ProCLIP: Progressive Vision-Language Alignment via LLM-based Embedder

Xiaoxing Hu[1,2] **Kaicheng Yang**[3†] **Ziqi Ye**[4,2] **Ziyang Gong**[1] **Qi Ming**[5]
Zonghao Guo[6] **Yu Tian**[6] **Xiang An**[3] **Ziyong Feng**[3] **Xue Yang**[1✉]

[1]SAIS & SCS, Shanghai Jiao Tong University [2]Shanghai Innovation Institute
[3]DeepGlint [4]Fudan University [5]Beijing University of Technology [6]Tsinghua University

[†]Project lead [✉]Corresponding author

Code: https://github.com/VisionXLab/ProCLIP

Abstract

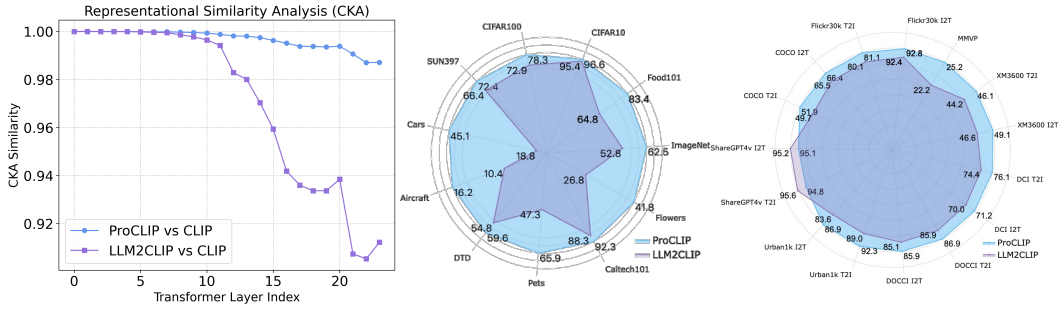
Contrastive Language-Image Pre-training (CLIP) is fundamentally limited by its 77-token restriction, lack of multilingual capabilities, and coarse-grained semantic representations. While replacing CLIP's native text encoder with a Large Language Model (LLM)-based embedder offers a promising solution, direct, from-scratch alignment can disrupt the semantic structures learned during pre-training. This often leads to degradation of cross-modal knowledge and particularly impairs recognition capabilities in zero-shot scenarios. In this paper, we propose **ProCLIP**, a curriculum-learning-inspired progressive alignment framework designed to systematically bridge the LLM embedder and CLIP's visual space, unlocking CLIP's potential for long-text, multilingual, and fine-grained understanding. Our framework operates in two stages: (1) **Representation Inheritance**, which distills CLIP's original text-space knowledge into an LLM adapter to establish a robust initial vision-language prior, and (2) **Contrastive Tuning**, which refines the cross-modal alignment with self-distillation regularization on the image encoder to further prevent catastrophic forgetting. To maintain semantic and geometric consistency, we introduce instance-level semantic and global structural alignment constraints throughout both stages. Extensive experiments demonstrate that ProCLIP improves zero-shot classification accuracy by 6.8%–13.5% over existing LLM-augmented baselines and achieves state-of-the-art performance in diverse long-text, multilingual, and fine-grained cross-modal retrieval tasks under comparable settings.

1 Introduction

CLIP [36] has demonstrated strong vision-language representations via large-scale contrastive learning, serving as a foundation for diverse downstream tasks, including retrieval [47], generation [53], and detection [49]. Despite its success, CLIP's native text encoder is constrained by a strict 77-token limit and primarily relies on English-only supervision [55, 44]. Moreover, the lack of fine-grained textual supervision limits its ability to capture nuanced semantic details, restricting applicability in complex, real-world scenarios [19].

To overcome these limitations, recent works have sought to enhance CLIP's textual capabilities. Long-CLIP [55] interpolates positional embeddings to accommodate longer inputs, yet still struggles with fine-grained and multilingual understanding. Other methods, such as FLAME [5] and LLM2CLIP [20], replace CLIP's text encoder with Large Language Model (LLM)-based embedders [3, 25] to leverage richer linguistic priors. However, these approaches typically employ a

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) CKA similarity across layers **(b) Classification performance** **(c) Retrieval performance**

Figure 1: (a) CKA analysis of visual representations. (b) Comparison of classification performance and (c) retrieval performance between ProCLIP and our baseline LLM2CLIP. The figure uses a ViT-L base model trained on 3M samples.

from-scratch alignment, directly forcing the pre-trained image encoder to align with the LLM space via standard contrastive learning. By ignoring the valuable cross-modal knowledge embedded in CLIP, this naive strategy can disrupt pretrained semantic structures and degrade cross-modal representations, particularly affecting performance in zero-shot scenarios.

This motivates the following question: *Can we leverage CLIP's pretrained knowledge to achieve effective cross-modal alignment with LLMs?* To address this challenge, we propose **ProCLIP**, a curriculum-learning-inspired progressive alignment framework that treats pretrained weights as foundational priors to bridge LLM embedders and CLIP's visual space. Rather than enforcing an abrupt semantic shift, ProCLIP adopts a two-stage curriculum:

(1) **Representation Inheritance**: We distill knowledge from the original CLIP text encoder into a lightweight adapter for the LLM. Instance-level semantic and global structural alignment constraints anchor the frozen LLM to CLIP's textual space, establishing a robust initial prior. (2) **Contrastive Tuning**: Building upon this foundation, we refine the representation space via contrastive tuning. This process preserves pretrained vision-language knowledge while improving cross-modal alignment.

We validate the preservation of pretrained knowledge via Centered Kernel Alignment (CKA) analysis on vision encoder features (Fig. 1 (a)). ProCLIP demonstrates superior retention of pre-trained knowledge compared to existing baselines, confirming that progressive alignment effectively facilitates cross-modal learning without sacrificing generalization. The framework offers multiple advantages: 1) *Simplicity*: an intuitive, easy-to-implement design using pretrained weights as a semantic bridge; 2) *Superiority*: state-of-the-art performance under identical data constraints (Fig. 1 (b) and (c)); 3) *Versatility*: supports long-text, multilingual, and fine-grained scenarios; 4) *Potential*: a robust paradigm for modular upgrades in multi-encoder foundation models.

Our main contributions are summarized as follows:

- We **identify a critical limitation** in current LLM-augmented CLIP models: the from-scratch alignment neglects pretrained priors, leading to degraded cross-modal representations and impaired zero-shot performance.
- We **propose ProCLIP**, a curriculum-learning-inspired progressive alignment framework that effectively leverages pretrained knowledge to gradually build robust cross-modal representations, substantially preserving zero-shot recognition capabilities.
- We **demonstrate ProCLIP's efficacy** through extensive experiments. Under identical training constraints, ProCLIP achieves 6.8%–13.5% absolute gains in zero-shot classification and shows superior performance across diverse short/long-text, multilingual, and fine-grained cross-modal retrieval tasks.

2 Related Work

Vision-Language Contrastive Learning. Vision-language contrastive learning, exemplified by CLIP [36], aims to align multimodal representations in a shared semantic space via large-scale pretraining. As a foundational bridge between modalities, CLIP enables diverse open-vocabulary

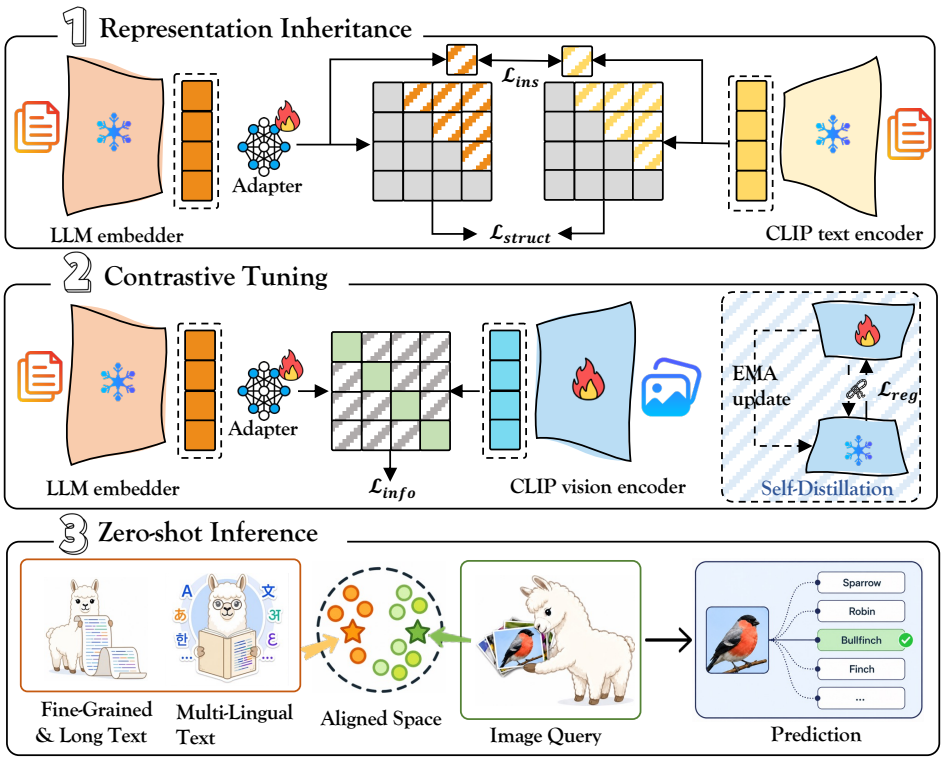


Figure 2: The pipeline of our proposed ProCLIP. It consists of representation inheritance via cross-architecture distillation and contrastive tuning integrated with self-distillation regularization.

tasks, including image classification [63, 62, 23], semantic segmentation [9, 27, 12, 51, 8, 24], and object detection [10, 22]. Despite its versatility, CLIP is fundamentally constrained by its text encoder's fixed input length and limited capacity, which hampers its performance in processing long-form, multilingual, and fine-grained semantics. To address these bottlenecks, Long-CLIP [55] extends input lengths through positional embedding interpolation, while LoTLIP [48] aggregates diverse textual information using corner tokens to enhance long-text understanding. However, these methods remain tethered to the original CLIP text encoder's architecture, preventing the integration of broader open-world knowledge and leaving multilingual support unresolved.

LLM-based Representation Learning. Large Language Models (LLMs) have demonstrated exceptional proficiency across diverse natural language processing tasks [43, 1, 2, 30]. Recent research has pivoted toward repurposing decoder-only architectures for robust representation learning. Notably, LLM2Vec [3] transforms LLMs into versatile encoders via bidirectional attention and contrastive alignment, while Qwen3-Embedding [58] leverages massive-scale pretraining to achieve state-of-the-art performance on the MTEB benchmark [33]. Inspired by these advancements, recent works [20, 5, 56] replace CLIP's original text encoder with LLM-based embedders to handle multilingual, long-form, and complex textual inputs. However, these methods often employ coarse alignment strategies that inadvertently degrade the model's inherent generalization. Addressing this gap by developing more refined and efficient alignment techniques remains a critical and open challenge.

3 Methodology

In this section, we first introduce the preliminary concepts (Sec. 3.1), including contrastive language-image pre-training and improving CLIP with an LLM-based embedder. We then present our proposed **ProCLIP** framework, which comprises two primary stages: 1) Representation Inheritance via Cross-Architecture Distillation (Sec. 3.2) and 2) Contrastive Tuning with Self-Distillation Regularization (Sec. 3.3). For a more detailed theoretical analysis, please refer to the Appendix.

3.1 Preliminary

Contrastive Language-Image Pre-training. CLIP [36] learns joint embeddings for images and text from large-scale image-text pairs using contrastive learning. Given a batch of image-text pairs $\{(I_i, T_i)\}_{i=1}^{\mathcal{B}}$, the pre-trained image encoder $\mathcal{E}_I$ and text encoder $\mathcal{E}_T$ map them into joint semantic representations $\{(v_i, t_i)\}_{i=1}^{\mathcal{B}}$. The dual-tower architecture is optimized with a symmetric InfoNCE loss:

$$\mathcal{L}_{\text{CLIP}} = -\sum_{i=1}^{\mathcal{B}} \left[\log \frac{\exp(v_i \cdot t_i/\tau)}{\sum_{j=1}^{\mathcal{B}} \exp(v_i \cdot t_j/\tau)} + \log \frac{\exp(t_i \cdot v_i/\tau)}{\sum_{j=1}^{\mathcal{B}} \exp(t_i \cdot v_j/\tau)} \right], \tag{1}$$

where τ is a learnable temperature. However, the native CLIP text encoder $\mathcal{E}_T$ is limited to sequences of 77 tokens.

Improving CLIP with an LLM-based Embedder. Recent approaches such as LLM2CLIP [20] replace the CLIP text encoder with a pre-trained LLM-based embedder $\mathcal{G}_T$ to improve long-text comprehension, multilingual understanding, and fine-grained semantic discrimination. Text inputs $\{T_i\}$ are encoded offline as embeddings $\{t_i' = \mathcal{G}_T(T_i)\}$. During fine-tuning, an MLP-based adapter projects these embeddings into the CLIP visual space.

