OneLIP: Unlocking and Improving Long-Text Representations of CLIP via One-Stage Adaptation

Renjie Pan[1], **Jiayan Song**[1], **Hua Yang**[1]*

[1]School of Information Science and Electronical Engineering & School of Integrated Circuits, Shanghai Jiao Tong University, Shanghai, China
{rjpan21, sjy1231, hyang}@sjtu.edu.cn

Abstract

Contrastive Language-Image Pretraining (CLIP) has demonstrated impressive generalization on vision-language tasks by aligning images and short texts. However, its inherent 77-token length limits the capacity of capturing complex semantics in long captions. Existing long-text adaptations for CLIP typically rely on either multi-stage training or truncation-based alignment, both inevitably resulting in semantic degradation and cumbersome tuning. Therefore, we propose OneLIP, a unified framework that extends CLIP to understand long captions within a single training stage, eliminating the need for brittle truncation or multi-stage pipelines. OneLIP addresses semantic degradation by introducing two key innovations: (1) Token Refinement and Importance-guided Modeling (TRIM) module, which selects and refines informative tokens via SVD-based contribution scoring and cross-modal relevance modeling; (2) Per-sample Online Hard Negative Mining (PO-HNM) strategy dynamically maintains sample-specific negatives based on dual-consistency difficulty tracking, which is superior in long-text scenarios where key semantics are distributed in scattered positions. Extensive experiments on long-text image retrieval, short-text image retrieval, zero-shot classification, and text-to-image generation demonstrate OneLIP's robustness and versatility across diverse input lengths, offering a faithful solution for long-text representation learning of CLIP.

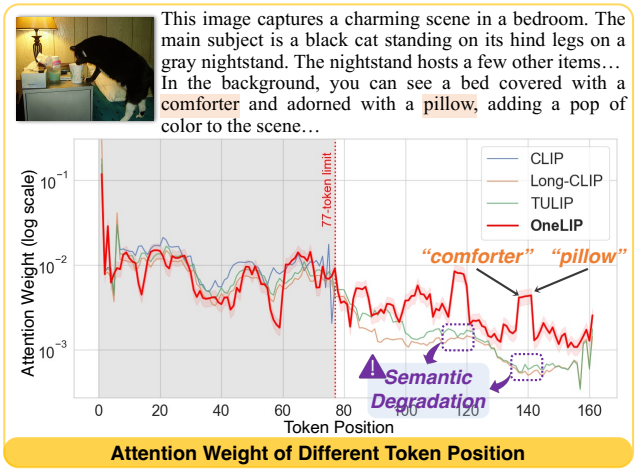

Figure 1: **Attention distribution collapse in long-text CLIP:** Semantic degradation appears due to deficient representation learning of back tokens, revealing poor long-range dependency modeling of several baselines. OneLIP addresses this challenge via a unified one-stage adaptation, focusing on both front and back tokens.

1 Introduction

Contrastive Language-Image Pretraining (CLIP) (Radford et al. 2021) has emerged as a cornerstone of vision-language models, powering both visual understanding (Pan, Dong, and Yang 2025; Lyu et al. 2025; Shao et al. 2024; Chen et al. 2023; Cai, Pan, and Yang 2025) and multimodal generation (Bai et al. 2025; Zhou et al. 2024; Liu et al. 2023; Li et al. 2022) via its contrastive learning paradigm. By aligning images with concise textual descriptions (captions or phrases), CLIP learns rich and transferable cross-modal representations, enabling strong zero-shot generalization across diverse downstream tasks(Tschannen et al. 2025; Sun et al. 2024; Song et al. 2024; Pan et al. 2023; Mu et al. 2022).

Despite its success, CLIP's text encoder is inherently limited to short input sequences (e.g., 77 tokens), which restricts its capacity to capture complex and fine-grained semantics in long-text scenarios. To overcome this, prior efforts have explored long-text adaptations, primarily by modifying positional encoding schemes and distilling knowledge from the base encoder into a long-text encoder (Chun and Yun 2025; Zhang et al. 2024; Jing et al. 2024; Fang et al. 2023), followed by contrastive alignment. While partially effective, these methods suffer from several limitations: **(1)** During distillation, existing approaches truncate long texts to a fixed length for an arbitrary alignment with pre-trained base encoder (Najdenkoska et al. 2025; Zheng et al. 2024), leading to inevitable semantic degradation and loss of critical information. As observed in Figure 1, the attention distribution of current long-text CLIP collapses towards front tokens as sequence length increases, indicating a deficiency to capture long-range dependencies. This collapse correlates with a stagnating performance at longer sequence lengths (see Figure 3), underscoring the bottleneck in current textual representation learning. **(2)** Long-text adaptation typi-

*Corresponding author.

cally requires multi-stage training pipelines involving careful initialization, distillation, and contrastive alignment (Najdenkoska et al. 2025; Gao et al. 2024), resulting in complex training procedures and heightened computation overhead. **(3)** Since contrastive learning is highly sensitive to negative sample quality, a dependence on either offline negative mining or static batch-shared negatives (Huynh et al. 2022; He et al. 2020) is indispensable, while both strategies lacking in the flexibility to track how sample difficulty evolves during training. This limitation is especially pronounced in long-text scenarios, where key semantics are distributed in more scattered positions and semantic drift is more common, rendering static negatives rapidly outdated. Together, these challenges not only hinder long-text representation learning, but also result in unstable optimization and non-trivial training overhead, ultimately limiting the effectiveness of long-text CLIP adaptation.

