HieRD: Hierarchical Relational Distillation for Vision-Language Embedding Models

Anonymous Authors[1]

Abstract

Knowledge distillation is crucial for compressing large Vision–Language Models (VLMs) into efficient architectures. While prior VLM research has primarily focused on reasoning tasks like visual question answering, multimodal embedding learning, a key component for large-scale retrieval, has received comparatively less attention. Existing distillation methods typically align static global representations, overlooking hierarchical feature structure and fine-grained cross-modal interactions. This leads to a structural gap where student models fail to inherit object-level semantics and spatial relationships from teachers. To address this limitation, we propose **HieRD**, a Hierarchical Representation Distillation framework that preserves hierarchical structure within and across modalities throughout the distillation process by leveraging clustered visual tokens and multi-granular alignment with phrase-level text. Experimental results on multimodal embedding and downstream tasks show that HieRD consistently outperforms strong baselines, reflecting the effectiveness of its fine-grained semantic and spatial modeling, while enabling compact and efficient embedding models.

1. Introduction

Recent advances in vision–language foundation models have substantially expanded the capabilities of multimodal AI, enabling strong performance on generative and reasoning-centric tasks such as visual question answering, image captioning, and multimodal instruction following across diverse benchmarks (Li et al., 2024; Team, 2025; Bai et al., 2023; Wang et al., 2024; Bai et al., 2025). Despite these successes, the large scale and computational demands of such models severely limit their deployment in real-time and resource-constrained settings. In contrast, multimodal embedding models, which encode images and text into a shared semantic space for efficient retrieval, remain comparatively underexplored despite growing practical demand (Zhang et al., 2025a; Li et al., 2026; Zhang et al., 2025b). These models are critical for applications such as cross-modal search, semantic indexing, and recommendation systems, where efficiency and scalability are paramount. While recent efforts such as VLM2Vec (Jiang et al., 2024) have begun to establish benchmarks for vision–language embeddings, state-of-the-art performance still relies on large, resource-intensive architectures that are impractical for deployment.

Knowledge distillation (KD) (Hinton et al., 2015) provides an effective framework for compressing large language and vision-language models into efficient student architectures. However, most existing KD methods assume architectural similarity and shared tokenization between teacher and student (Hinton et al., 2015; Gu et al., 2024; Feng et al., 2025a). Recent efforts relax these assumptions through cross-architecture alignment mechanisms such as optimal transport (Boizard et al., 2025) or fine-grained token correspondence modeling (Wan et al., 2024).

In vision-language models, prior distillation approaches have explored cross-modal knowledge transfer, but their primary focus remains on enhancing generative or reasoning capabilities rather than learning compact multimodal embeddings. For example, Align-KD (Feng et al., 2025a) distills cross-modal alignment knowledge by guiding the student to mimic teacher attention distributions and cross-modal projections to improve downstream task performance. In contrast, EM-KD (Feng et al., 2025b) addresses unbalanced vision tokens between teacher and student models by matching semantic and affinity relationships; however, it lacks an explicit and interpretable mechanism for identifying which visual tokens correspond to the same underlying semantic entity across models.

These limitations are further compounded by the fact that most existing frameworks implicitly treat all tokens uniformly, ignoring the hierarchical and relational structure inherent in both language and vision. Linguistic theory has

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

long established that natural language exhibits hierarchical organization, with words forming nested phrases and higher-level semantic constructs (Chomsky, 1965; Crain & Nakayama, 1987; Everaert et al., 2015). Similarly, interpretability studies of transformer models suggest that textual and visual representations progressively organize into semantically coherent clusters across layers (Jawahar et al., 2019), where lower layers capture fine-grained lexical or patch-level features and higher layers synthesize these into abstract and object-level concepts (Montavon et al., 2018). Ignoring this intrinsic hierarchy leads to student models that mimic isolated teacher behaviors without capturing the representational trajectories essential for semantic understanding and generalization (Kovalev & Tikhomirov, 2025).

To address these gaps, we propose **HieRD**, a **Hie**rarchical **R**epresentation **D**istillation framework tailored for vision–language embedding models. **HieRD** explicitly captures hierarchical cross-modal structure by grouping visual tokens into semantically coherent clusters that approximate object-level entities and aligning them with phrase-level textual spans. Rather than enforcing flat token-wise matching, our framework introduces multi-granular relational alignment objectives that preserve both intra-modality structure and cross-modal correspondence at multiple levels of abstraction. By abstracting token representations into cluster-level semantic units, HieRD naturally accommodates architectural differences between teacher and student models, enabling meaningful alignment even when token counts or layer structures differ.

We further provide theoretical motivation showing that transformer-based vision encoders naturally develop cluster-like representations as depth increases, which can serve as stable and interpretable anchors for hierarchical distillation. In summary, our contributions are as follows:

- We identify key limitations in existing vision–language distillation methods, including inadequate modeling of hierarchical structure and ambiguous token correspondence across heterogeneous architectures.
- We introduce **HieRD**, a hierarchical relational distillation framework that aligns clusters of visual tokens with phrase-level textual spans to preserve fine-grained cross-modal semantics.
- We provide theoretical motivation for exploiting cluster structures in transformer models as effective anchors for hierarchical distillation.
- Extensive experiments demonstrate that HieRD consistently outperforms strong distillation baselines on diverse multimodal embedding tasks, enabling compact and efficient models with high-quality semantic representations.

2. Background

2.1. Knowledge Distillation

Originally formalized by (Hinton et al., 2015), Knowledge Distillation is a widely used framework for model compression, aiming to transfer the predictive behavior of a high-capacity teacher to a more efficient student model (Wilf et al., 2025; Gu et al., 2024; Wan et al., 2024; Zhang et al., 2024). In multimodal embedding learning, KD is commonly adopted to construct a shared semantic space that aligns visual and textual representations, serving retrieval and similarity-based tasks.

Given a mini-batch of multimodal input which jointly constructed from visual and textual content, a contrastive loss $\mathcal{L}_{con}$ (Jiang et al., 2024) encourages representations of corresponding sentences to be more similar than those of non-corresponding ones, thereby promoting discriminative semantic separation in the embedding space.

However, distillation methods that focus solely on matching final embeddings often fail to capture the relational structure of the teacher's representation space. To mitigate this issue, Park et al. proposed Relational Knowledge Distillation (RKD), which replaces point-wise alignment with relational supervision over mini-batches. By preserving relative distances and angular relationships induced by the teacher, RKD enables the student to better maintain the global geometry of the embedding space.

Accordingly, we define the base training objective as

$$\mathcal{L}_{base} = \mathcal{L}_{con} + \alpha \mathcal{L}_{RKD}, \tag{1}$$

where α balances the contribution of relational supervision. The detailed formulation of RKD is provided in Appendix B.

2.2. Hierarchical Representations in Language and Vision

Natural language is inherently hierarchical, organizing discrete lexical units into nested syntactic and semantic structures. Linguistic theories and empirical studies have long demonstrated that phrases and clauses serve as fundamental units of meaning, capturing semantic compositionality beyond individual words (Chomsky, 1965; Socher et al., 2011; 2013). In recursive and compositional models, meaning is constructed in a bottom-up manner, progressively merging lower-level units into higher-level abstractions (Tai et al., 2015).

Recent analyses of Transformer-based architectures reveal that such hierarchical organization also emerges implicitly in deep neural networks. In language models, lower layers tend to encode surface-level lexical or syntactic information, while higher layers capture increasingly abstract semantic representations (Tenney et al., 2019; Rogers et al., 2020;

Clark et al., 2019). This layer-wise specialization is further amplified in large language models, where early layers emphasize lexical and factual retrieval and deeper layers support complex reasoning and abstraction (Wang & Choi, 2023; Jin et al., 2025). Collectively, these findings suggest that representation evolution across Transformer layers parallels the construction of a linguistic parse tree, progressing from fine-grained tokens to semantically coherent structures.

In the visual domain, Vision Transformers process images as sets of local patch tokens. Although these tokens are initially spatially localized, studies in representation analysis and interpretability indicate that tokens corresponding to the same object or semantic region tend to become increasingly correlated across layers, implicitly forming object-centric groupings (Caron et al., 2021). This emergent behavior highlights that both language and vision representations exhibit hierarchical and compositional structures (Naseer et al., 2021), which are not adequately captured by flat, token-level alignment strategies.

2.3. Clustering as Structural Abstraction in Representation Spaces

Clustering provides a natural mechanism for abstracting fine-grained representations into higher-level semantic units. By grouping elements that are closely related in the feature space, clustering enables the transition from local representations to more structured and interpretable abstractions.

Among various clustering techniques, density-based methods such as DBSCAN (Ester et al., 1996) are particularly well-suited for high-dimensional and irregular data distributions. Unlike partition-based algorithms, DBSCAN does not require specifying the number of clusters in advance and can adapt to varying cluster densities, making it robust to noise and outliers.

3. Methodology

3.1. Theoretical Analysis of Visual Token Clustering in Self-Attention

Problem Setup and Notation. We consider a VLM composed of a vision encoder followed by a multimodal projector. The vision encoder maps an input image into a sequence of visual patch tokens updated through stacked Transformer layers with unimodal self-attention, prior to any projection into a shared multimodal embedding space or cross-modal interaction with text.

Let $Z_t = \{z_1^t, \dots, z_N^t\} \subset \mathbb{R}^D$ denote the set of visual token embeddings at layer t, where N is the number of image patches and D is the hidden dimension. To analyze the geometric behavior of self-attention, we assume that the value matrix satisfies $\|W_V\|_2 \leq 1$. Under this assumption, each token at layer $t+1$ is updated as a convex combination of token embeddings from the previous layer, $z_j^{t+1} = \sum_{k=1}^{N} a_{k,j}^t \, W_V z_k^t$, where $a_{k,j}^t$ denotes the attention weight from token k to token j at layer t.

Motivation and Analytical Objective. Self-attention aggregates token representations through weighted averaging, encouraging tokens associated with the same object or coherent semantic region to move closer in the embedding space, while interactions across distinct regions are typically weaker. This observation raises a fundamental question: *does unimodal self-attention within the vision encoder induce implicit clustering among visual tokens, and how does this behavior evolve across layers*?

To formalize this intuition, we study the contraction properties of self-attention by analyzing how the diameter of the visual token set and intra-cluster distances evolve across layers. We define the diameter of a set of vectors $A \subset \mathbb{R}^D$ as

$$\mathbf{d}(A) = \max_{a,b \in A} \|a - b\|_2 \tag{2}$$

Theorem 3.1. *Let Z_t and Z_{t+1} denote the sets of visual token embeddings at layers t and $t+1$ of the vision encoder, respectively. The following statements hold*

(i) The diameter of the token set is non-increasing across layers, i.e., $\mathbf{d}(Z_{t+1}) \leq \mathbf{d}(Z_t)$.

