HiMo-CLIP: Modeling Semantic Hierarchy and Monotonicity in Vision-Language Alignment

Ruijia Wu[1,2†], **Ping Chen**[1,2†], **Fei Shen**[3], **Shaoan Zhao**[1,2], **Qiang Hui**[1,2], **Huanlin Gao**[1,2], **Ting Lu**[1,2], **Zhaoxiang Liu**[1,2], **Fang Zhao**[1,2*], **Kai Wang**[1,2], **Shiguo Lian**[1,2*]

[1]Data Science & Artificial Intelligence Research Institute, China Unicom
[2]Unicom Data Intelligence, China Unicom
[3]National University of Singapore
{wurj25, chenp181, zhaof50, liansg}@chinaunicom.cn, shenfei29@nus.edu.sg

Abstract

Contrastive vision-language models like CLIP have achieved impressive results in image-text retrieval by aligning image and text representations in a shared embedding space. However, these models often treat text as flat sequences, limiting their ability to handle complex, compositional, and long-form descriptions. In particular, they fail to capture two essential properties of language: **semantic hierarchy**, *which reflects the multi-level compositional structure of text*, and **semantic monotonicity**, *where richer descriptions should result in stronger alignment with visual content*. To address these limitations, we propose HiMo-CLIP, a representation-level framework that enhances CLIP-style models without modifying the encoder architecture. HiMo-CLIP introduces two key components: a hierarchical decomposition (HiDe) module that extracts latent semantic components from long-form text via in-batch PCA, enabling flexible, batch-aware alignment across different semantic granularities, and a monotonicity-aware contrastive loss (MoLo) that jointly aligns global and component-level representations, encouraging the model to internalize semantic ordering and alignment strength as a function of textual completeness. These components work in concert to produce structured, cognitively-aligned cross-modal representations. Experiments on multiple image-text retrieval benchmarks show that HiMo-CLIP consistently outperforms strong baselines, particularly under long or compositional descriptions.

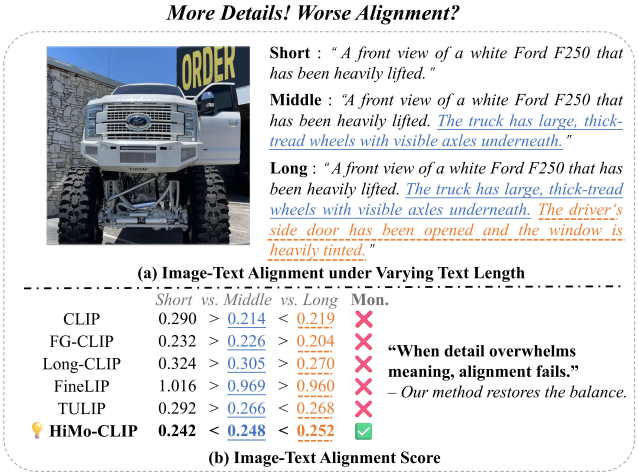


Figure 1: (a) Text descriptions become semantically richer as visual details increase from short to long. (b) However, existing models often break semantic monotonicity across description granularities, whereas HiMo-CLIP preserves consistent alignment. (*Note: FineLIP similarity* > 1 *is due to customized test-time scaling.*)

1 Introduction

Cross-modal contrastive learning has become a foundational approach for vision-language tasks, demonstrating strong performance in image-text retrieval, zero-shot classification, and image captioning. Models such as CLIP (Radford et al. 2021) achieve this by projecting images and texts into a shared embedding space, where semantically aligned pairs are drawn closer while mismatched pairs are pushed apart. However, these models typically treat textual inputs as flat, unstructured sequences, ignoring the rich semantic hierarchy and compositionality inherent in natural language (Lewis et al. 2022; Ji, Chen, and Wang 2021; Koukounas et al. 2024). This simplification limits their ability to fully exploit longer or more complex descriptions that contain multi-level semantics beyond short captions (Guo et al. 2023; Wu et al. 2021).

In practice, a single image may be described at multiple semantic levels. As shown in Figure 1, the same image of a white Ford F250 can be paired with either a short caption, "A front view of a white Ford F250 that has been heavily lifted", or a longer, more detailed description that elaborates on visual attributes such as the oversized wheels, visible axles, and tinted windows. A cognitively aligned model should not only assess whether a text matches the image, but also reason about how it matches, through object category, appearance, or fine-grained contextual cues (Zeng, Zhang, and Li 2021; Tian, Wu, and Yang 2025; Sun et al. 2024). This motivates two underexplored properties in vision-language contrastive learning: **semantic hierarchy**, *the ability to repre-*

*Corresponding authors.
†Equal contribution.
This work was supported by the National Natural Science Foundation of China Enterprise Innovation and Development Joint Fund Project U24B20181.

sent and align descriptions across varying levels of semantic granularity, and **semantic monotonicity**, *the expectation that more informative and complete descriptions should lead to stronger alignment with the corresponding image*.

Modeling these properties poses significant challenges. Existing methods often rely on fixed-length captions or handcrafted subphrases to approximate semantic granularity, yet such designs overlook the dynamic and context-dependent nature of semantic focus. As illustrated in Figure 1, the long-form description of the Ford F250 includes multiple visual cues, such as "oversized wheels," "visible axles," and "tinted windows." However, which part of the description should be most semantically aligned depends on the batch context. For instance, in a batch containing other trucks, "tinted windows" may offer the most discriminative cue, whereas in a batch with various vehicle types, "Ford F250" or "heavily lifted" may dominate the alignment. Static, substring-based decomposition, such as fixed truncation or manual segmentation, cannot flexibly adapt to such shifts, and may introduce semantic noise or supervision bias (Yuksekgonul et al. 2022). This calls for a data-driven, context-aware mechanism to decompose and align semantic components at varying levels of granularity.