Directly applying contrastive learning on the LLM embeddings and visual encoder is challenging because the LLM embedding space initially lacks alignment with the visual encoder. Unconstrained fine-tuning on limited data can lead to drift in the image encoder, reducing zero-shot generalization. ProCLIP addresses this with a progressive, two-stage alignment strategy.

3.2 Stage I: Representation Inheritance via Cross-Architecture Distillation

Given pre-trained CLIP encoders $\{\mathcal{E}_I, \mathcal{E}_T\}$ and a pre-trained LLM embedder $\mathcal{G}_T$, our goal is to replace $\mathcal{E}_T$ with $\mathcal{G}_T$ while preserving CLIP's established visual-textual relationships. We first extract textual embeddings offline: $t_i' = \mathcal{G}_T(T_i) \in \mathbb{R}^d$.

Cross-Architecture Distillation. We use the original CLIP text encoder $\mathcal{E}_T$ as a teacher to provide structural supervision to the LLM embeddings. This design leverages CLIP's pretrained visual-textual space as a **stable structural scaffold**, rather than relying on CLIP to process long or multilingual text. The LLM embedder, which already encodes long-text and multilingual semantics, inherits these visual alignment priors, allowing it to produce embeddings that respect CLIP's visual space while supporting long-text and multiple languages.

Specifically, we use an MLP adapter to project LLM embeddings to the CLIP dimension: $t_i^* = \text{Adapter}(t_i')$, and define two alignment losses:

Instance-level semantic alignment:

$$\mathcal{L}_{\text{ins}} = \sum_{i=1}^{\mathcal{B}} \|t_i^* - e_i\|_2 \tag{2}$$

aligning individual LLM embeddings t_i^* to the corresponding CLIP teacher embeddings $e_i = \mathcal{E}_T(T_i)$.
Global structural alignment:

$$\mathcal{L}_{\text{struct}} = \sum_{i<j}^{\mathcal{B}} |d(t_i^*, t_j^*) - d(e_i, e_j)| \tag{3}$$

preserving inter-sample geometric relationships. The overall Stage I loss is $\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{ins}} + \mathcal{L}_{\text{struct}}$.

3.3 Stage II: Contrastive Tuning Integrated with Self-Distillation Regularization.

After the inheritance phase, the adapter has been preliminarily anchored to CLIP's established vision-language embedding space. This significantly eases the subsequent optimization landscape for cross-modal contrastive fine-tuning. We utilize the standard InfoNCE loss to fully align the image embeddings $v_i = \mathcal{E}_I(I_i)$ with the projected LLM embeddings t_i^*:

$$\mathcal{L}_{\text{info}} = -\sum_{i=1}^{\mathcal{B}} \left[\log \frac{\exp(v_i \cdot t_i^*/\tau)}{\sum_{j=1}^{\mathcal{B}} \exp(v_i \cdot t_j^*/\tau)} + \log \frac{\exp(t_i^* \cdot v_i/\tau)}{\sum_{j=1}^{\mathcal{B}} \exp(t_i^* \cdot v_j/\tau)} \right]. \tag{4}$$

Table 1: Cross-modal retrieval performance Recall@1 on multiple datasets. **Bold values** indicate the better performance between LLM2CLIP and ProCLIP.

Method	#Data	Flickr30k		COCO		ShareGPT4V		Urban-1k		DOCCI		DCI		Avg.	
		I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I
Model Architecture: CLIP ViT-B/32															
LLM2CLIP	3M	83.5	70.1	55.6	41.1	94.2	**93.4**	78.2	84.2	76.2	77.1	62.2	64.4	75.0	71.7
ProCLIP	3M	**86.0**	**73.5**	**57.8**	**43.5**	**94.4**	92.6	**80.8**	**85.3**	**78.1**	**79.5**	**65.7**	**68.3**	**77.1**(+2.1)	**73.8**(+2.1)
LLM2CLIP	15M	86.2	72.2	58.5	43.2	**95.3**	**94.2**	80.6	**85.3**	**79.2**	**80.7**	64.3	67.6	77.4	73.9
ProCLIP	15M	**86.6**	**72.6**	**59.0**	**43.5**	94.5	93.9	**82.2**	**85.3**	78.4	80.6	**67.1**	**69.2**	**78.0**(+0.6)	**74.2**(+0.3)
CLIP	30M	85.6	69.5	59.3	42.8	76.8	67.3	62.5	43.2	57.2	54.8	48.1	46.3	64.9	54.0
DreamLIP	30M	87.2	66.4	58.3	41.1	78.2	69.1	66.5	47.4	59.3	56.1	49.1	47.9	66.4	54.7
LLM2CLIP	30M	87.8	72.4	61.1	44.3	96.7	**95.9**	86.6	88.8	**82.9**	82.9	67.9	69.5	80.5	75.7
ProCLIP	30M	**90.2**	**74.6**	**62.4**	**45.9**	**96.8**	**95.9**	**88.5**	**89.9**	**82.9**	**84.1**	**70.6**	**71.9**	**81.9**(+1.4)	**77.0**(+1.3)
Model Architecture: CLIP ViT-B/16															
Long-CLIP	1M	90.0	76.2	62.8	46.3	88.4	89.6	78.9	79.5	63.1	71.4	51.7	57.3	72.5	70.1
FG-CLIP	1.6B	90.7	76.4	64.1	45.4	89.5	90.0	79.0	80.3	65.9	75.4	61.8	60.6	75.2	71.4
LLM2CLIP	3M	88.0	75.3	60.5	44.8	**94.4**	**94.4**	80.6	86.0	**81.7**	82.2	67.2	69.1	78.7	75.3
ProCLIP	3M	**89.4**	**77.6**	**61.7**	**46.8**	94.3	93.3	**82.9**	**88.1**	81.0	**82.5**	**67.3**	**72.0**	**79.4**(+0.7)	**76.7**(+1.4)
LLM2CLIP	15M	88.9	76.6	62.4	46.5	**95.0**	**95.2**	84.5	88.4	**83.8**	**85.1**	69.3	72.4	80.7	77.3
ProCLIP	15M	**90.8**	**77.9**	**63.2**	**47.8**	94.2	94.9	**85.8**	**89.6**	82.5	84.6	**70.2**	**74.0**	**81.2**(+0.5)	**78.0**(+0.7)
CLIP	30M	90.5	75.9	63.2	48.2	77.9	68.3	63.9	45.4	59.3	55.4	49.6	48.7	67.4	57.0
DreamLIP	30M	89.9	73.3	62.3	44.9	79.9	70.3	67.9	50.1	60.4	57.3	50.2	48.8	68.4	57.5
LLM2CLIP	30M	90.2	78.1	65.4	48.5	**96.8**	**96.4**	89.7	91.3	**86.2**	86.8	73.1	74.8	83.6	79.3
ProCLIP	30M	**92.7**	**79.1**	**67.1**	**49.7**	96.0	**96.4**	**90.0**	**93.4**	85.1	**87.3**	**73.6**	**76.9**	**84.2**(+0.6)	**80.5**(+1.2)
Model Architecture: CLIP ViT-L/14															
Long-CLIP	1M	89.7	75.7	62.5	46.4	93.3	92.1	82.4	86.2	66.4	78.6	57.2	64.0	75.2	73.8
FG-CLIP	1.6B	93.7	81.5	68.9	50.9	94.0	93.2	83.0	87.1	69.3	80.4	66.7	66.1	79.3	76.5
LLM2CLIP	3M	92.4	80.1	65.5	49.7	**95.2**	**95.6**	83.6	89.0	85.1	85.9	70.0	74.4	82.0	79.1
ProCLIP	3M	**92.8**	**81.1**	**66.4**	**51.9**	95.1	94.8	**86.9**	**92.3**	**85.9**	**86.9**	**71.2**	**76.1**	**83.0**(+1.0)	**80.5**(+1.4)
LLM2CLIP	15M	91.3	80.6	67.0	50.6	**96.3**	95.3	86.4	90.5	**86.4**	**88.5**	71.7	75.3	83.2	80.1
ProCLIP	15M	**93.4**	**81.4**	**67.6**	**52.5**	96.1	**95.4**	**88.3**	**92.6**	86.2	88.4	**74.4**	**76.8**	**84.3**(+1.1)	**81.2**(+1.1)
LLM2CLIP	30M	93.1	81.0	68.2	52.0	**97.5**	**97.7**	92.7	93.9	**88.2**	89.6	74.9	78.3	85.8	82.1
ProCLIP	30M	**94.5**	**81.6**	**69.3**	**53.2**	96.8	97.0	**93.0**	**94.4**	87.5	**89.8**	**75.9**	**79.5**	**86.2**(+0.4)	**82.6**(+0.5)
MLLM-based method															
VLM2Vec (LLaVA Next 1.6)	1M	91.6	79.8	61.3	51.5	93.9	91.0	90.0	92.0	80.5	86.0	68.7	74.7	81.0	79.2

Beyond standard contrastive learning, it is imperative to mitigate the catastrophic forgetting of pre-trained visual knowledge during this adaptation. Therefore, we impose a **self-distillation regularization constraint** on the CLIP image encoder. Symmetrical to the knowledge transfer in Stage I, we apply both instance-level and structural constraints on the visual side. Let $v_i^* = \mathcal{E}_I^*(I_i)$ and $v_j^* = \mathcal{E}_I^*(I_j)$ denote the "teacher" vision embeddings derived from an EMA-updated image encoder. The regularization loss $\mathcal{L}_{\text{reg}}$ is formulated as:

$$\mathcal{L}_{\text{reg}} = \sum_{i=1}^{\mathcal{B}} \|v_i^* - v_i\|_2 + \sum_{i<j}^{\mathcal{B}} |d(v_i^*, v_j^*) - d(v_i, v_j)|. \tag{5}$$

The teacher model $\mathcal{E}_I^*$ is updated via EMA from the student image encoder $\mathcal{E}_I$ at each training step to provide stable, historical targets:

$$\mathcal{E}_I^* \leftarrow \alpha \mathcal{E}_I^* + (1-\alpha)\mathcal{E}_I, \tag{6}$$

where $\alpha \in [0, 1)$ controls the momentum update rate. The overall objective function is thus defined as:

$$\mathcal{L}_{\text{tune}} = \mathcal{L}_{\text{info}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}, \tag{7}$$

where λ_{reg} is a hyperparameter balancing the contrastive alignment and the self-distillation penalty.

4 Experiments

4.1 Experimental Setup

Datasets and Benchmarks. For the alignment pre-training phase, we utilize CC3M [6], CC12M [6], and YFCC15M [41], combined with curated high-quality captions from DreamLIP [59]. To systematically explore the effects of data scaling, we conduct experiments across three data volumes: 3M (CC3M), 15M (CC3M + CC12M), and 30M (CC3M + CC12M + YFCC15M). For downstream evaluation, we comprehensively assess the model across multiple dimensions: (1) zero-shot classification on 11 diverse datasets, (2) robustness evaluations on 5 ImageNet variants, (3) standard cross-modal retrieval on 6 benchmarks, (4) multilingual retrieval on XM3600 [40], and (5) fine-grained compositional understanding on SugarCrepe [18] and MMVP-VLM [42]. Additional experiments and evaluations are provided in the Appendix.

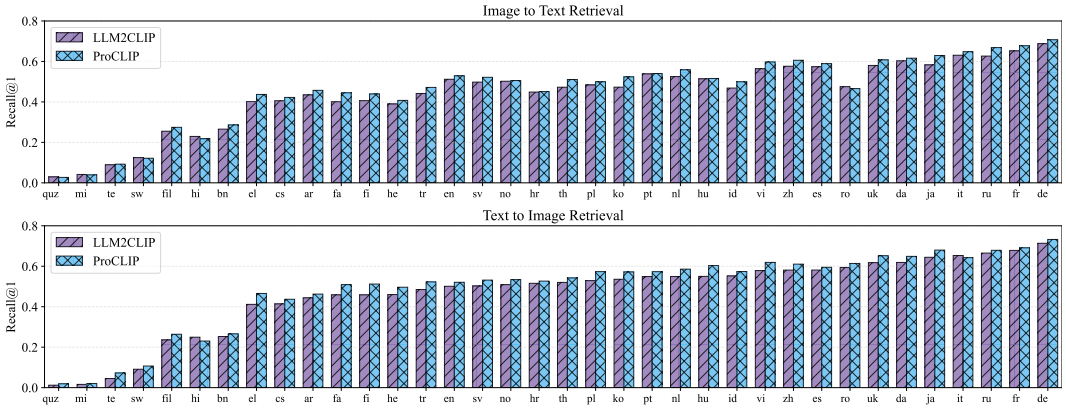


Figure 3: Per-language image-text retrieval performance on the XM3600 benchmark.