To address the above limitations, we propose **One-Stage Long-Text Adaptation for CLIP (OneLIP)**, a unified framework that enables CLIP to process arbitrarily long texts ***within a single training stage***, avoiding multi-stage tuning pipelines or brittle truncation-based aligning strategies that may cause semantic degradation. At the core of OneLIP are three synergistic components: **(1)** We replace the absolute positional encoding with relative positional encoding to unlock long-text position awareness and scalability. **(2)** To mitigate semantic degradation caused by truncation-based distillation, we introduce **Token Refinement and Importance-guided Modeling (TRIM)**, a lightweight select-then-refine module that first selects the informative and modality-relevant tokens via TRIM-Select, which consists of SVD-based structural importance and cross-modal relevance modeling. Then, TRIM-Refine is introduced to further refine the selected tokens through learnable embeddings and dynamic weighting, aligning with the representation space of the base encoder. In this way, TRIM enables a fine-grained token-wise distillation that preserves semantic integrity and captures sufficient long-range dependencies, in contrast to prior truncation-based or pooling-based approaches. Notably, TRIM uniquely combines both training-free and training-based mechanisms, striking a balance between effectiveness and efficiency without introducing significant overhead. **(3)** Motivated by the scarcity and volatility of high-quality negatives in long-text scenarios, we further introduce **Per-sample Online Hard Negative Mining (PO-HNM)**, which maintains sample-specific negative queues. Negatives are dynamically updated based on a dual-consistency difficulty score that jointly considers cross-modal and intra-modal consistency, ensuring sufficient exposure to challenging and informative negatives throughout the training process.

Altogether, OneLIP unifies long-text encoding, token selection, and contrastive alignment in a fully end-to-end manner, eliminating redundant tuning or counterintuitive truncation, as well as preserving semantically faithful adaptation to long texts. Extensive experiments on long-text image retrieval, short-text image retrieval, zero-shot image classification, and text-to-image generation demonstrate the robustness and versatility of OneLIP. Moreover, OneLIP remains competitive under a limited batch sizes, demonstrating practical advantages for real-world deployment.

Our key contributions are summarized as follows: **(1)** We propose OneLIP, a one-stage adaptation framework for long-text CLIP. Unlike prior studies that rely on truncation or multi-stage pipelines, OneLIP enables end-to-end adaptation via a lightweight token refinement module TRIM to support token-wise distillation. This effectively mitigates semantic degradation and strengthens contrastive alignment over long texts. **(2)** We introduce PO-HNM, a novel hard negative mining strategy that dynamically maintains sample-specific negatives using a dual-modality difficulty score. This design effectively addresses the challenges in long-text scenarios, i.e., scattered semantics and evolving training dynamics. **(3)** Extensive experiments demonstrate that OneLIP not only enhances long-text understanding, but also consistently improves short-text and general vision-language tasks, highlighting its scalability and semantic fidelity across diverse textual inputs.

2 Related Work

Long-Text Adaptation for CLIP Contrastive language-Image Pretraining (Radford et al. 2021) demonstrate strong generalization but are inherently limited by their short-text input length. To overcome this, some works (Yu et al. 2022; Alayrac et al. 2022) support longer texts via generative decoding or external cross-attention, though they diverge from CLIP's contrastive architecture. More closely related approaches (Chun and Yun 2025; Jing et al. 2024; Wu et al. 2024; Zheng et al. 2024; Zhang et al. 2024) extend CLIP for long-text scenarios using truncation-based supervision and multi-stage training. For example, (Zhang et al. 2024) interpolates absolute position embeddings for longer inputs, which enables extended input lengths but fails to explicitly model long-range dependencies. (Najdenkoska et al. 2025) follows a two-stage pipeline that first initializes a long-text encoder, then aligns it with base encoder using truncated inputs. However, this leads to semantic degradation for tokens beyond the truncation window, and the multi-stage design brings optimization complexity and sensitivity. Other methods (Wu et al. 2024; Zheng et al. 2024) attempt to enrich supervision by synthesizing multi-sentence captions and sample sub-captions for alignment, yet still avoiding full long-text modeling. In summary, balancing efficient and faithful long-range semantic understanding still remains an open challenge in long-text CLIP.

Hard Negative Mining in Contrastive Learning Hard negative mining (HNM) plays an important role in contrastive learning. CLIP and its variants (Li et al. 2023; Mu et al. 2022; Yang et al. 2022; Radford et al. 2021) primarily rely on large batch sizes to ensure diverse negatives per anchor. A few works (Yang et al. 2022; Zhou et al. 2022) further incorporates label-derived samples and near-synonym captions to enhance semantic discrimination. In broader contrastive learning literature beyond CLIP, SimCLR (Chen et al. 2020) and BYOL (Grill et al. 2020) demonstrate that increasing the pool of negatives can boost representation learning, but this effect saturates when the samples are se-