(ii) Assume that Z_t can be partitioned into M clusters $Z_t^m = \{z_t^j : j \in G^m\}$, where $\{G^m\}_{m=1}^M$ forms a partition of $\{1, \dots, N\}$. For any $j \in G^m$ and $j' \in G^{m'}$ with $m \neq m'$, the attention weights satisfy $\varepsilon_l < a_t^{j',j} < \varepsilon_u$ with $\varepsilon_u \geq \varepsilon_l \geq 0$, indicating limited cross-cluster attention. Let $N_m = |G^m|$ and $A_m = (N - N_m)/N$. Then, the intra-cluster diameter at the next layer satisfies

$$\mathbf{d}(Z_{t+1}^m) < (1 - A_m \varepsilon_l)^2 \mathbf{d}(Z_t^m) + A_m \varepsilon_u \mathbf{d}(Z_t) \big(A_m \varepsilon_u + 2\big) \tag{3}$$

Theorem 3.1 (proof in Appendix A.1) offers an asymptotic characterization of the clustering dynamics induced by unimodal self-attention in the vision encoder. Specifically, Part (i) establishes that the overall diameter of the visual token embeddings does not increase across layers, reflecting an inherent contraction property of attention-based aggregation. Moreover, Part (ii) shows that when attention between distinct clusters diminishes, i.e., as the inter-cluster attention bounds ε_u and ε_l approach zero, the intra-cluster diameter is guaranteed to decrease, leading to progressively more compact visual token clusters over successive layers.

3.2. Token Grouping and Group-Level Representation

Vision Token Clustering. Motivated by the clustering behavior of self-attention representations analyzed in Section 3.1, we construct object-level visual token clusters from

the *teacher* model. Given an input image, the teacher vision encoder produces a set of visual patch tokens $\{z_i^{\mathcal{T}}\}_{i=1}^{N_{\mathcal{T}}}$, where each token corresponds to an image patch and is associated with the normalized spatial coordinate of its patch center, $\mathbf{c}_i = (x_i, y_i) \in [0,1]^2$.

To jointly capture semantic similarity and spatial coherence, we define a pairwise distance matrix $\mathbf{D}^{\mathcal{T}} \in \mathbb{R}^{N_{\mathcal{T}} \times N_{\mathcal{T}}}$ as

$$D_{ij}^{\mathcal{T}} = 1 - \frac{\langle z_i^{\mathcal{T}}, z_j^{\mathcal{T}} \rangle}{\|z_i^{\mathcal{T}}\|_2 \, \|z_j^{\mathcal{T}}\|_2} + \lambda \, \|\mathbf{c}_i - \mathbf{c}_j\|_2,$$

where the first term measures cosine dissimilarity in representation space and the second term encodes spatial proximity in the image plane. By combining these two factors, the distance metric favors object-level grouping while avoiding the merging of semantically similar but spatially distant patches. We apply DBSCAN with this distance to $\mathbf{D}^{\mathcal{T}}$ to obtain teacher-side visual token clusters $\{C_m^{\mathcal{T}}\}_{m=1}^{M}$, while treating isolated or non-coherent tokens as noise.

Given the teacher clusters, we establish explicit correspondence to the *student* model based on spatial coverage. Each student visual token $z_j^{\mathcal{S}}$ is associated with a rectangular region $\mathcal{R}_j^{\mathcal{S}} \subset [0,1]^2$ in the normalized image plane, which represents the spatial extent of the image patch covered by the token, rather than a single point location. For each teacher cluster $C_m^{\mathcal{T}}$, the corresponding student cluster is defined as

$$C_m^{\mathcal{S}} = \{j \mid \exists\, i \in C_m^{\mathcal{T}} \text{ s.t. } \mathbf{c}_i \in \mathcal{R}_j^{\mathcal{S}}\}.$$

This region-based mapping allows multiple fine-grained teacher tokens to correspond to a single coarser student patch. Such a design is consistent with Theorem 3.1, which predicts increasing object-level compactness of visual token representations across layers. Further details of the teacher–student token mapping procedure are provided in the Appendix D.

Text Span Construction. For textual inputs, we construct linguistically grounded text spans rather than clustering tokens in representation space. We partition the selected Transformer layers into lower- and higher-level functional groups. At lower layers, contiguous tokens are grouped into complete word spans to preserve fine-grained lexical information, whereas at higher layers, spans correspond to phrase-level units such as noun phrases and verb phrases, capturing compositional semantics.

This hierarchical span construction yields progressively abstract textual representations and is shared by both the Teacher and Student. Figure 1 illustrates the resulting text spans and their interaction with visual clusters within the distillation objectives.

Span construction is implemented using a syntactic parser (e.g., spaCy) to extract spans from the input text. Each span represents a predefined set of tokens encoding a coherent semantic concept, enabling the Student to learn aggregated linguistic semantics rather than isolated word-level attributes.

Unified Attention-Weighted Group Representation. Given a pre-defined group of tokens $G_m = \{t_1, \ldots, t_{|G_m|}\}$, we construct a unified representation that summarizes the group semantics. Here, G_m corresponds to a visual token cluster in the image modality or a linguistically constructed text span in the text modality, as defined in the preceding sections.

Instead of simple averaging, we employ an attention-weighted aggregation to allow more informative tokens within the group to contribute more strongly. Formally, at layer l, the representation of group G_m is defined as

$$\mathrm{U}_{m,l} = \sum_{t \in G_m} w_{t,l} \, \mathbf{h}_{t,l}, \tag{4}$$

where $\mathbf{h}_{t,l}$ denotes the hidden state of token t at layer l, and $w_{t,l}$ is its normalized importance weight.

The importance weights are derived from the Teacher model's self-attention. For each token $t \in G_m$, we accumulate the incoming attention it receives from other tokens within the same group across all attention heads:

$$\mathrm{Att}_{t,l} = \frac{1}{H} \sum_{h=1}^{H} \sum_{\substack{s \in G_m \\ s \neq t}} \alpha_{s \to t, l}^{(h)}, \tag{5}$$

where $\alpha_{s \to t, l}^{(h)}$ denotes the attention weight from token s to token t at head h and layer l, computed with diagonal masking (and causal masking when applicable). The final weights are obtained by normalizing within the group:

$$w_{t,l} = \frac{\mathrm{Att}_{t,l}}{\sum_{r \in G_m} \mathrm{Att}_{r,l}}$$

3.3. Group-Level Distillation Objective

Intra-modal Structural Alignment. As illustrated in Figure 1, we enforce structural consistency between the Teacher and Student *within each modality* by preserving the relative geometry induced by the Teacher. For visual inputs, this geometry is defined over object-level image token clusters, whereas for textual inputs it is defined over linguistically constructed text spans.

Let $\mathrm{U}_{i,l}^{T,\mu} \in \mathbb{R}^{d_T}$ and $\mathrm{U}_{i,l}^{S,\mu} \in \mathbb{R}^{d_S}$ denote the representations of the i-th group at layer l for the Teacher and Student, respectively, where $\mu \in \{\text{image}, \text{text}\}$ indexes the modality. Here, a group corresponds to an image cluster when $\mu = \text{image}$, and a text span when $\mu = \text{text}$.

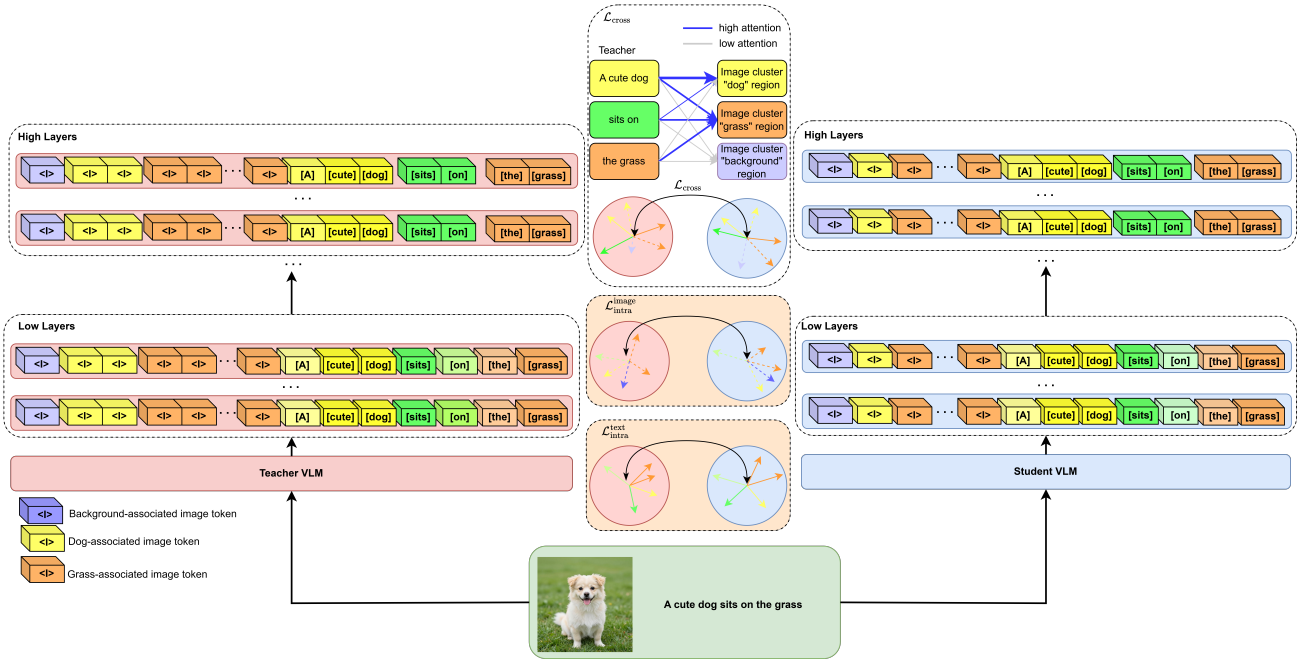

Figure 1. **Illustration of intra-modal and cross-modal distillation at the group level.** The Teacher guides the Student by preserving structural relationships within each modality (image clusters and text spans) as well as across modalities through cross-modal alignment.