Table 2: Zero-shot classification performance on 11 datasets. Bold numbers highlight the emphasized comparisons within each architecture block.

Method	#Data	Food101	CIFAR10	CIFAR100	SUN397	Cars	Aircraft	DTD	Pets	Caltech101	Flowers	ImageNet	Avg.
Model Architecture: CLIP ViT-B/32													
LLM2CLIP	3M	49.6	89.2	61.5	60.3	11.5	8.6	47.8	38.0	79.0	22.6	41.0	46.3
ProCLIP	3M	**64.5**	**90.7**	**65.8**	**65.0**	**21.2**	**11.6**	**52.0**	**51.7**	**83.3**	**30.8**	**47.9**	**53.1**(+6.8)
LLM2CLIP	15M	57.2	88.3	61.4	61.3	19.6	8.4	50.6	42.3	80.7	23.5	43.3	48.8
ProCLIP	15M	**74.9**	**90.0**	**66.5**	**65.1**	**39.6**	**13.9**	**53.7**	**68.5**	**86.7**	**35.5**	**53.3**	**58.9**(+10.1)
CLIP	30M	54.8	90.0	67.1	62.0	13.0	3.6	27.6	49.4	83.8	41.4	45.6	48.9
DreamLIP	30M	68.2	**91.8**	**69.2**	62.2	20.7	8.0	32.1	62.8	86.1	**48.5**	**55.7**	55.0
LLM2CLIP	30M	58.5	88.3	61.0	61.2	20.6	8.4	50.3	37.6	81.7	26.0	45.1	49.0
ProCLIP	30M	**74.4**	88.8	66.9	**65.9**	**38.0**	**16.2**	**53.0**	**64.5**	**86.8**	40.4	54.0	**59.0**(+10.0)
Model Architecture: CLIP ViT-B/16													
Long-CLIP	1M	86.8	88.7	65.3	62.3	62.8	18.8	43.2	86.7	88.9	65.4	66.9	66.9
FG-CLIP	1.6B	84.9	87.5	64.3	59.8	60.9	17.6	40.3	85.7	86.6	64.4	65.8	65.3
LLM2CLIP	3M	56.9	**92.6**	64.4	62.2	15.4	11.7	50.9	46.5	82.9	23.6	45.8	50.3
ProCLIP	3M	**73.1**	92.5	**68.9**	**67.9**	**32.3**	**13.5**	**54.1**	**59.8**	**87.0**	**35.8**	**54.8**	**58.2**(+7.9)
LLM2CLIP	15M	63.2	**90.8**	64.5	62.9	27.3	9.9	52.8	50.3	83.2	23.7	46.5	52.3
ProCLIP	15M	**80.3**	**90.8**	**69.7**	**67.4**	**44.3**	**16.5**	**56.7**	**75.8**	**88.4**	**40.8**	**58.6**	**62.7**(+10.4)
CLIP	30M	61.3	92.2	66.9	62.2	19.3	5.7	30.9	49.3	83.7	43.4	50.0	51.4
DreamLIP	30M	76.4	**92.7**	**71.7**	64.7	23.4	8.3	33.9	66.3	87.6	**59.1**	58.4	58.4
LLM2CLIP	30M	64.4	90.2	64.6	63.7	27.0	11.2	55.0	45.9	84.0	27.1	49.7	53.0
ProCLIP	30M	**81.0**	89.3	68.3	**68.2**	**48.5**	**17.9**	**57.3**	**70.2**	**88.8**	44.8	**59.2**	**63.0**(+10.0)
Model Architecture: CLIP ViT-L/14													
Long-CLIP	1M	90.8	96.3	78.7	71.3	70.3	28.1	56.0	90.4	93.6	71.1	72.0	74.5
FG-CLIP	1.6B	89.7	95.7	76.9	69.9	68.4	26.8	55.7	79.8	92.1	69.8	70.4	72.3
LLM2CLIP	3M	64.8	95.4	72.9	66.4	18.8	10.4	54.8	47.3	88.3	26.8	52.8	54.4
ProCLIP	3M	**83.4**	**96.6**	**78.3**	**72.4**	**45.1**	**16.2**	**59.6**	**65.9**	**92.3**	**41.8**	**62.5**	**64.9** (+10.5)
LLM2CLIP	15M	70.1	95.2	72.3	66.4	32.4	9.5	58.0	54.3	88.3	26.6	54.0	57.0
ProCLIP	15M	**87.1**	**95.4**	**77.6**	**72.3**	**59.8**	**21.1**	**62.1**	**77.0**	**92.4**	**48.8**	**66.0**	**69.3** (+12.3)
LLM2CLIP	30M	71.2	94.0	70.5	67.0	32.1	11.3	57.8	54.7	89.3	28.8	56.4	57.5
ProCLIP	30M	**88.9**	**94.1**	**77.7**	**72.5**	**61.1**	**25.2**	**62.8**	**81.5**	**92.9**	**57.2**	**67.8**	**71.0** (+13.5)

Implementation Details. For the LLM-based embedder, we primarily utilize LLaMA3-8B-CC, consistent with the LLM2CLIP baseline [20]. During our progressive alignment, the Representation Inheritance phase is trained for four epochs, followed by another four epochs of Contrastive Tuning. See the Appendix for detailed parameter configurations and evaluation settings.

4.2 Main Results

Cross-Modal Retrieval. As shown in Tab. 1, ProCLIP achieves stronger average retrieval performance than the competitive LLM2CLIP baseline across the evaluated model scales, while a few individual metrics remain comparable or slightly lower. In standard short-text scenarios (e.g., Flickr30k and COCO), ProCLIP yields substantial gains. Notably, using the ViT-L/14 architecture

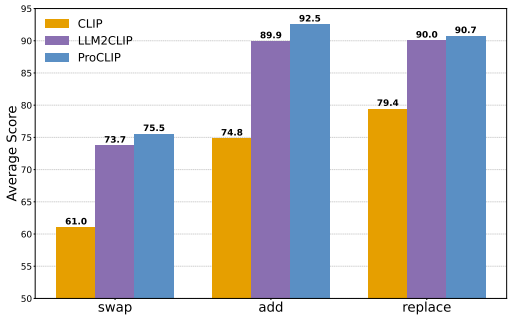


Figure 4: Performance on SugarCrepe.

Table 3: Robustness performance. The best results are marked in **bold**.

Method	#Data	Robustness IN-V2	IN-A	IN-O	IN-R	IN-S	Avg.
Model Architecture: CLIP ViT-L/14							
CLIP	400M	69.8	70.8	32.2	87.8	59.6	64.0
LLM2CLIP	3M	49.0	46.6	**32.4**	75.0	44.8	49.5
ProCLIP	3M	**58.3**	**63.3**	31.6	**84.0**	**52.3**	**57.9**
LLM2CLIP	15M	50.8	50.1	33.8	78.2	46.3	51.8
ProCLIP	15M	**62.1**	**66.4**	**34.2**	**86.4**	**55.3**	**60.9**
LLM2CLIP	30M	52.7	52.7	34.0	78.6	47.3	53.0
ProCLIP	30M	**63.4**	**68.0**	**34.1**	**86.8**	**55.7**	**61.6**

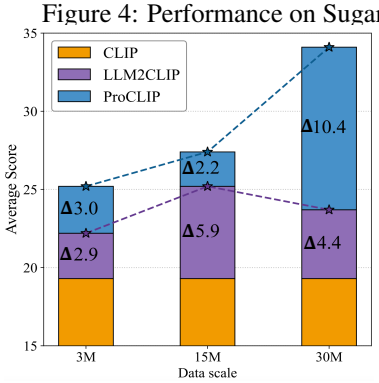


Figure 5: MMVP performance.

Table 4: Comparison with other methods across different model scales and LLM embedders.

Method	ViT Init	LLM Embedder	#Data	IN-1k	COCO I2T	COCO T2I	Flickr30k I2T	Flickr30k T2I
FLAME	random	Mistral-Nemo	3M	36.0	43.3	28.6	67.3	53.6
ShareLock	DINOv2 B/14	Llama3	3M	52.1	-	-	-	-
LIFT	random	NV-Embedv2	512M	43.6	34.6	36.0	69.1	72.9
LiT	CLIP B/16	Llama3-CC	3M	51.0	56.2	41.9	85.2	71.9
LLM2CLIP	CLIP B/16	Llama3-CC	3M	45.8	60.5	44.8	88.0	75.3
ProCLIP	CLIP B/16	Llama3-CC	3M	**54.8**	**61.7**	**46.8**	**89.4**	**77.6**
SAIL	DINOv2 L/14	NV-Embedv2	3M	54.0	45.4	32.9	-	-
LiT	CLIP L/14	Llama3-CC	3M	60.1	59.4	44.6	88.0	74.7
LLM2CLIP	CLIP L/14	Llama3-CC	3M	52.8	65.5	49.7	92.4	80.1
ProCLIP	CLIP L/14	NV-Embedv2	3M	61.4	64.8	51.7	91.9	**81.4**
ProCLIP	CLIP L/14	Llama3-CC	3M	**62.5**	**66.4**	**51.9**	**92.8**	81.1

trained on 30M data, it achieves 94.5% I2T R@1 on Flickr30k, surpassing LLM2CLIP by 1.4%. ProCLIP also shows clear improvements on several long-text benchmarks (e.g., DCI and Urban-1k). For instance, on the DCI dataset, ProCLIP (ViT-B/16, 30M) attains remarkable scores of 73.6% I2T and 76.9% T2I R@1. These results validate ProCLIP's versatility in handling both concise and complex textual descriptions.

Zero-Shot Classification. Tab. 2 summarizes the zero-shot classification results on 11 downstream tasks. Direct alignment methods like the baseline LLM2CLIP struggle to achieve optimal zero-shot generalization. In contrast, ProCLIP effectively addresses this issue, yielding consistent and significant absolute gains of 6.8%–13.5% over LLM2CLIP across all model architectures and data scales. Furthermore, when compared to the vanilla CLIP trained under the identical 30M data setting, ProCLIP demonstrates substantial superiority (e.g., achieving 59.0% vs. 48.9% on ViT-B/32, and 63.0% vs. 51.4% on ViT-B/16). This benefit mainly comes from our curriculum alignment strategy, where self-distillation mitigates representation collapse while maintaining robust cross-modal alignment. ProCLIP still lags behind Long-CLIP and FG-CLIP in zero-shot classification, though it is worth noting that FG-CLIP relies on a massively larger 1.6B dataset. Nevertheless, ProCLIP, similar to LLM2CLIP, aims at a broader CLIP extension, improving not only cross-modal retrieval but also multilingual and fine-grained understanding. It therefore offers a more comprehensive trade-off across diverse cross-modal capabilities.

Robustness and Multilingual Capabilities. Tab. 3 evaluates the robustness of our representations on the ViT-L/14 architecture. ProCLIP shows average gains of 8.4%–9.1% across different data scales in this setting. Most notably, on challenging out-of-distribution (OOD) benchmarks such as ImageNet-A and ImageNet-R, ProCLIP outperforms LLM2CLIP by over 8 percentage points. This highlights ProCLIP's superior capacity to handle distribution shifts and complex perturbations. Additionally, leveraging the semantic breadth of the LLM embedder, ProCLIP exhibits robust multilingual capabilities. As shown in Fig. 3, ProCLIP consistently outperforms LLM2CLIP on the XM3600 benchmark [40] across multiple languages (trained with 3M samples, using ViT-L base model). This superiority is primarily attributed to our progressive strategy, which effectively bridges the semantic gap between the original visual space and the vast, multilingual LLM space.

Fine-Grained Understanding. Fig. 4 demonstrates the significant potential of ProCLIP in complex compositional reasoning tasks, where all models are trained on 3M samples using a ViT-L base model. It surpasses the original CLIP baseline by an average margin of 11.3%–17.7% on SugarCrepe. These improvements demonstrate that anchoring an LLM embedder to the visual space vastly

Table 5: MLLM performance under 224^2 image resolution.

Method	SEED-Bench (image)	GQA	VizWiz	PoPE	TextVQA	MMBench	VQAv2
CLIP	65.3	62.0	44.0	85.7	54.2	65.5	77.4
ProCLIP (ViT-L-30M)	**66.4**	**62.3**	**44.7**	**85.9**	**54.5**	65.0	**78.1**

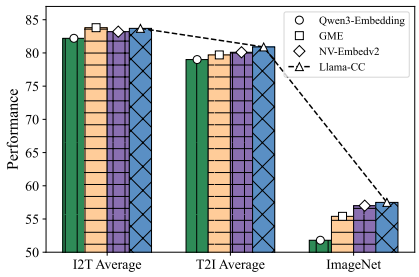


Figure 6: Ablation on LLM embedders.

Table 6: Ablation on different components.

