Unlocking Fine-Grained Perception in CLIP via Structurally-Aware Latent Masked Modeling

Juntong Li [*], **Lingwei Dang** [*†], **Haomin Wu** ,
Ziyan Qiu , **Qingxin Xiao** , **Qingyao Wu**[‡]
School of Software Engineering, South China University of Technology

Abstract

Vision-Language Models (VLMs) such as CLIP excel in global semantic alignment but often lack fine-grained perceptual capabilities. This hinders dense prediction tasks and bottlenecks the visual potential of Multimodal Large Language Models (MLLMs). Existing research has attempted to enhance CLIP's visual representations by incorporating geometric priors from vision-centric models. However, these strategies often struggle to achieve deep alignment for both local spatial structures and global semantics, potentially even distorting the original image-text space. To address these limitations, we propose SALM, an unsupervised embedding alignment framework based on structurally-aware latent mask modeling. SALM effectively synergizes local and global alignment via a dual-path design combining explicit and implicit mechanisms, without requiring external textual supervision. First, we introduce a dual-matrix alignment strategy that explicitly calibrates intra-sample spatial correlations and activation intensities, thereby effectively injecting local geometric priors. Based on this, we further design a latent mask modeling mechanism to guide CLIP to restore the missing semantic details of the target model, thereby implicitly aggregating fine-grained structures into the global semantic space. Furthermore, driven by the empirical observations that CLIP's shallow features inherently possess strong spatial observational capabilities, we naturally extend SALM to a highly efficient self-distillation paradigm, SALM-Self. This unlocks CLIP's intrinsic fine-grained potential without relying on any external models. Extensive experiments demonstrate that SALM not only significantly improves performance in dense prediction tasks but also boosts CLIP's zero-shot accuracy, effectively enhancing the fine-grained understanding capabilities of MLLMs. Project page at https://qzfm.github.io/salm_project_page/.

1 Introduction

Vision-Language Models (VLMs) have demonstrated remarkable generalization capabilities in multimodal understanding, content generation, and other downstream tasks. Among them, CLIP [45] and its variants [69, 54] rely on strong cross-modal alignment to maintain robust visual features even in zero-shot scenarios. However, constrained by contrastive learning's primary focus on global semantic consistency, CLIP often struggles to capture fine-grained information such as color, counting, and local geometric structure [55, 32]. This deficiency in fine-grained perception not only limits performance in tasks such as dense prediction [66, 65] but also constitutes a bottleneck for Multimodal Large Language Models (MLLMs) [37, 33, 2, 74] aiming for refined visual understanding [16, 38].

[*]Equal contributions. Email: qlzjftm@gmail.com, levondang@163.com.
[†]Project Leader.
[‡]Corresponding Author. Email: qyw@scut.edu.cn.

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

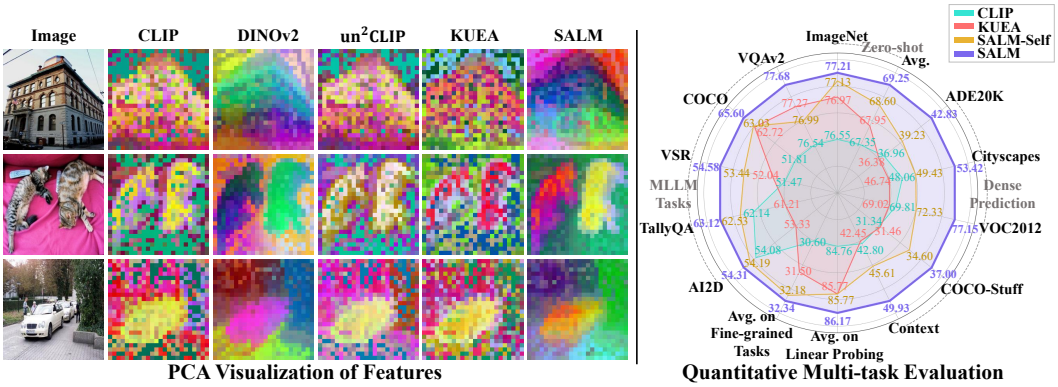

Figure 1: Comparison of feature visualizations and performance. Left: PCA visualization of features. In contrast to the fragmented and noisy outputs from CLIP and others, our method produces spatially coherent features with distinct semantic layouts. Right: The radar chart validates the superior performance of our method across multiple quantitative metrics.