For the visual modality, we align the intra-cluster geometry of the Student with that of the Teacher by matching pairwise distances between image clusters:

$$\mathcal{L}_{\text{intra}}^{\text{image}} = \sum_{l \in L_{\text{key}}} \frac{1}{|\mathcal{P}_l^{\text{image}}|} \sum_{(i,j) \in \mathcal{P}_l^{\text{image}}} \Delta_{i,j,l}^{\text{image}}, \tag{6}$$

where $\mathcal{P}_l^{\text{image}}$ represents all pairs (i, j) with $i \neq j$ of the clusters i and j, and $d(\cdot, \cdot)$ denotes cosine distance, and we define

$$\Delta_{i,j,l}^{\text{image}} = \left(d(\mathrm{U}_{i,l}^{S,\text{image}}, \mathrm{U}_{j,l}^{S,\text{image}}) - d(\mathrm{U}_{i,l}^{T,\text{image}}, \mathrm{U}_{j,l}^{T,\text{image}})\right)^2 \tag{7}$$

For the textual modality, we preserve the relative geometry among text spans induced by the Teacher in an analogous manner:

$$\mathcal{L}_{\text{intra}}^{\text{text}} = \sum_{l \in L_{\text{key}}} \frac{1}{|\mathcal{P}_l^{\text{text}}|} \sum_{(m,n) \in \mathcal{P}_l^{\text{text}}} \Delta_{m,n,l}^{\text{text}}, \tag{8}$$

where $\mathcal{P}_l^{\text{text}}$ represents all pairs (m, n) with $m \neq n$ of the spans m and n, and we define

$$\Delta_{m,n,l}^{\text{text}} = \left(d(\mathrm{U}_{m,l}^{S,\text{text}}, \mathrm{U}_{n,l}^{S,\text{text}}) - d(\mathrm{U}_{m,l}^{T,\text{text}}, \mathrm{U}_{n,l}^{T,\text{text}})\right)^2 \tag{9}$$

The final intra-modal structural alignment loss is defined as

$$\mathcal{L}_{\text{intra}} = \mathcal{L}_{\text{intra}}^{\text{image}} + \mathcal{L}_{\text{intra}}^{\text{text}} \tag{10}$$

To investigate the theoretical insights of the image-intra and text-intra losses in preserving some characteristics from the teacher model, we develop the following theorem.

Theorem 3.2. *Assume that we have two collections of vectors $X = [x_i]_{i=1}^n \subset \mathbb{R}^{d_x}$ and $Y = [y_i]_{i=1}^n \subset \mathbb{R}^{d_y}$, with $d_x \leq d_y$. Suppose that the pairwise cosine distance are preserved, i.e., for all i, j, $1 - \frac{x_i^\top x_j}{\|x_i\|_2 \|x_j\|_2} = 1 - \frac{y_i^\top y_j}{\|y_i\|_2 \|y_j\|_2}$. Then, there exists a matrix $Q \in St(d_x, d_y)$ such that $\tilde{y}_i = Q\tilde{x}_i$ for all i, where $\tilde{x}_i = \frac{x_i}{\|x_i\|_2}$ and $\tilde{y}_i = \frac{y_i}{\|y_i\|_2}$ denote the ℓ_2-normalized vectors. Here, $St(d_x, d_y) = \{R \in \mathbb{R}^{d_y \times d_x} \mid R^\top R = I_{d_x}\}$ denotes Stiefel manifold.*

We assume that the loss $\mathcal{L}_{\text{intra}}^{\text{image}}$ can be minimized to its true optimum of 0. In this case, the two sets of vectors $\left[U_{i,l}^{S,\text{image}}\right]_i$ and $\left[U_{i,l}^{T,\text{image}}\right]_i$ preserve pairwise cosine distances. By Theorem 3.2 (proof in Appendix A.2), there exists a matrix $Q^{\text{image}} \in \mathrm{St}(d_S, d_T)$ with orthonormal columns (i.e., $(Q^{\text{image}})^\top Q^{\text{image}} = I_{d_S}$), which defines an isometric linear embedding, such that

$$\tilde{U}_{i,l}^{T,\text{image}} = Q^{\text{image}} \tilde{U}_{i,l}^{S,\text{image}}, \quad \forall i,$$

where $\tilde{U}_{i,l}^{S,\text{image}}$ and $\tilde{U}_{i,l}^{T,\text{image}}$ denote the normalized vectors of $U_{i,l}^{S,\text{image}}$ and $U_{i,l}^{T,\text{image}}$. Similarly, if $\mathcal{L}_{\text{intra}}^{\text{text}}$ is minimized to 0, there exists a matrix $Q^{\text{text}} \in \mathrm{St}(d_S, d_T)$ such that

$$\tilde{U}_{m,l}^{T,\text{text}} = Q^{\text{text}} \tilde{U}_{m,l}^{S,\text{text}}, \quad \forall m,$$

with normalized version $\tilde{U}_{m,l}^{S,\text{text}}$ and $\tilde{U}_{m,l}^{T,\text{text}}$ of $U_{i,l}^{S,\text{text}}$ and $U_{i,l}^{T,\text{text}}$. Thus, minimizing both $\mathcal{L}_{\text{intra}}^{\text{image}}$ and $\mathcal{L}_{\text{intra}}^{\text{text}}$ enforces geometric alignment between teacher and student representations: the intrinsic relational structure of visual cluster

and text spans is preserved across models up to an isometric transformation.

Cross-modal Structural Alignment. To preserve the relative geometry between visual and textual representations, we introduce a cross-modal distillation loss that aligns the image–text similarity structure of the Student with that of the Teacher. Unlike intra-modal alignment, this objective explicitly operates across modalities between image clusters and text spans, as illustrated in Figure 1.

At layer l, let $\mathrm{U}_{i,l}^{T,\text{image}} \in \mathbb{R}^{d_T}$ and $\mathrm{U}_{i,l}^{S,\text{image}} \in \mathbb{R}^{d_S}$ denote the representations of the i-th visual cluster in the Teacher and Student, respectively. Similarly, let $\mathrm{U}_{j,l}^{T,\text{text}} \in \mathbb{R}^{d_T}$ and $\mathrm{U}_{j,l}^{S,\text{text}} \in \mathbb{R}^{d_S}$ denote the representations of the j-th text span.

For each image–text pair (i, j), we measure their cosine similarity in both models. The cross-modal structural alignment loss is defined as

$$\mathcal{L}_{\text{cross}} = \sum_{l \in L_{\text{key}}} \sum_{i \in \mathcal{C}_l^{\text{image}}} \sum_{j \in \mathcal{C}_l^{\text{text}}} \gamma_{i,j,l} \Big(d(\mathrm{U}_{i,l}^{S,\text{image}}, \mathrm{U}_{j,l}^{S,\text{text}}) - d(\mathrm{U}_{i,l}^{T,\text{image}}, \mathrm{U}_{j,l}^{T,\text{text}}) \Big)^2, \tag{11}$$

where $d(\cdot, \cdot)$ denotes cosine similarity.

The weight $\gamma_{i,j,l}$ reflects the semantic relevance of the image–text pair and is derived from the Teacher's cross-attention patterns. Specifically, we aggregate the cross-attention from all text tokens in text span j to all visual tokens in visual cluster i across attention heads, and normalize the result over all image–text pairs at layer l:

$$\gamma_{i,j,l} = \frac{\sum_{t \in C_j^{\text{text}}} \sum_{v \in C_i^{\text{img}}} \sum_h \alpha_{t \to v, l}^{(h)}}{\sum_{i' \in \mathcal{C}_l^{\text{img}}} \sum_{j' \in \mathcal{C}_l^{\text{text}}} \sum_{t \in C_{j'}^{\text{text}}} \sum_{v \in C_{i'}^{\text{img}}} \sum_h \alpha_{t \to v, l}^{(h)}} \tag{12}$$

This normalization emphasizes image-text pairs that are strongly associated in the Teacher's cross-modal geometry, while suppressing weak or noisy alignments.

We now investigate the role of cross-modal structural alignment in the following theorem.

Theorem 3.3. *Assume that the losses $\mathcal{L}_{intra}^{image}$ and $\mathcal{L}_{intra}^{text}$ can be optimized to 0. If the loss $\mathcal{L}_{cross}$ can be also optimized to 0 and span $\left\{ \tilde{U}_{i,l}^{S,image} \left(\tilde{U}_{m,l}^{S,text} \right)^T : i \in \mathcal{P}_l^{image}, m \in \mathcal{P}_l^{text} \right\} = \mathbb{R}^{d_S \times d_S}$, we then have $Q^{text} = Q^{image}$, where Q^{image} and Q^{text} are the linear isometries guaranteed by Theorem 3.2, with $Q^{image}, Q^{text} \in \mathrm{St}(d_S, d_T)$, which preserve pairwise cosine similarities when mapping student representations to the teacher space.*

It is evident that Theorem 3.3 (proof in Appendix A.3) encourages the two modalities (i.e., images and texts) to be aligned more consistently by sharing the transformation matrices between the student and teacher models.

Hidden Representation Alignment. To further align fine-grained semantic content, we apply a feature-level distillation loss between corresponding *group-level* representations of the Teacher and Student, where a group denotes an image cluster for visual inputs or a text span for textual inputs.

Since the Student typically has a smaller hidden dimension ($d_S < d_T$), we introduce a learnable linear projector $W_l \in \mathbb{R}^{d_S \times d_T}$ to map Student representations into the Teacher feature space:

$$\hat{\mathrm{U}}_{m,l}^S = \mathrm{U}_{m,l}^S W_l. \tag{13}$$

The hidden alignment loss is defined as

$$\mathcal{L}_{\text{hid}} = \sum_{l \in L_{\text{key}}} \sum_m \left(1 - \frac{(\hat{\mathrm{U}}_{m,l}^S)^\top \mathrm{U}_{m,l}^T}{\|\hat{\mathrm{U}}_{m,l}^S\|_2 \, \|\mathrm{U}_{m,l}^T\|_2} \right), \tag{14}$$

where m indexes image clusters or text spans and L_{key} denotes the selected layers used for distillation.

3.4. Combined Distillation Objective

We combine the standard multimodal embedding loss with the proposed hierarchical distillation objectives. Specifically, the student model is trained using the base loss together with our proposed structural alignment losses, and hidden-level alignment loss. The overall training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda_{\text{struct}} \mathcal{L}_{\text{struct}} + \lambda_{\text{hid}} \mathcal{L}_{\text{hid}}, \tag{15}$$

where $\mathcal{L}_{\text{struct}} = \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{cross}}$, and λ_{struct} and λ_{hid} control the contributions of structural and hidden-level distillation, respectively.

4. Experiments

4.1. Experimental Setup

Datasets. Following prior work, we evaluate our method on two multimodal tasks, image classification (CLS) and visual question answering (VQA) drawn from the MMEB (Massive Multimodal Embedding Benchmark) (Jiang et al., 2024). For classification, we report results on ImageNet-1K, N24News, HatefulMemes, VOC2007, and SUN397, spanning general-domain, news-related, and social-media scenarios. For VQA, we employ OK-VQA, A-OKVQA, DocVQA, InfographicsVQA, ChartQA, and Visual7W, which require grounding textual queries in visual content. Dataset statistics are provided in Appendix C.3.

Models. We adopt B3 Qwen2-2B (Thirukovalluru et al., 2025) as the Teacher model, a well-finetuned variant of

Table 1. Quantitative comparison on CLS and VQA benchmarks. The best and second best results are highlighted in **bold** and underline.