Method	Stage 1 $\mathcal{L}_{ins}$	Stage 1 $\mathcal{L}_{struct}$	Stage 2 $\mathcal{L}_{info}$	Stage 2 $\mathcal{L}_{reg}$	IN-1k	I2T Avg.	T2I Avg.
CLIP					74.5	66.9	55.5
LLM2CLIP					52.8	82.0	79.1
ProCLIP	✓				58.9	69.3	59.4
	✓	✓			59.5	70.3	61.2
	✓	✓	✓		59.2	82.9	80.2
	✓	✓	✓	✓	**62.5**	**83.0**	**80.5**

Table 7: Ablation study on different base model architectures. All methods use 3M training samples.

Base Model	Method	IN-1k	I2T Avg.	T2I Avg.
CLIP ViT-L	LLM2CLIP	52.8	82.0	79.1
	ProCLIP	**62.5**	**83.0**	**80.5**
EVA02-CLIP ViT-L	LLM2CLIP	56.4	83.6	80.9
	ProCLIP	**66.8**	**84.2**	**82.2**
SigLIP2 SO/14@224	LLM2CLIP	63.8	84.6	81.9
	ProCLIP	**67.7**	**86.3**	**83.9**

Table 8: Ablation of $\mathcal{L}_{reg}$.

λ_{reg}	IN-1k	I2T Avg.	T2I Avg.
0	59.2	82.9	80.2
2e-4	60.3	82.8	80.3
4e-4	62.5	**83.0**	**80.5**
6e-4	**62.6**	82.6	79.8
1e-3	62.5	81.8	79.5

enhances fine-grained semantic discrimination. Fig. 5 further presents the fine-grained understanding performance on the challenging MMVP benchmark [42], trained with ViT-L base model. While LLM2CLIP improves over CLIP by 2.9%–5.9%, ProCLIP dramatically advances these results, achieving additional gains of up to 10.4% relative to the baseline.

Comparison with other LLM-augmented Methods. To rigorously prove the effectiveness of ProCLIP, we provide a comprehensive comparison with recent state-of-the-art LLM-augmented CLIP models, including FLAME [5], ShareLock [38], LIFT [52], SAIL [56], LiT [54], and LLM2CLIP [20]. As detailed in Tab. 4, under identical or lower training budgets, ProCLIP consistently achieves superior performance. Benefiting from representation inheritance and self-distillation, ProCLIP not only excels in ImageNet-1k classification but also significantly enhances general retrieval capabilities.

MLLM Performance. As shown in Tab. 5, when integrating the fine-tuned vision encoder into the MLLM (using LLaVA1.5-7B [31] framework), we observe performance improvements over CLIP on most benchmarks. This can be mainly attributed to the alignment with high-quality data, which enhances the semantic representation capability of the vision encoder. This also validates the flexibility of the ProCLIP paradigm, which supports both cross-modal retrieval via a dual-encoder architecture and the utilization of standalone vision encoder representations for downstream tasks.

4.3 Ablation and Analysis

Ablation of Different LLM-based Embedder. Fig. 6 compares several LLM embedders (Qwen3 [58], GME [57], NV-Embedv2 [25], and Llama3-CC [20]) paired with ViT-L/14 (trained with 15M samples). Among them, Llama3-CC achieves the best overall results in classification and retrieval. As for its superior classification performance, we hypothesize that this is primarily attributable to the incorporation of textual information from image-text pairs during its text embedding training phase. This pre-adaptation to the text distribution in image-text pairs facilitates subsequent cross-modal alignment, enabling the model to fit the data distribution more effectively and converge more rapidly. Notably, while retrieval performance remains relatively consistent across embedders, ImageNet classification accuracy varies significantly. This suggests that the alignment discrepancy between each LLM's feature space and the original CLIP space differs, leading to varying degrees of generalization degradation after training.

Table 9: Ablation on progressive training.

Method	IN-1k	I2T Avg.	T2I Avg.	Training Hours ↓
One stage	58.4	82.6	79.8	**1.5**
Two stage	**62.5**	**83.0**	**80.5**	1.7

Table 10: Performance and training efficiency comparison.

Method	IN-1k	I2T Avg.	T2I Avg.	Training Hours↓
Adapter	**62.5**	**83.0**	80.5	**1.7**
LoRA(r=16)	61.8	82.6	**80.6**	∼15

Ablation of Different Base Model. Tab. 7 shows ProCLIP excels across architectures, surpassing LLM2CLIP in classification and retrieval. The remaining classification gap relative to vanilla CLIP arises from the latter's overwhelming data scale, a dominant factor in zero-shot tasks.

Ablation of λ_{reg}. Tab. 8 presents the sensitivity analysis of the regularization weight for λ_{reg}. The results indicate that 4e-4 offers an optimal balance and is thus selected as our final configuration. Furthermore, the empirical evidence suggests that excessively strong regularization hampers alignment, leading to suboptimal results.

Ablation of Different Components. Tab. 6 details the ablation study of our components. First, instance semantic distillation achieves 58.9% ImageNet-1k zero-shot accuracy using text only, confirming the successful transfer of CLIP's textual priors to the adapter. Adding structural alignment loss further boosts both classification and retrieval by enabling the LLM to capture CLIP's global geometric structure beyond point-wise semantics. Subsequent image-text contrastive tuning significantly enhances retrieval but compromises classification due to image encoder overfitting. This is mitigated by self-distillation, which recovers classification accuracy from 59.2% to 62.5%. Ultimately, structured self-distillation yields the best results by stabilizing the image representation space and preserving pretrained knowledge, effectively balancing fine-tuning gains with generalization.

One stage vs Multi stage. As shown in Tab. 9, collapsing our two-stage strategy into a single stage significantly degrades both generalization and retrieval metrics. This confirms that our curriculum learning-inspired approach is essential for achieving superior cross-modal alignment.

Table 11: Ablation on adapter.

Adapter	IN-1k	I2T Avg.	T2I Avg.
MLP	**62.5**	**83.0**	80.5
GLU	61.8	82.6	**80.6**
Transformer	61.6	82.5	**80.6**

Ablation of adapter selection. To ensure a fair evaluation of ProCLIP, we maintain baseline consistency in adapter selection and fix all configurations at 67M parameters. Our ablation studies compare Transformer (1-layer), MLP (4-layer), and GLU (4-layer) adapters. Results in Tab. 11 indicate that the MLP architecture offers the optimal balance between computational efficiency and accuracy.

Offline Training vs. LoRA Fine-tuning. LLM embedders can be trained using PEFT methods (e.g., LoRA) or by extracting features offline and training only the adapter. As shown in Tab. 10, using LoRA requires active LLM computation during training, which incurs substantial computational overhead without providing additional performance gains. Therefore, in our final setup, we train only the adapter following the LLM embedder, keeping the setting consistent with the baseline.

5 Conclusion

This paper presents **ProCLIP**, a progressive vision-language alignment framework that integrates CLIP image encoders with LLM-based embedders. ProCLIP employs a two-stage strategy: first, anchoring the LLM embedder to CLIP's textual space via knowledge distillation to inherit semantic priors; second, performing cross-modal contrastive tuning with self-distillation to prevent catastrophic forgetting. To maintain feature-space consistency, we introduce instance-level semantic and global structural alignment constraints. Extensive evaluations across various scales and architectures validate the efficacy and generality of **ProCLIP**. We will release all model weights and code.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (62506229), Natural Science Foundation of Shanghai under 25ZR1402268, and Shanghai QiYuan Innovation Foundation.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[3] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024.

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[5] Anjia Cao, Xing Wei, and Zhiheng Ma. Flame: Frozen large language models enable data-efficient language-image pre-training. In *CVPR*, pages 4080–4090, 2025.

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pages 3558–3568, 2021.

[7] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, pages 370–387. Springer, 2024.

[8] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*, pages 4113–4123, 2024.

[9] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022.

[10] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, pages 14084–14093, 2022.

[11] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[12] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, pages 540–557. Springer, 2022.

[13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[14] Tiancheng Gu, Kaicheng Yang, Xiang An, Ziyong Feng, Dongnan Liu, Weidong Cai, and Jiankang Deng. Rwkv-clip: a robust vision-language representation learner. *arXiv preprint arXiv:2406.06973*, 2024.

[15] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[16] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021.

[17] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021.

[18] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

[19] Xiaoxing Hu, Kaicheng Yang, Jun Wang, Haoran Xu, Ziyong Feng, and Yupei Wang. Decoupled global-local alignment for improving compositional understanding. *arXiv preprint arXiv:2504.16801*, 2025.

[20] Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Usman Naseem, Chunyu Wang, Qi Dai, Xiyang Dai, Dongdong Chen, et al. Llm2clip: Powerful language model unlocks richer cross-modality representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 5131–5139, 2026.

[21] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[22] Prannay Kaul, Weidi Xie, and Andrew Zisserman. Multi-modal classifiers for open-vocabulary object detection. In *ICML*, pages 15946–15969. PMLR, 2023.

[23] Gahyeon Kim, Sohee Kim, and Seokju Lee. Aapl: Adding attributes to prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1572–1582, June 2024.

[24] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *ECCV*, pages 143–160. Springer, 2024.

[25] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

[26] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[27] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

[28] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[30] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024.

[32] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[33] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

[34] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. Docci: Descriptions of connected and contrasting images. In *ECCV*, pages 291–309. Springer, 2024.

[35] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[37] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400. PMLR, 2019.

[38] Jona Ruthardt, Gertjan J Burghouts, Serge Belongie, and Yuki M Asano. Do better language models have crisper vision? 2024.

[39] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[40] Ashish V Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *arXiv preprint arXiv:2205.12522*, 2022.

[41] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[42] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, pages 9568–9578, 2024.

[43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[44] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[45] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26700–26709, 2024.

[46] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NIPS*, 32, 2019.

[47] Zihao Wang, Wei Liu, Qian He, Xinglong Wu, and Zili Yi. Clip-gen: Language-free training of a text-to-image generator with clip. *arXiv preprint arXiv:2203.00386*, 2022.

[48] Wei Wu, Kecheng Zheng, Shuailei Ma, Fan Lu, Yuxin Guo, Yifei Zhang, Wei Chen, Qingpei Guo, Yujun Shen, and Zheng-Jun Zha. Lotlip: Improving language-image pre-training for long text understanding. *NIPS*, 37:64996–65019, 2024.

[49] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *CVPR*, pages 7031–7040, 2023.

[50] Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment. *arXiv preprint arXiv:2505.05071*, 2025.

[51] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *ECCV*, pages 736–753. Springer, 2022.

[52] Jingfeng Yang, Ziyang Wu, Yue Zhao, and Yi Ma. Language-image alignment with fixed text encoders. *arXiv preprint arXiv:2506.04209*, 2025.

[53] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *ICCV*, pages 2922–2931, 2023.

[54] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133, 2022.

[55] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *ECCV*, pages 310–325. Springer, 2024.

[56] Le Zhang, Qian Yang, and Aishwarya Agrawal. Assessing and learning alignment of unimodal vision and language models. In *CVPR*, pages 14604–14614, 2025.

[57] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024.

[58] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

[59] Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *European Conference on Computer Vision*, pages 73–90. Springer, 2024.

[60] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[61] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European conference on computer vision*, pages 696–712. Springer, 2022.

[62] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022.

[63] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly explain the research background and motivation of ProCLIP, as well as the overall rationale behind its methodological design. They also show that ProCLIP significantly outperforms the baseline across multiple aspects and establishes a new state-of-the-art in the field of LLM-enhanced CLIP.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Appendix F, Future Work, we clearly discuss the current limitations of ProCLIP, including its increased computational overhead. We also point out several possible directions for addressing this issue.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: In Appendix C, Theoretical Analysis, we provide a thorough theoretical discussion of the design of ProCLIP, along with complete assumptions and proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In Appendix B, Experiment Details, we present all hyperparameters used in this work. We also clearly describe the datasets and base models used in the main paper.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Our codebase is entirely built upon the publicly available LLM2CLIP codebase. We will also fully open-source our own code in the near future, including the model weights.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: In Appendix B, Experiment Details, we clearly describe the training and evaluation details. The datasets used in this work are also clearly described in the main paper.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We did not report a statistical significance analysis due to the high computational cost of the experiments. However, it is worth noting that the training datasets, training code, and random seeds are fixed, which helps ensure the reproducibility of our experiments.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Appendix B, Experiment Details, we clearly state that the experiments were conducted using 8 $\times$H100 GPUs.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and confirm that the research conforms to it. The paper preserves submission anonymity, uses existing research assets with proper citation, and does not involve human subjects or private sensitive data

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The topic studied in this paper is purely within the field of machine learning and does not have any obvious additional positive or negative societal impact. Its primary value lies in the machine learning community, as this work provides a superior algorithm.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work focuses on CLIP-style vision-language representation learning rather than releasing high-risk generative models or scraped datasets. The released code and model weights are intended for research purposes, so this question is not directly applicable.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credit the original creators of the codebase, datasets, and base models used in this paper. We also follow their corresponding licenses and terms of use.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The newly introduced assets, including code and model weights, will be released with accompanying documentation, usage instructions, and evaluation details.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing experiments or research with human subjects, so participant instructions, screenshots, and compensation details are not applicable.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve human subjects or study participants, so potential participant risks and IRB approval are not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper describes how LLMs are used as a core component of ProCLIP to enhance CLIP-style vision-language representation learning.

