consistently outperforms state-of-the-art methods in heterogeneous model compression settings.

2 Background

2.1 Knowledge Distillation in VLM2Vec

Knowledge distillation (KD) in VLM2Vec (Jiang et al., 2025) transfers semantic alignment from a high-capacity teacher ($\mathcal{T}$) to a compact student ($\mathcal{S}$) using a joint objective.

Supervised Loss: The student is optimized via the contrastive InfoNCE loss (van den Oord et al., 2019) to maximize discriminative ability between query-positive pairs. For a batch of size B , let $\mathbf{h}_{q,i}^S$ and $\mathbf{h}_{p,i}^S$ denote the embeddings of the i -th query and its positive key. The loss with a learnable temperature τ is:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{e^{\langle \mathbf{h}_{q,i}^S, \mathbf{h}_{p,i}^S \rangle / \tau}}{\sum_{j=1}^B e^{\langle \mathbf{h}_{q,i}^S, \mathbf{h}_{p,j}^S \rangle / \tau}} \quad (1)$$

Distillation Loss: To match representations without explicit classification logits, VLM2Vec aligns the final token hidden states via Mean Squared Error (MSE). A learnable linear projector $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d_{\mathcal{T}} \times d_{\mathcal{S}}}$ resolves dimension mismatches ($d_{\mathcal{S}} \neq d_{\mathcal{T}}$):

$$\mathcal{L}_{\text{MSE}} = \frac{1}{B} \sum_{i=1}^B \|\mathbf{h}_i^S - \mathbf{h}_i^{\mathcal{T}} \mathbf{W}_{\text{proj}}\|_2^2 \quad (2)$$

The standard baseline combines both objectives

using a balancing coefficient λ_{MSE} :

$$\mathcal{L}_{\text{Base}} = \mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{MSE}} \mathcal{L}_{\text{MSE}} \quad (3)$$

While $\mathcal{L}_{\text{Base}}$ performs point-wise alignment, it treats the teacher as a black box, ignoring the broader batch geometry and cross-modal token interactions.

2.2 Geometric Topology Alignment

Rather than rigid point-wise feature matching, aligning the pairwise similarity structures (geometric topology) within a batch is more effective. We formalize this using Centered Kernel Alignment (CKA) (Kornblith et al., 2019) instead of relational metrics like RKD (Park et al., 2019) (see Appendix B.1 for comparison).

Let $\mathbf{K} = \mathbf{H}^S (\mathbf{H}^S)^T$ and $\mathbf{L} = \mathbf{H}^T (\mathbf{H}^T)^T$ be the $B \times B$ Gram matrices representing intra-batch similarities for student and teacher. CKA measures the statistical correlation between these structures:

$$\text{CKA}(\mathbf{K}, \mathbf{L}) = \frac{\text{tr}(\mathbf{KCLC})}{\sqrt{\text{tr}(\mathbf{KCKC}) \cdot \text{tr}(\mathbf{LCLC})}} \quad (4)$$

where $\mathbf{C} = \mathbf{I}_B - \frac{1}{B} \mathbf{1}\mathbf{1}^T$ is the centering matrix. CKA captures distance, angular, and higher-order geometry globally in a single step, avoiding the scaling and balancing issues of rigid element-wise metrics.

2.3 Mutual Information Maximization

Mutual Information (MI) quantifies the statistical dependence between two random variables X and Y :

$$I(X; Y) = \mathbb{E}_{p(x,y)} \left[\log \frac{p(x,y)}{p(x)p(y)} \right] \quad (5)$$

As direct high-dimensional MI computation is computationally infeasible, variational lower bounds are used. The InfoNCE loss serves as a computational surrogate to maximize the lower bound of MI (van den Oord et al., 2019):

$$I(X; Y) \geq \log(B) - \mathcal{L}_{\text{InfoNCE}} \quad (6)$$

We provide the complete information-theoretic proofs and detailed expansions of this bound in Appendix B.2. This theoretical connection serves as the backbone for our **Semantic Consistency Learning (SCL)**, ensuring the student preserves fine-grained semantic content across multi-modal pathways while filtering geometric noise.

3 Methodology

We propose **Principal Geometry Alignment (PGA-KD)**, a distillation framework to transfer representations from a high-capacity teacher VLM ($\mathcal{T}$) to a compact student VLM ($\mathcal{S}$). As illustrated in Figure 1, PGA-KD addresses two core limitations of standard distillation: overfitting to geometric noise and semantic inconsistency across modalities.

3.1 Principal Geometry Alignment (PGA)

Standard knowledge distillation (KD) methods typically treat all information from the teacher as equally important. However, when distilling a high-dimensional teacher ($d_{\mathcal{T}} = 1536$) into a compact student ($d_{\mathcal{S}} = 896$), this assumption becomes problematic.

The Signal vs. Noise Problem: In high-dimensional spaces, a teacher’s representation contains both principal components (major variance directions defining semantic clusters) and minor components (small, high-frequency fluctuations caused by architectural redundancy). Existing methods like RKD and CKD force the student to mimic both, wasting its limited capacity on modeling noise. PGA-KD solves this by decomposing the teacher’s geometry, filtering out the noise, and forcing the student to align solely with the principal semantic structure.

Capturing Geometry via Similarity Matrices:

We represent the geometric topology of a batch by analyzing mutual relations between samples. Let $\mathbf{H}^T \in \mathbb{R}^{B \times d_{\mathcal{T}}}$ and $\mathbf{H}^S \in \mathbb{R}^{B \times d_{\mathcal{S}}}$ be the batch embeddings for the teacher and student, respectively. We compute the similarity matrices $\mathbf{G}^T, \mathbf{G}^S \in \mathbb{R}^{B \times B}$ as:

$$\mathbf{G}_{ij}^T = \mathbf{h}_i^T \cdot \mathbf{h}_j^T, \quad \mathbf{G}_{ij}^S = \mathbf{h}_i^S \cdot \mathbf{h}_j^S \quad (7)$$

These matrices, $\mathbf{G}^T$ and $\mathbf{G}^S$, represent the raw geometric structures of the teacher and student within the current batch.

Geometric Denoising via Eigendecomposition:

To separate signal from noise, we perform an eigendecomposition on $\mathbf{G}^T$, breaking the geometry into a weighted sum of orthogonal components:

$$\mathbf{G}^T = \sum_{i=1}^B \lambda_i \mathbf{v}_i \mathbf{v}_i^T \quad (8)$$

where eigenvalues λ_i represent the *energy* of the i -th geometric direction, sorted such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_B$.

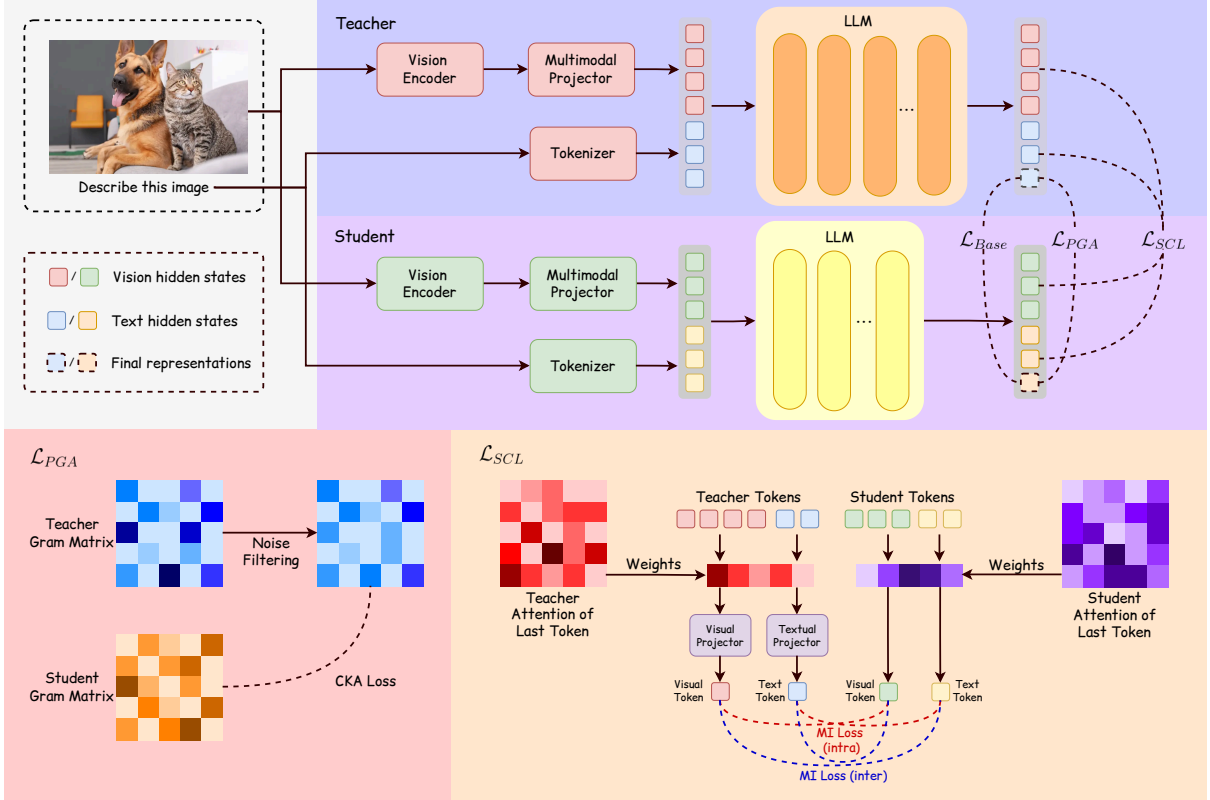


Figure 1: **The PGA-KD Distillation Framework.** Knowledge transfer is driven by three core losses: (1) $\mathcal{L}_{Base}$, a combination of InfoNCE and MSE for point-wise feature anchoring; (2) $\mathcal{L}_{PGA}$, a geometric loss based on CKA and eigendecomposition to align noise-filtered principal structures; and (3) $\mathcal{L}_{SCL}$, a semantic loss based on Mutual Information (MI) maximization to synchronize intra-modal and inter-modal dependencies.

In large VLMs, the spectrum of λ usually exhibits a *heavy tail* of tiny eigenvalues representing noise. To denoise the geometry, we keep only the top k components that contribute to $\eta\%$ (e.g., 85%) of the total information (energy):

$$\hat{\mathbf{G}}^T = \sum_{i=1}^k \lambda_i \mathbf{v}_i \mathbf{v}_i^\top, \quad \text{s.t.} \quad \frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^B \lambda_j} \geq \eta \quad (9)$$

The resulting low-rank approximation $\hat{\mathbf{G}}^T$ defines the Principal Geometry, retaining critical semantic structures while eliminating architectural noise. The mathematical justification for this spectral filtering approach is detailed in Appendix B.3.