Method	CLS						VQA						
	ImageNet-1K	N24News	HatefulMemes	VOC2007	SUN397	Avg	OK-VQA	A-OKVQA	DocVQA	I-VQA	ChartQA	Visual7W	Avg
B3-Qwen2-2B → FastVLM-0.5B													
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
SFT	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	52.4	49.8	74.3	34.3	**50.8**	46.4	51.3
RKD	53.6	70.3	60.8	79.9	66.2	66.2	51.1	51.2	74.1	34.9	50.2	46.5	51.3
CKD	53.2	70.6	**61.2**	78.6	66.7	66.1	51.8	51.0	74.7	35.0	49.6	46.2	51.4
EMO	52.4	68.4	59.1	80.4	59.3	63.9	50.9	50.1	73.1	33.8	49.0	**47.8**	50.8
EM-KD	53.4	67.3	59.4	78.0	63.2	64.3	51.1	49.3	77.2	36.7	47.6	47.3	51.5
HieRD	**56.0**	**71.7**	60.3	**80.6**	**67.2**	**67.2**	**52.4**	**51.8**	**77.6**	**38.3**	49.8	46.8	**52.8**
B3-Qwen2-2B → LLaVA-OneVision-0.5B													
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
SFT	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	47.9	42.5	37.9	21.0	24.7	42.9	36.2
RKD	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	57.0	66.9	58.6	84.1	68.2	66.9	51.2	43.9	47.2	22.9	30.9	45.6	40.3
EMO	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	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
HieRD	**63.3**	**73.6**	**59.3**	**85.7**	**68.7**	**70.1**	**52.5**	**46.9**	**49.3**	**23.6**	**34.5**	**47.2**	**42.3**

Qwen2VL-2B-Instruct (Wang et al., 2024), providing strong multimodal embedding quality and stable cross-modal alignment. For the Student, we consider two lightweight architectures: FastVLM-0.5B (Vasu et al., 2025) and LLaVA-OneVision-0.5B (Li et al., 2024).

Implementation and Evaluation Details. Both Teacher and Student models are fine-tuned using LoRA-based parameter-efficient training. All experiments follow the standard MMEB evaluation protocol (Jiang et al., 2024), and Precision@1 is used as the primary metric across all tasks. Detailed training configurations, LoRA settings, and hyperparameters are reported in Appendix C.2.

Baselines. We compare our method against a range of traditional and state-of-the-art knowledge distillation methods for vision–language models and large language models, including MSE, Relational KD (Park et al., 2019), Comparative KD (Wilf et al., 2025), EMO (Truong et al., 2025), and EM-KD (Feng et al., 2025b). Further details on these baselines are provided in Appendix C.1.

4.2. Main Results

Our experimental results in Table 1 indicate that HieRD consistently improves the performance of compact student models when distilled from a large-scale teacher across a broad range of multimodal benchmarks. We evaluate two representative settings, where the B3-Qwen2-2B teacher is distilled into FastVLM-0.5B and LLaVA-OneVision-0.5B students.

Distillation to FastVLM-0.5B. On classification tasks, HieRD achieves the highest average accuracy among student-side methods, reaching 67.2%. Compared with standard fine-tuning, HieRD yields consistent gains across datasets, including improvements of 2.9% on average, highlighting its effectiveness in transferring fine-grained and scene-level visual semantics. Relative to prior distillation approaches, HieRD also provides steady improvements, indicating better preservation of intermediate semantic structure. On the VQA benchmarks, HieRD attains an average score of 52.8%, with particularly clear gains on document-centric and image-specific tasks. For example, HieRD improves performance on DocVQA and I-VQA by 0.4% and 1.6%, respectively, over the strongest distillation baseline. These results suggest that hierarchical alignment facilitates more effective transfer of spatially grounded and modality-specific reasoning signals.

Distillation to LLaVA-OneVision-0.5B. When distilling into the LLaVA-OneVision-0.5B student, HieRD exhibits a similar trend. On classification tasks, it reaches an average accuracy of 70.1%, with strong results on ImageNet-1K (63.3%) and N24News (73.6%), demonstrating that hierarchical distillation remains effective across different student architectures. In the VQA setting, HieRD achieves an average score of 42.3%, consistently improving performance across all evaluated datasets. In particular, DocVQA accuracy increases substantially compared to standard fine-tuning (49.3% vs. 38.6%), while I-VQA also shows clear gains (23.6% vs. 21.3%). Overall, these results indicate that HieRD provides a robust and architecture-agnostic mechanism for enhancing multimodal reasoning in compact models through hierarchical representation alignment.

5. Ablation Study

To analyze the contribution of each component in our framework, we conduct ablation studies using the B3-Qwen2-2B teacher (Thirukovalluru et al., 2025) and the FastVLM-0.5B student (Vasu et al., 2025), focusing on (i) the effect of individual loss components, and (ii) alternative designs for group-level representation aggregation. The impact of text span granularity in hierarchical distillation is analyzed in

the Appendix E.1.

Beyond performance-oriented ablations, we further conduct an empirical analysis to examine whether the cross-modal alignment properties predicted by Theorem 3.3 manifest in practice.

Impact of Loss Components. Table 2 evaluates the contribution of each loss component in HieRD. Removing any single term leads to consistent performance drops across all datasets; for example, excluding $\mathcal{L}_{\text{intra}}$ reduces ImageNet-1K accuracy from 56.0 to 55.2, while removing $\mathcal{L}_{\text{hid}}$ decreases SUN397 accuracy from 67.2 to 66.8. When two components are removed, the degradation becomes more pronounced, with ImageNet-1K accuracy dropping to 54.2, close to the baseline student performance (52.9). In contrast, the full HieRD configuration achieves the best results across all benchmarks, indicating that the three loss components provide complementary supervision.

Table 2. Ablation result of HieRD under different loss component configurations.

Method	ImageNet-1K	VOC2007	SUN397
Student	52.9	77.7	64.7
HieRD (w/o $\mathcal{L}_{\text{intra}}$)	55.2	80.5	66.2
HieRD (w/o $\mathcal{L}_{\text{cross}}$)	54.7	78.9	67.0
HieRD (w/o $\mathcal{L}_{\text{hid}}$)	55.7	80.4	66.8
HieRD (w/o $\mathcal{L}_{\text{intra}}$, $\mathcal{L}_{\text{cross}}$)	54.2	79.2	65.1
HieRD (w/o $\mathcal{L}_{\text{intra}}$, $\mathcal{L}_{\text{hid}}$)	53.4	78.7	65.4
HieRD (w/o $\mathcal{L}_{\text{cross}}$, $\mathcal{L}_{\text{hid}}$)	53.6	80.0	64.8
HieRD (full)	**56.0**	**80.6**	**67.2**

Effectiveness of Attention-Weighted Aggregation. We validate the effectiveness of our proposed representation formulation. Specifically, we compare our **Attention-Weighted** aggregation strategy against a standard **Mean Pooling** baseline in Table 3. We observe that the Attention-Weighted mechanism consistently outperforms Mean Pooling. This suggests that the attention mechanism successfully allows the model to prioritize informative tokens while suppressing noise, leading to more robust representations.

Table 3. Comparison of aggregation strategies. We evaluate the impact of using Mean Pooling versus Attention-Weighted aggregation.

Aggregation Strategy	ImageNet-1K	VOC2007	SUN397
Mean	53.7	80.1	63.7
Attention	**56.0**	**80.6**	**67.2**

Empirical Analysis of Theorem 3.3 In this subsection, we empirically validate the prediction of Theorem 3.3, which states that under sufficiently optimized intra-modal and cross-modal objectives, the learned linear isometries Q^{image} and Q^{text} should coincide. To examine this behavior, we track the orthogonality measure $(Q^{\text{image}})^\top Q^{\text{text}}$ during training.

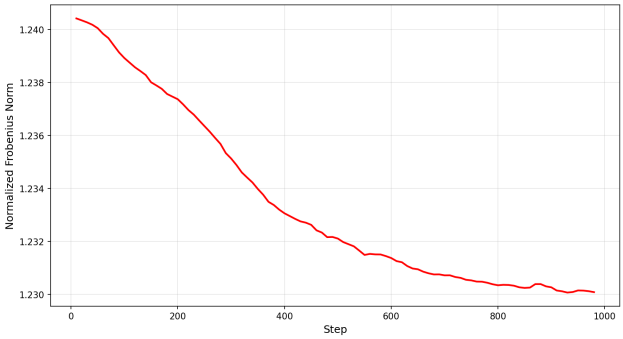

Figure 2. Visualization of the orthogonality measure between the learned isometries Q^{image} and Q^{text} across training. Lower values indicate closer alignment to the identity matrix.

As shown in Figure 2, the proposed orthogonality measure exhibits a consistent decreasing trend during training. This behavior indicates that the two learned transformations Q^{image} and Q^{text} become increasingly aligned, suggesting convergence toward a shared isometric mapping. While not constituting a formal proof, this empirical observation provides supportive evidence for the hypothesis underlying Theorem 3.3, namely that sufficiently optimized intra-modal and cross-modal objectives encourage a unified cross-modal geometry.

Additional experimental details and analyses are provided in Appendix E.2.

6. Conclusion

We propose HieRD, a hierarchical distillation framework that overcomes the limitations of flat token-matching in vision–language models by aligning visual token clusters with phrase-level textual spans. This structure-aware approach effectively bridges architectural gaps between heterogeneous models, enabling robust knowledge transfer across different token counts and layer depths. Experimental results show that HieRD consistently outperforms existing baselines, producing efficient student models with high-quality semantic representations. By leveraging emergent cluster structures in transformers as stable alignment anchors, HieRD provides a scalable and comprehensive solution for distilling complex cross-modal correspondences.

Impact Statement

This paper presents work whose goal is to advance research in vision–language modeling, particularly in the area of cross-modal distillation. Potential societal benefits include enabling more efficient multimodal models for deployment in resource-constrained environments. We do not identify

any specific negative ethical consequences that must be highlighted here.

References

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Boizard, N., Haddad, K. E., Hudelot, C., and Colombo, P. Towards cross-tokenizer distillation: the universal logit distillation loss for llms. *arXiv preprint arXiv:2402.12030*, 2025.

Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021.

Chomsky, N. *Aspects of the Theory of Syntax*. MIT Press, 1965.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://aclanthology.org/W19-4828/.

Crain, S. and Nakayama, M. Structure dependence in grammar formation. *Language*, 63(1):522–543, 1987.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., and et al. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd, volume 96, pages 226–231*, 1996.

Everaert, M. B., Huybregts, M. A., Chomsky, N., Berwick, R. C., and Bolhuis, J. J. Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, 19(12):729–743, 2015. ISSN 1364-6613. URL https://www.sciencedirect.com/science/article/pii/S1364661315002326.