To address the above challenges, we propose HiMo-CLIP, a representation-level framework that enhances CLIP-style models by explicitly modeling semantic hierarchy and monotonicity, without modifying the underlying encoder architecture. HiMo-CLIP introduces two key components: a Hierarchical Decomposition (HiDe) module and a Monotonicity-aware contrastive Loss (MoLo). The HiDe module leverages in-batch Principal Component Analysis (PCA) (Jolliffe 2011) to extract latent semantic components from long-form text, enabling flexible, batch-aware alignment across varying semantic granularities. This allows the same sentence to emphasize different aspects, such as object type, attributes, or context-depending on the distribution of the current batch. Unlike manual subphrases or truncated captions, these components are derived in a self-supervised and context-adaptive manner, ensuring both semantic consistency and scalability during training. The MoLo loss complements HiDe by introducing a dual-branch alignment objective: one branch aligns global image-text pairs, while the other aligns image features with each semantic component independently. This formulation encourages the model to internalize semantic ordering, i.e., more detailed and informative texts should yield stronger alignment signals. Notably, all components operate purely in the embedding space, requiring no changes to the pretrained encoder or additional supervision. Extensive experiments across multiple image-text retrieval benchmarks demonstrate that HiMo-CLIP consistently outperforms strong baselines, particularly under long-form or compositional descriptions.

Our contributions are summarized as follows:

- We identify and formally define **semantic hierarchy** and **semantic monotonicity** as two fundamental yet underexplored properties of cross-modal contrastive learning, essential for modeling multi-granular and completeness-sensitive text-image alignment.
- We propose **HiMo-CLIP**, a self-supervised and encoder-agnostic framework that contains a Hierarchical Decomposition module (HiDe) and a Monotonicity-aware Contrastive loss (MoLo). Together, they enable structured, context-aware, and monotonic alignment without requiring architectural modifications or external annotations.
- We validate HiMo-CLIP on multiple image-text retrieval benchmarks and demonstrate consistent improvements over strong CLIP-style baselines, particularly in scenarios involving long-form or compositional descriptions.

2 Related Work

2.1 Vision-Language Pretraining with CLIP

Vision–language pretraining has progressed rapidly with models such as CLIP (Radford et al. 2021), which employ contrastive learning on large-scale image–text pairs to learn transferable cross-modal representations. CLIP uses a dual-encoder architecture with a visual backbone (e.g., ViT) and a text encoder (e.g., Transformer), aligned in a shared embedding space via an InfoNCE loss (Oord, Li, and Vinyals 2018). Despite strong zero-shot and retrieval performance, CLIP is constrained by its reliance on short textual inputs (Alper and Averbuch-Elor 2024; Xu et al. 2023). The 77-token limit, with most information concentrated in the first 20 tokens, restricts its capacity to encode rich and structured semantics (Zhang et al. 2024). As a result, long-form descriptions are truncated, and compositional cues such as attributes and spatial relations are weakened or entangled, yielding flattened embeddings that fail to capture fine-grained semantics (Tschannen et al. 2025; Li et al. 2022; Yamada et al. 2022; Fang et al. 2024; Zhai et al. 2023). Although several datasets evaluate hierarchical structure, they largely focus on short-text alignment and do not fully capture the challenges posed by long-form descriptions (Vulić et al. 2017; Santurkar, Tsipras, and Madry 2020; Alper and Averbuch-Elor 2024).

2.2 Long-Form Text Modeling

To extend contrastive models beyond short captions, recent works have introduced architectural modifications and alignment strategies for handling long-form text. LongCLIP (Zhang et al. 2024) increases token capacity via positional interpolation and aligns global image features with long-text backbones through principal component matching. FineLIP (Asokan, Wu, and Albreiki 2025) refines token-level alignment via adaptive modulation, while DreamLIP (Zheng et al. 2024) generates subcaptions to facilitate hierarchical alignment between local regions and text fragments. FG-CLIP (Xie et al. 2025) introduces region-aware hard negatives to boost fine-grained discrimination. Despite these advances, most methods focus on simplifying the visual side, through downsampling, cropping, or sparse selection, to accommodate longer texts, often degrading visual fidelity. More critically, few address the semantic redundancy and structural diffusion intrinsic to long-form language. This results in unstable performance when input text becomes more detailed. LoTLIP (Wu et al. 2024) introduces clause-aware corner tokens to better capture long-text semantics,

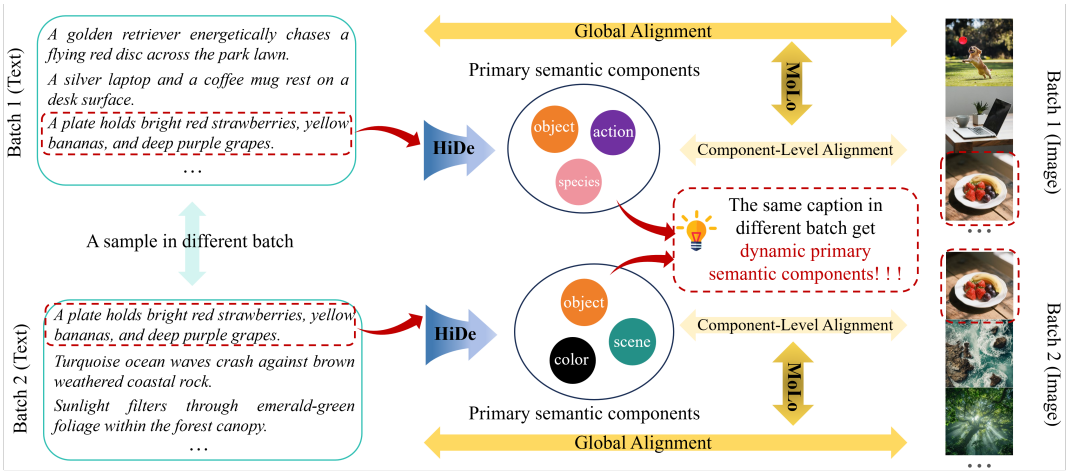


Figure 2: HiMo-CLIP Framework. Our method enhances CLIP with two encoder-agnostic modules: (1) **HiDe**, which applies in-batch PCA to extract discriminative semantic components and reveal a dynamic hierarchy conditioned on batch context; and (2) **MoLo**, which aligns images with both full-text embeddings (*global*) and primary semantic components (*component-level*), encouraging semantic monotonicity.

while TULIP (Najdenkoska et al. 2024) extends token limits via relative position encoding. However, both lack mechanisms for structured semantic compression, leaving redundancy and irrelevant details insufficiently filtered.