The appendix is organized as follows:

- **Appendix A: Methodological Discussion.**
  We discuss the differences between ProCLIP and representative CLIP extension methods, including Long-CLIP, FG-CLIP, and LLM2CLIP.
- **Appendix B: Experimental Details.**
  We provide detailed training settings, evaluation protocols, benchmark descriptions, dataset statistics, and prompt templates.
- **Appendix C: Theoretical Analysis.**
  We theoretically analyze why progressive alignment can incorporate LLM-based textual representations into CLIP while preserving its pre-trained generalization ability.
- **Appendix D: More Results and Analysis.**
  We report additional quantitative results, including fine-grained understanding, multilingual retrieval, linear probing, robustness, data/model scaling, open-vocabulary semantic segmentation, and efficiency analysis.
- **Appendix E: Visualization.**
  We present qualitative visualizations to demonstrate the advantages of ProCLIP in long-text cross-modal retrieval.
- **Appendix F: Future Work.**
  We discuss several potential directions for improving ProCLIP, including training efficiency, fine-grained alignment, and architecture extension.

A Methodological Discussion

Although several recent methods have attempted to enhance CLIP from different perspectives, ProCLIP differs from them in both motivation and capability coverage. Long-CLIP mainly focuses on extending the text length supported by CLIP, while FG-CLIP emphasizes fine-grained visual-textual alignment and recognition. These methods are effective in preserving or even improving zero-shot classification performance, but their improvements are relatively specialized. In contrast, ProCLIP aims to provide a more comprehensive extension of CLIP by aligning it with stronger LLM-based textual representations. Beyond maintaining competitive recognition ability, ProCLIP further improves cross-modal retrieval, multilingual understanding, long-text comprehension, and fine-grained semantic alignment.

Another important distinction lies in training efficiency. Compared with LLM2CLIP, which directly aligns CLIP with LLM representations and may severely damage the original visual generalization ability, ProCLIP introduces a curriculum alignment strategy with self-distillation regularization. This design enables stable cross-modal alignment while preserving the pre-trained knowledge of CLIP. As a result, ProCLIP achieves a better trade-off between enhancing advanced language-side capabilities and retaining the generalization ability of the original vision-language model. Therefore, ProCLIP should not be viewed as a method solely optimized for zero-shot classification, but rather as a general-purpose CLIP enhancement framework for diverse cross-modal scenarios.

Table 12: Comparison between ProCLIP and representative CLIP extension methods. ProCLIP provides a more comprehensive enhancement across multilingual, long-text, fine-grained scenarios while maintaining data-efficient training.

Method	Training Data	Training Method	Long Text	Multilingual	Fine-grained	Zero-shot Recognition
CLIP	Large-scale	Pre-train	Limited	Limited	Moderate	Strong
Long-CLIP	Small-scale	Post-train	✓	Limited	Moderate	Strong
FG-CLIP	Large-scale	Continue Pre-train	Limited	Limited	✓	Strong
LLM2CLIP	Small-scale	Post-train	✓	✓	✓	Moderate
ProCLIP	Small-scale	Post-train	✓	✓	✓	Strong

B Experiment Details

B.1 Training Details

Details of the hyperparameter configurations used for two-stage training of ProCLIP are presented in Tab. 13. Under the default setting, our MLP layers are consistent with the baseline LLM2CLIP, both consisting of four linear layers.

Table 13: Detailed hyperparameters for training ProCLIP.

Hyperparameters of stage1	
Batch size	1024 (8×128)
Optimizer	AdamW
Weight decay	0.05
Adam β	(0.9,0.98)
Adam ϵ	1e-6
Learning rate	1e-5
Learning rate schedule	cosine decay
Epochs	4
Training GPUs	8×H100

Hyperparameters of stage2	
Batch size	4096 (8×512)
Optimizer	AdamW
Weight decay	0.05
Adam β	(0.9,0.98)
Adam ϵ	1e-6
Learning rate	1e-5
Learning rate schedule	cosine decay
Ema α	0.999
λ_{reg}	0.0004
Epochs	4
Training GPUs	8×H100

B.2 Evaluation Details

For the LLM2CLIP baseline, we reproduce all results with the official LLM2CLIP codebase, since LLM2CLIP does not provide a complete set of checkpoints from ViT-B/32 to ViT-L/14 and released checkpoints may be affected by different hardware and training environments. For a fair comparison, all LLM2CLIP results reported in the main paper are reproduced by ourselves using the same training data as ProCLIP and training for 4 epochs. We will release all checkpoints in the future, including both our reproduced baselines and ProCLIP.

For the ShareGPT4V benchmark, since no fixed evaluation set is provided, we randomly sample 1,000 images from COCO2017, following the common practice in prior works.

For comparison methods, we mainly compare ProCLIP with representative works, including CLIP [36], LLM2CLIP [20], DreamLIP [59], FG-CLIP [50], and Long-CLIP [55]. Among them, LLM2CLIP serves as our primary baseline. Notably, LLM2CLIP received the Outstanding Paper Award at AAAI 2026. Building upon it, we propose ProCLIP, a more comprehensive and effective method.

B.3 Details of Benchmarks.

Zero-Shot Classification & Linear Probe. Following the previous works [53, 14], we evaluate the zero-shot classification and linear probe performance of the models on 11 datasets. The detailed information about these datasets and the prompt used in zero-shot classification are presented in Tab. 14 and Tab. 17.

Robustness. We evaluated the robustness of our model on five out-of-distribution datasets, including ImageNet-v2 [37], ImageNet-A [17], ImageNet-O [17], ImageNet-R [16], and ImageNet-Sketch [46].

Cross-Modal Retrieval. Following the previous works [20, 5], we evaluate the cross-modal retrieval performance of the models on 6 datasets: Flickr30k [35], COCO [29], ShareGPT4V [7], Urban-1k [55], DOCCI [34], and DCI [45]. The details information about these dataset are present on Tab. 15.

Multilingual Retrieval. We evaluated the multilingual capabilities of our model on XM3600 [40]. XM3600 contains 3,600 images covering a total of 36 languages, including Arabic (ar), Bengali(bn), Chinese-Simplified (zh), Croatian (hr), Czech (cs), Danish (da), Dutch (nl), English (en),Farsi (fa), Filipino (fil), Finnish (fi), French (fr), German (de), Greek (el), Hebrew (he), Hindi (hi), Hungarian (hu), Indonesian (id), Italian (it), Japanese (ja), Korean (ko),Maori(mi), Norwegian (no), Persian (fa),

Table 14: List of zero-shot datasets with the data distribution and evaluation metrics.

Dataset	Classes	Train size	Test size	Evaluation metric
Food101	101	75,750	25,250	accuracy
CIFAR10	10	50,000	10,000	accuracy
CIFAR100	100	50,000	10,000	accuracy
SUN397	397	19,850	19,850	accuracy
Cars	196	8,144	8,041	accuracy
Aircraft	100	6,667	3,333	mean per class
DTD	47	3,760	1,880	accuracy
Pets	37	3,680	3,669	mean per class
Caltech101	101	3,000	5,677	mean-per-class
Flowers	102	2,040	6,149	mean per class
ImageNet	1000	1,281,167	50,000	accuracy

Table 15: Zero-shot image-text retrieval evaluation settings.

Dataset	**Test Images**	**Evaluation Protocol**	**Text type**
MSCOCO	5,000	Image-to-Text & Text-to-Image	short
Flickr30k	1,000	Image-to-Text & Text-to-Image	short
ShareGPT4V	1000	Image-to-Text & Text-to-Image	long
Urban-1k	1000	Image-to-Text & Text-to-Image	long
DOCCI	5000	Image-to-Text & Text-to-Image	long
DCI	7805	Image-to-Text & Text-to-Image	long

Table 16: Number of hard negative captions of all types in SugarCrepe.

	REPLACE			**SWAP**		**ADD**	
	Object	Attribute	Relation	Object	Attribute	Object	Attribute
# negative captions	1652	788	1406	246	666	2062	692

Polish (pl), Portuguese (pt), Romanian (ro), Russian (ru), Spanish (es), Swedish (sv), Swahili(sw), Thai (th), Turkish (tr), Telugu (te), Ukrainian (uk), and Vietnamese (vi).

Fine-Grained Understanding. We evaluated the fine-grained understanding capability of the VLM on MMVP-VLM [42] and SugarCrepe. Here, fine-grained understanding is conceptually similar to compositional understanding, focusing primarily on the model's ability to distinguish subtle textual differences in cross-modal matching. MMVP-VLM consists of 150 samples in total, testing 9 patterns:

- **Orientation and Direction**: Questions about the direction something is facing or moving, such as the direction the dog or duck is facing, or the orientation of the school bus.
- **Presence of Specific Features**: Questions that focus on the existence or non-existence of certain elements or features in the image.
- **State and Condition**: Questions that pertain to the state or condition of an object, such as whether a flag is blowing in the wind or if the ground is wet.
- **Quantity and Count**: Questions about the number of objects or features present in the image.
- **Positional and Relational Context**: This aspect refers to the model's ability to understand the position and relationship of objects or elements within an image in relation to each other and their surroundings.
- **Color and Appearance**: Questions regarding the color of certain objects or elements.
- **Structural and Physical Characteristics**: This category involves the model's ability to identify and analyze the physical attributes and structural features of objects in an image.
- **Text**: Questions related to text or symbols present in the image.

Table 17: Full list of prompts to evaluate the performance of zero-shot classification on 11 visual recognition datasets.

CIFAR 10 & CIFAR 100			
a photo of a {label}.	a blurry photo of a {label}.	a black and white photo of a {label}.	a low contrast photo of a {label}.
a high contrast photo of a {label}.	a bad photo of a {label}.	a good photo of a {label}.	a photo of a small {label}.
a photo of a big {label}.	a photo of the {label}.	a blurry photo of the {label}.	a black and white photo of the {label}.
a low contrast photo of the {label}.	a high contrast photo of the {label}.	a bad photo of the {label}.	a good photo of the {label}.
a photo of the small {label}.	a photo of the big {label}.		
Food101			
a photo of {label}, a type of food.			
Caltech101			
a photo of a {label}.	a painting of a {label}.	a plastic {label}.	a sculpture of a {label}.
a sketch of a {label}.	a tattoo of a {label}.	a toy {label}.	a rendition of a {label}.
a embroidered {label}.	a cartoon {label}.	a {label} in a video game.	a plushie {label}.
an origami {label}.	art of a {label}.	graffiti of a {label}.	a drawing of a {label}.
a doodle of a {label}.	a photo of the {label}.	a painting of the {label}.	the plastic {label}.
a sculpture of the {label}.	a sketch of the {label}.	a tattoo of the {label}.	the toy {label}.
a rendition of the {label}.	the embroidered {label}.	the cartoon {label}.	the {label} in a video game.
the plushie {label}.	the origami {label}.	art of the {label}.	graffiti of the {label}.
a drawing of the {label}.	a doodle of the {label}.		
Stanford Cars			
a photo of a {label}.	a photo of the {label}.	a photo of my {label}.	i love my {label}!
a photo of my dirty {label}.	a photo of my clean {label}.	a photo of my new {label}.	a photo of my old {label}.
DTD			
a photo of a {label} texture.	a photo of a {label} pattern.	a photo of a {label} thing.	a photo of a {label} object.
a photo of the {label} texture.	a photo of the {label} pattern.	a photo of the {label} thing.	a photo of the {label} object.
FGVC Aircraft			
a photo of a {label}, a type of aircraft.	a photo of the {label}, a type of aircraft.		
Flowers102			
a photo of a {label}, a type of flower.			
Pets			
a photo of a {label}, a type of pet.			
SUN39			
a photo of a {label}.	a photo of the {label}.		
ImageNet			
a bad photo of a {label}.	a photo of many {label}.	a sculpture of a {label}.	a photo of the hard to see {label}.
a low resolution photo of the {label}.	a rendering of a {label}.	graffiti of a {label}.	a bad photo of the {label}.
a cropped photo of the {label}.	a tattoo of a {label}.	the embroidered {label}.	a photo of a hard to see {label}.
a bright photo of a {label}.	a photo of a clean {label}.	a photo of a dirty {label}.	a dark photo of the {label}.
a drawing of a {label}.	a photo of my {label}.	the plastic {label}.	a photo of the cool {label}.
a close-up photo of a {label}.	a black and white photo of the {label}.	a painting of the {label}.	a painting of a {label}.
a pixelated photo of the {label}.	a sculpture of the {label}.	a bright photo of the {label}.	a cropped photo of a {label}.
a plastic {label}.	a photo of the dirty {label}.	a jpeg corrupted photo of a {label}.	a blurry photo of the {label}.
a photo of the {label}.	a good photo of the {label}.	a rendering of the {label}.	a {label} in a video game.
a photo of one {label}.	a doodle of a {label}.	a close-up photo of the {label}.	a photo of a {label}.
the origami {label}.	the {label} in a video game.	a sketch of a {label}.	a doodle of the {label}.
an origami {label}.	a low resolution photo of a {label}.	the toy {label}.	a rendition of the {label}.
a photo of the clean {label}.	a photo of a large {label}.	a rendition of a {label}.	a photo of a nice {label}.
a photo of a weird {label}.	a blurry photo of a {label}.	a cartoon {label}.	art of a {label}.
a sketch of the {label}.	a embroidered {label}.	a pixelated photo of a {label}.	itap of the {label}.
a jpeg corrupted photo of the {label}.	a good photo of a {label}.	a plushie {label}.	a photo of the nice {label}.
a photo of the small {label}.	a photo of the weird {label}.	the cartoon {label}.	art of the {label}.
a drawing of the {label}.	a photo of the large {label}.	a black and white photo of a {label}.	the plushie {label}.
a dark photo of a {label}.	itap of a {label}.	graffiti of the {label}.	a toy {label}.
itap of my {label}.	a photo of a cool {label}.	a photo of a small {label}.	a tattoo of the {label}.