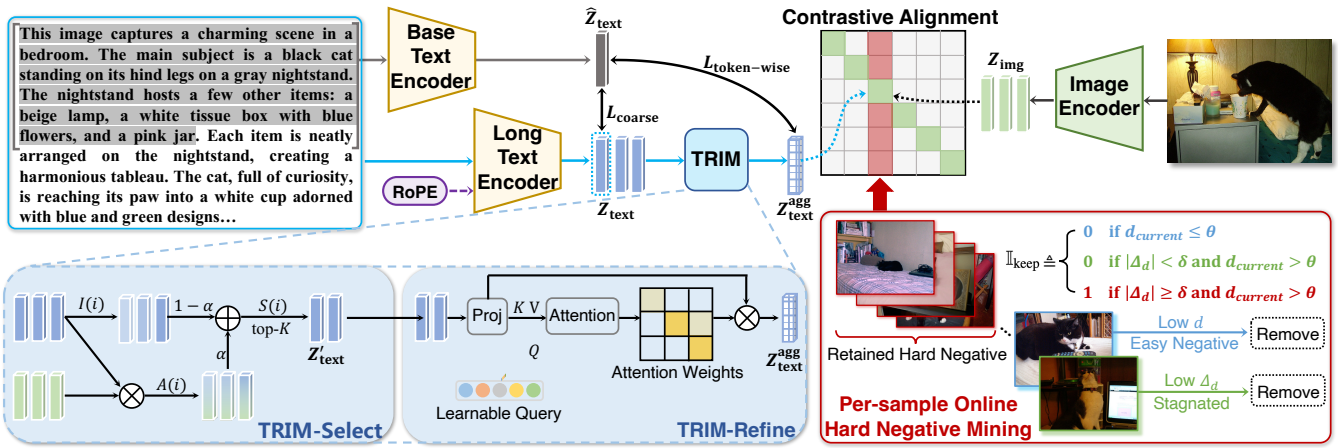

Figure 2: Overview of **OneLIP**. First, TRIM is proposed to facilitate token-wise distillation, which largely mitigates potential semantic degradation. Then, Per-sample Online Hard Negative Mining (PO-HNM) continually ensures the quality of hard negatives during training. The auxiliary warm-up ($\mathcal{L}_{\text{coarse}}$), token-wise distillation ($\mathcal{L}_{\text{token-wise}}$), and contrastive alignment ($\mathcal{L}_{\text{i-t}}$ and $\mathcal{L}_{\text{t-i}}$) are trained in an end-to-end manner.

mantically trivial (He et al. 2020). To address this, other approaches introduce memory banks (He et al. 2020), inter-class nearest sampling (Khosla et al. 2020), or pseudo-labeling (Chen and He 2021; Gao, Yao, and Chen 2021; Pan, Yang, and Zhao 2025) to improve the diversity and difficulty of negatives. Nonetheless, these methods often focus on a single modality, and rely on static, batch-shared negatives without modeling sample-specific difficulty or adapting to the evolving training dynamics (Xu et al. 2023). This limitation is particularly pronounced in long-text scenarios, where key semantics are more dispersed and the effectiveness of static negatives quickly deteriorates.

3 Methods

3.1 Preliminary

Given N image-text pairs $\{(I_i, T_i)\}_{i=1}^{N}$, CLIP encodes each image I_i into a visual embedding $\mathbf{v}_i$, and each text T_j into a textual embedding $\mathbf{t}_j$ using separate vision and text encoders. A symmetric contrastive objective is employed to construct a shared embedding space:

$$\mathcal{L}_{\text{i-t}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\cos\left(\mathbf{v}_i, \mathbf{t}_i\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\cos\left(\mathbf{v}_i, \mathbf{t}_j\right)/\tau\right)}, \quad (1)$$

where $\cos(\cdot)$ denotes cosine similarity and τ is a learnable temperature parameter. Text-to-image loss $\mathcal{L}_{\text{t-i}}$ is formulated analogously by swapping images and texts. Equation 1 captures semantic correspondences between two modalities, form the main objective of contrastive alignment.

3.2 One-Stage Long-Text Adaptation for CLIP

Task Formulation OneLIP aims at extending CLIP to long texts and mitigating semantic degradation through one training stage, which is composed of three synergistic components. First, we initialize a long-text encoder $f_{\text{text}}^{\text{long}}$ and unlock its capability of processing long texts; Next, we introduce TRIM, a select-then-refine module that first identifies informative and modality-relevant tokens, then refine the selected tokens via dynamic weighting, facilitating subsequent token-wise distillation to preserve semantic integrity; Finally, PO-HNM maintains a sample-specific queue of challenging negatives, which is updated via a dual-consistency difficulty score. Each component is detailed below.

From Absolute to Relative: Unlock Long-text Position Modeling To process sequences beyond 77-token limit, we replace the base encoder's absolute positional encoding with Rotary Positional Embeddings (RoPE) (Su et al. 2024), which inject relative positional information by rotating the query and key vectors in self-attention. Formally, for a query-key pair $(\mathbf{q}, \mathbf{k})$ at position i and j, RoPE defines:

$$\text{Att}\left(\mathbf{q}_i, \mathbf{k}_j\right) = \cos\left(\theta_{i-j}\right)\cdot\left\langle\mathbf{q}_i, \mathbf{k}_j\right\rangle + \sin\left(\theta_{i-j}\right)\cdot\left\langle\mathbf{q}_i^{\perp}, \mathbf{k}_j\right\rangle, \quad (2)$$

where θ_{i-j} encodes the relative distance, and $\mathbf{q}_i^{\perp}$ denotes the orthogonal rotation of $\mathbf{q}_i$. Equation 2 encodes relative distance and directional information among tokens. Given a long-text caption $x = [x_1, x_2, x_3, \cdots, x_L]$, the output of long-text encoder is $Z_{\text{text}} = f_{\text{text}}^{\text{long}}(x) \in \mathbb{R}^{L_1 \times d}$, where d is the embedding dimension. However, extended sequences lead to instability due to rapid phase rotations. To address this, we apply Neural Tangent Kernel (NTK) scaling (bloc97 2023) to proportionally stretch positional frequencies with respect to sequence length, resulting in smoother rotation trajectory and improved long-range token interaction. In addition, we also set an auxiliary synchronous warm-up by aligning the first 77 tokens from long-text encoder with the base text encoder:

$$\mathcal{L}_{\text{coarse}} = \mathbb{E}_{x\sim\mathcal{D}}\left[1 - \frac{Z_{\text{text}}\left(x_{\leq 77}\right)\cdot\hat{Z}_{\text{text}}\left(x_{\leq 77}\right)}{\left\|Z_{\text{text}}\left(x_{\leq 77}\right)\right\|\cdot\left\|\hat{Z}_{\text{text}}\left(x_{\leq 77}\right)\right\|}\right], \quad (3)$$

where $\hat{Z}_{\text{text}} = f_{\text{text}}^{\text{base}}(x) \in \mathbb{R}^{L_2 \times d}$ represents the base encoder output. Unlike prior methods that rely on multi-stage tuning, $\mathcal{L}_{\text{coarse}}$ is optimized jointly with other learning objectives.