Feng, Q., Li, W., Lin, T., and Chen, X. Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language model. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Feng, Z., Yang, S., Duan, B., Yang, W., and Wang, J. Emkd:distilling efficient multimodal large language model with unbalanced vision tokens. In *Association for the Advancement of Artificial Intelligence*, 2025b.

Gu, Y., Dong, L., Wei, F., and Huang, M. Minillm: Knowledge distillation of large language models. In *The International Conference on Learning Representations*, 2024.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jawahar, G., Sagot, B., and Seddah, D. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1356/.

Jiang, Z., Meng, R., Yang, X., Yavuz, S., Zhou, Y., and Chen, W. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *International conference on machine learning*, 2024.

Jin, M., Luo, W., Cheng, S., Wang, X., Hua, W., Tang, R., Wang, W. Y., and Zhang, Y. Disentangling memory and reasoning ability in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1681–1701, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.84/.

Kovalev, G. and Tikhomirov, M. Iterative layer-wise distillation for efficient compression of large language models, 2025. URL https://arxiv.org/abs/2511.05085.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., and Li, C. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., Zhou, J., and Lin, J. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv*, 2026.

Montavon, G., Samek, W., and Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, February 2018. ISSN 1051-2004. URL http://dx.doi.org/10.1016/j.dsp.2017.10.011.

Naseer, M. M., Ranasinghe, K., Khan, S. H., Hayat, M., Shahbaz Khan, F., and Yang, M.-H. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23296–23308. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c404a5adbf90e09631678b13b05d9d7a-Paper.pdf.

Park, W., Kim, D., Lu, Y., and Cho, M. Relational knowledge distillation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Rogers, A., Kovaleva, O., and Rumshisky, A. A primer in bertology: What we know about how bert works, 2020. URL https://arxiv.org/abs/2002.12327.

Socher, R., Lin, C. C., Ng, A. Y., and Manning, C. D. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2011.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Tai, K. S., Socher, R., and Manning, C. D. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1556–1566, Beijing, China, July 2015. Association for Computational Linguistics. URL https://aclanthology.org/P15-1150/.

Team, Q. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Tenney, I., Das, D., and Pavlick, E. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1452/.

Thirukovalluru, R., Meng, R., Liu, Y., K, K., Su, M., Nie, P., Yavuz, S., Zhou, Y., Chen, W., and Dhingra, B. Breaking the batch barrier (b3) of contrastive learning via smart batch mining. *arXiv preprint arXiv:2505.11293*, 2025.

Truong, M.-P., Vu, H. A., Vu, T., Diep, N. T. N., Van, L. N., Nguyen, T. H., and Le, T. 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*, 2025.

Vasu, P. K. A., Faghri, F., Li, C.-L., Koc, C., True, N., Antony, A., Santhanam, G., Gabriel, J., Grasch, P., Tuzel, O., and Pouransari, H. Fastvlm: Efficient vision encoding for vision language models. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Wan, F., Huang, X., Cai, D., Quan, X., Bi, W., and Shi, S. Knowledge fusion of large language models. In *International Conference on Learning Representations*, 2024.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wang, R. and Choi, M. Large language models on lexical semantic change detection: An evaluation, 2023. URL https://arxiv.org/abs/2312.06002.

Wilf, A., Xu, A. T., Liang, P. P., Obolenskiy, A., Fried, D., and Morency, L.-P. Comparative knowledge distillation. In *Winter Conference on Applications of Computer Vision*, 2025.

Zhang, S., Zhang, X., Sun, Z., Chen, Y., and Xu, J. Dual-space knowledge distillation for large language models. In *Empirical Methods in Natural Language Processing*, 2024.

Zhang, X., Zhang, Y., Xie, W., Li, M., Dai, Z., Long, D., Xie, P., Zhang, M., Li, W., and Zhang, M. Gme: Improving universal multimodal retrieval by multimodal llms, 2025a. URL https://arxiv.org/abs/2412.16855.

Zhang, X., Zhang, Y., Xie, W., Li, M., Dai, Z., Long, D., Xie, P., Zhang, M., Li, W., and Zhang, M. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9274–9285, 2025b.

A. Appendix

A.1. Proof of Theorem 3.1

Throughout the proof, we follow the same notation and assumptions as in Section 3.1. For clarity, we first restate the theorem and the core definitions used in our analysis.

We define the diameter of a set of vectors $A \subset \mathbb{R}^D$ as the maximum Euclidean distance between any two elements in the set:

$$\mathbf{d}(A) = \max_{a,b \in A} \|a - b\|_2. \tag{16}$$

Theorem A.1. *Let $Z_t = \{\mathbf{z}_t^k\}_{k=1}^N$ and $Z_{t+1} = \{\mathbf{z}_{t+1}^k\}_{k=1}^N$ denote the sets of visual token embeddings at layers t and $t+1$ of the vision encoder, respectively. Under the unimodal self-attention update with $W_V = I$. The following statements hold*

(i) The diameter of the token set is non-increasing across layers, i.e., $\mathbf{d}(Z_{t+1}) \leq \mathbf{d}(Z_t)$.

(ii) Assume that Z_t can be partitioned into M clusters $Z_t^m = \{z_t^j : j \in G^m\}$, where $\{G^m\}_{m=1}^M$ forms a partition of $\{1, \ldots, N\}$. For any $j \in G^m$ and $j' \in G^{m'}$ with $m \neq m'$, the attention weights satisfy $\varepsilon_l < a_t^{j',j} < \varepsilon_u$ with $\varepsilon_u \geq \varepsilon_l \geq 0$, indicating limited cross-cluster attention. Let $N_m = |G^m|$ and $A_m = (N - N_m)/N$. Then, the intra-cluster diameter at the next layer satisfies

$$\mathbf{d}(Z_{t+1}^m) < (1 - A_m\varepsilon_l)^2 \mathbf{d}(Z_t^m) + A_m\varepsilon_u \mathbf{d}(Z_t)\big(A_m\varepsilon_u + 2\big). \tag{17}$$

Proof of Part (i). With $h(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_2$, we first note that

$$h\left(W_V a, W_V b\right) = \|W_V\left(a - b\right)\|_2 \leq \|W_V\|_2 \|a - b\|_2 \leq h\left(a, b\right).$$

Because the distance $h(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_2$ is a convex function, we can apply Jensen's inequality. For any pair of tokens j, j', we have:

$$\begin{aligned}
\left\|\mathbf{z}_{t+1}^j - \mathbf{z}_{t+1}^{j'}\right\|_2 &= \left\|\sum_{k=0}^N \mathbf{a}_t^{k,j} W_V \mathbf{z}_t^k - \sum_{k'=0}^N \mathbf{a}_t^{k',j'} W_V \mathbf{z}_t^{k'}\right\|_2 \\
&= h\left(W_V \sum_{k=0}^N \mathbf{a}_t^{k,j} \mathbf{z}_t^k, W_V \sum_{k'=0}^N \mathbf{a}_t^{k',j'} \mathbf{z}_t^{k'}\right) \\
&\leq h\left(\sum_{k=0}^N \mathbf{a}_t^{k,j} \mathbf{z}_t^k, \sum_{k'=0}^N \mathbf{a}_t^{k',j'} \mathbf{z}_t^{k'}\right)
\end{aligned}$$

By applying Jensen's inequality for the convex function $h(\cdot, \cdot)$ twice, first for the first argument and then for the second, we obtain:

$$\begin{aligned}
\left\|\mathbf{z}_{t+1}^j - \mathbf{z}_{t+1}^{j'}\right\|_2 &\leq \sum_{k=0}^N \mathbf{a}_t^{k,j} h\left(\mathbf{z}_t^k, \sum_{k'=0}^N \mathbf{a}_t^{k',j'} \mathbf{z}_t^{k'}\right) \\
&\leq \sum_{k=0}^N \sum_{k'=0}^N \mathbf{a}_t^{k,j} \mathbf{a}_t^{k',j'} h\left(\mathbf{z}_t^k, \mathbf{z}_t^{k'}\right).
\end{aligned}$$

Then, using the definition of the diameter $\mathbf{d}(Z_t) = \max_{k,k'} h(\mathbf{z}_t^k, \mathbf{z}_t^{k'})$ and the fact that $\sum_{k=0}^N \mathbf{a}_t^{k,j} = 1$, we have:

$$\begin{aligned}
\left\|\mathbf{z}_{t+1}^j - \mathbf{z}_{t+1}^{j'}\right\|_2 &\leq \sum_{k=0}^N \sum_{k'=0}^N \mathbf{a}_t^{k,j} \mathbf{a}_t^{k',j'} \mathbf{d}\left(Z_t\right) \\
&= \left(\sum_{k=0}^N \mathbf{a}_t^{k,j}\right)\left(\sum_{k'=0}^N \mathbf{a}_t^{k',j'}\right) \mathbf{d}\left(Z_t\right) \\
&= \mathbf{d}\left(Z_t\right).
\end{aligned}$$

Therefore, we reach $\mathbf{d}\left(Z_{t+1}\right) \leq \mathbf{d}\left(Z_t\right)$.

Proof of Part (ii). For any $j, j' \in G^m$, we derive as

$$\begin{aligned}
\left\|\mathbf{z}_t^j - \mathbf{z}_t^{j'}\right\|_2 &= \left\|\sum_{k=0}^{N} \mathbf{a}_t^{k,j} W_V \mathbf{z}_t^k - \sum_{k=0}^{N} \mathbf{a}_t^{k,j'} W_V \mathbf{z}_t^k\right\|_2 = h\left(\sum_{k=0}^{N} \mathbf{a}_t^{k,j} \mathbf{z}_t^k, \sum_{k=0}^{N} \mathbf{a}_t^{k,j'} \mathbf{z}_t^k\right) \\
&\leq h\left(W_V \sum_{k=0}^{N} \mathbf{a}_t^{k,j} \mathbf{z}_t^k, W_V \sum_{k=0}^{N} \mathbf{a}_t^{k,j'} \mathbf{z}_t^k\right) \\
&\leq \sum_{k=0}^{N}\sum_{k'=0}^{N} \mathbf{a}_t^{k,j} \mathbf{a}_t^{k',j'} h\left(\mathbf{z}_t^k, \mathbf{z}_t^{k'}\right) \\
&\leq \sum_{k\in G^m}\sum_{k'\in G^m} \mathbf{a}_t^{k,j} \mathbf{a}_t^{k',j'} h\left(\mathbf{z}_t^k, \mathbf{z}_t^{k'}\right) + \sum_{k\in G^m}\sum_{k'\notin G^m} \mathbf{a}_t^{k,j} \mathbf{a}_t^{k',j'} h\left(\mathbf{z}_t^k, \mathbf{z}_t^{k'}\right) \\
&\quad + \sum_{k\notin G^m}\sum_{k'\in G^m} \mathbf{a}_t^{k,j} \mathbf{a}_t^{k',j'} h\left(\mathbf{z}_t^k, \mathbf{z}_t^{k'}\right) + \sum_{k\notin G^m}\sum_{k'\notin G^m} \mathbf{a}_t^{k,j} \mathbf{a}_t^{k',j'} h\left(\mathbf{z}_t^k, \mathbf{z}_t^{k'}\right) \\
&\leq \sum_{k\in G^m}\sum_{k'\in G^m} \mathbf{a}_t^{k,j} \mathbf{a}_t^{k',j'} \mathbf{d}\left(Z_t^m\right) + \sum_{k\in G^m}\sum_{k'\notin G^m} \mathbf{a}_t^{k,j} \mathbf{a}_t^{k',j'} \mathbf{d}\left(Z_t\right) \\
&\quad + \sum_{k\notin G^m}\sum_{k'\in G^m} \mathbf{a}_t^{k,j} \mathbf{a}_t^{k',j'} \mathbf{d}\left(Z_t\right) + \sum_{k\notin G^m}\sum_{k'\notin G^m} \mathbf{a}_t^{k,j} \mathbf{a}_t^{k',j'} \mathbf{d}\left(Z_t\right).
\end{aligned}$$