- 📷 **Viewpoint and Perspective**: Questions concerning the perspective from which the photo was taken.

SugarCrepe contains three types of negative samples—Add, Replace, and Swap—designed to probe whether vision-language models (VLMs) can recognize text that differs only in subtle structural changes yet conveys a completely different overall meaning in image-text matching tasks.

MLLM benchmarks. We further integrate the fine-tuned vision encoder into LLaVA and evaluate its performance on several MLLM downstream benchmarks, including SEED-Bench [26], GQA [21], VizWiz [15], PoPE [28], TextVQA [39], MMBench [32], and VQAv2 [13].

Open-Vocabulary Semantic Segmentation. We further evaluate the model performance on a pixel-level task, namely open-vocabulary semantic segmentation, across three major datasets: ADE20K [60], COCO-Stuff164k [4], and VOC20 [11]. We adopt the same evaluation protocol as MaskCLIP [61].

C Theoretical Analysis

In this section, we provide a theoretical interpretation of why the proposed progressive alignment framework can effectively incorporate LLM representations into CLIP while preserving the pre-trained generalization ability of CLIP. Our analysis focuses on four key questions: (1) why directly

aligning an uncalibrated LLM space with the visual space may cause representation degradation; (2) how instance-level and structural distillation jointly provide local and global semantic constraints across the two training stages; (3) why short-text distillation can benefit long-text understanding; and (4) why multilingual capability can emerge even without multilingual image-text pairs during training.

Let $\mathcal{V}$ denote the visual embedding space and $\mathcal{T}$ denote the CLIP text embedding space. For each training instance, let $v_i \in \mathcal{V}$ be the visual embedding, $e_i \in \mathcal{T}$ be the original CLIP text embedding, and $t_i^* \in \mathcal{T}$ be the projected LLM embedding produced by the adapter. We assume that all embeddings are normalized, as commonly adopted in CLIP-like contrastive learning.

C.1 Why Direct Alignment May Degrade CLIP Representations

Existing direct alignment methods, such as LLM2CLIP, typically optimize the contrastive objective between visual embeddings and projected LLM embeddings without first ensuring that the LLM space is compatible with the original CLIP text space. In this case, the training process needs to simultaneously discover a proper language-to-CLIP transformation and perform cross-modal alignment. This may introduce unstable gradients, especially when the projected LLM embeddings are initially far away from the CLIP semantic manifold.

Let $\mathcal{L}_{\text{info}}$ denote the InfoNCE loss. For a positive image-text pair (v_i, t_i^*), the contrastive gradient encourages the visual embedding v_i to move toward the projected LLM embedding t_i^* while pushing it away from other text embeddings. However, when t_i^* has not yet inherited the geometry of the original CLIP text space, this update direction may deviate from the well-structured visual-textual alignment learned by CLIP. Consequently, the visual encoder may be forced to fit an uncalibrated or geometrically incompatible target space, which can damage its pre-trained zero-shot generalization ability.

Formally, suppose the original CLIP model has learned a well-aligned cross-modal structure satisfying

$$\mathbb{E}\left[d(v_i^{(0)}, e_i)\right] \leq \epsilon_C, \tag{8}$$

where $v_i^{(0)}$ denotes the visual embedding extracted by the original CLIP encoder and $d(\cdot, \cdot)$ is a metric distance. In direct alignment, the projected LLM embedding t_i^* is optimized mainly through the image-text contrastive loss. Before the adapter becomes well calibrated, the discrepancy

$$d(v_i^{(0)}, t_i^*) \tag{9}$$

is not explicitly controlled. Therefore, the gradient of $\mathcal{L}_{\text{info}}$ may drive the visual representation away from its original CLIP neighborhood.

This explains why direct alignment may lead to representation degradation: the visual encoder is updated toward an LLM-induced target space that has not yet been aligned with the CLIP text manifold. As a result, the optimization may partially overwrite the original CLIP representation rather than merely refining it. This phenomenon is particularly harmful under limited training data, where the model can easily overfit to the alignment data and lose the broad zero-shot transfer ability acquired during CLIP pre-training.

C.2 Dual Distillation as Inheritance and Preservation

ProCLIP avoids the above issue through a progressive alignment strategy. Importantly, both the instance-level distillation loss $\mathcal{L}_{\text{ins}}$ and the structural distillation loss $\mathcal{L}_{\text{struct}}$ are used in both stages, but they play different roles at different stages.

In Stage I, the two distillation losses serve as inheritance constraints. The instance-level distillation loss provides point-wise semantic anchoring by pulling each projected LLM embedding t_i^* toward its corresponding CLIP text embedding e_i. The structural distillation loss provides global geometric anchoring by preserving the pairwise semantic relations among CLIP text embeddings. Together, they calibrate the LLM representation space into a CLIP-compatible text manifold before cross-modal contrastive learning.

In Stage II, the same two losses serve as preservation constraints. While the contrastive loss refines the visual-language alignment, $\mathcal{L}_{\text{ins}}$ restricts instance-level representation drift and $\mathcal{L}_{\text{struct}}$ prevents

the global semantic topology from being distorted. Therefore, instance distillation controls local semantic consistency, whereas structural distillation controls global geometric consistency throughout the entire progressive alignment process.

C.3 Instance Distillation Provides Point-wise Semantic Anchoring

We first analyze the role of instance-level distillation. In Stage I, $\mathcal{L}_{\text{ins}}$ directly constrains the projected LLM embedding to approach the corresponding CLIP text embedding. This provides a bounded initialization for the subsequent cross-modal contrastive learning.

Theorem 1 (Bounded Initial Cross-Modal Discrepancy)**.** *Assume that the embedding space is equipped with a metric distance $d(\cdot,\cdot)$. Let ϵ_C denote the original CLIP cross-modal alignment error, such that*

$$\mathbb{E}\left[d(v_i^{(0)}, e_i)\right] \leq \epsilon_C. \tag{10}$$

If the instance-level distillation in Stage I learns an adapter such that

$$\mathbb{E}\left[d(t_i^*, e_i)\right] \leq \epsilon_A, \tag{11}$$

then the expected discrepancy between the original visual embedding and the projected LLM embedding before Stage II is bounded by

$$\mathbb{E}\left[d(v_i^{(0)}, t_i^*)\right] \leq \epsilon_C + \epsilon_A. \tag{12}$$

Proof. By the triangle inequality, for each image-text pair (I_i, T_i), we have

$$d(v_i^{(0)}, t_i^*) \leq d(v_i^{(0)}, e_i) + d(e_i, t_i^*). \tag{13}$$

Taking the expectation over the data distribution gives

$$\mathbb{E}\left[d(v_i^{(0)}, t_i^*)\right] \leq \mathbb{E}\left[d(v_i^{(0)}, e_i)\right] + \mathbb{E}\left[d(e_i, t_i^*)\right] \leq \epsilon_C + \epsilon_A. \tag{14}$$

□

This result shows that instance-level distillation provides an explicitly bounded initialization for Stage II. In contrast, without such point-wise anchoring, the distance between the projected LLM embedding and the CLIP visual embedding is not controlled by the original CLIP alignment structure and can be as large as the diameter of the representation space. Therefore, $\mathcal{L}_{\text{ins}}$ reduces the initial cross-modal gap and stabilizes the subsequent contrastive optimization.

In Stage II, $\mathcal{L}_{\text{ins}}$ further acts as an instance-level preservation constraint. It prevents each trainable representation from drifting excessively away from its CLIP-compatible reference. Thus, the role of instance distillation is two-fold: it first anchors the projected LLM embeddings to the CLIP text space, and then preserves local semantic consistency during contrastive refinement.

C.4 Structural Distillation Preserves Semantic Topology

Instance-level distillation aligns each projected LLM embedding with its corresponding CLIP text embedding. However, point-wise matching alone does not explicitly ensure that the relative semantic structure among different samples is preserved. To address this, ProCLIP introduces a structural distillation loss that encourages the projected LLM embeddings to inherit the pairwise geometry of the original CLIP text embeddings.

Let $\mathcal{E} = \{e_i\}_{i=1}^N$ denote the empirical CLIP text embedding set, and $\mathcal{T}^* = \{t_i^*\}_{i=1}^N$ denote the corresponding projected LLM embedding set. The structural distillation objective can be formulated as

$$\mathcal{L}_{\text{struct}} = \sum_{i<j} \left| d(t_i^*, t_j^*) - d(e_i, e_j) \right|. \tag{15}$$

Proposition 1 (Empirical Semantic Topology Preservation)**.** *If the structural alignment error satisfies*

$$\left| d(t_i^*, t_j^*) - d(e_i, e_j) \right| \leq \delta, \quad \forall i, j, \tag{16}$$

then the projected LLM embedding set $\mathcal{T}^*$ forms a δ-approximate isometry of the empirical CLIP text embedding set $\mathcal{E}$.

Proof. The definition of additive approximate isometry requires that pairwise distances are preserved up to an additive error δ. Since for all pairs (i, j),

$$\left|d(t_i^*, t_j^*) - d(e_i, e_j)\right| \leq \delta, \tag{17}$$

the pairwise geometry of $\mathcal{T}^*$ approximates that of $\mathcal{E}$ within error δ. Therefore, the projected LLM embeddings preserve the empirical semantic topology of the original CLIP text space. □

This structural constraint is important because CLIP's zero-shot ability relies not only on image-text instance matching, but also on the global organization of semantic concepts. For example, semantically related concepts such as "dog" and "cat" should remain closer than unrelated concepts such as "dog" and "airplane". By preserving the pairwise structure of the original CLIP text space, $\mathcal{L}_{\text{struct}}$ prevents the adapter from learning only isolated point-wise correspondences and encourages the projected LLM space to inherit the global semantic geometry of CLIP.

In Stage II, structural distillation further prevents topology distortion during visual-language contrastive tuning. Even when the contrastive loss updates the representations to improve image-text matching, $\mathcal{L}_{\text{struct}}$ constrains the relative organization among samples. Therefore, the model can acquire improved LLM-based language understanding while maintaining the semantic structure of the original CLIP manifold.

C.5 Why Short-Text Distillation Benefits Long-Text Understanding

A natural question is why distillation from short CLIP-style texts can improve the final model's ability to understand longer and richer descriptions. The key reason is that Stage I does not aim to teach the LLM new language knowledge. Instead, it learns a geometric interface that maps the LLM representation space into the CLIP-compatible embedding space.

Let $g_\phi(\cdot)$ denote the adapter that projects an LLM representation into the CLIP text space. The LLM encoder itself already contains rich compositional and contextual knowledge learned from large-scale language pre-training. For a short text s_i, Stage I minimizes the discrepancy between $g_\phi(\mathrm{LLM}(s_i))$ and the original CLIP text embedding e_i. This process teaches the adapter how to translate LLM representations into the CLIP-compatible coordinate system.

The important point is that the learned mapping g_ϕ is shared across inputs. Once it learns how to align the local semantic directions of the LLM space with those of the CLIP text space using short texts, the same mapping can be applied to longer texts. Long texts can be regarded as compositions of multiple semantic units, such as objects, attributes, relations, actions, and contextual descriptions. If the adapter preserves the semantic directions of these basic units, then the projected representation of a long text can still lie in a CLIP-compatible region.

We can illustrate this with a simple compositional view. Suppose a long-text representation can be approximately decomposed into several semantic components:

$$h_{\text{long}} \approx \sum_{k=1}^{K} \alpha_k h_k, \tag{18}$$

where h_k denotes the LLM representation of a semantic unit and α_k denotes its contextual weight. If the adapter is locally smooth around these semantic components, then

$$g_\phi(h_{\text{long}}) \approx g_\phi\left(\sum_{k=1}^{K} \alpha_k h_k\right). \tag{19}$$

For a linear adapter, this further becomes

$$g_\phi(h_{\text{long}}) \approx \sum_{k=1}^{K} \alpha_k g_\phi(h_k). \tag{20}$$

Even for a nonlinear adapter, this approximation holds locally when the adapter is smooth. Therefore, short-text distillation can align the basic semantic directions, while the LLM provides the compositional capability needed for long-text understanding.