Token Refinement and Importance-guided Modeling (TRIM) TRIM is designed to address semantic degradation and enable token-wise distillation. It selects and refines a subset of informative and modality-relevant tokens from long texts, supporting efficient and faithful alignment in the one-stage training of OneLIP. TRIM consists of two sub-modules: TRIM-Select, which identifies the informative tokens through a combination of SVD-based token contribution scoring and cross-modal relevance modeling; and TRIM-Refine, which refines the selected tokens via a set of learnable queries and dynamic weighting.

TRIM-Select. Given the embedding of a long text $Z_{\text{text}} \in \mathbb{R}^{L_1 \times d}$, TRIM-Select evaluates an importance score for each token by considering structural importance and cross-modal relevance. First, we apply Singular Value Decomposition (SVD): $Z_{\text{text}} = U\Sigma V^\top$, where $U \in \mathbb{R}^{L_1 \times r}$ contains the left singular vectors, $\Sigma \in \mathbb{R}^{r \times r}$ the top-r singular values, and $V^\top$ the semantic basis. The structural importance of token x_i is defined as:

$$I(i) = \sum_{j=1}^{r} |U_{ij}| \cdot \sigma_j, \tag{4}$$

which reflects its contribution to dominant semantic components. Next, to measure cross-modal relevance of token x_i, we compute $A(i) = \langle Z_{\text{text}}[i], Z_{\text{img}} \rangle$, where $Z_{\text{img}} \in \mathbb{R}^d$ is the pooled visual embedding, and $\langle \cdot \rangle$ denotes inner product. $A(i)$ quantifies token-image alignment and provides a cross-modal grounding signal. The final importance-guided score for token x_i contains a weighted combination:

$$S(i) = (1 - \alpha) \cdot I(i) + \alpha \cdot A(i). \tag{5}$$

Top-K tokens with the highest $S(i)$ are selected to constitute the compact sequence $Z'_{\text{text}} \in \mathbb{R}^{K \times d}$, filtering out redundant or low-contribution tokens.

TRIM-Refine. Intuitively, the choice of K in TRIM-Select should be kept within a moderate range to ensure performance, especially due to the substantial variation in token lengths across long texts. To improve the robustness and decouple sensitivity from K, we further introduce TRIM-Refine to further condenses Z'_{text} while preserving key semantics. Specifically, TRIM-Refine aggregates the selected tokens using a set of lightweight learnable queries $Q \in \mathbb{R}^{L_2 \times d}$, which extract contextually aligned representations via attention-weighted aggregation over Z'_{text}. The refined representation can be expressed as:

$$Z_{\text{text}}^{\text{agg}} = \text{Attn}(Q, Z'_{\text{text}}, Z'_{\text{text}}), \tag{6}$$

where $\text{Attn}(\cdot)$ represents scaled dot-product attention, and $Z_{\text{text}}^{\text{agg}} \in \mathbb{R}^{L_2 \times d}$ serves as the final textual representation, which is aligned in shape with the base encoder output $\hat{Z}_{\text{text}}$. TRIM-Refine avoids a full reliance on parameter-free SVD and decouples K from token dimensions, greatly enhancing the model's robustness.

Token-wise Distillation. As mentioned before, a fine-grained supervision can mitigate semantic degradation and boost one-stage adaptation. Thus, we introduce a token-wise distillation in long-text CLIP adaptation, which can be expressed as:

$$\mathcal{L}_{\text{token-wise}} = \frac{1}{L_2} \sum_{i=1}^{L_2} \left[1 - \cos\left(Z_{\text{text}}^{\text{agg}}[i],\ \hat{Z}_{\text{text}}[i] \right) \right], \tag{7}$$

where $\hat{Z}_{\text{text}}$ denotes the base encoder output, and $Z_{\text{text}}^{\text{agg}}$ represents the most informative set of tokens learned through both TRIM-Select and TRIM-Refine. This enables $\mathcal{L}_{\text{token-wise}}$ to operate on compact yet informative latent tokens, yielding a more robust and semantically faithful distillation process.

Per-sample Online Hard Negative Mining (PO-HNM) Most CLIP variants typically adopt hard negative mining (HNM) via either offline HNM (e.g., via pre-computed similarities) during pre-processing, or batch-shared negatives. However, such static strategies become suboptimal in long-text scenarios, which is primarily due to the inherently diverse and evolving semantic structures of long texts. In these cases, negatives that are effective in earlier stages may become trivial as the training evolves, leading to less informative contrastive updates. To address this issue, we propose PO-HNM, a dynamic and fine-grained strategy that maintains a dedicated negative queue for each training sample and updates it online throughout training. This design allows the model to continuously track the difficulty of negatives based on evolving representations across both modalities. Concretely, for each training sample, we select M candidate negatives and compute their difficulty score using a dual-consistency similarity function:

$$d = \beta \cos(Z_{\text{img}}, \hat{Z}_{\text{text}}^{\text{neg}}) + (1 - \beta) \cos(Z_{\text{text}}^{\text{agg}}, \hat{Z}_{\text{text}}^{\text{neg}}), \tag{8}$$