PGA Objective & Spectral Mismatch: Minimizing the standard Euclidean distance (MSE) between $\mathbf{G}^S$ and $\hat{\mathbf{G}}^T$ via the Frobenius norm $\|\mathbf{G}^S - \hat{\mathbf{G}}^T\|_F^2$ forces the student to match both geometric directions and exact eigenvalue magnitudes ($\lambda_i^S \approx \lambda_i^T$). However, the compact student lacks the dimensional capacity to support the teacher’s high-energy spectrum. Forcing this spectral regression distorts the student’s latent space.

To resolve this, we employ Centered Kernel

Alignment (CKA) (Kornblith et al., 2019). CKA normalizes the objective, making it invariant to the isotropic scaling of eigenvalues. This effectively decouples geometric *structure* (correlations) from *energy* (magnitudes), allowing the student to align its semantic directions without being penalized for its smaller scale of variance.

Formally, let $\mathbf{C} = \mathbf{I} - \frac{1}{B} \mathbf{1}\mathbf{1}^\top$ be the centering matrix. We maximize the CKA correlation between the centered Gram matrices $\tilde{\mathbf{G}}^S = \mathbf{C}\mathbf{G}^S\mathbf{C}$ and $\tilde{\mathbf{G}}^T = \mathbf{C}\hat{\mathbf{G}}^T\mathbf{C}$:

$$\mathcal{L}_{PGA} = 1 - \frac{\text{Tr}(\tilde{\mathbf{G}}^S \tilde{\mathbf{G}}^T)}{\sqrt{\text{Tr}(\tilde{\mathbf{G}}^S \tilde{\mathbf{G}}^S) \cdot \text{Tr}(\tilde{\mathbf{G}}^T \tilde{\mathbf{G}}^T)}} \quad (10)$$

The Frobenius inner product $\text{Tr}(\tilde{\mathbf{G}}^S \tilde{\mathbf{G}}^T)$ maximizes the structural consensus, ensuring both models agree on relative sample similarities regardless of absolute feature scales.

3.2 Semantic Consistency Learning (SCL)

While Principal Geometry Alignment (PGA) preserves the global structural topology, it treats data points as atomic units, potentially overlooking

fine-grained internal dependencies between modalities. High-capacity teachers maintain precise cross-modal synergy, whereas compact students often suffer from modality collapse, where one modality dominates the representation. To bridge this gap, we introduce **Semantic Consistency Learning (SCL)**, which enforces internal reasoning pathway consistency via Mutual Information (MI) maximization across two intra-modal and two inter-modal propagation paths.

Feature Aggregation and Projection: To capture modality-specific information, we perform attention-based aggregation using the last token’s attention map from each model’s own final layer. Let $\mathbf{A} \in \mathbb{R}^{H \times L \times L}$ be the attention for H heads and sequence length L . We take the attention weights of the final token, average them over the H heads, and renormalize them within each modality to obtain visual weights α_v and textual weights α_t . The aggregated features $\mathbf{v}, \mathbf{t} \in \mathbb{R}^d$ are computed as weighted sums of their respective token embeddings:

$$\mathbf{v} = \sum_{i \in \text{Img}} \alpha_{v,i} \mathbf{h}_i, \quad \mathbf{t} = \sum_{j \in \text{Txt}} \alpha_{t,j} \mathbf{h}_j \quad (11)$$

where Img and Txt denote the indices of visual and textual tokens, respectively.

To bridge the dimensionality gap between the teacher (d_T) and student (d_S), we employ modality-specific projection heads, ϕ_v and ϕ_t , to map aggregated teacher features into the student’s latent space:

$$\hat{\mathbf{v}}^T = \phi_v(\mathbf{v}^T), \quad \hat{\mathbf{t}}^T = \phi_t(\mathbf{t}^T) \quad (12)$$

Mutual Information Maximization: We maximize the MI between the student’s features and the teacher’s projections using the InfoNCE loss (van den Oord et al., 2019) as a variational lower bound. For a batch size B , the alignment loss between a student variable $\mathbf{x}^S$ and a teacher variable $\mathbf{y}^T$ is defined as:

$$\mathcal{L}_{\text{MI}}(\mathbf{x}^S, \mathbf{y}^T) = -\frac{1}{B} \sum_{i=1}^B \log \frac{f(\mathbf{x}_i^S, \mathbf{y}_i^T)}{\sum_{j=1}^B f(\mathbf{x}_i^S, \mathbf{y}_j^T)} \quad (13)$$

where $f(\mathbf{x}, \mathbf{y}) = \exp(\text{sim}(\mathbf{x}, \mathbf{y})/\tau)$, $\text{sim}(\cdot)$ denotes cosine similarity, and τ is a temperature hyperparameter.

The Four-Way Consistency Objective: SCL decomposes distillation into granular alignment

tasks across both intra-modal and inter-modal pathways:

1) *Intra-Modality Consistency* ($\mathcal{L}_{\text{Intra}}$): Preserves the semantic fidelity of individual modalities, forcing the student’s encoders to align with the teacher’s respective visual and textual representations:

$$\mathcal{L}_{\text{Intra}} = \mathcal{L}_{\text{MI}}(\mathbf{v}^S, \hat{\mathbf{v}}^T) + \mathcal{L}_{\text{MI}}(\mathbf{t}^S, \hat{\mathbf{t}}^T) \quad (14)$$

2) *Inter-Modality Consistency* ($\mathcal{L}_{\text{Inter}}$): Transfers cross-modal reasoning by aligning the student’s image features with the teacher’s text features and vice versa, capturing the critical cross-modal bridges established by the teacher:

$$\mathcal{L}_{\text{Inter}} = \mathcal{L}_{\text{MI}}(\mathbf{v}^S, \hat{\mathbf{t}}^T) + \mathcal{L}_{\text{MI}}(\mathbf{t}^S, \hat{\mathbf{v}}^T) \quad (15)$$

The final SCL objective integrates all pathways to ensure robust reconstruction of the underlying semantic dependencies:

$$\mathcal{L}_{\text{SCL}} = \mathcal{L}_{\text{Intra}} + \mathcal{L}_{\text{Inter}} \quad (16)$$

3.3 Total Optimization Objective

The final training objective integrates the geometric and semantic regularizers with the baseline loss:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Base}} + \lambda_{\text{PGA}} \mathcal{L}_{\text{PGA}} + \lambda_{\text{SCL}} \mathcal{L}_{\text{SCL}} \quad (17)$$

where $\mathcal{L}_{\text{Base}} = \mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{MSE}} \mathcal{L}_{\text{MSE}}$, and $\lambda_{\text{PGA}}, \lambda_{\text{SCL}}$ are balancing hyperparameters.

4 Experiments

4.1 Experimental Setup

Datasets & Evaluation Tracks: Per VLM2Vec (Jiang et al., 2025), we evaluate across four meta-tasks: Image Classification, Visual Question Answering (VQA), Cross-Modal Retrieval, and Visual Grounding, under ID and OOD tracks. Dataset statistics are in Appendix C.1.

Architectures & Baselines: We employ B3-Qwen2-2B (Thirukovalluru et al., 2025) ($d_T = 1536$) as teacher. To verify robustness, distillation is evaluated on two 0.5B student variants ($d_S = 896$): FastVLM-0.5B (Vasu et al., 2025) and Llava-Onevision-0.5B (Li et al., 2025). Baselines include MSE regression, RKD (Park et al., 2019), CKD (Xu et al., 2025), EMO (Truong et al., 2025), and EM-KD (Feng et al., 2026). Training hyperparameters, architecture details, and baseline configurations are deferred to Appendix C.2, C.3, and C.4, respectively.

Table 1: **In-Distribution Benchmark Results across Four Different Meta-Tasks using B3-Qwen2-2B (Teacher) and FastVLM-0.5B (Student)**. The table presents a comprehensive evaluation of knowledge distillation methods. For each task, bold numbers indicate the best performance among student variants, and underlined numbers indicate the second-best performance. Notably, EMO exhibits severe performance degradation on retrieval meta-tasks, the underlying causes of which are detailed in Appendix D.

Method	Classification						VQA						
	IN-1K	N24News	Hateful	VOC07	SUN397	Avg	OK	A-OK	Doc	I-VQA	Chart	Vis7W	Avg
Teacher	82.9	79.2	56.2	88.0	80.9	77.4	63.0	53.5	92.3	58.9	53.1	53.1	62.3
Student	52.9	68.4	60.0	77.7	64.7	64.7	50.7	50.2	74.0	34.1	49.4	46.1	50.8
MSE	53.0	70.1	58.4	78.2	65.2	65.0	<u>52.4</u>	49.8	74.3	34.3	50.8	46.4	51.3
RKD (Park et al., 2019)	<u>53.6</u>	70.3	<u>60.8</u>	79.9	66.2	<u>66.2</u>	51.1	<u>51.2</u>	74.1	34.9	<u>50.2</u>	46.5	51.3
CKD (Xu et al., 2025)	53.2	70.6	61.2	78.6	<u>66.7</u>	66.1	51.8	51.0	<u>74.7</u>	35.0	49.6	46.2	51.4
EMO (Truong et al., 2025)	52.4	68.4	59.1	<u>80.4</u>	59.3	63.9	50.9	50.1	73.1	33.8	49.0	<u>47.8</u>	50.8
EM-KD (Feng et al., 2026)	53.4	67.3	59.4	78.0	63.2	64.3	51.1	49.3	77.2	<u>36.7</u>	47.6	47.3	<u>51.5</u>
PGA-KD (Ours)	62.6	<u>70.4</u>	56.9	81.9	75.1	69.4	59.5	54.6	74.6	38.4	44.6	52.3	54.0

Method	Retrieval									Grounding	
	VisDial	CIRR	VN-t2i	VN-i2t	COCO-t2i	COCO-i2t	NIGHTS	WebQA	Avg	COCO	Avg
Teacher	83.4	60.3	76.0	79.9	76.6	49.7	68.7	88.6	72.9	70.5	70.5
Student	61.2	32.9	46.8	49.6	65.2	58.6	62.6	69.4	55.8	65.0	65.0
MSE	<u>64.2</u>	27.7	46.8	<u>51.6</u>	69.6	59.0	<u>64.0</u>	<u>72.5</u>	56.9	68.3	68.3
RKD (Park et al., 2019)	62.1	32.1	45.7	50.9	64.8	58.3	63.0	70.6	55.9	68.1	68.1
CKD (Xu et al., 2025)	61.5	28.9	45.0	47.7	66.0	<u>59.2</u>	63.0	69.8	56.2	68.7	68.7
EMO (Truong et al., 2025)	41.7	26.7	0.5	22.6	3.9	33.3	62.4	3.0	24.3	66.8	66.8
EM-KD (Feng et al., 2026)	61.5	36.5	<u>47.3</u>	49.8	<u>68.4</u>	58.5	63.5	73.0	<u>57.3</u>	<u>69.2</u>	<u>69.2</u>
PGA-KD (Ours)	73.3	<u>35.1</u>	50.8	55.6	66.1	63.5	64.3	<u>72.5</u>	60.2	69.5	69.5

4.2 Main Results

In-Distribution Evaluation with FastVLM Student: As shown in Table 1, PGA-KD consistently outperforms existing baselines, establishing a new state-of-the-art across all meta-tasks. On Classification, PGA-KD achieves an average of 69.4%, surpassing the best baseline (RKD) by +3.2%. In VQA, it yields a top average score of 54.0% (+2.5% over EM-KD), driven by remarkable gains on hard reasoning tasks like OK-VQA (59.5%, +7.1% over MSE) and A-OK (54.6%, +3.4% over RKD). For Retrieval and Grounding, PGA-KD maintains its superior performance with averages of 60.2% and 69.5% respectively.