Self-supervised vision models [9, 52, 18] possess inherent advantages in comprehending local structures and spatial relationships. Effectively leveraging these priors to mitigate CLIP's limitations remains a challenge. Some approaches [25, 23, 72, 63] employ multi-experts as enhanced visual encoders; however, they often lack the capacity to fundamentally enhance fine-grained perception at the perceptual level and introduce doubled computational overhead. Other methods attempt to distill diverse encoders into a single aggregated model [47, 19, 48, 73] or fine-tune CLIP via generative reconstruction [62, 39, 60, 59, 35, 17]. Nevertheless, these approaches either incur immense training costs or disrupt CLIP's original embedding distribution, severely compromising zero-shot capabilities or downstream compatibility, thereby limiting their practical applicability.

Recent works have explored kernel-based methods [24, 13] to achieve efficient embedding alignment. Although kernel-based alignment can elegantly preserve the structure of the original space, it struggles to guarantee the spatial consistency of local dense features (*e.g.*, KUEA [13] has no improvement over vanilla CLIP in segmentation tasks). This is because they primarily focus on the global alignment of embedding clusters between samples, neglecting the spatial positioning and intrinsic correlations of local features within samples, which restricts CLIP's fine-grained local perception. Furthermore, their performance is highly sensitive to the choice of the kernel function [13], making it difficult to adaptively capture complex nonlinear correspondences between heterogeneous feature spaces.

Our key insight is that simultaneously achieving adaptive alignment of local spatial structures and global semantics can comprehensively address CLIP's fine-grained perception limitations, while avoiding disruption to the integrity of the original embedding space, preserving its zero-shot capability. We propose **SALM**, a novel framework for unsupervised embedding alignment via structurally-aware latent masked modeling. Despite requiring zero text supervision, SALM achieves adaptive alignment of local spatial consistency and fine-grained global semantics between CLIP and vision-centric models without breaking pre-trained image-text alignment.

We first introduce a Dual-Matrix Alignment (DMA) strategy to explicitly constrain the local spatial geometry at the manifold level. Specifically, we model the pairwise relationships between patches within the feature map by decoupling them into angular and magnitude components, constructing two key matrices: the Spatial Relation Matrix and the Energy Difference Matrix. By aligning these matrices, we effectively capture local semantic correlations and calibrate relative activation magnitudes, respectively. This mechanism explicitly injects fine-grained geometric priors from the target model into CLIP's space, enhancing local detail expression while maximizing the preservation of the original pre-trained distribution.

Building on explicit spatial constraints, we further aim to aggregate structural information into global semantics. This objective demands that the representation implicitly and intrinsically encodes and aligns with the detailed semantics of vision-centric models. In other words, these features must possess the capability to restore the fine-grained latents of these models. Inspired by masked modeling paradigm [18, 70, 31, 4, 41, 61], we propose an implicit alignment method based on latent masked

modeling to bridge these distinct encoders. Specifically, we freeze the vision-centric encoder and apply high-ratio random masking to its dense features. Utilizing a lightweight Cross-Guided Adapter (CGA), we reconstruct the latents conditioned on both the visible context and the CLIP features. The unmasked patches serve as anchors, guiding the encoder to adaptively fill the missing fine-grained semantic gaps, effectively establishing a deep semantic alignment with the target feature space.

Furthermore, our empirical observations indicate that CLIP's shallow layers inherently capture rich local spatial details, which gradually diminish during deep feature aggregation. Inspired by this, we pose a natural question: can we unlock CLIP's intrinsic fine-grained potential directly within the SALM framework, without relying on any external vision-centric models? However, directly utilizing shallow features presents two major challenges: (1) they lack high-level semantics, meaning direct dual-matrix alignment would severely degrade the original deep semantics; and (2) they contain substantial low-level noise, making feature reconstruction difficult to optimize. To overcome these challenges, we introduce a small adjustment to our framework, extending it into a highly efficient self-distillation paradigm termed **SALM-Self**. Specifically, we discard the reconstruction loss and apply the DMA constraint on the reconstructed shallow features. This asymmetric design elegantly filters out low-level noise while strictly preserving the deep semantic integrity, enhancing CLIP's fine-grained perception and dense prediction capabilities with zero reliance on external models.