Overall, existing models tend to overlook the modality asymmetry: images typically exhibit spatial density and local coherence, whereas long-form text is verbose and hierarchically structured, highlighting the need for structured semantic compression on the language side.

3 Method

We propose HiMo-CLIP, a representation-level framework that enhances CLIP-style models to capture semantic hierarchy and monotonicity without modifying encoder architectures (see Fig. 2). The framework introduces two core components: a Hierarchical Decomposition (HiDe) module that extracts multi-granular semantic components via in-batch PCA, and a Monotonicity-aware contrastive Loss (MoLo) that leverages these components to enforce natural semantic ordering. Operating entirely in the embedding space, HiMo-CLIP enables self-supervised learning of structured, cognitively aligned cross-modal representations.

3.1 Overview

HiMo-CLIP builds upon CLIP's dual-encoder paradigm, where images and texts are projected into a shared embedding space. For an image I, its visual embedding is denoted by $v = f_v(I) \in \mathbb{R}^d$, and for a text T, the textual embedding is $u = f_t(T) \in \mathbb{R}^d$. While CLIP effectively aligns these global representations, it overlooks two essential linguistic properties: semantic hierarchy, which reflects the multi-level compositional structure inherent in natural language, and semantic monotonicity, which implies that more complete textual descriptions should result in stronger alignment with the corresponding image. HiMo-CLIP addresses these limitations by introducing two lightweight and encoder-agnostic modules. The Hierarchical Decomposition (HiDe) module dynamically extracts latent semantic components from text embeddings using in-batch Principal Component Analysis (PCA), capturing varying levels of semantic granularity. The Monotonicity-aware Contrastive Loss (MoLo) jointly aligns images with both full-text embeddings and their semantic components, implicitly encouraging alignment scores to increase as the text becomes more complete. Both modules operate entirely at the representation level, requiring no modification to the pretrained encoders, and enable structured and cognitively consistent cross-modal alignment.

3.2 Hierarchical Decomposition (HiDe) Module

Motivation. Natural language descriptions, particularly long-form ones, often express semantics across multiple levels, such as object categories, attributes, and contextual details. Capturing this inherent hierarchy is crucial for precise vision-language alignment. However, existing methods typically rely on static subphrases or fixed truncation strategies, which are inadequate for adapting to batch-dependent semantic relevance. For example, in a batch dominated by various vehicle types, category-level cues like "Ford F250" may be most salient, whereas in a batch of similar trucks, fine-grained attributes like "tinted windows" may become more informative. A decomposition strategy that is static across batches fails to reflect such context shifts, resulting in suboptimal alignment. To address this, HiDe introduces a dynamic and context-aware mechanism that leverages in-batch PCA to extract latent semantic components from text embeddings. This allows the model to capture semantically meaningful structures at varying levels of granularity, adaptively shaped by the composition of the current batch.

Architecture. Given a mini-batch $\mathcal{B} = \{(I_i, T_i)\}_{i=1}^N$ of N image-text pairs, HiDe decomposes textual embeddings into context-aware semantic components via PCA. Each text input T_i is encoded as $u_i = f_t(T_i) \in \mathbb{R}^d$. The batch mean

embedding is computed as $\bar{u} = \frac{1}{N}\sum_{j=1}^{N} u_j$, and each embedding is centered by $\hat{u}_i = u_i - \bar{u}$. PCA is then performed on the set $\{\hat{u}_i\}_{i=1}^{N}$ through Singular Value Decomposition (SVD) to extract the top m principal components, organized as $\mathbf{P} = [p_1, \ldots, p_m]^\top \in \mathbb{R}^{m \times d}$, where m is chosen such that the cumulative explained variance exceeds a predefined threshold (e.g., 0.9). Each centered embedding is projected onto these principal components and reconstructed by

$$u_i' = \mathbf{P}^\top(\mathbf{P}\hat{u}_i) + \bar{u}. \tag{1}$$

This procedure produces a semantic component vector u_i' for each original embedding u_i, serving as a compact sub-semantic representation that underpins the hierarchical alignment within the HiMo-CLIP framework.

3.3 Monotonicity-Aware Contrastive Loss (MoLo)

Motivation. Semantic monotonicity captures the intuition that more complete textual descriptions should align more strongly with their images than partial ones. Standard contrastive models treat each input as an independent unit, lacking this ordering property. To address this, we propose a dual-level objective jointly optimizing global text embeddings and their semantic components. Extracted via PCA from the full text, these components inherently represent partial semantics. Aligning both levels encourages the model to learn that full-text embeddings, containing all semantic substructures, yield the highest alignment scores, thereby achieving monotonicity without extra supervision.