Architecture-Agnostic Efficacy with Llava-Onevision Student: Table 2 demonstrates that the benefits of PGA-KD generalize seamlessly to alternative architectures. Using the Llava-Onevision-0.5B student, PGA-KD advances the average Classification performance to 70.4% (+3.5% higher than the next-best CKD), highlighted by a substantial +9.5% gain on ImageNet-1K (66.5% vs. 57.0% for CKD). Similarly, it establishes a peak average of 41.0% in VQA, validating that its formulation is structurally agnostic and robust across diverse Vision-Language Models (VLMs).

Generalization under Distribution Shifts: Ta-

ble 3 reports OOD benchmarks to evaluate robustness against distribution shifts. PGA-KD mitigates overfitting and leads OOD Classification with a 33.7% average (+3.4% over EMO), showing strong localized gains on Place365 (38.1%, +3.3% over EMO) and ImageNet-R (62.6%, +9.9% over MSE). On OOD VQA, it scores the highest average of 51.8% (+1.6% over MSE), driven by a +4.9% improvement on TextVQA (66.4%), proving it fosters generalizable multimodal representations.

Summary of Framework Transferability:

Across all core capabilities, PGA-KD consistently bridges the capacity gap between the 2B generative teacher and the 0.5B compact embedding students while maintaining a highly competitive alignment overhead (see Appendix E for detailed computational complexity and training overhead). By combining geometric topological robustness ($\mathcal{L}_{PGA}$) with rigorous multi-modal information synchronization ($\mathcal{L}_{SCL}$), our approach proves the feasibility of deploying highly capable, low-footprint multimodal engines directly in resource-constrained environments.

Where PGA-KD Helps and Where It Does Not:

The gains are not uniform across tasks. PGA-KD improves most on reasoning and retrieval (OK-VQA +8.8, VisDial +12.1 over the student), but

Table 2: **In-Distribution Benchmark Results across Classification and VQA Tasks using B3-Qwen2-2B (Teacher) and Llava-Onevision-0.5B (Student).** The table presents a comprehensive evaluation of knowledge distillation methods. For each task, bold numbers indicate the best performance among student variants, and underlined numbers indicate the second-best performance.

Method	Classification						VQA						
	IN-1K	N24News	Hateful	VOC07	SUN397	Avg	OK	A-OK	Doc	I-VQA	Chart	Vis7W	Avg
Teacher	82.9	79.2	56.2	88.0	80.9	77.4	63.0	53.5	92.3	58.9	53.1	53.1	62.3
Student	55.6	67.1	57.0	83.9	66.0	65.9	48.2	42.8	38.6	21.3	25.1	43.8	36.6
MSE	55.3	66.8	56.2	84.1	66.8	65.8	47.8	42.5	37.9	21.0	24.7	42.9	36.1
RKD (Park et al., 2019)	55.1	66.8	57.6	83.7	67.0	66.0	47.8	43.7	38.6	21.6	25.5	43.5	36.8
CKD (Xu et al., 2025)	<u>57.0</u>	<u>66.9</u>	<u>58.6</u>	84.1	<u>68.2</u>	<u>66.9</u>	<u>51.2</u>	<u>43.9</u>	47.2	<u>22.9</u>	30.9	45.6	<u>40.3</u>
EMO (Truong et al., 2025)	56.1	66.4	56.4	83.6	66.2	65.7	46.4	43.6	41.1	20.7	23.1	43.0	36.3
EM-KD (Feng et al., 2026)	53.2	63.3	53.6	82.6	65.2	63.6	44.7	37.9	36.4	20.5	22.6	39.8	33.7
PGA-KD (Ours)	66.5	68.9	59.5	83.9	73.3	70.4	53.6	47.2	<u>46.0</u>	23.9	<u>29.7</u>	<u>45.4</u>	41.0

Table 3: **Out-of-Distribution (OOD) Benchmark Results across Classification and VQA Tasks using B3-Qwen2-2B (Teacher) and FastVLM-0.5B (Student).** The table presents a comprehensive evaluation of knowledge distillation methods under distribution shift. For each task, bold numbers indicate the best performance among student variants, and underlined numbers indicate the second-best performance.

Method	Classification (OOD)						VQA (OOD)				
	Place365	IN-A	IN-R	ObjectNet	Country211	Avg	ScienceQA	VizWiz	GQA	TextVQA	Avg
Teacher	44.7	49.7	90.4	71.7	25.7	56.4	40.7	48.4	68.1	80.0	59.3
Student	30.0	<u>25.3</u>	48.2	28.4	7.8	27.9	35.8	35.5	63.4	56.7	47.9
MSE	30.4	<u>25.9</u>	<u>52.7</u>	25.5	<u>8.0</u>	28.5	37.9	37.7	63.5	<u>61.5</u>	<u>50.2</u>
RKD (Park et al., 2019)	31.1	<u>25.9</u>	47.5	27.8	8.4	28.1	36.0	36.2	64.6	57.0	48.5
CKD (Xu et al., 2025)	30.3	25.4	47.9	31.6	7.5	28.5	36.3	35.8	64.1	56.6	48.2
EMO (Truong et al., 2025)	<u>34.8</u>	21.8	52.5	<u>37.4</u>	5.2	<u>30.3</u>	36.6	36.1	<u>65.5</u>	57.7	49.0
EM-KD (Feng et al., 2026)	32.5	27.4	50.4	32.2	7.6	30.0	36.5	36.1	65.4	57.6	48.9
PGA-KD (Ours)	38.1	24.1	62.6	40.7	7.5	33.7	<u>36.9</u>	<u>36.8</u>	67.0	66.4	51.8

is flat on DocVQA (74.6, below EM-KD’s 77.2) and weaker on ChartQA (44.6, below the student’s 49.4). This pattern is systematic rather than random: filtering the low-energy tail trades some fine-grained textual detail for a cleaner global structure, which text-dense tasks such as DocVQA and ChartQA rely on more heavily. The threshold η controls this balance, and can be raised (e.g. $\eta = 0.95$ recovers the best DocVQA, 75.3, in Table 5) when the target workload is OCR- or document-oriented.

Finally, to confirm that these gains reflect a genuine improvement rather than seed variance, we repeat the OOD VQA evaluation across three random seeds; PGA-KD remains ahead of the strongest baselines throughout, as detailed in Appendix F.

4.3 Ablation Analysis

4.3.1 Contribution of PGA and SCL

Table 4 shows the impact of each loss component in our final objective, $\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Base}} + \lambda_{\text{PGA}}\mathcal{L}_{\text{PGA}} + \lambda_{\text{SCL}}\mathcal{L}_{\text{SCL}}$. We observe two key findings:

First, $\mathcal{L}_{\text{PGA}}$ is critical for transferring structural knowledge. Removing it causes the average VQA

score to drop by 2.4% (from 54.0% to 51.6%). This shows that aligning final representations alone is insufficient; mimicking the teacher’s geometric structure is essential for student performance.

Second, $\mathcal{L}_{\text{SCL}}$ effectively preserves multimodal understanding. Both intra-modal and inter-modal alignments are necessary, as removing $\mathcal{L}_{\text{Intra}}$ and $\mathcal{L}_{\text{Inter}}$ reduces performance by 0.9% and 1.5%, respectively. This confirms that multi-pathway information transfer helps the student learn complex cross-modal relationships.

Table 4: **Component Ablation.** Isolating the impact of Principal Geometry Alignment (PGA) and Semantic Consistency Learning (SCL) on FastVLM-0.5B.

Configuration	OK	A-OK	Doc	I-VQA	Avg. VQA
Student	50.7	50.2	74.0	34.1	50.8
PGA-KD (w/o $\mathcal{L}_{\text{PGA}}$)	57.1	54.2	67.7	37.6	51.6
PGA-KD (w/o $\mathcal{L}_{\text{Intra}}$)	58.0	54.3	73.9	35.7	53.1
PGA-KD (w/o $\mathcal{L}_{\text{Inter}}$)	58.1	54.7	73.3	39.3	52.5
PGA-KD (w/o $\mathcal{L}_{\text{PGA}}, \mathcal{L}_{\text{Intra}}$)	56.6	53.1	68.4	<u>38.1</u>	51.5
PGA-KD (w/o $\mathcal{L}_{\text{PGA}}, \mathcal{L}_{\text{Inter}}$)	57.8	54.6	67.9	37.0	51.6
PGA-KD (w/o $\mathcal{L}_{\text{Intra}}, \mathcal{L}_{\text{Inter}}$)	<u>58.5</u>	53.6	74.8	37.1	<u>53.2</u>
PGA-KD (Full)	59.5	<u>54.6</u>	<u>74.6</u>	38.4	54.0

4.3.2 Does Spectral Filtering Matter?

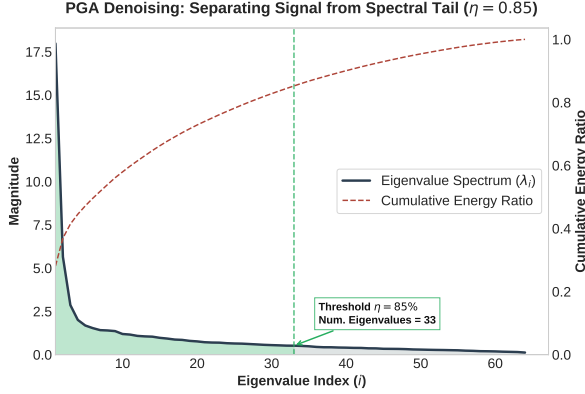


Figure 2: **Spectral Decay** ($B = 64$). The top 33 out of 64 eigenvalues capture $\eta = 85\%$ of the total energy, representing the Principal Geometry (green) versus the spectral noise (gray).

Figure 2 visualizes the eigenvalue distribution (λ_i) and cumulative energy of the teacher’s Gram matrix $\mathbf{G}^T$. As the heavy-tailed spectrum reveals, the top 33 out of 64 eigenvalues capture 85% of the total variance ($\eta = 0.85$). This indicates that nearly half of the representation space is dominated by low-magnitude components, which we treat as a low-energy *spectral tail* of lower semantic density. By filtering out this noise and retaining only the Principal Geometry, we distill a compact, semantically dense signal. This targeted alignment prevents the student from overfitting to geometric noise, a denoising process formally detailed in Appendix G.1.