Z_{img} and $Z_{\text{text}}^{\text{agg}}$ are the anchor image/long-text embedding, $\hat{Z}_{\text{text}}^{\text{neg}}$ is the negative token embedding from the base encoder, and $\beta \in [0, 1]$ balances inter-modal and intra-modal difficulty. These candidates are inserted into a per-sample queue, which is periodically updated every T_{update} steps. For each stored negative, we re-compute its difficulty score d_{current} and compare it with the previously stored d_{prev} to obtain a difficulty change metric as: $\Delta d = |d_{\text{prev}} - d_{\text{current}}|$. To determine whether a negative should remain in the queue, we introduce a binary indicator $\mathbb{I}_{\text{keep}} \in \{0, 1\}$ based on both difficulty magnitude and evolution dynamics:

$$\mathbb{I}_{\text{keep}} = \begin{cases} 0 & \text{if } d_{\text{current}} \leq \theta, \\ 0 & \text{if } |\Delta d| < \delta \text{ and } d_{\text{current}} > \theta, \\ 1 & \text{if } |\Delta d| \geq \delta \text{ and } d_{\text{current}} > \theta, \end{cases} \tag{9}$$

where θ is a difficulty threshold distinguishing "easy" from "hard" negatives, and δ controls the minimum required dynamics for a hard negative to be considered still informative. A negative is retained ($\mathbb{I}_{\text{keep}} = 1$) if and only if it is both difficult ($d_{\text{current}} > \theta$) and still evolving ($|\Delta d| \geq \delta$). In contrast, negatives that are either too easy or have stagnated are removed and replaced with new candidates, ensuring a continually challenging and informative negative pool throughout training. We provide additional pseudocode and efficiency analysis with other HNM strategies in the Appendix.

3.3 Learning Objective

As illustrated in Figure 2, the learning objective of OneLIP consists of three parts: (1) $\mathcal{L}_{\text{token-wise}}$ in Equation 7 enforces fine-grained semantic alignment between the long-text encoder and the base encoder at the token level, serving as the foundation for faithful long-text representation; (2) Bi-directional contrastive alignment loss ($\mathcal{L}_{\text{i-t}}$ in Equation 1) aligns $Z_{\text{text}}^{\text{agg}}$ and Z_{img} through a symmetric Info-NCE based objective; (3) auxiliary warm-up $\mathcal{L}_{\text{coarse}}$ in Equation 3. Notably, the contrastive alignment is enhanced by PO-HNM, thus negatives will be updated throughout training. The overall learning objective is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{coarse}} + \lambda_2 \mathcal{L}_{\text{token-wise}} + \lambda_3 (\mathcal{L}_{\text{i-t}} + \mathcal{L}_{\text{t-i}}), \quad (10)$$

where a specific weight is added to each objective balance the contribution of each component. Importantly, Equation 10 is optimized end-to-end, enabling OneLIP to seamlessly support long-text adaptation in a single training stage. This unified formulation not only mitigates semantic degradation but also avoids the inefficiency of multi-stage tuning.

4 Experiments

4.1 Experimental Settings

Datasets OneLIP is trained on ShareGPT4V (Chen et al. 2024) dataset, including 1M image-text pairs with instruction-style long captions. Evaluation is conducted across four tasks: long-text image retrieval, short-text image retrieval, zero-shot image classification, and text-to-image generation. For long-text retrieval, we test on ShareGPT4V test set (1,000 samples) and Urban-1K (Zhang et al. 2024), including 1000 fine-grained urban scene descriptions. For short-text image retrieval, we follow Long-CLIP to validate on full Flickr30K (Young et al. 2014) (31,783 images) and MSCOCO (Lin et al. 2014) (5,000 images). Each image within both datasets is described by 5 different captions. For zero-shot image classification, we evaluate on ImageNet (Deng et al. 2009), ImageNet-O (Hendrycks et al. 2021), ImageNet-V2 (Ramesh et al. 2022), CIFAR-10 (Krizhevsky, Hinton et al. 2009), and CIFAR-100 (Krizhevsky, Hinton et al. 2009).

Implementation Details OneLIP is implemented on CLIP-ViT-B/16 and CLIP-ViT-L/14. We use a batch-size of 256 and train for 20 epochs. For long-text encoding, RoPE is extended using NTK scaling with factor $\left(\gamma \cdot \frac{T_g}{T_f}\right) - (\gamma - 1)$ to accommodate new higher token positions, where γ is set to 8.0. In TRIM-Select, we set $\alpha = 0.5$ and select top-K=100 tokens. TRIM-Refine reduces them to $L_2 = 77$ via learnable attention. In PO-HNM, the per-sample negative query size is $M = 1024$, which are updated every $T_{\text{update}} = 2000$ steps. Difficulty score balancing uses $\beta = 0.6$, with pruning thresholds $\theta = 0.4$, and $\delta = 0.05$. Loss weights are initialized as $\lambda_1 = 0.2$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.7$ for the first 5,000 warm-up steps, then adjusted to 0.05, 0.25, and 0.7, respectively for the rest of the training. All experiments are conducted on NVIDIA A100 GPU.

	Method	ShareGPT4V I→T	ShareGPT4V T→I	Urban-1K I→T	Urban-1K T→I
ViT-B/16	CLIP	78.2	79.6	68.1	53.6
	Fine-tuned CLIP	94.1	93.6	80.4	79.8
	Long-CLIP	94.6	93.3	78.9	79.5
	TULIP	<u>98.6</u>	<u>98.6</u>	**88.1**	<u>86.6</u>
	OneLIP (ours)	**99.1**	**98.7**	<u>88.0</u>	**87.2**
ViT-L/14	CLIP	81.8	84.0	68.7	52.8
	Fine-tuned CLIP	95.3	95.4	78.0	76.5
	Long-CLIP	95.8	95.6	82.7	86.1
	TULIP	<u>99.0</u>	<u>99.0</u>	<u>90.1</u>	<u>91.1</u>
	OneLIP (ours)	**99.1**	**99.5**	**92.3**	**93.0**

Table 1: R@1 performance comparison for long-text image retrieval on ShareGPT4V and Urban-1K. The best results are highlighted in bold, and the second-best results are underlined.