Architecture. The global objective preserves CLIP's contrastive alignment between full-image and full-text embeddings. Let v_i and u_i be the visual and textual embeddings for the i-th image-text pair. The global loss is defined as:

$$\mathcal{L}_{\text{global}} = \frac{1}{2N}\sum_{i=1}^{N}\left[\mathcal{L}_{\text{info}}(v_i, u_i) + \mathcal{L}_{\text{info}}(u_i, v_i)\right], \tag{2}$$

where $\mathcal{L}_{\text{info}}(a, b)$ denotes the cosine-similarity-based contrastive loss, instantiated as the widely adopted InfoNCE loss (Oord, Li, and Vinyals 2018; Radford et al. 2021). To incorporate semantic granularity, each image embedding v_i is aligned with its semantic component u_i' obtained via PCA. The component-level loss is formulated as:

$$\mathcal{L}_{\text{comp}} = \frac{1}{2N}\sum_{i=1}^{N}\left[\mathcal{L}_{\text{info}}(v_i, u_i') + \mathcal{L}_{\text{info}}(u_i', v_i)\right]. \tag{3}$$

The final MoLo loss $\mathcal{L}_{\text{MoLo}}$ combines both objectives:

$$\mathcal{L}_{\text{MoLo}} = \mathcal{L}_{\text{global}} + \lambda \cdot \mathcal{L}_{\text{comp}}, \tag{4}$$

where λ balances the influence of the component-level term. This formulation implicitly enforces semantic monotonicity by leveraging the natural inclusion property of PCA components. As the components are subsets of the full-text embedding, the model learns to associate greater semantic completeness with stronger alignment, thereby producing more cognitively consistent cross-modal representations.

4 Experiments

4.1 Training Details

Following LongCLIP (Zhang et al. 2024), we train HiMo-CLIP on ShareGPT4V (Chen et al. 2024), which has 1.2M image–caption pairs with multi-sentence annotations averaging 143.6 words. The model is initialized from CLIP and fine-tuned 10 epochs on 8 NVIDIA H100s (global batch 1024) using AdamW (1e$-$6 LR, 0.01 weight decay, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with 200-step warm-up. Images are resized to 224×224, and texts padded/truncated to 248 tokens with interpolated positional embeddings. For HiDe, the explained-variance threshold is set to $\tau = 0.9$ and the component loss weight is $\lambda = 1.0$.

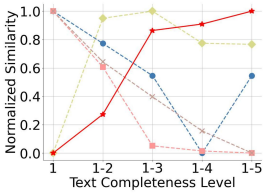
Method	BackBone	Train Data	Urban1k I2T/T2I	Docci I2T/T2I	Long-DCI I2T/T2I
CLIP	ViT-B/16	400M	68.1/53.6	58.5/58.2	35.0/33.1
Long-CLIP	ViT-B/16	1M	78.9/79.5	63.2/71.5	42.2/48.4
TULIP	ViT-B/16	1M	88.1/86.6	75.5/75.8	50.2/50.6
FineLIP*	ViT-B/16	1M	88.2/88.2	75.8/77.3	55.8/53.1
LoTLIP	ViT-B/16	100M	88.8/84.8	73.2/71.6	54.5/53.3
SigLIP	ViT-B/16	10B	63.0/62.3	70.2/70.6	45.4/43.0
BLIP	ViT-B/16	14M	45.5/48.5	50.5/53.5	-/-
EVA-02-CLIP	ViT-B/16	400M	67.0/60.8	67.7/68.0	-/-
jina-clip-v2	ViT-B/16	1.7B	80.4/78.0	77.6/78.2	-/-
MetaCLIP	ViT-B/16	400M	68.9/63.3	70.9/71.5	-/-
Ours	ViT-B/16	1M	**89.2/89.6**	**77.8/79.9**	**58.6/57.1**
CLIP	ViT-L/14	400M	68.7/52.8	57.5/60.7	33.4/31.3
Long-CLIP	ViT-L/14	1M	82.7/86.1	66.5/78.6	46.5/54.3
TULIP	ViT-L/14	1M	90.1/91.1	75.5/75.8	55.7/56.4
FineLIP*	ViT-L/14	1M	92.3/91.2	81.5/82.6	59.6/58.5
MetaCLIP	ViT-L/14	400M	73.4/70.0	76.5/76.7	-/-
SAIL-L-NV2	ViT-L/14	23M	81.5/80.2	76.5/78.9	-/-
EVA-02-CLIP	ViT-L/14	400M	73.3/68.5	73.5/75.0	47.3/47.9
Ours	ViT-L/14	1M	**93.0/93.1**	**82.4/84.4**	**62.2/61.9**

Table 1: Long-caption retrieval results. $*$ denotes reimplemented methods. **Bold** and underlined indicate best and second-best scores.

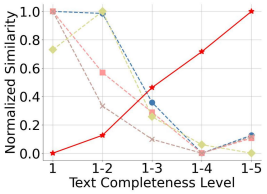
4.2 Evaluation Details

Text-Image Retrieval. We evaluate HiMo-CLIP on standard benchmarks covering image-text matching and compositional reasoning. For long-text retrieval, we use Urban1k (Zhang et al. 2024), Docci (Onoe et al. 2024), and Long-DCI (Najdenkoska et al. 2024), which contain captions averaging 128, 136, and 200 words, respectively. Evaluation follows each dataset's official protocol using standard metrics such as Recall@K. For compositional reasoning, we adopt COLA (Ray et al. 2023), which contains 210 human-validated multi-object queries with fine-grained attribute variations, and we evaluate using COLA-multi accuracy. For short-text retrieval, we report results on Flickr30k (Young et al. 2014) and COCO (Lin et al. 2014), using the same evaluation settings (please see supplementary materials).

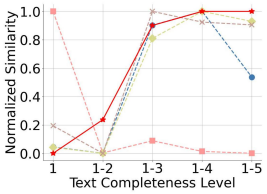
Hierarchical Monotonic Alignment. To evaluate how alignment strength scales with textual completeness, we introduce *HiMo@K*, a hierarchical monotonicity metric applicable to both standard benchmarks and long-form datasets. Given a caption with n sentences ($n \geq K$), we divide it into K contiguous segments of roughly equal length. Let t_k denote the concatenation of the first k segments, and s_{t_k} the matching score between the paired image and t_k. For each image-caption pair, we compute $\{s_{t_1}, s_{t_2}, \ldots, s_{t_K}\}$ and as-

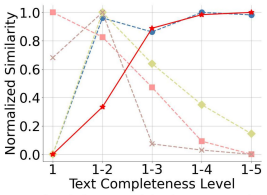
1) A green-eyed black cat lies on the grass. 2) Its body faces bottom-left in a three-quarters position. 3) Its front paws extend slightly forward. 4) Its hind legs stretch to the left. 5) Its head turns towards the bottom-right.