To validate this noise hypothesis, we vary the cumulative energy threshold η , where $\eta = 1.0$ denotes full-spectrum alignment (standard CKA) and $\eta < 1.0$ applies spectral filtering.

Table 5: **Impact of Spectral Filtering.** Isolating the effect of varying the energy threshold η on performance.

Configuration	OK	A-OK	Doc	I-VQA	Avg. VQA
Student	50.7	50.2	74.0	34.1	50.8
$\eta = 1.00$ (No Filtering)	56.9	54.2	73.8	37.0	52.7
$\eta = 0.95$	<u>58.3</u>	<u>55.7</u>	<u>75.3</u>	<u>38.2</u>	<u>53.9</u>
$\eta = 0.85$ (Ours)	59.5	54.6	74.6	38.4	54.0
$\eta = 0.75$	57.9	53.5	73.6	36.8	52.5
$\eta = 0.50$	56.2	51.6	66.5	36.2	50.0

As shown in Table 5, the results follow an inverted-U curve. Under-filtering ($\eta = 1.0$) degrades performance, confirming that trailing eigenvalues contain non-transferable noise that disrupts distillation. Conversely, over-filtering ($\eta = 0.75$

or $\eta = 0.50$) removes vital structural information, lowering accuracy. The optimal balance is achieved at $\eta = 0.85$, which successfully isolates the Principal Geometry for stable alignment (see Appendix G.2 for more details).

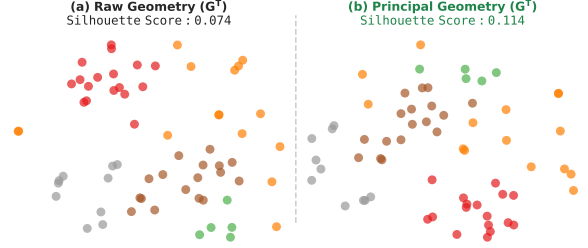


Figure 3: Geometric Denoising Effects ($B = 64$, $\eta = 0.85$). Spectral filtering enhances cluster purity and improves the Silhouette Score from 0.074 to 0.114 (+54%).

Figure 3 qualitatively supports this denoising effect via t-SNE algorithm (van der Maaten and Hinton, 2008). We map the embedding space by converting Gram matrices into Euclidean distances via a kernel-to-distance transform. To quantify cluster purity, we apply K-Means ($k = 5$) to generate pseudo-labels and compute the Silhouette Score (Rousseeuw, 1987). Filtering the spectral tail improves this score from 0.074 to 0.114 (+54%), demonstrating that our method effectively removes geometric noise for a cleaner structural signal (see Appendix G.3 for more details).

4.3.3 Sensitivity Analysis of Loss Weights

Table 6 evaluates hyper-parameter configurations on the FastVLM-0.5B track. To minimize computational costs, we bypassed a full grid search in favor of gradient scale balancing, constraining λ_{SCL} to 0.01 to prevent its four sub-losses from oversaturating the objective with large gradients.

As observed, increasing λ_{PGA} from 0.1 to 2.0 significantly boosts performance from 52.3 to 54.0, particularly in OK-VQA and DocVQA. While larger λ_{SCL} (0.05, 0.10) shows marginal task-specific shifts, it fails to improve average accuracy, validating our gradient-balancing constraint. Finally, a stronger feature alignment constraint ($\lambda_{\text{MSE}} = 1.0$) proves crucial for cross-task stability. Ultimately, the combination of $\lambda_{\text{PGA}} = 2.0$, $\lambda_{\text{SCL}} = 0.01$, and $\lambda_{\text{MSE}} = 1.0$ yields the optimal balance. An extended sensitivity analysis on architectural LoRA hyperparameters is provided in Appendix H.

Table 6: **Sensitivity Analysis of Loss Weights.** Evaluation of different combinations for hyper-parameters λ_{PGA} , λ_{SCL} , and λ_{MSE} on FastVLM-0.5B.

λ_{PGA}	λ_{SCL}	λ_{MSE}	OK	A-OK	Doc	I-VQA	Avg. VQA
0.1	0.01	1.0	57.3	54.7	69.4	37.5	52.3
0.5	0.01	1.0	58.8	54.6	70.9	37.1	52.8
2.0	0.05	1.0	58.9	55.4	75.1	36.8	54.0
2.0	0.10	1.0	58.2	55.5	74.2	36.9	53.6
2.0	0.01	0.2	57.4	54.3	75.8	37.9	53.4
2.0	0.01	0.5	58.9	54.1	75.8	38.4	53.9
2.0	0.01	1.0	59.5	54.6	74.6	38.4	54.0

5 Conclusion

This paper presents PGA-KD, a novel distillation framework that addresses geometric noise and semantic inconsistency in VLM2Vec distillation. By introducing Principal Geometry Alignment (PGA), we effectively filter out spectral noise to preserve the teacher’s core semantic structure. In addition, our Semantic Consistency Learning (SCL) utilizes multi-pathway mutual information maximization to stabilize cross-modal dependencies. Experimental results on the MMEB benchmark demonstrate that PGA-KD outperforms state-of-the-art baselines, achieving superior accuracy across all four meta-tasks in both In-Distribution (ID) and Out-of-Distribution (OOD) settings. These findings validate that aligning structural topology and semantic logic is key to developing robust, compact multimodal embedding models.

Limitations

While PGA-KD significantly reduces model size and memory footprint for multimodal embedding tasks, we acknowledge certain boundaries in our current scope. First, our empirical evaluations focus strictly on aligning vision and language modalities. Extending Principal Geometry Alignment to synchronize continuous data structures, such as video streams or 3D point clouds, remains an unexplored direction. Second, our current experiments validate the distillation process using up to 2B-parameter teacher models. Scaling this framework to compress extreme-scale architectures (e.g., 70B+ parameters) is currently bounded by computational resources. Consequently, expanding the framework to encompass broader modalities and massive-scale teacher models is left as an open work for future research.

Broader Impact

PGA-KD compresses large vision–language embedding models into compact students deployable on resource-constrained edge devices, lowering the compute, energy, and cost barriers to multimodal retrieval and VQA and broadening access to such systems, including for accessibility applications on low-resource hardware. The same efficiency, however, can lower the cost of large-scale image–text search, surveillance, and user profiling. Distillation also transfers whatever biases and failure modes the teacher and its training data (here, MMEB) have learned, and compression may degrade performance unevenly across rare classes, languages, and underrepresented visual concepts. Our experiments use only publicly available benchmark datasets and introduce no new data collection. We encourage practitioners to evaluate distilled students for fairness and robustness on their target domain before deployment.

Acknowledgments

Linh Ngo Van and Duc Anh Nguyen were funded by Vingroup Innovation Foundation (VINIF) under project code VINIF.2026.DA117. Thien Huu Nguyen was partially supported by NSF Grant #2239570. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government.

References

- Antonio Almod var, Jos  Miguel Hern ndez-Lobato, Sameer Khurana, Ricard Marxer, and Alfonso Ortega. 2025. [Aligning multimodal representations through an information bottleneck](#). In *Forty-second International Conference on Machine Learning*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sib  Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.
- 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.

- 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, San Diego, California, United States. Association for Computational Linguistics.
- 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, San Diego, California, United States. Association for Computational Linguistics.
- 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.
- Carl Eckart and Gale Young. 1936. [The approximation of one matrix by another of lower rank](#). *Psychometrika*, 1(3):211–218.
- Davide Ettori. 2026. [Spectral geometry for deep learning: Compression and hallucination detection via random matrix theory](#). *Preprint*, arXiv:2601.17357.
- Qianhan Feng, Wenshuo Li, Tong Lin, and Xinghao Chen. 2025. [Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language large model enhancement](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4178–4188.
- Ze Feng, Sen Yang, Boqiang Duan, Wankou Yang, and Jingdong Wang. 2026. [Em-kd: Distilling efficient multimodal large language model with unbalanced vision tokens](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 21111–21119.
- Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. [Knowledge distillation: A survey](#). *International journal of computer vision*, 129(6):1789–1819.
- 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*, 351:116606.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. [Distilling the knowledge in a neural network](#). *Preprint*, arXiv:1503.02531.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- 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, San Diego, California, United States. Association for Computational Linguistics.
- 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*.
- Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. [Similarity of neural network representations revisited](#). In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.
- Vinh Le, Nguyen Hong Dang, Tu Vu, Linh Ngo Van, Duc Anh Nguyen, and Trung Le. 2026. [HieRD: Hierarchical relational distillation for vision-language embedding models](#). In *Forty-third International Conference on Machine Learning*.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. [LLaVA-onevision: Easy visual task transfer](#). *Transactions on Machine Learning Research*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. [BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.
- Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2022b. [Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.
- Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, Yuxuan Han, Haijun Li, Wanying Chen, Junke Tang, Chengkun Hou, Zhixing

- Du, Tianli Zhou, Wenjie Zhang, Huping Ding, and 23 others. 2025. [Ovis2.5 technical report](#). *Preprint*, arXiv:2508.11737.
- 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.
- Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. [Relational knowledge distillation](#). In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3962–3971.
- 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.
- Peter J Rousseeuw. 1987. [Silhouettes: a graphical aid to the interpretation and validation of cluster analysis](#). *Journal of computational and applied mathematics*, 20:53–65.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. [Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter](#). *Preprint*, arXiv:1910.01108.
- Raghuveer Thirukovalluru, Rui Meng, Ye Liu, Karthikeyan K, Mingyi Su, Ping Nie, Semih Yavuz, Yingbo Zhou, Wenhui Chen, and Bhuwan Dhingra. 2025. [Breaking the batch barrier \(b3\) of contrastive learning via smart batch mining](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 23218–23241. Curran Associates, Inc.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- 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 7594–7606, Suzhou, China. Association for Computational Linguistics.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. [Representation learning with contrastive predictive coding](#). *Preprint*, arXiv:1807.03748.
- Laurens van der Maaten and Geoffrey Hinton. 2008. [Visualizing data using t-sne](#). *Journal of Machine Learning Research*, 9(86):2579–2605.
- Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Gransch, Oncel Tuzel, and Hadi Pouransari. 2025. [Fastvlm: Efficient vision encoding for vision language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (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 2026*, pages 3513–3527, Rabat, Morocco. Association for Computational Linguistics.
- 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*, 343:116001.
- 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.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025. [Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency](#). *Preprint*, arXiv:2508.18265.
- Alex Tianyi Xu, Alex Wilf, Paul Pu Liang, Alexander Obolenskiy, Daniel Fried, and Louis-Philippe Morency. 2025. [Comparative knowledge distillation](#). In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 7690–7699.
- Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. [Tinyllama: An open-source small language model](#). *Preprint*, arXiv:2401.02385.