Instance-level and structural distillation play complementary roles in this transfer. The instance-level loss anchors individual short-text semantics to the CLIP text space, while the structural loss preserves the relative relations among semantic concepts. Since long texts are often formed by composing multiple concepts and relations, preserving both local semantic anchors and global semantic topology makes the projected long-text representation more likely to remain meaningful in the CLIP-compatible space.

This explains why short-text supervision does not limit ProCLIP to short-text semantics. Instead, it calibrates the LLM representation into the CLIP embedding space, allowing the model to exploit the LLM's inherent long-context and compositional understanding ability during inference. In other words, distillation learns the interface, while the LLM supplies the linguistic expressiveness.

C.6 Why Multilingual Ability Emerges without Multilingual Image-Text Data

Another important property of ProCLIP is that it can improve multilingual understanding even though the training process does not require multilingual image-text pairs. This can be explained by separating language knowledge from visual alignment.

The original CLIP text encoder is usually limited in its language coverage and often performs best on English prompts. In contrast, modern LLMs encode different languages into a relatively shared semantic space through large-scale multilingual or cross-lingual pre-training. For semantically equivalent texts in different languages, their LLM representations tend to be closer than those produced by a monolingual CLIP text encoder. Let x^{en} and x^{ml} denote an English text and its multilingual counterpart with similar meaning. The LLM encoder naturally encourages

$$\text{LLM}(x^{\text{en}}) \approx \text{LLM}(x^{\text{ml}}) \tag{21}$$

in the semantic representation space.

During training, ProCLIP learns a language-agnostic adapter g_ϕ that maps LLM semantic representations into the CLIP-compatible space. Importantly, this adapter is not tied to a specific language tokenization pattern, but operates on the LLM representation. Therefore, if two texts in different languages are close in the LLM semantic space, their projected embeddings will also be close after the adapter, provided that g_ϕ is smooth:

$$\left\| g_\phi(\text{LLM}(x^{\text{en}})) - g_\phi(\text{LLM}(x^{\text{ml}})) \right\| \leq L_g \left\| \text{LLM}(x^{\text{en}}) - \text{LLM}(x^{\text{ml}}) \right\|, \tag{22}$$

where L_g is the Lipschitz constant of the adapter.

This inequality shows that the cross-lingual semantic proximity encoded by the LLM can be preserved after projection. As a result, although ProCLIP may use only image-text alignment data in a single language, the learned adapter can transfer the visual alignment to other languages through the multilingual semantic space already embedded in the LLM.

The two distillation losses further support this transfer. Instance-level distillation maps the LLM semantic representation into the CLIP-compatible space, while structural distillation preserves the relative semantic geometry among concepts. If multilingual expressions with similar meanings are already close in the LLM space, then the learned adapter can project them into similar CLIP-compatible regions. Therefore, the multilingual ability of ProCLIP does not come from multilingual image-text supervision. Instead, it comes from the combination of two factors: (1) the LLM provides a multilingual semantic representation space, and (2) ProCLIP learns a CLIP-compatible projection from this semantic space to the visual-language space. In this sense, ProCLIP acts as a bridge that connects the multilingual knowledge of the LLM with the visual grounding ability of CLIP.

C.7 Stage-II Distillation Mitigates Representation Drift

During Stage II, ProCLIP optimizes the contrastive objective to refine cross-modal alignment. However, if this objective is used alone, both the visual representation and the projected LLM representation may drift away from the original CLIP semantic space. To prevent this, ProCLIP continues to use instance-level and structural distillation losses as preservation constraints.

Let z_i denote a trainable representation during Stage II, and let $z_i^{(0)}$ denote the corresponding CLIP-compatible reference representation. The instance-level distillation loss penalizes local deviations:

$$\mathcal{L}_{\text{ins}} = \sum_i d(z_i, z_i^{(0)}). \tag{23}$$

This term prevents each individual sample from drifting excessively.

The structural distillation loss further preserves the global semantic geometry:

$$\mathcal{L}_{\text{struct}} = \sum_{i<j} \left| d(z_i, z_j) - d(z_i^{(0)}, z_j^{(0)}) \right|. \tag{24}$$

Unlike instance-level regularization, which constrains each point independently, structural distillation constrains the relative organization among samples. Therefore, even when contrastive learning updates the representation for better image-text matching, the global semantic topology inherited from CLIP is still preserved.

We further analyze the effect of instance-level drift regularization. For simplicity, consider the induced optimization in the visual representation space. Let $v_i^{(0)}$ denote the visual embedding extracted by the original frozen CLIP encoder, and let v_i denote the visual embedding during training. The Stage-II objective for one instance can be written as

$$\mathcal{L}_{\text{tune}}(v_i) = \mathcal{L}_{\text{info}}(v_i, t_i^*) + \lambda_{\text{ins}} \left\| v_i - v_i^{(0)} \right\|_2^2. \tag{25}$$

Theorem 2 (Bounded Tangential Representation Drift)**.** *Assume that the embeddings are normalized and the temperature of InfoNCE is fixed. Then the gradient norm of the InfoNCE loss is bounded by a constant G. At a constrained stationary point of $\mathcal{L}_{tune}$ on the unit sphere, the tangential component of the instance-level representation drift is bounded by*

$$\left\| \Pi_{v_i}(v_i - v_i^{(0)}) \right\|_2 \leq \frac{G}{2\lambda_{ins}}, \tag{26}$$

where $\Pi_{v_i} = I - v_i v_i^\top$ projects a vector onto the tangent space at v_i.

Proof. Taking the Euclidean gradient of the objective with respect to v_i, we have

$$\nabla_{v_i} \mathcal{L}_{\text{tune}} = \nabla_{v_i} \mathcal{L}_{\text{info}} + 2\lambda_{\text{ins}}(v_i - v_i^{(0)}). \tag{27}$$

Because v_i is constrained to the unit sphere, stationarity should be evaluated in the tangent space rather than by setting the Euclidean gradient to zero. Thus, at a constrained stationary point,

$$\Pi_{v_i} \nabla_{v_i} \mathcal{L}_{\text{tune}} = 0. \tag{28}$$

Therefore,

$$2\lambda_{\text{ins}} \Pi_{v_i}(v_i - v_i^{(0)}) = -\Pi_{v_i} \nabla_{v_i} \mathcal{L}_{\text{info}}. \tag{29}$$

Taking the L_2 norm on both sides gives

$$\left\| \Pi_{v_i}(v_i - v_i^{(0)}) \right\|_2 \leq \frac{1}{2\lambda_{\text{ins}}} \left\| \Pi_{v_i} \nabla_{v_i} \mathcal{L}_{\text{info}} \right\|_2. \tag{30}$$

Since projection cannot increase the norm, and the gradient norm of InfoNCE is bounded by G under normalized embeddings and fixed temperature, we obtain

$$\left\| \Pi_{v_i}(v_i - v_i^{(0)}) \right\|_2 \leq \frac{G}{2\lambda_{\text{ins}}}. \tag{31}$$

□

This bound shows that λ_{ins} explicitly controls the trade-off between learning new LLM-based alignment and preserving the original CLIP representation. A larger λ_{ins} imposes a stronger constraint on local representation drift. Meanwhile, $\mathcal{L}_{\text{struct}}$ constrains the global pairwise geometry. Therefore, the two distillation losses jointly mitigate representation degradation from both local and global perspectives.

C.8 Discussion: From Direct Alignment to Progressive Alignment

The above analysis reveals the essential difference between direct alignment and ProCLIP. Direct alignment attempts to align the visual encoder with an uncalibrated LLM embedding space. This makes the visual encoder vulnerable to unstable contrastive gradients and may degrade its pre-trained semantic structure. In contrast, ProCLIP first aligns the LLM representation with the CLIP text manifold and then performs visual-language refinement under continuous distillation constraints.

This progressive design has three advantages. First, instance-level distillation provides point-wise semantic anchoring, reducing the initial cross-modal gap and preventing the optimization from starting from an uncalibrated LLM space. Second, structural distillation preserves the global semantic topology of CLIP, preventing the learned representation from collapsing into isolated instance-level matches. Third, because the adapter operates on LLM semantic representations rather than language-specific tokens, the long-text and multilingual knowledge already encoded in the LLM can be projected into the CLIP-compatible visual-language space.

Together, these mechanisms explain why ProCLIP can leverage the richer linguistic, compositional, and multilingual knowledge of LLMs while preserving the robust visual generalization ability.

D More results and analysis.

Detailed comparison on fine-grained and multilingual benchmarks. Table 18 compares the fine-grained and multilingual capabilities of different methods. S.C. in the table is the abbreviation for SugarCrepe. ProCLIP consistently surpasses LLM2CLIP across SugarCrepe, XM3600, and MMVP under different CLIP architectures and data scales. The gains on SugarCrepe demonstrate that ProCLIP better captures fine-grained compositional relations, including replacement, swapping, and addition. Meanwhile, its improvements on XM3600 I2T/T2T indicate stronger multilingual cross-modal alignment. ProCLIP also achieves notable gains on MMVP, suggesting enhanced visual discrimination in challenging fine-grained understanding tasks. These consistent improvements verify the effectiveness of ProCLIP in both compositional reasoning and multilingual generalization.

Detailed comparison of linear probe performance. Table 19 summarizes the detailed linear probe performance across 11 downstream classification datasets. ProCLIP consistently outperforms LLM2CLIP in average accuracy across different architectures and data scales, demonstrating stronger transferable visual representations. The improvements are particularly evident on fine-grained datasets such as Food101, Cars, Aircraft, DTD, Pets, and Flowers, indicating that ProCLIP better preserves discriminative visual features. Notably, ProCLIP also remains competitive with the original CLIP baseline, suggesting that it enhances language-guided adaptation while avoiding degradation of the general-purpose representation capability.

Analysis of Data Scale and Model Scale. We further analyze the effects of data scale and model scale, as shown in Tab. 20. For data scale, we observe that model performance improves with increasing data size. For example, when trained on 3M samples, ProCLIP achieves a zero-shot IN-1k accuracy of 62.5%, which rises to 67.8% when the dataset size increases to 30M. Under the same data scale, ProCLIP consistently outperforms LLM2CLIP. Notably, when we randomly sample 1M images from CC3M for training, ProCLIP achieves comparable or even superior zero-shot retrieval performance compared with LLM2CLIP, reaching 61.8% on zero-shot IN-1k. This highlights the data efficiency of ProCLIP. For model scale, we further expand the linear layers by three times, using $12\times$ layers in total, which leads to additional performance gains. This suggests that ProCLIP can continue to benefit from simple parameter scaling.

Detailed comparison of robustness performance. Table 21 compares the robustness performance on ImageNet distribution-shift benchmarks. ProCLIP consistently surpasses LLM2CLIP across all CLIP architectures and training data scales, showing stronger generalization under distribution shifts. The gains are particularly notable on IN-A, IN-R, and IN-S, where robust visual recognition is more challenging. Compared with LLM2CLIP, ProCLIP better preserves the pre-trained robustness of CLIP while improving adaptation effectiveness. These results demonstrate that ProCLIP can effectively mitigate robustness degradation and maintain more reliable visual representations under shifted distributions.

Generalized to pixel-level tasks. Table 22 reports the open-vocabulary semantic segmentation performance on ADE20K, COCO-Stuff164k, and VOC20. Compared with the orig-

Table 18: Detailed comparison of fine-grained and multilingual capability on SugarCrepe, XM3600, and MMVP across different CLIP architectures and data scales.