1) A red RAM truck with a black underside and grille is parked in a spot. 2) "POWER WAGON" is written sideways on the back. 3) Lights are on top and a net is in the back. 4) A word "RAM" is written in small letters on the side. 5) Another "RAM" is written in larger font on the grille.

1) A small brown smiling porcelain pig sits on a red Lego Ferrari. 2) Light reflects slightly on the pig's back. 3) Both cast shadows on the wooden floor. 4) A white wall is in the background. 5) A white shelf is to the right of the wall.

1) An outdoor medium view shows a wired fence with three wooden posts. 2) Each post has a dark brown cowboy boot on top. 3) The boots feature white patterns and yellow circles. 4) The left boot is upside down, toe forward. 5) The middle boot is upside down, facing left.

Figure 3: Semantic monotonicity on HiMo-Docci. Each image is paired with five increasingly complete subtexts (HiMo@5). HiMo-CLIP shows consistent score increases, unlike other methods. All scores are sample-normalized for fair comparison.

Method	**BackBone**	**Train Data**	**Flickr30k** I2T/T2I	**COCO** I2T/T2I
CLIP	ViT-B/16	400M	83.3/61.9	51.8/32.7
EVA-02-CLIP	ViT-B/16	-	85.7/71.2	58.7/42.2
Long-CLIP	ViT-B/16	1M	87.3/70.7	57.6/40.4
FineLIP*	ViT-B/16	1M	85.4/65.6	53.1/36.2
LoTLIP	ViT-B/16	100M	86.9/65.2	59.6/38.1
SigLIP	ViT-B/16	10B	**89.8/75.1**	**65.4/47.2**
Action-CLIP	ViT-B/16	0.7M	88.1/74.7	58.4/43.2
Ours	ViT-B/16	1M	88.4/72.1	60.6/40.8
CLIP	ViT-L/14	400M	86.1/66.0	56.1/35.4
Long-CLIP	ViT-L/14	1M	89.0/76.7	62.8/46.3
TULIP	ViT-L/14	1M	89.0/73.5	62.6/46.1
FineLIP*	ViT-L/14	1M	90.5/74.7	60.3/43.5
Action-CLIP	ViT-L/14-336	0.7M	91.5/74.0	62.5/44.1
Ours	ViT-L/14	1M	**92.5/78.2**	**65.1/47.2**

Table 2: Results on short-caption cross-modal retrieval. $*$ indicates reimplementation.

Method	H2↑				H3↑				HK↑	Ma.↑
	Ur.	Doc.	LD.	Avg	Ur.	Doc.	LD.	Avg		
CLIP	77.0	74.9	65.5	72.5	40.7	33.9	31.0	35.2	0.43	27.6
EVA-02-CLIP	78.0	70.0	70.8	72.9	24.4	26.7	31.3	27.5	0.28	25.7
Long-CLIP	34.8	16.7	53.5	35.0	39.4	28.5	41.9	36.6	-0.55	32.4
TULIP	97.2	89.2	83.9	90.1	62.1	46.6	45.1	51.3	0.67	34.8
FineLIP*	97.9	**98.5**	92.8	96.4	64.9	60.6	53.7	59.7	0.83	34.3
FG-CLIP†	96.6	94.1	91.3	94.0	62.1	58.5	53.8	58.1	0.75	30.0
Ours	**99.3**	98.0	**96.4**	**97.9**	**70.9**	**62.3**	**59.3**	**64.2**	**0.88**	**38.6**

Table 3: HiMo@2/3/K and COLA-multi accuracy on Ur., Doc. and LD. $*$ is reimplementation due to unavailable official models. FG-CLIP† uses ViT-L/14-336, others use ViT-L/14-224. H2/3: HiMo@2/3, HK: HiMo@K (Pearson on HiMo-Docci), Ma.: COLA-multi accuracy. Ur.: Urban1k, Doc.: Docci, LD.: Long-DCI.

sess whether similarity increases as more context is added. For example, with $n = 5$ and $K = 3$, the caption is split into segments as follows: the first contains sentence 1, the second sentences 2–3, and the third sentences 4–5. HiMo@3 is computed over t_1, t_2, and t_3 (the full caption). When $n = 6$ and $K = 2$, the two segments contain the first and second halves, respectively.

(1) *General Definition ($K > 3$).* For deeper hierarchies, HiMo@K is defined as the Pearson correlation between subtext index k and similarity score s_{t_k}:

$$\text{HiMo@K} = \rho(k, s_{t_k}) = \frac{\sum_{k=1}^{K}(k-\bar{k})(s_{t_k}-\bar{s})}{\sqrt{\sum_{k=1}^{K}(k-\bar{k})^2}\sqrt{\sum_{k=1}^{K}(s_{t_k}-\bar{s})^2}}, \quad (5)$$

where $\bar{k}$ and $\bar{s}$ denote the mean of the segment indices and similarity scores.

(2) *Shallow Cases ($K = 2, 3$).* For smaller K, correlation is unstable. We thus define HiMo@K as strict monotonic accuracy:

$$\text{HiMo@K} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[s_{t_1}^{(i)} < s_{t_2}^{(i)} (< s_{t_3}^{(i)})] \times 100\%, \quad (6)$$

where $\mathbb{I}[\cdot]$ is the indicator function, and N is the number of image-caption pairs.