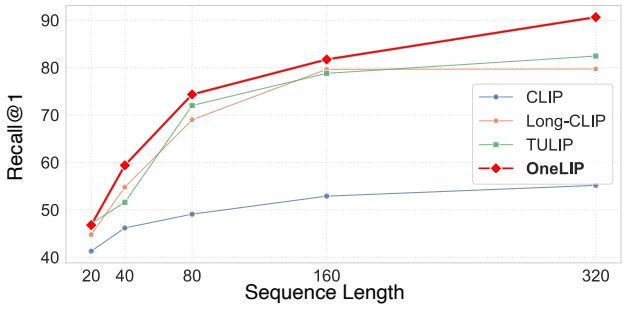

Figure 3: Average R@1 performance trained with different max sequence length. Three baselines stagnate early as sequence length increases, indicating that extended semantics in longer captions suffer from degradation.

Evaluation Metrics and Baselines We use Recall@K (R@K) top-1 accuracy (Acc@1) to evaluate retrieval tasks and image classification, respectively. The baselines are composed of the vanilla CLIP (Radford et al. 2021), Fine-tuned CLIP (CLIP trained with long texts following the positional interpolation strategy of Long-CLIP (Zhang et al. 2024)), Long-CLIP, and TULIP (Najdenkoska et al. 2025).

4.2 Results and Analysis

Long-Text Image Retrieval Table 1 reports the R@1 performance on long-text image retrieval. Across both backbones, OneLIP consistently outperforms all baselines. Notably, on ShareGPT4V, OneLIP achieves 99.1% (I-T) and 98.7% (T-I) under the ViT-B/16, slightly surpassing TULIP (Najdenkoska et al. 2025) while offering a more efficient end-to-end training pipeline. On Urban-1K, OneLIP achieves an overall improvement, demonstrating its superiority to capture long-range semantic dependencies of dense urban scenes. While Long-CLIP (Zhang et al. 2024) and Fine-tuned CLIP show improvements over vanilla CLIP, their gains plateau for longer captions, echoing the attention collapse observed in Figure 1. Moreover, we visualize the average R@1 performance across different maximum se-

	Method	MS-COCO						Flickr30K					
		I→T			T→I			I→T			T→I		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ViT-B/16	CLIP	51.8	76.8	84.3	32.7	57.7	68.2	44.1	68.2	77.0	24.7	45.1	54.6
	Fine-tuned CLIP	37.4	62.3	72.1	21.8	43.4	54.5	25.7	45.8	55.4	17.9	34.5	43.1
	Long-CLIP	57.6	81.1	**87.8**	40.4	65.8	75.2	46.8	71.4	79.8	34.1	56.3	65.7
	TULIP	56.8	80.3	-	40.7	66.1	-	46.1	70.8	-	35.2	57.4	-
	OneLIP (ours)	**58.1**	**81.6**	**87.8**	**42.1**	**66.4**	**75.7**	**47.0**	**72.2**	**81.2**	**35.6**	**57.5**	**66.3**
ViT-L/14	CLIP	56.1	79.5	86.8	35.4	60.1	70.2	48.5	72.6	80.2	28.0	49.3	58.7
	Fine-tuned CLIP	37.9	63.1	72.2	23.1	45.1	55.9	26.0	46.3	55.6	17.9	34.9	43.5
	Long-CLIP	62.8	85.1	91.2	46.3	70.8	79.8	53.4	77.5	85.3	41.2	64.1	72.6
	TULIP	62.6	84.7	-	46.1	71.1	-	56.7	79.5	-	41.6	64.3	-
	OneLIP (ours)	**63.4**	**85.5**	**91.6**	**49.0**	**71.9**	**81.3**	**56.8**	**80.0**	**85.5**	**44.0**	**66.1**	**73.4**

Table 2: Performance comparison for short-text image retrieval on MS-COCO 5K and full Flickr30K. The best results are highlighted in bold, and the second-best results are underlined.

	Method	Datasets					Average
		IN	IN-O	IN-V2	C-10	C-100	
ViT-B/16	CLIP	**68.4**	42.2	61.9	90.8	67.3	66.12
	Fine-tuned CLIP	55.1	31.7	44.8	83.9	59.2	54.94
	Long-CLIP	66.8	42.7	61.2	90.7	**69.3**	66.14
	OneLIP (ours)	68.1	**43.0**	**62.5**	**91.4**	69.2	**66.84**
ViT-L/14	CLIP	**75.5**	31.9	69.9	95.5	76.8	69.92
	Fine-tuned CLIP	58.4	29.2	52.7	92.7	68.7	60.30
	Long-CLIP	73.5	33.7	67.9	95.3	78.5	69.78
	OneLIP (ours)	75.0	**35.2**	**72.9**	**96.7**	**79.3**	**71.82**

Table 3: Acc@1 performance comparison for zero-shot image classification. "IN", "IN-O","IN-V2","C-10" and "C-100" are the abbreviations of ImageNet, ImageNet-O, ImageNet-V2, CIFAR-10 and CIFAR-100, respectively. The best results are highlighted in bold, and the second-best results are underlined.

quence lengths in Figure 3. It is observed that baseline methods stagnate early as sequence length increases, confirming that critical semantics in later tokens are poorly leveraged and suffer from degradation. OneLIP continues to improve with longer sequences, validating that our design mitigates both truncation-induced information loss and insufficient token-level supervision, leading to more faithful long-text representation learning.

Short-Text Image Retrieval To evaluate whether OneLIP retains the performance of CLIP in conventional short-text settings, we conduct experiments on two widely used benchmarks: MS-COCO 5K and full Flickr30K. As shown in Table 2, OneLIP achieves consistently superior results across both backbones and retrieval directions, demonstrating strong robustness and transferability beyond the long-text regime. Unlike fine-tuned CLIP, which shows degraded performance likely due to overfitting to long texts, OneLIP maintains high accuracy across all short-text metrics. For instance, on MS-COCO with ViT-B/16, OneLIP achieves 58.1% R@1 (I-T) and 42.1% R@1 (T-I), surpassing all baselines. On Flickr30K, OneLIP obtains 44.0% R@1 (T-I) under ViT-L/14, outperforming strong multi-stage baselines.