A Related Work

Knowledge distillation (KD) (Hinton et al., 2015) aims to transfer knowledge from a large teacher model to a smaller student model while preserving the teacher’s performance and representational capacity. Early KD research primarily focused on compressing unimodal models by matching softened logits, intermediate representations, or attention patterns (Gou et al., 2021). In natural language processing, successful examples such as (Sanh et al., 2020) and (Zhang et al., 2024) demonstrated that compact students can retain strong reasoning and language understanding abilities distilled from large pretrained models like (Devlin et al., 2019) and (Touvron et al., 2023).

Recently, distillation has been extended to vision–language models (VLMs), where the challenge is substantially greater due to the need to jointly preserve both visual semantics and cross-modal alignment. Large VLMs such as (Radford et al., 2021), (Liu et al., 2023), and (Li et al., 2022a) learn a shared embedding space that aligns images and text through rich multimodal interactions. Distilling such models therefore requires the student not only to mimic the teacher’s representations, but also to maintain the structural relationships and semantic correspondences across modalities. As a result, VLM distillation has emerged as an active research direction with growing interest in preserving multimodal representation quality under compact architectures (Li et al., 2022b; Jiang et al., 2025).

Existing work on multimodal distillation mainly focuses on two aspects: preserving the geometric structure of the embedding space and maintaining cross-modal consistency. To preserve representation geometry, RKD (Park et al., 2019) aligns pairwise distances and triplet-wise angles between samples, while CKD (Xu et al., 2025) further models higher-order group relations through aggregation-based comparisons. These methods highlight the importance of preserving relational structures rather than directly matching individual features.

Another line of work focuses on retaining cross-modal interactions between visual and textual representations. In multimodal embedding frameworks such as VLM2Vec, the student must capture how textual concepts correspond to visual regions. Align-KD (Feng et al., 2025) addresses this through attention mimicking at early transformer layers, whereas EM-KD (Feng et al., 2026) aligns affinity

matrices that encode cross-modal correlations. Together, these approaches show that effective VLM distillation depends on preserving both geometric relations and multimodal alignment.

Most closely related to our setting, HieRD (Le et al., 2026) also distills vision-language embedding models, aligning relational structure within and across modalities through clustered visual tokens and phrase-level text spans; in contrast, PGA-KD operates on the batch-level teacher Gram geometry and filters its low-energy spectral tail before alignment through CKA. More broadly, recent work on distilling large language and embedding models explores cross-tokenizer alignment (Vu et al., 2026a), preference distillation (Nguyen et al., 2026), optimal-transport-based transfer (Vuong et al., 2026; Hai et al., 2026), span- and trajectory-level alignment (Dao et al., 2026a; Chi et al., 2026), and embedding-model representation alignment (Vu et al., 2026b; Tran et al., 2026; Dao et al., 2026b; Huy et al., 2026), several of which, like ours, rely on Centered Kernel Alignment to match relational structure rather than individual features. Our work extends this line to heterogeneous VLM2Vec distillation, combining principal-geometry alignment with multi-pathway semantic consistency.

B Theoretical Background and Foundations

This section provides the comprehensive mathematical derivations and theoretical formulations for the baseline relational distillation metrics and information-theoretic principles omitted from the main text.

B.1 Relational Distillation Metrics (RKD)

Relational Knowledge Distillation (RKD) (Park et al., 2019) decomposes the geometric structure of the representation space into distinct geometric primitives: distances and angles.

Let $d_{ij} = \|\mathbf{h}_i - \mathbf{h}_j\|_2$ denote the Euclidean distance between the embeddings of samples i and j . RKD constructs distance matrices for both teacher and student, normalizes them by their respective mean distances to handle scaling variances, and minimizes the element-wise difference using Huber loss l_δ :

$$\mathcal{L}_{dist} = \sum_{(i,j) \in \mathcal{R}^2} l_\delta \left(\frac{d_{ij}^S}{\mu^S}, \frac{d_{ij}^T}{\mu^T} \right) \quad (18)$$

where $\mu = \frac{1}{|\mathcal{R}^2|} \sum_{(i,j)} d_{ij}$ represents the mean distance within the batch.

To capture higher-order geometric structures beyond pairwise relationships, RKD introduces an angular loss over triplets of samples (i, j, k) :

$$\mathcal{L}_{angle} = \sum_{(i,j,k) \in \mathcal{R}^3} l_{\delta}(\cos \angle_{ijk}^{\mathcal{S}}, \cos \angle_{ijk}^{\mathcal{T}}) \quad (19)$$

where the angle cosine is calculated using the vectors between the samples:

$$\cos \angle_{ijk} = \frac{\langle \mathbf{h}_i - \mathbf{h}_j, \mathbf{h}_k - \mathbf{h}_j \rangle}{\|\mathbf{h}_i - \mathbf{h}_j\|_2 \|\mathbf{h}_k - \mathbf{h}_j\|_2} \quad (20)$$

Limitations of RKD vs. Selection of CKA: Although RKD effectively aligns relative magnitudes and orientations, the separation of distance and angle optimization requires careful weight balancing. Because these two metrics operate on different scales and converge at different rates, improper balancing causes the model to over-prioritize one primitive at the expense of the overall geometric structure.

In contrast, Centered Kernel Alignment (CKA) treats the Gram matrices $\mathbf{K}$ and $\mathbf{L}$ as holistic representations of the geometry. By computing the statistical dependence between these structures via normalized Hilbert-Schmidt Independence Criterion (HSIC), CKA captures distance, angular, and higher-order relations implicitly in a unified step without requiring rigid local constraints.

B.2 Information-Theoretic Foundations of InfoNCE

Mutual Information $I(X; Y)$ measures the reduction in uncertainty about a random variable X given knowledge of Y , accommodating complex non-linear dependencies. In contrastive learning frameworks, the InfoNCE loss serves as a variational lower bound on this quantity.

Following [van den Oord et al. \(2019\)](#), consider a batch of size B containing one positive pair (x_i, y_i) drawn from the joint distribution $p(x, y)$ and $B - 1$ negative samples drawn from the marginal distribution $p(y)$. The optimal critic $f(x, y) \propto \frac{p(x, y)}{p(y)}$ maximizes the objective. The standard InfoNCE loss is formulated as:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E} \left[\log \frac{f(x_i, y_i)}{\sum_{j=1}^B f(x_i, y_j)} \right] \quad (21)$$

By evaluating the expectation over the batch distributions, it can be mathematically shown that the

loss is bounded by the true mutual information:

$$\mathcal{L}_{\text{InfoNCE}} \geq -\mathbb{E} \left[\log \frac{\frac{p(x_i, y_i)}{p(y_i)}}{\frac{p(x_i, y_i)}{p(y_i)} + \sum_{j \neq i} \frac{p(x_i, y_j)}{p(y_j)}} \right] \quad (22)$$

$$\approx \log(B) - I(X; Y) \quad (23)$$

Rearranging this inequality yields the foundational lower bound used in our framework:

$$I(X; Y) \geq \log(B) - \mathcal{L}_{\text{InfoNCE}} \quad (24)$$

This relationship guarantees that minimizing $\mathcal{L}_{\text{InfoNCE}}$ across our intra- and inter-modal pathways directly maximizes the semantic information retained by the student model.

B.3 Theoretical Justification of Spectral Filtering

The efficacy of Principal Geometry Alignment (PGA) is predicated on a fundamental hypothesis regarding the spectral distribution of high-capacity Vision-Language Models: the representation space is composed of a low-rank intrinsic semantic manifold embedded within a high-dimensional ambient space dominated by architectural noise. In this section, we provide a rigorous justification for our spectral filtering mechanism by analyzing the signal-to-noise decomposition of the teacher’s geometry through the lens of matrix approximation theory.

Signal-Noise Decomposition Hypothesis. Let $\mathbf{H}^{\mathcal{T}} \in \mathbb{R}^{B \times d_{\mathcal{T}}}$ denote the teacher’s representation of a batch, where $d_{\mathcal{T}}$ is large (e.g., 1536). The geometric topology of the batch is encapsulated in the Gram matrix $\mathbf{G}^{\mathcal{T}} = \mathbf{H}^{\mathcal{T}} (\mathbf{H}^{\mathcal{T}})^{\top}$. We posit that $\mathbf{G}^{\mathcal{T}}$ can be decomposed into a semantic signal component $\mathbf{G}_{sem}$ and a nuisance noise component $\mathbf{E}_{noise}$, such that $\mathbf{G}^{\mathcal{T}} = \mathbf{G}_{sem} + \mathbf{E}_{noise}$. Here, $\mathbf{G}_{sem}$ is assumed to be low-rank ($k \ll B$), capturing the true underlying cluster structure and semantic correlations of the data, whereas $\mathbf{E}_{noise}$ represents high-frequency spectral fluctuations arising from the teacher’s over-parameterization and stochastic training artifacts.

In the context of heterogeneous distillation, the student model $\mathcal{S}$ is constrained by a significantly lower dimensionality ($d_{\mathcal{S}} \ll d_{\mathcal{T}}$). Consequently, the student lacks the requisite degrees of freedom to faithfully reconstruct the full spectrum of $\mathbf{G}^{\mathcal{T}}$. A standard distillation objective that minimizes the distance to the raw Gram matrix forces the student

to allocate its limited capacity to fit both $\mathbf{G}_{sem}$ and $\mathbf{E}_{noise}$. This results in a *capacity dilution* phenomenon, where the student’s representation becomes contaminated by non-transferable architectural noise, thereby degrading its generalization capability on downstream tasks.

Optimality via Eckart-Young-Mirsky Theorem. To resolve this capacity mismatch, PGA employs a spectral filtering strategy. Let the eigendecomposition of the teacher’s Gram matrix be given by $\mathbf{G}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, with eigenvalues sorted in descending order $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_B$. We construct the filtered principal geometry $\hat{\mathbf{G}}_k^T$ by retaining only the top- k components corresponding to the cumulative energy threshold η . According to the Eckart-Young-Mirsky Theorem (Eckart and Young, 1936), the truncated matrix $\hat{\mathbf{G}}_k^T$ is the unique optimal rank- k approximation of the original geometry in terms of the Frobenius norm:

$$\hat{\mathbf{G}}_k^T = \arg \min_{\mathbf{M} \in \mathbb{R}^{B \times B}, \text{rank}(\mathbf{M})=k} \|\mathbf{G}^T - \mathbf{M}\|_F \quad (25)$$

The approximation error incurred by this truncation is explicitly quantified by the tail energy of the spectrum:

$$\|\mathbf{G}^T - \hat{\mathbf{G}}_k^T\|_F^2 = \sum_{i=k+1}^B \lambda_i^2 \quad (26)$$

In the regime of Large VLMs, the eigenspectrum typically exhibits a heavy tail where the trailing eigenvalues $(\lambda_{k+1}, \dots, \lambda_B)$ carry non-negligible magnitude but low energy relative to the principal components, so that a compact student cannot faithfully absorb them. By minimizing the divergence between the student’s geometry and this optimal approximation $\hat{\mathbf{G}}_k^T$, PGA concentrates the student’s limited optimization budget on the principal directions $(\lambda_1, \dots, \lambda_k)$ rather than the low-energy spectral tail. We stress the scope of this result: the Eckart-Young-Mirsky theorem certifies only that $\hat{\mathbf{G}}_k^T$ is the Frobenius-optimal rank- k approximation of $\mathbf{G}^T$; it does not by itself establish that the retained subspace is semantically superior or that aligning to it improves generalization. Identifying the top- k subspace with the transferable semantic signal is a modeling hypothesis, and the claim that aligning to it benefits a compact student is empirical, supported by the inverted-U response to the energy threshold η (Table 5), the improved silhouette score after filtering (Figure 3), and the OOD gains in Table 3.