We further manipulate as

$$\begin{aligned}
\sum_{k\in G^m}\sum_{k'\in G^m} \mathbf{a}_t^{k,j} \mathbf{a}_t^{k',j'} \mathbf{d}\left(Z_t^m\right) &= \mathbf{d}\left(Z_t^m\right) \sum_{k\in G^m} \mathbf{a}_t^{k,j} \sum_{k'\in G^m} \mathbf{a}_t^{k',j'} \\
&= \mathbf{d}\left(Z_t^m\right)\left(1 - \sum_{k\notin G^m} \mathbf{a}_t^{k,j}\right)\left(1 - \sum_{k'\notin G^m} \mathbf{a}_t^{k',j'}\right) \\
&< \mathbf{d}\left(Z_t^m\right)(1 - A_m\epsilon_l)^2 \mathbf{d}\left(Z_t^m\right).
\end{aligned}$$

$$\begin{aligned}
\sum_{k\in G^m}\sum_{k'\notin G^m} \mathbf{a}_t^{k,j} \mathbf{a}_t^{k',j'} \mathbf{d}(Z_t) &= \mathbf{d}(Z_t) \sum_{k\in G^m} \mathbf{a}_t^{k,j} \sum_{k'\notin G^m} \mathbf{a}_t^{k',j'} \\
&< A_m\epsilon_u \mathbf{d}(Z_t).
\end{aligned}$$

$$\begin{aligned}
\sum_{k\notin G^m}\sum_{k'\in G^m} \mathbf{a}_t^{k,j} \mathbf{a}_t^{k',j'} \mathbf{d}(Z_t) &= \mathbf{d}(Z_t) \sum_{k\notin G^m} \mathbf{a}_t^{k,j} \sum_{k'\in G^m} \mathbf{a}_t^{k',j'} \\
&< A_m\epsilon_u \mathbf{d}(Z_t).
\end{aligned}$$

$$\begin{aligned}
\sum_{k\notin G^m}\sum_{k'\notin G^m} \mathbf{a}_t^{k,j} \mathbf{a}_t^{k',j'} \mathbf{d}(Z_t) &= \mathbf{d}(Z_t) \sum_{k\notin G^m} \mathbf{a}_t^{k,j} \sum_{k'\notin G^m} \mathbf{a}_t^{k',j'} \\
&< A_m^2\epsilon_u^2 \mathbf{d}(Z_t).
\end{aligned}$$

Finally, we arrive at

$$\left\|\mathbf{z}_t^j - \mathbf{z}_t^{j'}\right\|_2^2 < (1 - A_m\epsilon_l)^2 \mathbf{d}(Z_t^m) + A_m\epsilon_u \mathbf{d}(Z_t)(A_m\epsilon_u + 2).$$

$$\mathbf{d}\left(Z_{t+1}^m\right) < (1 - A_m\epsilon_l)^2 \mathbf{d}(Z_t^m) + A_m\epsilon_u \mathbf{d}(Z_t)(A_m\epsilon_u + 2).$$

A.2. Proof of Theorem 3.2

Under the assumption, we have

$$1 - \frac{x_i^\top x_j}{\|x_i\|_2 \|x_j\|_2} = 1 - \frac{y_i^\top y_j}{\|y_i\|_2 \|y_j\|_2}$$
$$\tilde{x}_i^T \tilde{x}_j = \tilde{y}_i^T \tilde{y}_j, \forall i, j$$

Let denote $G_X = \left[\tilde{x}_i^T \tilde{x}_j\right]_{i,j=1}^n$ and $G_Y = \left[\tilde{y}_i^T \tilde{y}_j\right]_{i,j=1}^n$ as two Gram matrices, we then have $G_X = G_Y$. Moreover, we have $rank\left(G_X\right) = dim\left(span\left\{\tilde{x}_1, ..., \tilde{x}_n\right\}\right) = m$ and $rank\left(G_Y\right) = dim\left(span\left\{\tilde{y}_1, ..., \tilde{y}_n\right\}\right) = m$. Without loss of generality, we assume that $\{\tilde{x}_1, ..., \tilde{x}_m\}$ and $\{\tilde{y}_1, ..., \tilde{y}_m\}$ form basis of $span\left\{\tilde{x}_1, ..., \tilde{x}_n\right\}$ and $span\left\{\tilde{y}_1, ..., \tilde{y}_n\right\}$ respectively.

Now assume that $\tilde{x}_k = \sum_{i=1}^m \alpha_i \tilde{x}_i$ and $\tilde{y}_k = \sum_{i=1}^m \beta_i \tilde{y}_i$. It appears that

$$\left[\begin{array}{cccc} G_X[k,1] & ... & G_X[k,m] \end{array}\right] = \tilde{x}_k^T \left[\begin{array}{ccc} \tilde{x}_1 & ... & \tilde{x}_m \end{array}\right] = \alpha \left[\begin{array}{ccc} \tilde{x}_1^T \tilde{x}_1 & ... & \tilde{x}_1^T \tilde{x}_m \\ ... & ... & ... \\ \tilde{x}_m^T \tilde{x}_1 & ... & \tilde{x}_m^T \tilde{x}_m \end{array}\right]$$

$$\left[\begin{array}{cccc} G_Y[k,1] & ... & G_Y[k,m] \end{array}\right] = \tilde{y}_k^T \left[\begin{array}{ccc} \tilde{y}_1 & ... & \tilde{y}_m \end{array}\right] = \beta \left[\begin{array}{ccc} \tilde{y}_1^T \tilde{y}_1 & ... & \tilde{y}_1^T \tilde{y}_m \\ ... & ... & ... \\ \tilde{y}_m^T \tilde{y}_1 & ... & \tilde{y}_m^T \tilde{y}_m \end{array}\right]$$

Therefore, we have

$$\alpha \left[\begin{array}{ccc} \tilde{x}_1^T \tilde{x}_1 & ... & \tilde{x}_1^T \tilde{x}_m \\ ... & ... & ... \\ \tilde{x}_m^T \tilde{x}_1 & ... & \tilde{x}_m^T \tilde{x}_m \end{array}\right] = \beta \left[\begin{array}{ccc} \tilde{y}_1^T \tilde{y}_1 & ... & \tilde{y}_1^T \tilde{y}_m \\ ... & ... & ... \\ \tilde{y}_m^T \tilde{y}_1 & ... & \tilde{y}_m^T \tilde{y}_m \end{array}\right],$$

leading to $\alpha = \beta$ because two Gram matrices are identical and non-singular.

We extend the orthonormal set $\{e_j\}_{j=1}^m$ to a full orthonormal basis $\{e_1, \ldots, e_{d_x}\}$ of $\mathbb{R}^{d_x}$. Similarly, since $d_y \geq d_x$, we extend $\{f_j\}_{j=1}^m$ to an orthonormal set $\{f_1, \ldots, f_{d_x}\} \subset \mathbb{R}^{d_y}$, which is a subset of a full orthonormal basis of $\mathbb{R}^{d_y}$.

Define the matrices

$$E = [e_1, \ldots, e_{d_x}] \in \mathbb{R}^{d_x \times d_x}, \qquad F = [f_1, \ldots, f_{d_x}] \in \mathbb{R}^{d_y \times d_x}.$$

Since the vectors $\{e_k\}$ and $\{f_k\}$ are orthonormal, we have

$$E^\top E = I_{d_x}, \qquad F^\top F = I_{d_x}.$$

We define

$$Q = FE^\top \in \mathbb{R}^{d_y \times d_x}.$$

We first verify that Q lies on the Stiefel manifold $St(d_x, d_y)$. Indeed,

$$Q^\top Q = (FE^\top)^\top (FE^\top) = EF^\top FE^\top = EI_{d_x} E^\top = EE^\top = I_{d_x}.$$

Thus, $Q \in St(d_x, d_y)$.

By construction of $Q = FE^\top$, we have

$$QE = FE^\top E = F,$$

which implies $Qe_k = f_k$ for all k. Therefore, the linear map Q sends each basis vector e_k to its corresponding vector f_k. We now prove that $Q\tilde{x}_k = \tilde{y}_k$.