These results suggest that instead of degrading performance, OneLIP also adapts gracefully to short texts by preserving salient token semantics. This confirms that OneLIP not only enhances long-text representations, but also generalizes robustly to short captions, achieving a universal adaptation through a single-stage training pipeline.

Dataset	Baseline (CLIP)	TRIM	TRIM +HNM (M=1024)	TRIM +PO-HNM (M=512)	TRIM +PO-HNM (M=1024)
SG4V (I-T)	78.2	93.2	94.3	96.9	**99.1**
U1K (T-I)	53.6	80.0	83.9	87.0	**87.2**
COCO (I-T)	51.8	55.5	56.3	56.4	**58.1**
F30K (T-I)	24.7	32.0	32.3	34.9	**35.6**
CIFAR10	90.8	90.4	90.7	90.9	**91.4**
CIFAR100	67.3	68.1	68.5	69.2	**69.2**

Table 4: Ablation study of components in OneLIP, where vanilla CLIP is set as the baseline. Dataset abbreviations: SG4V for ShareGPT4V, U1K for Urban-1K, COCO for MS-COCO, F30K for Flickr30K, M for negative query size.

Zero-shot Image Classification To assess the generalization ability of OneLIP in open-world settings, we further evaluate its zero-shot classification performance. As shown in Table 3, OneLIP achieves competitive performance across most datasets. Notably, with ViT-B/16, OneLIP attains the best average accuracy (66.84%), indicating its robustness to diverse image domains, where similar trends are observed under ViT-L/14. While OneLIP exhibits a slight drop on ImageNet under ViT-L/14 (75.0% vs. 75.5% for vanilla CLIP), we attribute this to minor representational shifts introduced by long-text adaptation, which may slightly affect class name grounding. Importantly, OneLIP surpasses all baselines in average accuracy, and maintains strong performance on out-of-distribution benchmarks such as ImageNet-O and ImageNet-V2.

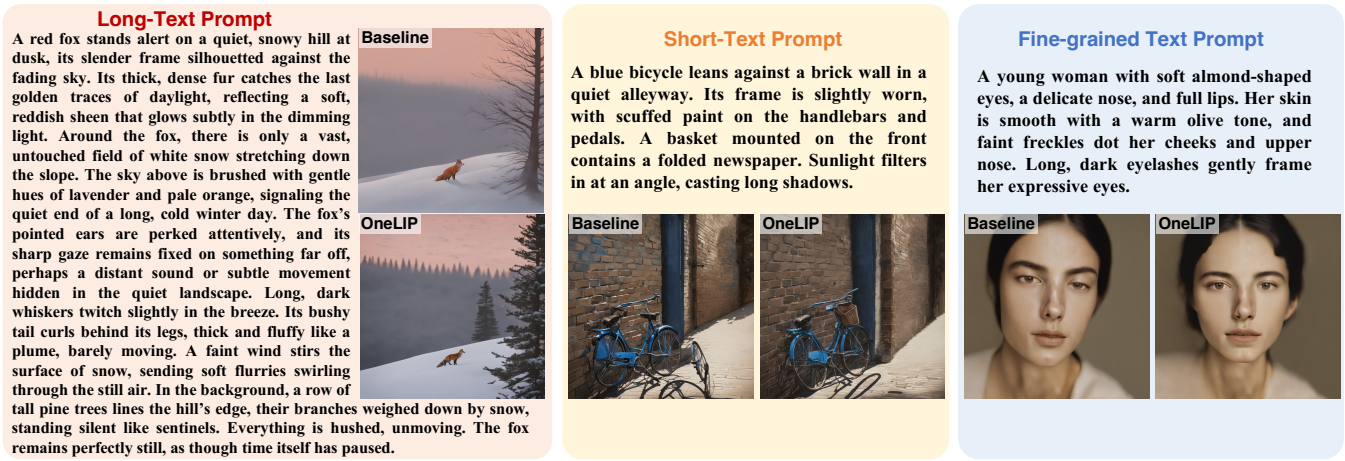

Figure 4: Plug-and-play text-to-image generation results, which qualitatively validate the text representations of OneLIP. Three types of prompts are compared: long texts, short texts, and fine-grained texts.

4.3 Ablation Study

We perform ablation studies to quantify the contributions of each component in OneLIP across the above three tasks using ViT-B/16 as the backbone. Starting from the CLIP baseline, we incrementally add TRIM and different hard negative mining strategies for comparison, as shown in Table 4.

Introducing TRIM alone leads to significant gains across all tasks. On ShareGPT4V (I-T), R@1 improves from 78.2 to 93.2, and on Urban-1K (T-I), from 53.6 to 80.0, indicating that selecting informative and modality-relevant tokens substantially enhances long-text alignment. TRIM also improves performance on short-text datasets (+4.6 on Flickr30K T-I), showing adaptability to various sequence lengths without overfitting to long inputs. Next, we assess hard negative mining. Compared to offline HNM, PO-HNM yields notable improvements (+4.8 on SG4V and +3.3 on Urban-1K) while requiring only half the number of negatives per sample ($M = 512$ vs. $M = 1024$), demonstrating both effectiveness and efficiency in long-text scenarios where semantic drift is common. For zero-shot classification, improvements are generally milder but still consistent, reflecting the inherently difference between classification and retrieval. Nonetheless, performance improvements on CIFAR-10 and CIFAR-100 confirm that OneLIP preserves generalization without sacrificing classification capability.

In summary, the ablations validate the contribution of each component in OneLIP. More studies on TRIM submodules, sequence length sensitivity, and update frequency in PO-HNM (T_{update}) are in the Appendix.

4.4 Plug-and-play Text-to-Image Generation

To evaluate OneLIP's long-text representation beyond retrieval and classification, we further conduct studies on text-to-image generation. Specifically, we replace the original CLIP-L text encoder in Stable Diffusion XL (Podell et al. 2023) with the long-text encoder in OneLIP. Importantly, we keep all settings unchanged and no additional training is introduced to ensure a fair comparison, reflecting the encoder's ability to capture textual semantics.