C Implementation Details

In this section, we provide comprehensive details regarding the model architectures, training configurations, baseline implementations, and dataset statistics to ensure the reproducibility of our results.

C.1 Dataset Specifications and Task Formulations

We evaluate the performance of our framework across a comprehensive suite of datasets from the MMEB benchmark (Jiang et al., 2025), extended with specialized grounding and robustness tracks. These evaluation benchmarks are structured into four unified meta-tasks:

Image Classification (10 Tasks): ImageNet-1K (general-domain object classification), N24News (multimodal news category classification), Hateful-Memes (binary multimodal hate speech detection), VOC2007 (multi-label object category classification), SUN397 (fine-grained scene recognition), Place365 (environment and scene context recognition), ImageNet-A (natural adversarial examples), ImageNet-R (abstract renditions and art styles), ObjectNet (robustness controls for rotation, background, and viewpoint shifts), and Country-211 (geographic distribution shift and visual culture analysis).

Visual Question Answering (VQA) (10 Tasks): OK-VQA (visual QA requiring external knowledge), A-OKVQA (augmented outside-knowledge QA with multiple choice), DocVQA (document text visual comprehension), Infographic VQA (infographic-based structural reasoning), ChartQA (structured chart and plot analysis), Visual7W (spatial grounding and visually-grounded QA), ScienceQA (multimodal scientific question answering), VizWiz (real-world QA using low-quality images from visually impaired users), GQA (compositional and structured reasoning over real-world scenes), and TextVQA (visual reading and reasoning in natural environments).

Cross-Modal Retrieval (12 Tasks): VisDial (dialogue-based multi-turn visual retrieval), CIRR (composed image retrieval via text modifiers), VisualNews-t2i (news-oriented text-to-image retrieval), VisualNews-i2t (news-oriented image-to-text retrieval), MSCOCO-t2i (text-to-image instance search), MSCOCO-i2t (image-to-text instance search), NIGHTS (retrieval under low-light/nocturnal environments), WebQA (multimodal knowledge-grounded web retrieval), OVEN

Table A1: **Dataset Statistics.** Comprehensive breakdown of meta-tasks, data splits (In-Distribution vs. Out-of-Distribution), query/target modalities, training sizes, evaluation samples, and candidate spaces across all evaluation benchmarks.

Meta-Task	Dataset	Query	Target	Split	# Train	# Eval	# Candidates
Classification (10 Tasks)	ImageNet-1K	I	T	ID	100K	1,000	1,000
	N24News	I + T	I	ID	49K	1,000	24
	HatefulMememes	I	T	ID	8K	1,000	2
	VOC2007	I	T	ID	8K	1,000	20
	SUN397	I	T	ID	20K	1,000	397
	Place365	I	T	OOD	-	1,000	365
	ImageNet-A	I	T	OOD	-	1,000	1,000
	ImageNet-R	I	T	OOD	-	1,000	200
	ObjectNet	I	T	OOD	-	1,000	313
	Country-211	I	T	OOD	-	1,000	211
VQA (10 Tasks)	OK-VQA	I + T	T	ID	9K	1,000	1,000
	A-OKVQA	I + T	T	ID	17K	1,000	1,000
	DocVQA	I + T	T	ID	40K	1,000	1,000
	InfographicVQA	I + T	T	ID	24K	1,000	1,000
	ChartQA	I + T	T	ID	28K	1,000	1,000
	Visual7W	I + T	T	ID	70K	1,000	1,000
	ScienceQA	I + T	T	OOD	-	1,000	1,000
	VizWiz	I + T	T	OOD	-	1,000	1,000
	GQA	I + T	T	OOD	-	1,000	1,000
	TextVQA	I + T	T	OOD	-	1,000	1,000
Retrieval (12 Tasks)	VisDial	T	I	ID	123K	1,000	1,000
	CIRR	I + T	I	ID	26K	1,000	1,000
	VisualNews-t2i	T	I	ID	100K	1,000	1,000
	VisualNews-i2t	I	T	ID	100K	1,000	1,000
	MSCOCO-t2i	T	I	ID	100K	1,000	1,000
	MSCOCO-i2t	I	T	ID	113K	1,000	1,000
	NIGHTS	I	I	ID	16K	1,000	1,000
	WebQA	T	I + T	ID	17K	1,000	1,000
	OVEN	I + T	I + T	OOD	-	1,000	1,000
	FashionIQ	I + T	I	OOD	-	1,000	1,000
	EDIS	T	I + T	OOD	-	1,000	1,000
	Wiki-SS-NQ	T	I	OOD	-	1,000	1,000
Visual Grounding (4 Tasks)	MSCOCO	I + T	I	ID	100K	1,000	1,000
	Visual7W-Pointing	I + T	I	OOD	-	1,000	1,000
	RefCOCO	I + T	I	OOD	-	1,000	1,000
	RefCOCO-Matching	I + T	I + T	OOD	-	1,000	1,000

(open-vocabulary visual embedding and retrieval), FashionIQ (interactive fashion image retrieval with relative text feedback), EDIS (early disaster identification and structural multi-modal search), and Wiki-SS-NQ (wikipedia-scale structured snippet natural question retrieval).

Visual Grounding (4 Tasks): MSCOCO (phrase-to-region localization and object referring expression comprehension), Visual7W-Pointing (visually-grounded point-and-click localization), RefCOCO (referring expression bounding box grounding), and RefCOCO-Matching (multi-modal phrase-region context matching and verification).

To provide a clear mathematical and structural understanding of our evaluation protocol, which strictly follows the comprehensive multi-task evaluation framework established by VLM2Vec (Jiang

et al., 2025), Table A1 details the comprehensive statistical breakdown of these datasets. Depending on the underlying meta-task, the query formulations and candidate space constraints are defined as follows:

For Image Classification, the input consists of an image, which may optionally be accompanied by a task instruction prompt. The target output is the corresponding class label (or a representative visual category, depending on the dataset). Under this setup, the candidate label space is explicitly defined by, and directly limited to, the total number of semantic classes present within that specific dataset.

For Visual Question Answering (VQA), the query comprises a multimodal combination of an instruction, an image, and a textual question, with

the target being the correct textual answer. During evaluation, the model is tasked with ranking the ground-truth answer against a standardized pool of 999 distractor candidates, yielding a total candidate space of 1,000 options.

For Cross-Modal Retrieval, the task encompasses both text-to-image ($T \rightarrow I$) and image-to-text ($I \rightarrow T$) sub-tasks. The input query can be a pure image, a pure textual description, or a combined image-text pair, whereas the target is the corresponding match in the destination modality. Performance is measured by ranking the correct target against a fixed evaluation pool of 1,000 candidate items.

For Visual Grounding, the query consists of a combined image-text pair containing localized instructions or referring expressions. The model is optimized to correctly identify and match the grounded visual component (e.g., specific image regions or bounding boxes). Mirroring the VQA evaluation setup, the model ranks the true target out of a standardized 1,000-candidate retrieval space.

C.2 Training Configurations and Hyperparameters

All distillation training pipelines are implemented in PyTorch and executed across a distributed computing infrastructure consisting of $8 \times$ NVIDIA A100 (80GB) GPUs. To maximize computational throughput and minimize memory footprints, we leverage distributed data-parallel training under bfloat16 mixed-precision execution. Parameter-Efficient Fine-Tuning (PEFT) is consistently applied to all student architectures via Low-Rank Adaptation (LoRA) (Hu et al., 2022). Every configuration is trained for a single epoch over the unified cross-modal alignment sequence using the AdamW optimizer with a decoupled weight decay parameter of 0.01. The base model layers are optimized with a peak learning rate of 1×10^{-4} , whereas the specialized projection heads ($\mathbf{W}_{\text{proj}}, \phi_v, \phi_t$) utilize a higher learning rate of 5×10^{-4} governed by a constant learning rate scheduler.

To ensure stability across heterogeneous task distributions, optimization properties are customized for each evaluation track and meta-task. Specifically, the scaling hyperparameters for the mean squared error (λ_{MSE}), Principal Geometry Alignment (λ_{PGA}), and Semantic Consistency Learning (λ_{SCL}) are calibrated alongside the spectral energy threshold (η) to protect the student model from

overfitting to the teacher’s architectural noise. A comprehensive specification of these hyperparameter configurations, including global batch sizes, learning rate constraints, and task-specific LoRA ranks, is detailed in Table A2.

C.3 Model Architectures and Projector Formulations

To ensure full reproducibility, we detail the specific architectural specifications of the teacher and student models derived from their official configurations, alongside the mathematical formulations of our learning framework’s projection layers. In this work, we employ a fixed, high-capacity Teacher model to distill knowledge into two distinct Student architectures, corresponding to the two primary experimental tracks. A parametric comparison detailing the architectural gaps between the teacher and both student models is provided in Table A3.

Teacher Model (B3-Qwen2-2B): This state-of-the-art multimodal embedding model is built upon the Qwen2-VL-2B architecture designed for comprehensive visual understanding (Thirukovalluru et al., 2025). It integrates a Qwen2-VL Vision Tower (a ViT-like architecture with 600M parameters, depth 32, and an embedding dimension of 1280) with a Qwen2-1.5B LLM backbone (28 layers, 12 attention heads, and a hidden dimension $d_T = 1536$). The architecture utilizes Multimodal Rotary Positional Embeddings (M-RoPE) to handle unified spatio-temporal positions and implements a final token pooling strategy to generate fixed-size embeddings.

Student 1 (FastVLM-0.5B): An efficiency-focused architecture optimized for resource-constrained edge deployment (Vasu et al., 2025). It replaces standard large ViTs with a compact MobileCLIP-L (mobileclip_l_1024) CNN-Transformer hybrid vision encoder (embedding dimension 1024). This vision tower is paired with a Qwen2-0.5B language backbone (24 layers, 14 attention heads, and a hidden dimension $d_S = 896$), using a 2-layer MLP (mlp2x_gelu) to map the concatenated intermediate visual features (dim 3072) to the LLM space.