Dataset for Deep Hierarchies. HiMo@2 and HiMo@3 can be computed on existing datasets like Flickr30k, COCO, and

τ	Urban1k	Docci	Long-DCI	HiMo@2↑	HiMo@K↑
	I2T/T2I	I2T/T2I	I2T/T2I		
0.60	85.2/84.3	74.3/75.1	51.4/49.4	<u>98.2</u>	<u>0.86</u>
0.70	91.7/92.1	81.3/82.8	60.1/59.3	**98.3**	<u>0.86</u>
0.80	<u>93.3</u>/**93.6**	<u>82.3</u>/<u>83.7</u>	**62.5/61.9**	94.8	0.79
0.85	**93.5**/<u>93.4</u>	81.6/83.3	61.7/61.3	96.4	0.85
0.90	93.0/93.1	**82.4/84.4**	<u>62.2</u>/**61.9**	97.9	**0.88**
0.95	91.7/92.8	80.9/83.0	60.8/60.4	94.6	0.81

Table 4: Ablation of threshold τ in HiDe. HiMo@2: average monotonic accuracy on Urban1k, Docci, and Long-DCI. HiMo@K: Pearson correlation on HiMo-Docci.

Docci. However, reliable evaluation for $K > 3$ requires high-quality clause-level structure. To this end, we construct HiMo-Docci, a curated subset of 1,000 Docci samples with human-authored captions reannotated into semantically valid subtexts. Each caption is manually segmented and verified to ensure hierarchical granularity. Prior datasets with hierarchical structure (Vulić et al. 2017; Santurkar, Tsipras, and Madry 2020; Alper and Averbuch-Elor 2024) focus on short text and are not well suited for long-form alignment. HiMo-Docci thus serves as a necessary resource for evaluating deep hierarchi.

4.3 Main Results and Analysis

Long-form Text Retrieval Performance. Table 1 shows that HiMo-CLIP consistently outperforms state-of-the-art methods on all long-text benchmarks. With a *ViT-L/14* backbone, HiMo-CLIP achieves **93.0%/93.1%** (I2T/T2I) on Urban1k, **82.4%/84.4%** on Docci, and **62.2%/61.9%** on Long-DCI, surpassing the strongest baseline, FineLIP, by clear margins. Notably, these results are obtained using only **1M** training samples, compared to LoTLIP's **100M** samples. Existing methods such as TULIP and LoTLIP mainly extend text length without modeling semantic structure. TULIP's RoPE-based extension (*ViT-L/14*, Docci T2I: 75.8%) treats long texts as flat sequences, while LoTLIP's corner tokens (*ViT-B/16*, Docci T2I: 71.6%) rely on static clause bindings. In contrast, HiDe's batch-adaptive PCA (**84.4%**) dynamically isolates context-critical semantics, enabling more effective long-text alignment.

Short-text Retrieval Compatibility. Contrary to methods that sacrifice short-text performance for long-text gains (e.g., Long-CLIP's regression on Flickr30k T2I), Table 2 shows HiMo-CLIP maintains competitive results on COCO and Flickr30k. With ViT-L/14, we achieve new SOTA on Flickr30k I2T R@1 (92.5%) and COCO T2I R@1 (47.2%), proving our hierarchical approach *generalizes* across textual granularities without overfitting to long descriptions.

Semantic Hierarchy and Monotonicity. Table 3 shows that HiMo-CLIP effectively models semantic hierarchy and monotonicity. It achieves near-perfect HiMo@2 (97.9%) and strong HiMo@3 (64.2%), outperforming FineLIP (96.4%, 59.7%) and TULIP (90.1%, 51.3%). On deeper hierarchies (HiMo@K on HiMo-Docci), HiMo-CLIP attains a Pearson correlation of 0.88, indicating consistent similarity growth with richer text, whereas LongCLIP lags behind and exhibits a negative correlation (-0.55). On the COLA multi-object retrieval task, HiMo-CLIP achieves the highest top-1 accuracy (38.6%), surpassing TULIP, FineLIP, and LongCLIP. These results highlight the effectiveness of HiDe and MoLo in capturing fine-grained, hierarchical semantics beyond flat sequence representations.

Qualitative Results. Figure 3 shows HiMo@5 results on HiMo-Docci, where HiMo-CLIP consistently exhibits monotonic similarity growth, unlike CLIP and Long-CLIP which show frequent drops, supporting the assumption that richer subtexts yield stronger alignment. Figures 4 and 5 further illustrate HiMo@2/3/4/7 with qualitative examples, demonstrating that HiMo-CLIP reliably preserves correct score ordering under deeper hierarchies. HiMo-CLIP achieves the highest qualitative HiMo@4 (0.93) and HiMo@7 (0.97), while FineLIP and TULIP exhibit score reversals, and Long-CLIP shows strong negative correlations (-0.94, -0.95). Even on shallow settings, HiMo-CLIP maintains correct ordering throughout, whereas FineLIP, TULIP, and FG-CLIP violate semantic monotonicity. These results highlight the robustness and scalability of HiMo-CLIP in modeling hierarchical semantic consistency.

4.4 Ablation Studies

HiDe Threshold Sensitivity. Table 4 shows that τ=0.9 achieves the best balance between retrieval and hierarchy, yielding peak Urban1k/Docci recall, HiMo@2 of 97.9%, and HiMo@K of 0.88. Smaller thresholds drop essential semantics, while larger ones introduce noise, degrading discrimination. These results highlight the importance of a compact, well-chosen decomposition threshold for preserving hierarchical semantics and semantic monotonicity.

Component Alignment Strategy. Table 5 shows that only the full objective ($\mathcal{L}_{\text{global}} + \mathcal{L}_{\text{comp}}$) achieves optimal performance. The 0.19 HiMo@K gap over $\mathcal{L}_{\text{global}}$ alone (0.69) demonstrates that joint global–component alignment is critical for preserving semantic monotonicity, validating MoLo's dual-branch design and the role of component-level supervision in maintaining hierarchical consistency.

Loss Weight Ablation. As shown in Table 6, setting λ=1 yields the best trade-off, achieving strong global alignment (Urban1k T2I 93.1%) and monotonicity (HiMo@2 97.9%). Larger λ (2) overweights components and degrades retrieval (Long-DCI I2T drops to 61.6%), while smaller λ (0.5) weakens monotonic structure (HiMo@2 97.1% vs. 97.9%). These findings emphasize the importance of balancing global and local alignment objectives to optimize both retrieval quality and hierarchical structure. Overall, the results demonstrate that λ=1 provides a balanced supervision signal, allowing the model to simultaneously retain strong retrieval performance and preserve semantic monotonicity across levels.