Indeed, according to the Gram–Schmidt orthogonalization, we have $f_j^T \tilde{y}_i = e_j^T \tilde{x}_i$ for $i, j \leq m$. For $k \leq m$, we have $\tilde{x}_k = \sum_{i=1}^m e_i \left(\tilde{x}_k^T e_i\right)$. This follows

$$Q\tilde{x}_k = \sum_{i=1}^m Qe_i \left(\tilde{x}_k^T e_i\right) = \sum_{i=1}^m f_i \left(\tilde{y}_k^T f_i\right) = \tilde{y}_k$$

Finally, for $k > m$, we have

$$Q\tilde{x}_k = \left[\begin{array}{ccc} Q\tilde{x}_1 & ... & Q\tilde{x}_m \end{array}\right] \alpha^T = \left[\begin{array}{ccc} \tilde{y}_1 & ... & \tilde{y}_m \end{array}\right] \beta^T = \tilde{y}_k.$$

A.3. Proof of Theorem 3.3

Under the assumptions, we have

$$d\left(U_{i,l}^{T,\text{image}}, U_{m,l}^{T,\text{text}}\right)^2 = d\left(U_{i,l}^{S,\text{image}}, U_{m,l}^{S,\text{text}}\right)^2$$
$$\left(\tilde{U}_{i,l}^{T,\text{image}}\right)^T \tilde{U}_{m,l}^{T,\text{text}} = \left(\tilde{U}_{i,l}^{S,\text{image}}\right)^T \tilde{U}_{m,l}^{S,\text{text}}.$$

This means that

$$\left(\tilde{U}_{i,l}^{S,\text{image}}\right)^T \left(Q^{\text{image}}\right)^T Q^{\text{text}} \tilde{U}_{m,l}^{S,\text{text}} = \left(\tilde{U}_{i,l}^{S,\text{image}}\right)^T \tilde{U}_{m,l}^{S,\text{text}}$$
$$\left(\tilde{U}_{i,l}^{S,\text{image}}\right)^T \left(\left(Q^{\text{image}}\right)^T Q^{\text{text}} - \mathbf{I}\right) \tilde{U}_{m,l}^{S,\text{text}} = 0.$$

Using the property of the trace, we gain

$$\text{trace}\left(\left(\left(Q^{\text{image}}\right)^T Q^{\text{text}} - \mathbf{I}\right)^T \tilde{U}_{i,l}^{S,\text{image}} \left(\tilde{U}_{m,l}^{S,\text{text}}\right)^T\right) = 0$$

$$\left\langle \left(Q^{\text{image}}\right)^T Q^{\text{text}} - \mathbf{I}, \tilde{U}_{i,l}^{S,\text{image}} \left(\tilde{U}_{m,l}^{S,\text{text}}\right)^T \right\rangle = 0.$$

This further implies that $\left(Q^{\text{image}}\right)^T Q^{\text{text}} - \mathbf{I}$ is orthogonal to span $\left\{\tilde{U}_{i,l}^{S,\text{image}} \left(\tilde{U}_{m,l}^{S,\text{text}}\right)^T : i, m\right\} = \mathbb{R}^{d\times d}$, implying that $\left(Q^{\text{image}}\right)^T Q^{\text{text}} - \mathbf{I} = \mathbf{0}$.

B. Theoretical background of Relational KD

For a pair of samples, the distance-wise relational metric ψ_{D} measures the Euclidean distance between their representations. The corresponding distillation loss $\mathcal{L}_{\text{RKD-D}}$ is defined over the set of all pairs of distinct examples $\mathcal{X}^2$ in a mini-batch:

$$\mathcal{L}_{\text{RKD-D}} = \sum_{(x_i,x_j)\in\mathcal{X}^2} l_\delta(\psi_{\text{D}}(t_i,t_j), \psi_{\text{D}}(s_i,s_j)),$$

where t and s denote the teacher and student representations, respectively. Furthermore, to capture higher-order geometric structures, an angle-wise relational potential ψ_{A} is computed for each triplet of examples as the cosine of the angle formed at the j-th example:

$$\psi_{\text{A}}(t_i,t_j,t_k) = \cos\angle t_i t_j t_k = \langle \mathbf{e}^{ij}, \mathbf{e}^{kj}\rangle,$$

where the normalized direction vectors are given by $\mathbf{e}^{ij} = \frac{t_i - t_j}{\|t_i - t_j\|_2}$ and $\mathbf{e}^{kj} = \frac{t_k - t_j}{\|t_k - t_j\|_2}$. The angle-wise distillation loss is then formulated as:

$$\mathcal{L}_{\text{RKD-A}} = \sum_{(x_i,x_j,x_k)\in\mathcal{X}^3} l_\delta(\psi_{\text{A}}(t_i,t_j,t_k), \psi_{\text{A}}(s_i,s_j,s_k)).$$

In both objectives, l_δ denotes the Huber loss, which provides robustness to outliers and is defined as:

$$l_\delta(x,y) = \begin{cases} \frac{1}{2}(x-y)^2 & \text{for } |x-y| \le 1, \\ |x-y| - \frac{1}{2} & \text{otherwise.} \end{cases}$$

The overall loss is defined as $L_{RKD} = L_{RKD-A} + L_{RKD-D}$.

C. Experimental Details

C.1. Baselines

We compare our method against several representative knowledge distillation baselines, which are summarized below.

- **MSE**: A point-wise distillation baseline where teacher and student embeddings are first projected into the same dimensional space, followed by alignment using Mean Squared Error.
- **Relational KD (RKD)** (Park et al., 2019): A structural distillation method that preserves inter-instance relationships by matching distance-wise and angle-wise potentials induced by the teacher.
- **Comparative KD (CKD)** (Wilf et al., 2025): A structural distillation approach designed for reduced-teacher-inference (RTI) settings. It replaces low-dimensional relations with high-dimensional comparisons between data points, enabling robust structural knowledge transfer under limited teacher supervision.
- **EMO** (Truong et al., 2025): A cross-tokenizer distillation framework originally proposed for text-only models. In our experiments, we follow the original training and distillation settings described in (Truong et al., 2025) and adapt them to the multimodal setting accordingly.
- **EM-KD** (Feng et al., 2025b): A method for improving efficient multimodal large language models by addressing unbalanced vision tokens. Our implementation follows the pseudo-code and training procedure provided in the original paper.

C.2. Implementation Details

Models Our student models include **FastVLM-0.5B** (Vasu et al., 2025) and **LLaVA-OneVision-0.5B** (Li et al., 2024), two lightweight vision-language models. For the distillation process, we employ **B3** (Thirukovalluru et al., 2025) as the teacher model, which currently ranks among the state-of-the-art methods on the MMEB leaderboard. The teacher model (B3) is kept fixed during training. All models adopt the `eos` pooling strategy to produce fixed-size embeddings.

The detailed training configurations for each model are summarized in Table 4 and Table 5.

Training and Evaluation The distillation process is conducted on two different benchmarks from the **MMEB** training benchmark, namely classification (CLS) and vision question answering (VQA). Parameter-efficient fine-tuning is applied via LoRA for all student models.

Table 4. Training configurations for all baseline methods, except EM-KD method.

Settings	FastVLM-0.5B	LLaVA-OneVision-0.5B
Epoch	1	1
Learning Rate	1×10^{-4}	1×10^{-4}
Projector LR	5×10^{-4}	5×10^{-4}
Batch Size	16	8
LR Scheduler	Cosine	Cosine
Warmup Ratio	0.03	0.03
Weight Decay	0.01	0.01
LoRA Rank	64	64
LoRA Alpha	64	64
Image Resolution	448	336

All experiments are implemented in **Python 3.11** using **PyTorch**. Training is performed on a **NVIDIA A100 SXM 80GB VRAM GPU**, and the `bf16` precision format is adopted to accelerate training and reduce memory consumption. To accommodate the high memory overhead of the EM-KD baseline within our available computational environment, we adapted its training configuration by reducing the effective batch size and the input image resolution.

Table 5. Training configurations for EM-KD method.

Settings	FastVLM-0.5B	LLaVA-OneVision-0.5B
Epoch	1	1
Learning Rate	1×10^{-4}	1×10^{-4}
Projector LR	5×10^{-4}	5×10^{-4}
Batch Size	8	4
LR Scheduler	Cosine	Cosine
Warmup Ratio	0.03	0.03
Weight Decay	0.01	0.01
LoRA Rank	64	64
LoRA Alpha	64	64
Image Resolution	448	128

Hyperparameters. The hyperparameters, α, λ_{struct} and λ_{hid}, are reported in Table 6. All reported values are selected based on preliminary validation experiments.

Table 6. Loss weighting coefficients used for different methods and model sizes.

Datasets	FastVLM-0.5B			LLaVA-OneVision-0.5B		
	α	λ_{struct}	λ_{hid}	α	λ_{struct}	λ_{hid}
CLS	0.25	2.5	0.25	0.25	2.5	0.25
VQA	0.25	2.5	0.25	0.25	2.5	0.25

Layer Mapping Configuration. Table 7 summarizes the selected intermediate layers used for hierarchical distillation at the word and phrase levels. To accommodate differences in network depth between the Student and Teacher models, we align intermediate representations using a fixed proportional mapping rather than performing full-depth layer-wise matching. This design choice is motivated by the well-documented inter-layer redundancy in Transformer architectures, where adjacent layers often encode highly similar features.

Moreover, prior studies such as EMKD (Feng et al., 2025b) and AlignKD (Feng et al., 2025a) observe that, in multimodal Transformers, the most pronounced modality-specific transformations for visual representations injected into the LLM backbone occur at the earliest layers. These initial layers are responsible for projecting image features into the shared language-centric representation space, while deeper layers primarily refine already aligned semantic representations. Motivated by this observation, we explicitly include the first Transformer layer in our distillation set to ensure that early-stage cross-modal alignment is effectively transferred from the Teacher to the Student.

Concretely, let N_S and N_T denote the number of layers in the Student and Teacher models, respectively. Each selected Student layer index l_S is mapped to its corresponding Teacher layer via proportional scaling, $l_T = \left\lfloor l_S \times \frac{N_T}{N_S} \right\rfloor$. Instead of exhaustively aligning all layers, we select a compact set of *key layers* $\mathcal{L}_{\text{key}}$ using a strided top-down strategy that prioritizes high-level semantic transitions closer to the output. Given a stride k and a layer budget M, the selected Student indices are defined as $\mathcal{I}_S = \{N_S, N_S - k, \dots\}$ with $|\mathcal{I}_S| = M$. In practice, we use small strides (typically $k \in \{3, 4\}$), which effectively reduce redundancy while preserving a coherent representation trajectory for distillation.

Table 7. Selected intermediate layers for hierarchical distillation under different tasks. CLS and VQA denote classification-oriented and question-answering-oriented supervision, respectively.

Model	CLS		VQA	
	Word	Phrase	Word	Phrase
FastVLM 0.5B	0	18, 21, 24	0	18, 21, 24
LLaVA-OneVision 0.5B	0	18, 21, 24	0	18, 21, 24

C.3. Dataset statistic

Table 8 presents the number of instances of each dataset from each task for the training and testing processes.

Table 8. The statistics of the training datasets (Jiang et al., 2024)

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

Classification datasets: The query consists of an instruction, an image, optionally accompanied by related text, while the target is the class label. The number of candidates equals the number of classes.

Visual Question Answering datasets: The query consists of an instruction, an image, and a piece of text as the question, while the target is the answer. Each query has 1 ground truth and 999 distractors as candidates.