Student 2 (LLaVA-OneVision-0.5B): A robust student configuration built following the LLaVA-OneVision paradigm (Li et al., 2025). It features a SigLIP-SO400M visual encoder (siglip_vision_model, depth 26, patch size 14, and a hidden dimension of 1152) matched with a Qwen2-0.5B-Instruct language backbone (24 lay-

Table A2: **Training Hyperparameters.** Detailed configurations for Classification (CLS), Visual Question Answering (VQA), Cross-Modal Retrieval (RET), and Visual Grounding (GRD) tasks across the FastVLM track, and CLS/VQA tasks for the OneVision track.

Hyperparameter	FastVLM Track				OneVision Track	
	CLS	VQA	RET	GRD	CLS	VQA
Global Batch Size	256	256	128	128	128	64
Learning Rate (Student)	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
Learning Rate (Projector)	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}
LR Scheduler	Constant	Constant	Constant	Constant	Constant	Constant
Warmup Ratio	1.0	0.03	0.03	0.03	0.03	0.03
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
Weight Decay	0.01	0.01	0.01	0.01	0.01	0.01
Epochs	1	1	1	1	1	1
LoRA Configuration						
Rank (r)	64	128	128	128	16	32
Alpha (α)	128	512	256	128	32	128
Dropout	0.05	0.05	0.05	0.05	0.05	0.05
Loss Weights						
λ_{MSE}	1.0	1.0	1.0	1.0	1.0	1.0
λ_{PGA}	1.0	1.0	1.0	1.0	1.0	1.0
λ_{SCL}	0.5	1.0	1.0	1.0	0.5	1.0
Spectral Threshold (η)	0.85	0.85	0.85	0.85	0.85	0.85

Table A3: **Architecture Specification Comparison.** The structural and dimensional gaps between the generative Teacher (Qwen2-VL) and the compact Students (FastVLM and OneVision) highlight the non-trivial nature of the distillation task.

Specification	Teacher	Student 1	Student 2
Base Arch	Qwen2-VL-2B	FastVLM	OneVision
Vision Encoder	Qwen2-VL-ViT	MobileCLIP-L	SigLIP-SO400M
Vision Layers	32	—	26
Vision Dim	1280	1024	1152
LLM Backbone	Qwen2-1.5B	Qwen2-0.5B	Qwen2-0.5B
LLM Layers	28	24	24
LLM Hidden (d)	1536 (d_T)	896 (d_S)	896 (d_S)
Interm. Size	8960	4864	4864

ers, 14 attention heads, and a hidden dimension $d_S = 896$). It employs an AnyRes dynamic gridding strategy with 2D pooling to process varying image aspect ratios.

Alignment Projectors: To bridge the dimensional and semantic gaps between the heterogeneous teacher and student representation spaces during training, we introduce learnable linear projection heads. These components are strictly frozen after the distillation phase and remain inactive during downstream inference. Specifically, a geometric Feature Projector defined as a linear layer $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d_T \times d_S}$ maps the teacher’s final embedding to the student’s dimension for the $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{PGA}}$ objectives. Additionally, two separate Multi-Layer Perceptrons (MLPs), denoted as Modality Projectors (ϕ_v, ϕ_t), are utilized to map the teacher’s

aggregated visual and textual channels independently into the student’s corresponding space for the $\mathcal{L}_{\text{SCL}}$ objective.

C.4 Baselines

We compare PGA-KD against a comprehensive set of distillation baselines, categorized by their alignment strategy:

MSE (Point-wise): A standard distillation baseline where the student’s embeddings are regressed directly to the teacher’s projected embeddings using Mean Squared Error loss.

Relational KD (RKD) (Park et al., 2019): A structural distillation method that preserves the geometric topology of the batch. It aligns the pairwise Euclidean distances and triplet angles between the teacher and student representations.

Comparative KD (CKD) (Xu et al., 2025): A generalized structural approach that aligns higher-order relationships. It defines a comparative geometry by computing the relative similarity of each sample against a set of anchors within the batch.

EMO (Truong et al., 2025): Originally designed for cross-tokenizer LLM distillation, this method is adapted for VLM2Vec by aligning embedding distributions.

EM-KD (Feng et al., 2026): A recent multi-modal knowledge distillation method designed to address the imbalance between visual and textual tokens. It achieves this by constructing and aligning cross-modal affinity matrices to improve fea-

ture synchronization.

D Mathematical Analysis of EMO Performance Degradation in Cross-Modal Retrieval

As observed in Table 1, the EMO distillation method exhibits a paradoxical performance profile. While it achieves competitive results on generative meta-tasks (Classification, VQA, and Grounding), it suffers from catastrophic degradation on Retrieval tasks. This failure is particularly severe in Text-to-Image ($T \rightarrow I$) retrieval benchmarks, where performance collapses to near zero (e.g., VN-t2i drops to 0.5%, COCO-t2i to 3.9%). This section provides a step-by-step mathematical analysis of the underlying mechanisms driving this discrepancy.

D.1 The Text-Centric Masking Mechanism

The core issue stems from EMO’s optimization objective. EMO was originally designed for Large Language Models (LLMs) and strictly optimizes the textual representation space.

Formally, let us define the representations of a Vision-Language Model:

- $\mathcal{S}$ and $\mathcal{T}$: The Student and Teacher models, respectively.
- $H^{\mathcal{S}} = [H_v^{\mathcal{S}}, H_t^{\mathcal{S}}]$: The full sequence of hidden states generated by the student, where $H_v^{\mathcal{S}}$ represents the visual tokens and $H_t^{\mathcal{S}}$ represents the textual tokens.
- $H^{\mathcal{T}} = [H_v^{\mathcal{T}}, H_t^{\mathcal{T}}]$: The corresponding hidden states generated by the teacher.

During distillation, EMO applies a strict masking mechanism to filter out all visual tokens (H_v). The distillation loss ($\mathcal{L}_{\text{EMO}}$), which consists of Optimal Transport (OT) and Centered Kernel Alignment (CKA) for attention maps, is exclusively computed on the textual tokens:

$$\mathcal{L}_{\text{EMO}} = \mathcal{L}_{\text{OT}}(H_t^{\mathcal{S}}, H_t^{\mathcal{T}}) + \mathcal{L}_{\text{CKA}}(A_t^{\mathcal{S}}, A_t^{\mathcal{T}}) \quad (27)$$

where A_t denotes the attention weights associated with the text tokens. Consequently, the visual representations ($H_v^{\mathcal{S}}$) receive absolutely no direct alignment supervision from the teacher.

D.2 Why EMO Succeeds in Generative Tasks ($I + T \rightarrow T$)

Despite ignoring visual tokens in the loss function, EMO succeeds in Classification, VQA, and Grounding because these are inherently Autoregressive Text Generation tasks.

In a Transformer architecture, tokens do not exist in isolation. During the forward pass, the textual tokens $H_t^{\mathcal{S}}$ continuously aggregate information from the visual tokens $H_v^{\mathcal{S}}$ via self-attention mechanisms. By forcing $H_t^{\mathcal{S}} \approx H_t^{\mathcal{T}}$, EMO essentially imposes a strict constraint on the output text representation. To minimize $\mathcal{L}_{\text{EMO}}$, the gradients must flow backward through the cross-attention weights to the vision encoder. The student is mathematically forced to indirectly adjust how it processes the image ($H_v^{\mathcal{S}}$) so that its final text tokens perfectly match the teacher’s. This implicit alignment allows the student to excel at generating text conditioned on images.

D.3 The Modality Gap Collapse in Retrieval Tasks

Retrieval tasks fundamentally differ from generative tasks. They rely on a Global Contrastive Space, requiring the model to measure the cosine similarity between a globally aggregated text vector ($\bar{t}$) and a globally aggregated image vector ($\bar{v}$). Both vectors must reside in a unified, aligned latent space.

When EMO applies massive gradient updates exclusively to the textual space (H_t), it creates a severe asymmetric shift:

Text-to-Image ($T \rightarrow I$) Catastrophe: The query is text, and the target is an image. Because EMO forcefully shifts the student’s text space towards the teacher’s space while leaving the image space behind, the modality gap is widened. A text query is displaced to a completely new coordinate in the latent space, meaning its mathematical dot product with the unshifted target image vector yields random noise. This explains the catastrophic drop to 0.5%.

Image-to-Text ($I \rightarrow T$) Degradation: When the image acts as the query, its vector representation remains relatively stable. It projects into the newly shifted text space. Because the text space is organized sensibly (inherited from the teacher), the image query can still occasionally land near semantically relevant text clusters by chance. Thus, performance degrades severely (e.g., dropping from

58.6% to 33.3% on COCO-i2t) but does not hit absolute zero.

D.4 The Necessity of Semantic Consistency Learning (SCL)

The failure of EMO underscores a critical principle: asymmetrically distilling a single modality irreversibly shatters the cross-modal contrastive space.

This highlights the architectural necessity of our proposed Semantic Consistency Learning ($\mathcal{L}_{\text{SCL}}$). Rather than masking modalities, $\mathcal{L}_{\text{SCL}}$ actively prevents modality collapse by enforcing Mutual Information maximization across all four logical pathways: intra-modal (Text-to-Text, Image-to-Image) and inter-modal (Text-to-Image, Image-to-Text). By explicitly regularizing the cross-modal bridges ($\mathcal{L}_{\text{Inter}}$), SCL guarantees that both modalities shift in unison during distillation, preserving the tightly coupled joint embedding space required for robust retrieval performance.

E Computational Complexity and Efficiency Analysis

To evaluate the operational viability of the proposed PGA-KD framework in practical training environments, we conduct a rigorous computational complexity and training overhead analysis.

E.1 Theoretical Time Complexity

The mathematical core of Principal Geometry Alignment (PGA) involves an eigendecomposition executed on the $B \times B$ teacher similarity matrix $\mathbf{G}^T$. Theoretically, the time complexity of performing eigendecomposition on a square matrix of size B scales at $\mathcal{O}(B^3)$. Because distillation pipelines in Vision-Language Models typically leverage small to moderate batch sizes (e.g., $B = 64$), this operation is highly efficient and computes in less than 1ms per training step. Concurrently, the Semantic Consistency Learning (SCL) module relies entirely on heavily parallelized, highly optimized GPU matrix multiplications, which introduces minimal computational friction.

E.2 Empirical Training Overhead Table

To quantify the concrete runtime impact, we benchmark the exact duration of a single training step (measured in milliseconds, ms) across various distillation configurations on the FastVLM-0.5B track. The empirical performance profiling is structured and detailed in Table A4.

Table A4: **Computational Overhead Comparison.** Evaluation of training step durations and relative overhead metrics across different distillation strategies, measured during Visual Grounding task training with a batch size of 16.