5 Conclusion

We introduced HiMo-CLIP, a representation-level framework enhancing CLIP-style models by explicitly modeling semantic hierarchy and enforcing monotonic alignment between text and image. Our method combines a hierarchical decomposition (HiDe) module with a monotonicity-aware

An outdoor side view of two weathered metal ore carts on a section of rail track. Noon, Daytime. The shadows of the ore carts are small and directly below them. Large gray rocks cover the ground all around the section of track.

Methods	1		2
CLIP	0.298	>	0.292
EVA-02-CLIP	0.304	>	0.302
FG-CLIP	0.266	<	0.272
Long-CLIP	0.288	>	0.273
FineLIP*	1.174	>	1.170
TULIP	0.310	<	0.339
Ours	0.269	<	0.273

(a)

A high angled close up view of a gymnetini beetle that's orientated to the left. It's horizontally visible across a surface of small pebbles and rocks. The overall view is very visible with light shadows on the left and right sides of the image.

Methods	1		2		3
CLIP	0.211	<	0.228	<	0.243
EVA-02-CLIP	0.262	<	0.288	>	0.286
FG-CLIP	0.188	<	0.233	>	0.222
Long-CLIP	0.271	<	0.285	>	0.265
FineLIP*	0.753	<	1.034	<	1.125
TULIP	0.227	<	0.322	>	0.307
Ours	0.245	<	0.262	<	0.267

(b)

Figure 4: Semantic monotonicity in HiMo@2 and HiMo@3. (a) **HiMo@2**: Each image is paired with two texts of increasing completeness, where alignment should strengthen with richer descriptions. (b) **HiMo@3**: Images are matched with three hierarchical texts (short, medium, long), following Fig. 1. Green/red denote correct/incorrect monotonicity. FineLIP scores may exceed 1 due to its score fusion strategy.

A low-angle view of a ceiling light inside a white ceiling that is centered in the frame. The light has a crevice all around that is not reached by light, making it appear black. In the center is the light bulb that isn't turned on, with a blurred reflection of the opposite view. On the top right of the frame is a piece of measuring tape on the centimeter side, which the numbers go from 1-7 and a large 3 can be seen next to a number that is mostly off the frame.

Methods	1		2		3		4	HiMo@4
FG-CLIP	0.178	<	0.187	<	0.188	>	0.254	0.84
Long-CLIP	0.277	>	0.268	>	0.254	<	0.255	-0.94
FineLIP*	0.923	<	0.928	<	0.951	<	1.212	0.82
TULIP	0.207	<	0.211	>	0.207	<	0.258	0.77
Ours	0.234	<	0.243	<	0.247	<	0.275	**0.93**

(a)

This is an indoor scene. It's a front view of a taxidermied deer in a museum display. The deer is a light brown color with its body facing forward, and its head turned slightly to the left. It has a dark nose and medium-sized pointy antlers on the top of its head. The deer's back hooves are on top of a log that stretches from the bottom right to the center left of the frame. A tree is to the left of the deer and placed vertically, and to the right of the deer is a concrete pillar that has a rectangular shape and goes to the top edge of the frame. The deer has light casting onto it, with a warm yellow glow on its left, and a white artificial light to its left.

Methods	1		2		3		4		5		6		7	HiMo@7
FG-CLIP	0.220	<	0.226	>	0.214	<	0.244	<	0.280	>	0.275	>	0.271	0.87
Long-CLIP	0.371	>	0.351	>	0.331	>	0.290	>	0.266	>	0.265	<	0.267	-0.95
FineLIP*	1.302	<	1.488	>	1.478	<	1.527	<	1.665	>	1.661	<	1.666	0.93
TULIP	0.306	<	0.330	<	0.336	<	0.357	<	0.371	>	0.365	>	0.361	0.89
Ours	0.266	<	0.277	<	0.280	<	0.286	<	0.295	<	0.296	<	0.297	**0.97**

(b)

Figure 5: Semantic monotonicity results on extended HiMo@K tasks. (a) **HiMo@4**: Each image is paired with 4 subtexts of increasing semantic detail for finer-grained monotonicity testing. (b) **HiMo@7**: Each image is paired with 7 increasingly rich captions, enabling stricter evaluation of hierarchical alignment than Fig. 4. Green/red markers indicate correct/incorrect semantic orderings. FineLIP scores may exceed 1 due to its fusion strategy.

Method	Urban1k	Docci	Long-DCI	HiMo@2↑	HiMo@K↑
	I2T/T2I	I2T/T2I	I2T/T2I		
$\mathcal{L}_{\text{global}}$	91.7/92.8	80.9/83.0	60.8/60.4	91.0	0.69
$\mathcal{L}_{\text{comp}}$	92.6/92.9	80.9/82.5	60.1/60.7	96.2	0.84
$\mathcal{L}_{\text{global}} + \mathcal{L}^{u,v}_{\text{comp}}$	87.5/91.4	78.8/82.1	57.7/57.8	91.7	0.69
$\mathcal{L}_{\text{global}} + \mathcal{L}_{\text{comp}}$	**93.0/93.1**	**82.4/84.4**	**62.2/61.9**	**97.9**	**0.88**

Table 5: Ablation of loss and HiDe variants. $\mathcal{L}_{\text{global}}$: no HiDe. $\mathcal{L}_{\text{comp}}$: HiDe on text only. $\mathcal{L}^{u,v}_{\text{comp}}$: HiDe on both modalities. HiMo@2: average on three long-form text datasets; HiMo@K: Pearson correlation on HiMo-Docci.