Experiments demonstrate that our method outperforms other alignment methods on dense prediction tasks, verifying that the model genuinely possesses fine-grained perception capabilities. Meanwhile, on multiple benchmarks, we achieve higher zero-shot accuracy without fine-tuning the text encoder. MLLMs integrated with our visual encoder exhibit superior performance over other methods.

In summary, the contributions of this paper are as follows:

- We propose SALM, a novel unsupervised embedding alignment framework bridging CLIP's global consistency and vision-centric models' fine-grained perception. It achieves simultaneous adaptive alignment of local geometry and global semantics.
- We propose a dual-matrix alignment strategy to explicitly constrain the local spatial geometry at the manifold level. By aligning the spatial relation matrix and energy difference matrix, we effectively capture intrinsic spatial correlations, ensuring the transfer of spatial structures.
- We design an implicit alignment based on latent masking modeling. Using the Cross-Guided Adapter, we guide the encoder to recover missing fine-grained details, establishing a deep and robust semantic alignment with the target feature space.
- We extend the SALM to a highly efficient self-distillation paradigm, SALM-Self. It successfully unlocks CLIP's intrinsic fine-grained potential without relying on any external vision-centric models.

2 Related Work

Vision-Language Models. As a milestone in Vision-Language Models (VLMs), CLIP [45] and its variants (e.g., SigLIP [69] and EVA-CLIP [54]) demonstrate remarkable generalization capabilities in zero-shot tasks, attributed to robust cross-modal alignment established through large-scale contrastive learning. However, driven by a contrastive learning objective that overly prioritizes global consistency, CLIP often overlooks fine-grained details such as color, texture, and local geometric structures [55, 32]. This deficiency in fine-grained perception not only constrains performance in tasks like dense prediction [66, 65] but also serves as a primary bottleneck hindering MLLMs from achieving higher visual understanding precision [16, 38].

Enhancement of Visual Representations. To address CLIP's limitations in fine-grained perception, researchers have attempted to incorporate prior knowledge from self-supervised vision models (e.g., DINOv2 [43, 9], MAE [18], DINOv3 [52]). Some attempts predominantly adopted mixture-of-experts strategies [25, 72, 63]. While introducing additional information, this approach incurs multiplicative computational overheads and lacks deep feature fusion. Another category of methods explores distillation [47, 19, 48, 73], generative reconstruction [62, 39, 60, 59, 35, 17] or architectural modifications [51] to finetune CLIP. However, these approaches typically demand substantial data and computational resources and risk compromising its zero-shot capabilities. Alignment schemes based on kernel-based methods [24, 13] have recently garnered attention for their efficiency, attempting to

align feature matrices within a kernel space. However, kernel methods typically neglect spatial positional relationships and intrinsic correlations among local features within individual samples. Several works [58, 64, 30] have explored enhancing CLIP with vision-centric priors for open-vocabulary dense perception. In contrast, our goal is to develop a general CLIP enhancement framework that improves fine-grained representation while maintaining the global image-text alignment inherited from pre-trained CLIP.

Masked Modeling. Masked Image Modeling (MIM) [18] facilitates the learning of local structures and spatial relationships by predicting masked image patches. Some works [31, 4, 41, 61] have leveraged masked modeling to learn rich semantic representations and temporal dependencies. Recent theoretical studies [70] have provided insights into the effectiveness of MIM, suggesting that MIM essentially performs an implicit semantic alignment across samples. Inspired by this, we propose establishing a connection between frozen source and target features via latent masked modeling.

3 Method

3.1 Empirical Observations in CLIP

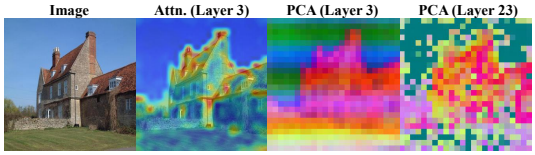

Figure 2: Visualization of CLIP's shallow attention versus deep feature representation. CLIP demonstrates the capability to localize object textures (*e.g.*, building edges) and detailed components (*e.g.*, doors and windows) in its early layers, even though these details may not be effectively extracted in representations.