D. Details of Vision Token Clustering and Teacher–Student Mapping

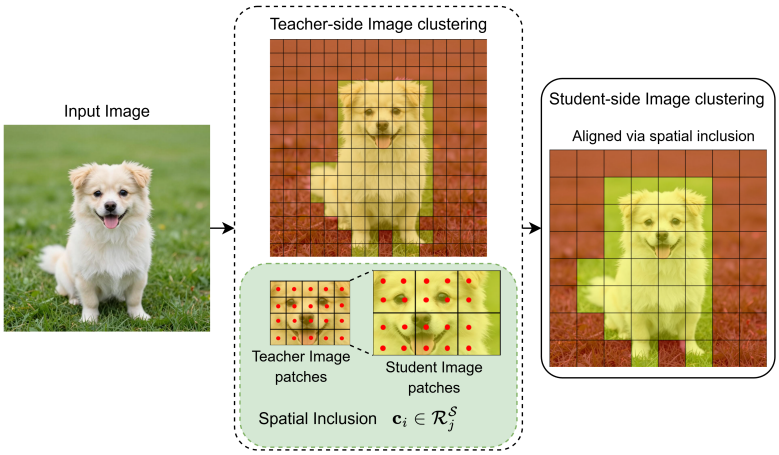

Figure 3. Illustration of the vision token clustering and teacher–student mapping procedure. Teacher-side visual tokens, represented by their patch centers $\mathbf{c}_i$, are first grouped into object-level clusters using DBSCAN. Each cluster is then associated with student visual tokens according to spatial inclusion within the corresponding patch regions $\mathcal{R}_j^{\mathcal{S}}$.

Figure 3 provides a schematic overview of the proposed vision token clustering and the subsequent teacher–student mapping procedure. In this section, we formally define the geometric quantities involved and describe how the mapping is computed in practice.

Teacher token coordinates. Let the teacher vision encoder process an input image of resolution $H_{\mathcal{T}} \times W_{\mathcal{T}}$, producing a set of visual patch tokens $\{z_i^{\mathcal{T}}\}_{i=1}^{N_{\mathcal{T}}}$. Each token corresponds to a patch centered at pixel location $(u_i^{\mathcal{T}}, v_i^{\mathcal{T}})$ in the teacher input space. We define its normalized spatial coordinate as

$$\mathbf{c}_i = (x_i, y_i) = \left(\frac{u_i^{\mathcal{T}}}{W_{\mathcal{T}}}, \frac{v_i^{\mathcal{T}}}{H_{\mathcal{T}}} \right) \in [0,1]^2.$$

Student patch regions. Due to differences in vision preprocessors between the teacher and student models, the two encoders may operate on input images with different spatial resolutions. As a result, the student vision encoder processes images of resolution $H_{\mathcal{S}} \times W_{\mathcal{S}}$, which may differ from the teacher resolution $H_{\mathcal{T}} \times W_{\mathcal{T}}$. Each student visual token $z_j^{\mathcal{S}}$ corresponds to a rectangular region in the normalized image plane, defined as

$$\mathcal{R}_j^{\mathcal{S}} = [x_j^{\min}, x_j^{\max}) \times [y_j^{\min}, y_j^{\max}) \subset [0,1]^2,$$

where

$$x_j^{\min} = \frac{p_j^x}{W_{\mathcal{S}}}, \quad x_j^{\max} = \frac{p_j^x + w_j}{W_{\mathcal{S}}}, \quad y_j^{\min} = \frac{p_j^y}{H_{\mathcal{S}}}, \quad y_j^{\max} = \frac{p_j^y + h_j}{H_{\mathcal{S}}}.$$

Here (p_j^x, p_j^y) denotes the top-left pixel coordinate of the student patch, and (w_j, h_j) its width and height. By construction, the set of regions $\{\mathcal{R}_j^{\mathcal{S}}\}$ forms a partition of the normalized image plane, independent of the teacher resolution.

Region-based cluster mapping. Given a teacher-side visual token cluster $C_m^{\mathcal{T}}$, the corresponding student-side cluster is defined by spatial inclusion:

$$C_m^{\mathcal{S}} = \left\{ j \;\middle|\; \exists\, i \in C_m^{\mathcal{T}} \text{ such that } \mathbf{c}_i \in \mathcal{R}_j^{\mathcal{S}} \right\}.$$

That is, a student token is associated with a teacher cluster if the center of at least one teacher patch belonging to that cluster falls within the spatial region represented by the student token. This formulation naturally supports many-to-one correspondence, where multiple fine-grained teacher tokens are mapped to a single coarser student patch, as illustrated in Figure 3.

E. Additional Experimental Results

E.1. Impact of Text Span Granularity

HieRD adopts a hierarchical span construction that assigns word-level spans to lower layers and phrase-level spans to higher layers, reflecting the progressive abstraction in Transformer representations. We compare this strategy with two static alternatives that apply a fixed granularity across all layers: word-level only and phrase-level only. As reported in Table 9, the hierarchical design consistently yields the strongest performance across all benchmarks. In particular, HieRD achieves accuracies of 56.0 on ImageNet-1K, 80.6 on VOC2007, and 67.2 on SUN397, exceeding the word-level baseline on all three datasets. While the phrase-level strategy performs competitively on VOC2007, it shows weaker results on ImageNet-1K and SUN397. These findings suggest that adapting text span granularity to layer depth provides a more effective supervision signal than using a single, fixed abstraction level.

Table 9. Ablation study on text span granularity using static and hierarchical strategies.

Granularity Strategy	ImageNet-1K	VOC2007	SUN397
Word-level only	54.9	79.7	66.5
Phrase-level only	55.2	80.6	66.4
HieRD (Ours)	**56.0**	**80.6**	**67.2**

E.2. Additional Empirical Evidence for Theorem 3.3

This subsection provides additional empirical evidence supporting the assumptions and conclusions of Theorem 3.3. In particular, we examine whether the intra-modal and cross-modal objectives can be effectively optimized in practice, and whether their optimization leads to the geometric behaviors predicted by Theorems 3.2 and 3.3.

To enable rapid and controlled verification, we conduct a lightweight experiment using B3-Qwen2-2B (Thirukovalluru et al., 2025) as the teacher model and FastVLM-0.5B (Vasu et al., 2025) as the student model, evaluated on the VOC2007 dataset. All optimization settings and loss formulations are kept identical to those used in the main experiments.

We first examine the behavior of the intra-modal objectives. Figure 4 plots the evolution of the intra-image and intra-text losses during training. Both losses exhibit a clear decreasing trend and converge toward zero, indicating that the student representations increasingly preserve pairwise relationships within each modality. This empirical observation supports the

assumption of Theorem 3.2, suggesting the existence of approximately orthogonal linear mappings Q^{image} and Q^{text} that align student representations with those of the teacher within each modality.

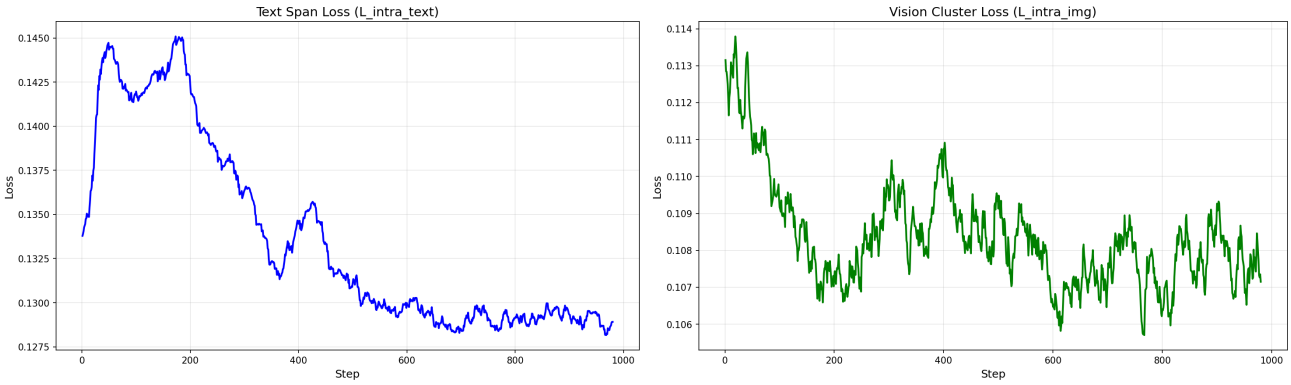

Figure 4. Evolution of intra-modal losses on VOC2007. Both intra-image and intra-text losses converge toward zero during training.

We further analyze the cross-modal objective by monitoring the cross-modal alignment loss. As shown in Figure 5, this loss also gradually decreases and converges over the course of training. Together with the convergence of the intra-modal losses, this behavior is consistent with the conditions of Theorem 3.3, suggesting that the learned transformations Q^{image} and Q^{text} become increasingly aligned during optimization. This observation aligns with the orthogonality analysis reported in the main paper, where $(Q^{\text{image}})^\top Q^{\text{text}}$ approaches the identity matrix during training.

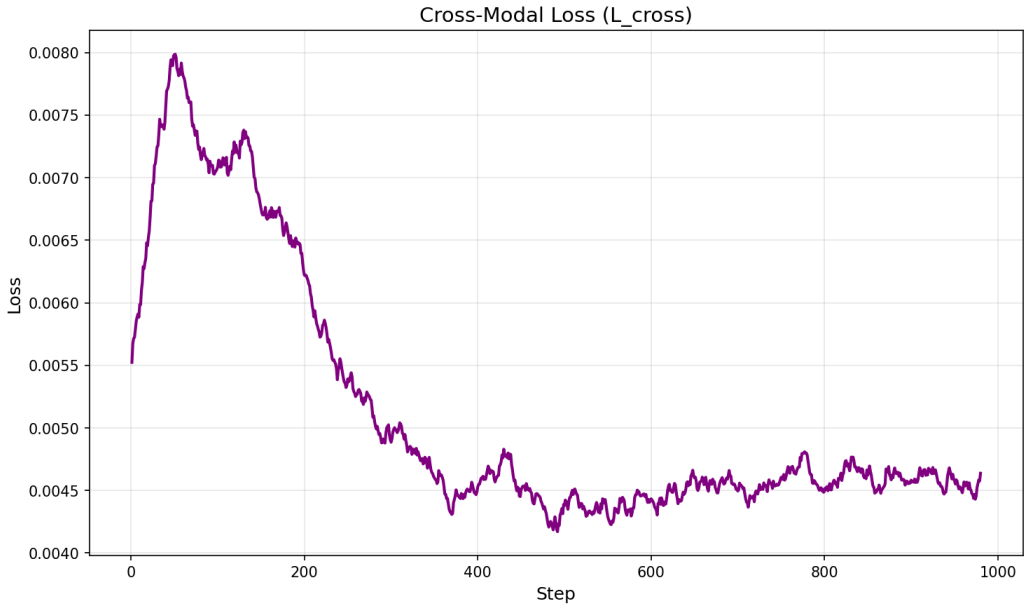

Figure 5. Evolution of the cross-modal loss on VOC2007. The loss decreases steadily, indicating improved alignment between image and text representations.

Although these results do not constitute a formal proof, they empirically validate the key assumptions underlying Theorem 3.3. In particular, they demonstrate that under sufficiently optimized intra-modal and cross-modal objectives, the learning dynamics naturally encourage convergence toward a unified cross-modal geometry, making the proposed theoretical framework practically meaningful.