Method	Training Step (ms)	Overhead
Student (Baseline)	1976	1.00×
+ MSE	2475	1.25×
+ RKD (Park et al., 2019)	2332	1.18×
+ CKD (Xu et al., 2025)	2768	1.40×
+ EM-KD (Feng et al., 2026)	2473	1.25×
+ PGA-KD (Ours)	2424	1.23×

The empirical benchmarks demonstrate that PGA-KD is highly efficient, introducing only a 23% computational overhead relative to the unregularized student baseline training loop (1976ms $\rightarrow$ 2424ms). Crucially, our framework achieves a lower runtime profile than traditional feature-level regression baselines like MSE (1.25×) and group-level relational methods like CKD (1.40×). This efficiency highlights that filtering out the spectral tail reduces the dimensional load during downstream alignment calculations, proving that PGA-KD secures state-of-the-art accuracy gains without forcing a trade-off in training throughput.

F Variance Analysis under Multiple Seeds

Method	SciQA	GQA	TextVQA	Avg
Teacher	40.7	68.1	80.0	62.9
Student	35.8	63.4	56.7	52.0
MSE	37.0 ± 1.1	64.0 ± 1.6	60.1 ± 1.5	53.7 ± 0.8
EM-KD	33.8 ± 2.4	64.3 ± 1.1	56.2 ± 1.2	51.4 ± 1.5
PGA-KD	37.0 ± 0.5	66.9 ± 0.8	67.0 ± 1.4	57.0 ± 0.4

Table A5: OOD VQA across three random seeds (mean $\pm$ std) on FastVLM-0.5B, over the SciQA/GQA/TextVQA subset; Avg is the mean of these three tasks. Teacher and Student are single-run references.

To confirm that our gains are stable rather than seed-dependent, we repeat the OOD VQA evaluation with three random seeds ($\{42, 1234, 2026\}$) for PGA-KD and the two strongest distillation baselines, MSE and EM-KD. We choose the OOD track because it fluctuates more than the in-distribution setting, making it a stricter test of stability. Table A5 reports mean and standard deviation across seeds over a representative subset of tasks. Note that the averages here are computed over this three-task subset and are therefore not directly compara-

ble to the four-task OOD VQA averages in Table 3. PGA-KD’s three-seed mean (57.0) exceeds every baseline’s mean (MSE 53.7, EM-KD 51.4), and even its lowest single run (56.8) stays above them. PGA-KD leads on GQA and TextVQA, ties the best baseline on ScienceQA, and has the smallest variance among the distillation methods, indicating that its advantage is not driven by favorable initialization.

G Geometric Analysis and Visualizations

This section provides the mathematical foundations and implementation details for the geometric visualizations used to validate our spectral denoising approach.

G.1 Principal Geometry Analysis (PGA) Denoising

The objective of this analysis is to isolate the structural signal from the spectral noise in the teacher’s representation space. We utilize the Gram matrix $\mathbf{G}^\mathcal{T} \in \mathbb{R}^{B \times B}$, which encodes the pairwise geometric relationships between B data samples within a batch.

Mathematical Foundation: Given the Gram matrix $\mathbf{G}^\mathcal{T}$, we perform eigen-decomposition as follows:

$$\mathbf{G}^\mathcal{T} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top \quad (28)$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_B)$ contains the eigenvalues sorted in descending order ($\lambda_1 \geq \lambda_2 \dots \geq \lambda_B$). To distinguish between the Principal Geometry and the *spectral tail*, we define the Cumulative Energy Ratio (CER) at index k :

$$\text{CER}(k) = \frac{\sum_{i=1}^k \lambda_i}{\sum_{j=1}^B \lambda_j} \quad (29)$$

Implementation Details: As illustrated in Figure 2, we set an energy threshold $\eta = 0.85$. The threshold k is determined by finding the smallest k such that $\text{CER}(k) \geq 0.85$.

- **Spectrum Extraction:** Eigenvalues are computed and clipped at $\max(\lambda, 10^{-10})$ to ensure numerical stability.
- **Signal Separation:** In our experiments, the first $k = 33$ eigenvalues constitute the principal geometry, while the remaining spectral tail is treated as noise.

G.2 Spectral Filtering Effect on Gram Matrix

To visually assess how spectral filtering alters the teacher model’s representations, we analyze the reconstructed Gram matrices and their corresponding removed noise across various cumulative energy thresholds η . Figure A1 reveals the impact of varying the filtering aggressiveness:

Aggressive Over-filtering ($\eta = 0.5$): Retaining only 3 eigenvalues heavily distorts the Gram matrix structure. The bottom row indicates that the removed component is not merely background noise, but contains critical geometric structures (reflected by high-intensity patterns in the noise matrix), leading to a severe loss of semantic information.

Optimal Denoising ($\eta = 0.85$): At our chosen threshold, the reconstructed matrix perfectly retains the crisp cluster definitions and blocking patterns of the original ground truth. Crucially, the removed noise matrix contains uniform, low-intensity residual patterns with minimal structural correlation, confirming that our method effectively isolates non-transferable architectural noise.

Under-filtering ($\eta = 0.95$): Keeping 24 eigenvalues preserves almost all elements, meaning the remaining noise matrix is virtually empty. Consequently, this allows the student to absorb the non-essential structural fluctuations present in the full spectrum.

G.3 Representation Visualization and Cluster Purity

To quantify the improvement in representation quality after spectral filtering, we project the high-dimensional space into a 2D space and measure clustering quality.

Distance Mapping: We convert the Gram matrices (Raw $\mathbf{G}^\mathcal{T}$ and Principal $\hat{\mathbf{G}}^\mathcal{T}$) into distance matrices $\mathbf{D}$ using the kernel-to-distance transform:

$$\mathbf{D}_{ij} = \sqrt{\mathbf{G}_{ii}^\mathcal{T} + \mathbf{G}_{jj}^\mathcal{T} - 2\mathbf{G}_{ij}^\mathcal{T}} \quad (30)$$

This matrix $\mathbf{D}$ serves as the input for the t-SNE algorithm (van der Maaten and Hinton, 2008).

Clustering and Quantification: We utilize the Silhouette Score (Rousseeuw, 1987) (S) to measure cluster separation, calculated as:

$$S = \frac{b - a}{\max(a, b)} \quad (31)$$

where a represents the mean intra-cluster distance and b is the mean nearest-cluster distance. Pseudo-labels for this calculation are generated via K-Means clustering ($n = 5$) on the denoised space.

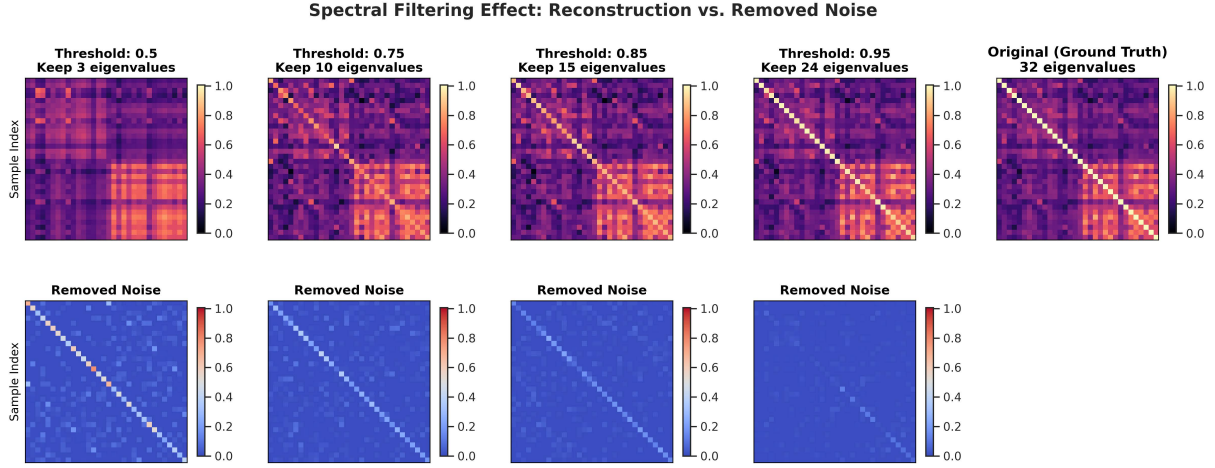


Figure A1: **Spectral Filtering Effect.** The top row displays the reconstructed Gram matrices ($\hat{G}^T$) under different energy thresholds ($\eta \in \{0.5, 0.75, 0.85, 0.95\}$) compared to the original ground truth (32 eigenvalues). The bottom row visualizes the isolated spectral noise removed during filtering at each respective threshold.

Observations: Comparing the two panels in Figure 3:

- Raw Geometry (G^T): Displays significant overlap and diffusion between semantic groups, yielding a Silhouette Score of 0.074.
- Principal Geometry ($\hat{G}_k^T$): After removing the spectral tail, the clusters become significantly more condensed and separated, with the Silhouette Score improving to 0.114.

H Sensitivity Analysis on LoRA Hyperparameters

Table A6: **Effect of LoRA Rank on Distillation.** Evaluation of subset benchmarks and overall average performance.

Rank (r)	OK	A-OK	Doc	I-VQA	Avg. VQA
2	0.1	0.2	0.1	0.0	0.1
4	30.4	27.4	15.5	10.8	22.3
8	48.8	43.4	50.3	22.0	40.3
16	56.0	50.8	69.5	31.8	49.2
64	56.9	53.1	74.4	<u>37.5</u>	52.5
128	59.5	54.6	<u>74.6</u>	38.4	54.0
256	<u>57.7</u>	53.4	74.9	37.3	<u>53.5</u>
512	57.7	<u>54.5</u>	74.0	37.4	53.2

Distilling a high-capacity generative teacher into a compact student architecture (such as FastVLM-0.5B) demands a critical balance of representational capacity within the parameter-efficient adapters. As illustrated by the empirical results in Table A6, the student’s downstream performance

exhibits a strong sensitivity to the scaling of the rank r and its corresponding scaling factor α .

When configured with restricted ranks ($r \leq 8$), the student encounters a strict capacity bottleneck. Under these constraints, the adapter layer lacks the sufficient degrees of freedom required to capture the rich structural knowledge transferred from the teacher, causing the average VQA scores to stall below the 41% threshold. However, a substantial performance inflection occurs as the rank scales to $r = 16$, where the average score rapidly surpasses 49%, indicating that a baseline model capacity is quickly established.

The framework achieves its optimal convergence and peak representational accuracy at $r = 128$ ($\alpha = 512$), yielding an average VQA score of 54.0%. Beyond this optimal configuration, the model reaches a state of capacity saturation. While expanding the parameter space to $r = 256$ maintains a stable performance of 53.5%, a further aggressive increase to $r = 512$ ($\alpha = 896$) induces a notable performance degradation, with the average score dropping to 53.2%. This trajectory suggests that allocating excessive capacity within the LoRA adapters risks overparameterization, which overfits the student to the specific cross-modal transfer distribution and compromises its generalization capabilities across unseen VQA benchmarks.