Motivated by CLIP's limited fine-grained performance, we visualized the attention maps within its shallow layers. As shown in Figure 2, the results reveal that in the shallow layers, CLIP actually exhibits excellent observational capability, accurately covering various local parts of the target. However, it ultimately fails to effectively aggregate and preserve the texture and geometric details of these regions. This indicates that the bottleneck of CLIP's fine-grained perception lies in the final expressiveness of the representation rather than its initial observational capability.

This finding underscores the limitations of existing strategies: multi-expert strategies do not fundamentally improve the quality of CLIP's intrinsic fine-grained representation, while kernel-based methods tend to neglect the expressiveness of local features. And generative methods may distract the model with pixel-level redundancies. Therefore, we do not intervene in the model's observation mechanism; instead, by introducing a vision-centric model to provide geometric priors, we drive the model at the feature manifold level to internalize fine-grained structural information through dual-matrix alignment and latent masked modeling. We additionally explore a self-distillation strategy, utilizing shallow features to unlock CLIP's fine-grained potential without relying on external experts.

3.2 Overview

Considering two distinct embedding spaces: a *source space* defined by the visual encoder $f_\theta(\cdot)$ of a pre-trained VLM, and a *reference space* characterized by a vision-centric model $g_\phi(\cdot)$ that encapsulates rich local structural priors. For a given image x, the extracted token sequences are denoted as $\mathbf{Z}^{\text{src}} = [\mathbf{z}^{\text{src}}_{\text{cls}}; \mathbf{Z}^{\text{src}}_{\text{p}}] \in \mathbb{R}^{(N+1)\times d}$ and $\mathbf{Z}^{\text{ref}} = [\mathbf{z}^{\text{ref}}_{\text{cls}}; \mathbf{Z}^{\text{ref}}_{\text{p}}] \in \mathbb{R}^{(N+1)\times d'}$, respectively. To address potential resolution mismatches, we spatially interpolate $\mathbf{Z}^{\text{src}}$ to the size of $\mathbf{Z}^{\text{ref}}$.

As illustrated in Figure 3, our framework consists of a trainable source visual encoder f_θ, a frozen reference encoder g_ϕ, and a regularization encoder f_{θ_0}. The Dual-Matrix Alignment (DMA) strategy (Sec. 3.3) operates on the patch sequences of both encoders. Subsequently, we employ a latent space masked modeling approach (Sec. 3.4) to further aggregate fine-grained semantics. We apply random masking to the reference features of the vision-centric model. A lightweight Cross-Guided Adapter (CGA) is tasked with reconstructing the reference tokens $\mathbf{Z}^{\text{ref}}$ by leveraging source space features and the visible reference features. Finally, we incorporate a regularization (Sec. 3.5) to ensure the stability of the adaptation process within the original multimodal latent space. Additionally, we extend it into a self-distillation variant that unlocks fine-grained perception without external models (Sec. 3.6).