Method	#Data	S.C. Replace	S.C. Swap	S.C. Add	XM3600 I2T	XM3600 T2T	MMVP
Model Architecture: CLIP ViT-B/32							
CLIP	400M	80.2	64.0	72.7	-	-	15.6
LLM2CLIP	3M	87.3	72.5	85.1	40.3	39.9	21.5
ProCLIP	3M	**88.0**	**72.9**	**87.5**	**43.0**	**43.8**	**25.9**
LLM2CLIP	15M	88.4	74.7	86.1	42.1	40.6	**25.9**
ProCLIP	15M	**89.0**	**75.0**	**87.9**	**43.7**	**44.8**	24.4
LLM2CLIP	30M	88.3	**75.4**	85.6	43.9	41.8	25.2
ProCLIP	30M	**89.1**	75.3	**88.0**	**44.5**	**45.2**	**30.4**
Model Architecture: CLIP ViT-B/16							
CLIP	400M	81.6	65.2	82.5	-	-	12.6
LLM2CLIP	3M	88.6	71.4	87.9	43.1	43.7	20.7
ProCLIP	3M	**89.3**	**73.4**	**89.0**	**44.0**	**45.6**	**29.6**
LLM2CLIP	15M	89.3	75.1	89.3	44.3	44.9	22.2
ProCLIP	15M	**89.7**	**75.9**	**89.9**	**45.0**	**45.6**	**24.4**
LLM2CLIP	30M	89.4	73.9	89.5	44.7	45.9	28.2
ProCLIP	30M	**89.5**	**74.8**	**91.0**	**45.9**	**46.3**	**30.4**
Model Architecture: CLIP ViT-L/14							
CLIP	400M	79.5	61.3	74.9	-	-	19.3
LLM2CLIP	3M	90.0	73.7	89.9	44.2	46.6	22.2
ProCLIP	3M	**90.7**	**75.5**	**92.5**	**46.1**	**49.1**	**25.2**
LLM2CLIP	15M	90.6	75.4	91.9	44.8	47.8	25.2
ProCLIP	15M	**91.9**	**77.2**	**93.1**	**46.7**	**51.0**	**27.4**
LLM2CLIP	30M	89.7	74.2	91.5	46.2	48.3	23.7
ProCLIP	30M	**90.3**	**75.4**	**92.8**	**47.8**	**52.6**	**34.1**

Table 19: Detailed linear probe performance on 11 downstream classification datasets across different CLIP architectures and training data scales.

Method	#Data	Food101	CIFAR10	CIFAR100	SUN397	Cars	Aircraft	DTD	Pets	Caltech101	Flowers	ImageNet	Avg.
Model Architecture: CLIP ViT-B/32													
CLIP	400M	88.6	95.1	80.1	73.4	80.8	44.9	76.3	89.3	92.7	94.7	74.3	80.9
LLM2CLIP	3M	87.9	95.7	**83.1**	74.1	78.0	**44.9**	77.7	**90.4**	92.4	94.6	74.2	81.2
ProCLIP	3M	**88.4**	**95.9**	**83.1**	**74.3**	**79.5**	44.1	**78.2**	90.3	**92.6**	**95.0**	**74.4**	**81.4**
LLM2CLIP	15M	87.7	95.7	82.7	74.0	77.5	44.2	**78.3**	**90.2**	92.5	94.4	74.2	81.0
ProCLIP	15M	**88.7**	**95.9**	**82.8**	**74.8**	**80.8**	**44.9**	78.1	**90.2**	**92.8**	**95.1**	**74.4**	**81.7**
LLM2CLIP	30M	87.6	95.9	83.0	74.1	76.3	43.5	77.6	**90.1**	**92.8**	93.8	74.3	80.8
ProCLIP	30M	**88.2**	**96.0**	**83.1**	**75.1**	**79.0**	**43.8**	**77.8**	89.8	92.6	**94.9**	**74.5**	**81.4**
Model Architecture: CLIP ViT-B/16													
CLIP	400M	92.7	96.0	82.5	75.7	85.9	52.8	78.9	93.1	93.9	96.4	79.6	84.4
LLM2CLIP	3M	91.6	**97.0**	84.5	76.0	82.1	50.1	80.3	92.3	93.6	95.7	79.6	83.9
ProCLIP	3M	**92.8**	96.8	**84.6**	**76.4**	**85.6**	**52.0**	**80.6**	**93.3**	**94.2**	**97.0**	**79.7**	**84.8**
LLM2CLIP	15M	91.9	**97.0**	**84.9**	75.6	83.7	50.7	80.4	92.9	93.8	96.6	79.6	84.3
ProCLIP	15M	**92.6**	96.7	84.3	**76.6**	**85.6**	**51.4**	**80.8**	**93.6**	**94.3**	**96.7**	**79.8**	**84.8**
LLM2CLIP	30M	91.3	**96.6**	84.8	75.3	80.6	48.2	80.3	92.5	93.4	95.0	79.7	83.4
ProCLIP	30M	**92.3**	**96.6**	**85.7**	**77.0**	**84.7**	**50.1**	**81.2**	**93.1**	**94.0**	**96.7**	**79.5**	**84.6**
Model Architecture: CLIP ViT-L/14													
CLIP	400M	95.3	89.1	87.2	79.4	90.7	63.0	81.8	95.3	96.9	98.8	82.9	88.1
LLM2CLIP	3M	94.5	**98.6**	**89.2**	79.6	86.7	57.7	83.4	94.1	96.4	97.1	**82.5**	87.2
ProCLIP	3M	**95.3**	98.5	88.8	**80.3**	**90.3**	**61.0**	**83.6**	**95.2**	**96.9**	**98.7**	81.9	**88.2**
LLM2CLIP	15M	94.4	**98.5**	**88.8**	78.5	86.0	55.0	82.7	93.9	95.9	97.1	82.6	86.7
ProCLIP	15M	**95.2**	98.4	88.6	**79.7**	**90.5**	**61.4**	**83.3**	**95.3**	**96.8**	**98.7**	**83.0**	**86.7**
LLM2CLIP	30M	94.1	98.2	88.4	78.7	84.6	54.8	82.4	93.7	95.8	96.5	82.2	86.3
ProCLIP	30M	**95.1**	**98.4**	**89.0**	**80.3**	**90.0**	**60.0**	**83.9**	**95.2**	**96.8**	**98.5**	**82.7**	**88.2**

Table 20: Comparison of data and model scales under ViT-L architecture.

Method	#Data	Adapter	IN-1k	I2T Avg.	T2I Avg.
CLIP	400M	-	74.5	66.9	55.5
ProCLIP	1M	4×linear	61.8	81.9	79.7
LLM2CLIP	3M	4×linear	52.8	82.0	79.1
ProCLIP	3M	4×linear	**62.5**	**83.0**	**80.5**
LLM2CLIP	15M	4×linear	54.0	83.2	80.1
ProCLIP	15M	4×linear	**66.0**	**84.3**	**81.2**
LLM2CLIP	30M	4×linear	56.4	85.8	82.1
ProCLIP	30M	4×linear	67.8	86.2	82.6
ProCLIP	30M	12×linear	**71.5**	**86.8**	**82.8**

Table 22: Comparison of open-vocabulary semantic segmentation performance.

Method	#Data	ADE20K	COCO-Stuff164k	VOC20
Model Architecture: CLIP ViT-L/14				
CLIP	400M	1.4	2.8	16.3
LLM2CLIP	30M	2.4	3.0	17.5
ProCLIP	30M	**3.8**	**4.8**	**18.7**

Table 23: Training efficiency comparison between ProCLIP and LLM2CLIP.

Method	Data	VRAM Usage (GB) ↓	Training Hours ↓
Model Architecture: CLIP ViT-L/14			
LLM2CLIP	3M	**32.6**	**1.2**
ProCLIP	3M	43.2	1.7

inal CLIP, LLM2CLIP brings moderate improvements on all three datasets, indicating that language-side enhancement can benefit dense prediction transfer to some extent. In contrast, ProCLIP achieves the best results across all benchmarks, improving over LLM2CLIP by 1.4, 1.8, and 1.2 points on ADE20K, COCO-Stuff164k, and VOC20, respectively. These consistent gains suggest that ProCLIP better preserves and transfers CLIP's visual-language alignment to pixel-level recognition, leading to stronger open-vocabulary segmentation capability.

Efficiency analysis. Table 23 compares the training efficiency of ProCLIP and LLM2CLIP under the same 8×H100 GPU setting. ProCLIP requires slightly higher VRAM usage and training time than LLM2CLIP, increasing from 32.6 GB to 43.2 GB and from 1.2 to 1.7 hours, respectively. However, the overall overhead is moderate, suggesting that ProCLIP maintains good training efficiency while achieving stronger performance.

Table 21: Detailed comparison of robustness performance on ImageNet distribution-shift benchmarks.

Method	#Data	Robustness				
		IN-V2	IN-A	IN-O	IN-R	IN-S
Model Architecture: CLIP ViT-B/32						
CLIP	400M	55.9	31.6	47.8	69.3	42.3
LLM2CLIP	3M	37.0	18.8	38.0	54.0	30.7
ProCLIP	3M	**43.6**	**25.5**	**40.6**	**62.9**	**35.3**
LLM2CLIP	15M	38.8	20.1	40.5	58.3	33.4
ProCLIP	15M	**48.3**	**28.3**	**44.0**	**67.4**	**38.9**
LLM2CLIP	30M	41.0	22.9	41.8	59.5	34.0
ProCLIP	30M	**48.9**	**29.5**	**46.3**	**67.8**	**39.0**
Model Architecture: CLIP ViT-B/16						
CLIP	400M	61.9	49.9	42.3	77.7	48.2
LLM2CLIP	3M	42.4	30.2	35.3	61.6	35.2
ProCLIP	3M	**50.8**	**41.4**	**38.5**	**71.3**	**41.7**
LLM2CLIP	15M	44.8	33.9	37.3	65.4	38.1
ProCLIP	15M	**54.8**	**45.3**	**40.5**	**76.0**	**45.2**
LLM2CLIP	30M	46.0	35.6	38.3	67.0	38.7
ProCLIP	30M	**55.2**	**46.0**	**41.9**	**75.9**	**44.9**
Model Architecture: CLIP ViT-L/14						
CLIP	400M	69.8	70.8	32.2	87.8	59.6
LLM2CLIP	3M	49.0	46.6	32.4	75.0	44.8
ProCLIP	3M	**58.3**	**63.3**	**31.6**	**84.0**	**52.3**
LLM2CLIP	15M	50.8	50.1	33.8	78.2	46.3
ProCLIP	15M	**62.1**	**66.4**	**34.2**	**86.4**	**55.3**
LLM2CLIP	30M	52.7	52.7	34.0	78.6	47.3
ProCLIP	30M	**63.4**	**68.0**	**34.1**	**86.8**	**55.7**

E Visualization.

As illustrated in Fig. 7, ProCLIP effectively extends long-text processing capabilities, enabling precise image-text alignment even within highly detailed descriptions. Moreover, ProCLIP extends CLIP's capabilities to include multilingual

support and enhanced fine-grained understanding, which are absent in the original model.

F Future Works.

Training Efficiency. ProCLIP is subject to the limitation of increased computational overhead. We consider the following directions to potentially reduce computational overhead:

- Adopting a PEFT-based approach to fine-tune the vision encoder in the second stage
- Fine-tuning only part of the vision encoder parameters in the second stage, such as the last few Transformer blocks
- Replacing online distillation with offline distillation, which would substantially reduce the additional computational cost introduced in the second stage

Fine-grained Visual-Text Alignment. Future work explores the integration of a dedicated local alignment loss to explicitly strengthen the model's fine-grained visual-textual alignment capabilities, thereby enhancing performance on tasks that require detailed spatial understanding.

More Model Architecture. Our approach replaces the original CLIP text encoder with an LLM-based embedder to enhance multiple capabilities. From another perspective, we consider whether the vision encoder in the dual-tower architecture can also be replaced to address limitations in visual representation.

The image captures a vibrant and appetizing bento box, a traditional Japanese meal served in a compartmentalized box. The box is divided into four distinct sections, each filled with a variety of food items. In the top left section, there's a slice of bread. Adjacent to it, in the top right section, are chunks of pineapple, their bright yellow color adding a tropical touch to the meal. The bottom left section is filled with a serving of broccoli, its green color standing out against the yellow of the box. Next to it, in the bottom right section, is a serving of meatballs, their brown color hinting at a savory flavor.



✓ ProCLIP



✗ CLIP

In the tranquil expanse of a grassy field, a zebra, adorned with a mesmerizing pattern of black and white stripes, is captured in the act of grazing. The zebra's head is lowered towards the earth, its mouth busy nibbling on the lush green grass beneath it. The field, a vibrant canvas of green, is speckled with patches of white flowers that add a touch of contrast to the scene. The zebra's mane, standing upright, adds a dynamic element to the otherwise serene landscape. The image is a beautiful representation of nature's simplicity and elegance.



✓ ProCLIP



✗ CLIP

In the heart of a verdant field, a spectacle unfolds. Two majestic horses, one as white as a cloud and the other as black as night, are caught in a moment of pure athleticism. They are rearing up on their hind legs, their front legs reaching for the sky in a display of power and grace. The riders, clad in protective helmets, hold on tightly, their bodies leaning forward in sync with the horses' movements. The field is a vibrant green, a stark contrast to the horses and riders. In the distance, trees stand tall and proud, their leaves rustling gently in the breeze.



✓ ProCLIP



✗ CLIP

In the image, there are two main objects placed on a wooden surface against a beige wall. The first object is a white owl-shaped candle holder. The owl, made of a white material, is sitting on the left side of the image. It has a small candle inside it, which is lit, casting a warm glow. The second object is a gold-colored clock. The clock is standing upright the image. It has a white face with black numbers and hands. The clock and the owl are positioned in such a way that they seem to be in conversation with each other, creating an interesting visual dynamic.



✓ ProCLIP



✗ CLIP

Figure 7: The visualization of ProCLIP vs CLIP in long-text cross-modal retrieval. From top to bottom: Case 1, Case 2, Case 3, and Case 4.