We compare against LongCLIP (Zhang et al. 2024) under three prompting scenarios: (1) long-text prompt with detailed descriptions; (2) short-text prompt with around 50 words; and (3) fine-grained prompts focusing on subtle details and precise modeling of attributes typically located at the end of the input.

As shown in Figure 4, OneLIP consistently generates images that better reflect text semantics. In long-text prompt case, OneLIP produces a more vivid *"red fox"* with finer texture and improved scene composition such as *"a line of pine trees stands along the slope" and "dark green needles heavy with snow"*, which are entirely absent in the baseline, suggesting OneLIP's superiority in late-positioned semantics. Under short prompts, OneLIP more accurately captures localized objects like the *"basket mounted on the front"* and *"pedals"*, indicating stronger token grounding. In fine-grained prompts, OneLIP preserves more realistic and subtle facial cues like the *"almond-shaped eyes"*, *"faint freckles"*, and *"long dark eyelashes"*, demonstrating that OneLIP not only improves long-text understanding but also enhances semantic fidelity and detail preservation in generative tasks.

5 Conclusion

We present OneLIP, a one-stage framework for adapting CLIP to long texts without truncation or multi-stage tuning. OneLIP introduces two key innovations: (1) TRIM selects and refines informative tokens for token-wise distillation to mitigate semantic degradation; and (2) PO-HNM strategy dynamically maintains sample-specific negatives via dual-modality difficulty tracking, which is crucial for addressing dispersed semantics in long captions. Extensive experiments across retrieval, classification, and generation tasks demonstrate that OneLIP provides an efficient, scalable, and semantically faithful solution for long-text CLIP adaptation.

Acknowledgments

This research was partly supported by grants of National Natural Science Foundation of China (NSFC, GrantNo.62171281), Science and Technology Commission of Shanghai Municipality (STCSM, Grant Nos. 25GA3200103, 2021SHZDZX0102).

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

bloc97. 2023. NTK-aware scaled RoPE allows LLaMA models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation. Accessed: June 2023.

Cai, X.; Pan, R.; and Yang, H. 2025. LoKi: Low-dimensional KAN for Efficient Fine-tuning Image Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 14869–14880.

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

Chen, R.; Liu, Y.; Kong, L.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; Qiao, Y.; and Wang, W. 2023. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7020–7030.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15750–15758.

Chun, S.; and Yun, S. 2025. LongProLIP: A probabilistic vision-language model with long context text. *arXiv preprint arXiv:2503.08048*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Fang, A.; Jose, A. M.; Jain, A.; Schmidt, L.; Toshev, A.; and Shankar, V. 2023. Data filtering networks. *arXiv preprint arXiv:2309.17425*.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595.

Gao, T.; Yao, X.; and Chen, D. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15262–15271.

Huynh, T.; Kornblith, S.; Walter, M. R.; Maire, M.; and Khademi, M. 2022. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2785–2795.

Jing, D.; He, X.; Luo, Y.; Fei, N.; Wei, W.; Zhao, H.; Lu, Z.; et al. 2024. Fineclip: Self-distilled region-based clip for better fine-grained understanding. *Advances in Neural Information Processing Systems*, 37: 27896–27918.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33: 18661–18673.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

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.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Lyu, W.; Lin, J.; Ren, W.; Xia, R.; Qian, F.; and Tang, Y. 2025. DidSee: Diffusion-Based Depth Completion for Material-Agnostic Robotic Perception and Manipulation. *arXiv preprint arXiv:2506.21034*.

Mu, N.; Kirillov, A.; Wagner, D.; and Xie, S. 2022. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, 529–544. Springer.

Najdenkoska, I.; Derakhshani, M. M.; Asano, Y. M.; Noord, N. V.; Worring, M.; and Snoek, C. G. M. 2025. TULIP: Token-length Upgraded CLIP. In *The Thirteenth International Conference on Learning Representations*.

Pan, R.; Dong, J.; and Yang, H. 2025. Discovering clone negatives via adaptive contrastive learning for image-text matching. In *The Thirteenth International Conference on Learning Representations*.

Pan, R.; Yang, H.; Li, C.; and Yang, J. 2023. Joint Intra & Inter-Grained Reasoning: A New Look Into Semantic Consistency of Image-Text Retrieval. *IEEE Transactions on Multimedia*, 26: 4912–4925.

Pan, R.; Yang, H.; and Zhao, X. 2025. ReAL: Improving Image-Text Retrieval with Authentic Negative Repository Learning. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

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.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Shao, T.; Tian, Z.; Zhao, H.; and Su, J. 2024. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *European Conference on Computer Vision*, 139–156. Springer.

Song, J.; Pan, R.; Zhou, J.; and Yang, H. 2024. M-RAT: a Multi-grained Retrieval Augmentation Transformer for Image Captioning. In *Proceedings of the Asian Conference on Computer Vision*, 3865–3882.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

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.

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*.

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.

Xu, X.; Wu, C.; Rosenman, S.; Lal, V.; Che, W.; and Duan, N. 2023. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10637–10647.

Yang, J.; Li, C.; Zhang, P.; Xiao, B.; Liu, C.; Yuan, L.; and Gao, J. 2022. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19163–19173.

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.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

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.

Zhou, D.; Chen, P.; Wang, Q.; Chen, G.; and Heng, P.-A. 2022. Acknowledging the unknown for multi-label learning with single positive labels. In *European Conference on Computer Vision*, 423–440. Springer.

Zhou, D.; Huang, J.; Bai, J.; Wang, J.; Chen, H.; Chen, G.; Hu, X.; and Heng, P.-A. 2024. Magictailor: Component-controllable personalization in text-to-image diffusion models. *arXiv preprint arXiv:2410.13370*.