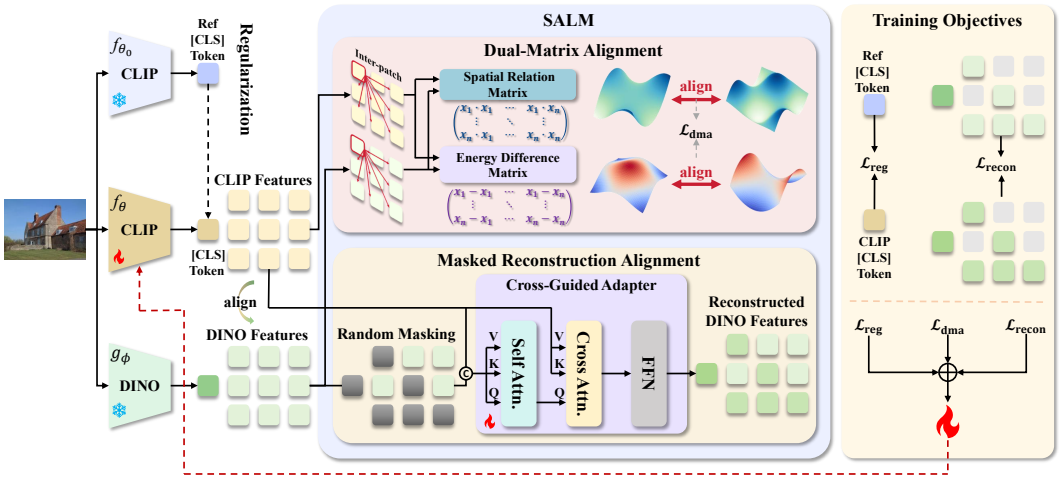

Figure 3: Overview of SALM. First, the Dual-Matrix Alignment strategy ensures relative consistency in spatial relationships and magnitude distributions (Sec. 3.3). Subsequently, a latent masked modeling is employed to reconstruct the DINO features, thereby aggregating fine-grained information into global semantics (Sec. 3.4). Finally, reference regularization is applied to the source encoder to ensure the stability of the adaptation process within the original multimodal latent space (Sec. 3.5).

3.3 Dual-Matrix Alignment Strategy

Recent kernel-based methods [24, 13] attempt to mitigate feature space heterogeneity; they are primarily confined to aligning inter-sample distribution clusters, neglecting the intrinsic geometric structure of intra-sample features. We argue that this oversight regarding local spatial relationships and relative saliency ranking constitutes the core bottleneck hindering CLIP's dense prediction capabilities. To address this, we propose the Dual-Matrix Alignment strategy, designed to explicitly capture and transfer second-order geometric statistics between patches at the manifold level, thereby injecting fine-grained structural priors without compromising the original semantic space.

Specifically, given the source feature sequence $\mathbf{Z}_{\text{p}}^{\text{src}}$ and reference feature sequence $\mathbf{Z}_{\text{p}}^{\text{ref}}$ for an image sample, we construct two key affinity matrices to describe the manifold. To capture semantic dependencies among local features, we define the Spatial Relation Matrix $\mathbf{K} \in \mathbb{R}^{N\times N}$: $\mathbf{K}_{ij}(\mathbf{Z}) = \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2}$, where $\mathbf{K}_{ij}$ quantifies the semantic similarity between the i-th and j-th patches.

The magnitude of feature norms typically encodes the saliency or information entropy of local regions. To incorporate magnitude considerations while avoiding discrepancies caused by the varying magnitude scales of distinct models, we define the Energy Difference Matrix $\mathbf{D} \in \mathbb{R}^{N\times N}$, utilizing normalization to eliminate dimensional scale effects: $\mathbf{D}_{ij}(\mathbf{Z}) = \mathcal{Z}(\|\mathbf{z}_i\|_2 - \|\mathbf{z}_j\|_2)$, where $\mathcal{Z}$ denotes Z-score normalization within the matrix, and $\mathbf{D}_{ij}$ characterizes the relative saliency ranking.

The total objective of the DMA strategy is defined as the sum of the Frobenius norm distances of these two matrices: $\mathcal{L}_{\text{dma}} = \|\mathbf{K}(\mathbf{Z}_{\text{p}}^{\text{src}}) - \mathbf{K}(\mathbf{Z}_{\text{p}}^{\text{ref}})\|_F^2 + \|\mathbf{D}(\mathbf{Z}_{\text{p}}^{\text{src}}) - \mathbf{D}(\mathbf{Z}_{\text{p}}^{\text{ref}})\|_F^2$.

3.4 Latent Masked Modeling for Implicit Alignment

While explicit alignment rectifies source spatial geometry, second-order constraints fail to integrate fine details into global semantics. To address this, we adopt latent masked modeling to aggregate spatial structural information, leveraging source features as semantic anchors to reconstruct missing details and global semantics within reference features for deep semantic integration.

Specifically, given the reference sequence $\mathbf{Z}^{\text{ref}}$, we apply a high-ratio (*e.g.*, 75%) random masking $M \in \{0,1\}^N$ to partition it into visible context $x_1 = \mathbf{Z}^{\text{ref}}[M]$ and masked target $x_2 = \mathbf{Z}^{\text{ref}}[1-M]$, replacing dropped patches with a learnable [MASK] token.