λ	Urban1k	Docci	Long-DCI	HiMo@2↑	HiMo@K↑
	I2T/T2I	I2T/T2I	I2T/T2I		
2.0	**93.1**/92.3	81.9/83.9	61.6/61.4	97.8	**0.88**
1.0	93.0/**93.1**	**82.4/84.4**	62.2/**61.9**	**97.9**	**0.88**
0.5	92.8/**93.1**	82.2/84.3	**62.3**/61.8	97.1	0.87

Table 6: Loss weight λ ablation. λ balances global and compositional losses. HiMo@2: average on three datasets; HiMo@K: Pearson correlation on HiMo-Docci.

contrastive loss (MoLo), enabling fine-grained alignment between image features and structured subtexts. Without altering the encoder architecture, HiMo-CLIP outperforms prior methods on long-form text retrieval, compositional reasoning and newly proposed HiMo@K metrics, achieving consistent gains across qualitative and quantitative evaluations. These results show that aligning cross-modal representations at multiple semantic levels is essential for robust, cognitively aligned vision-language understanding. We hope this work inspires more structured, interpretable, hierarchy-aware multimodal learning.

References

Alper, M.; and Averbuch-Elor, H. 2024. Emergent visual-semantic hierarchies in image-text representations. In *European Conference on Computer Vision*, 220–238. Springer.

Asokan, M.; Wu, K.; and Albreiki, F. 2025. FineLIP: Extending CLIP's Reach via Fine-Grained Alignment with Longer Text Inputs. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14495–14504.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2024. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, 370–387. Springer.

Fang, Y.; Sun, Q.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2024. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 105171.

Guo, J.; Wang, M.; Zhou, Y.; Song, B.; Chi, Y.; Fan, W.; and Chang, J. 2023. HGAN: Hierarchical graph alignment network for image-text retrieval. *IEEE Transactions on Multimedia*, 25: 9189–9202.

Ji, Z.; Chen, K.; and Wang, H. 2021. Step-wise hierarchical alignment network for image-text matching. *arXiv preprint arXiv:2106.06509*.

Jolliffe, I. 2011. Principal component analysis. In *International encyclopedia of statistical science*, 1094–1096. Springer.

Koukounas, A.; Mastrapas, G.; Eslami, S.; Wang, B.; Akram, M. K.; Günther, M.; Mohr, I.; Sturua, S.; Wang, N.; and Xiao, H. 2024. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*.

Lewis, M.; Nayak, N. V.; Yu, P.; Yu, Q.; Merullo, J.; Bach, S. H.; and Pavlick, E. 2022. Does clip bind concepts? probing compositionality in large image models. *arXiv preprint arXiv:2212.10537*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Najdenkoska, I.; Derakhshani, M. M.; Asano, Y. M.; Van Noord, N.; Worring, M.; and Snoek, C. G. 2024. Tulip: Token-length upgraded clip. *arXiv preprint arXiv:2410.10034*.

Onoe, Y.; Rane, S.; Berger, Z.; Bitton, Y.; Cho, J.; Garg, R.; Ku, A.; Parekh, Z.; Pont-Tuset, J.; Tanzer, G.; et al. 2024. Docci: Descriptions of connected and contrasting images. In *European Conference on Computer Vision*, 291–309. Springer.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ray, A.; Radenovic, F.; Dubey, A.; Plummer, B.; Krishna, R.; and Saenko, K. 2023. Cola: A benchmark for compositional text-to-image retrieval. *Advances in Neural Information Processing Systems*, 36: 46433–46445.

Santurkar, S.; Tsipras, D.; and Madry, A. 2020. Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*.

Sun, Z.; Fang, Y.; Wu, T.; Zhang, P.; Zang, Y.; Kong, S.; Xiong, Y.; Lin, D.; and Wang, J. 2024. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13019–13029.

Tian, M.; Wu, X.; and Yang, S. 2025. LLM-enhanced Action-aware Multi-modal Prompt Tuning for Image-Text Matching. *arXiv preprint arXiv:2506.23502*.

Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M. F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Vulić, I.; Gerz, D.; Kiela, D.; Hill, F.; and Korhonen, A. 2017. Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4): 781–835.

Wu, P.; He, X.; Tang, M.; Lv, Y.; and Liu, J. 2021. Hanet: Hierarchical alignment networks for video-text retrieval. In *Proceedings of the 29th ACM international conference on Multimedia*, 3518–3527.

Wu, W.; Zheng, K.; Ma, S.; Lu, F.; Guo, Y.; Zhang, Y.; Chen, W.; Guo, Q.; Shen, Y.; and Zha, Z.-J. 2024. Lotlip: Improving language-image pre-training for long text understanding. *Advances in Neural Information Processing Systems*, 37: 64996–65019.

Xie, C.; Wang, B.; Kong, F.; Li, J.; Liang, D.; Zhang, G.; Leng, D.; and Yin, Y. 2025. FG-CLIP: Fine-Grained Visual and Textual Alignment. *arXiv preprint arXiv:2505.05071*.

Xu, H.; Xie, S.; Tan, X. E.; Huang, P.-Y.; Howes, R.; Sharma, V.; Li, S.-W.; Ghosh, G.; Zettlemoyer, L.; and Feichtenhofer, C. 2023. Demystifying clip data. *arXiv preprint arXiv:2309.16671*.

Yamada, Y.; Tang, Y.; Zhang, Y.; and Yildirim, I. 2022. When are Lemons Purple? The Concept Association Bias of Vision-Language Models. *arXiv preprint arXiv:2212.12043*.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2: 67–78.

Yuksekgonul, M.; Bianchi, F.; Kalluri, P.; Jurafsky, D.; and Zou, J. 2022. When and why vision-language models behave

like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

Zeng, Y.; Zhang, X.; and Li, H. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, B.; Zhang, P.; Dong, X.; Zang, Y.; and Wang, J. 2024. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, 310–325. Springer.

Zheng, K.; Zhang, Y.; Wu, W.; Lu, F.; Ma, S.; Jin, X.; Chen, W.; and Shen, Y. 2024. Dreamlip: Language-image pre-training with long captions. In *European Conference on Computer Vision*, 73–90. Springer.