To facilitate reconstruction within the latent space and mitigate the risk of the decoder dominating representation learning, we design a lightweight Cross-Guided Adapter(CGA). The CGA recovers the

target x_2 utilizing the complete source features $\mathbf{Z}^{\text{src}}$ and the sparse context x_1. It first concatenates the full source features with the visible reference context to complement the spatial positional basis. Subsequently, in the cross-attention layer, the fused features serve as the query, while the source features are explicitly designated as key and value. The process of CGA is defined as follows:

$$\begin{aligned} \mathbf{H}_{\text{mid}} &= \text{SelfAttn}\big([\mathbf{Z}^{\text{src}}; x_1]\big) + [\mathbf{Z}^{\text{src}}; x_1], \\ \mathbf{H}_{\text{out}} &= \text{CrossAttn}\big(\mathbf{H}_{\text{mid}}, \mathbf{Z}^{\text{src}}, \mathbf{Z}^{\text{src}}\big) + \mathbf{H}_{\text{mid}}, \\ \hat{x}_2 &= \mathcal{P}_{\text{target}}\big(\text{FFN}(\mathbf{H}_{\text{out}})\big), \end{aligned} \tag{1}$$

where $\mathcal{P}_{\text{target}}$ represents the slicing operator extracting features corresponding to the target regions. The optimization objective is to minimize the reconstruction error: $\mathcal{L}_{\text{recon}} = \|\hat{x}_2 - x_2\|_2^2$.

3.5 Regularization for Zero-shot Preservation

Following [50, 13], to maintain the pre-trained image-text alignment, we introduce a regularization term $\mathcal{L}_{\text{reg}} = \|f_\theta(x) - f_{\theta_0}(x)\|_2^2$, where f_{θ_0} denotes the frozen initial source encoder. This ensures that the language-image alignment is effectively preserved even without incorporating any textual data during the fine-tuning phase.

The final objective is a balanced combination of the aforementioned constraints: $\mathcal{L} = \alpha\mathcal{L}_{\text{reg}} + \beta\mathcal{L}_{\text{recon}} + \gamma\mathcal{L}_{\text{dma}} + \lambda\mathcal{L}_{\text{uni}}$,where $\mathcal{L}_{\text{uni}}$ denotes the uniformity loss [70]. α, β, γ, and λ are hyperparameters.

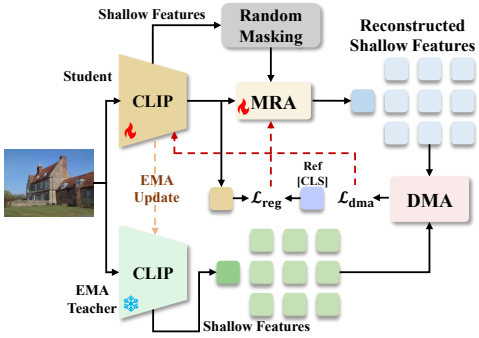

Figure 4: Overview of SALM-Self.

3.6 SALM-Self: Enhancing CLIP via Intra-Model Alignment

As discussed in Section 3.1, while CLIP's shallow layers inherently capture local geometric information, these details gradually diminish during deep feature aggregation. However, directly aligning or reconstructing these shallow features is impractical due to their lack of high-level semantics and the presence of substantial low-level noise. With minor adjustments to SALM, we adapt it into a CLIP-based self-distillation architecture, SALM-Self, as shown in Figure 4. Specifically, we set CLIP's shallow features as the reference features $\mathbf{Z}^{\text{ref}}$ and discard the reconstruction loss $\mathcal{L}_{\text{recon}}$. In its place, we perform dual-matrix alignment between the reconstructed shallow features and those of the teacher model, whose parameters are dynamically updated via an exponential moving average (EMA). The total objective loss is: $\mathcal{L} = \alpha\mathcal{L}_{\text{reg}} + \gamma\mathcal{L}_{\text{dma}} + \lambda\mathcal{L}_{\text{uni}}$.

4 Experiments

4.1 Experimental Setup

Implementation Details. We adopt OpenAI CLIP ViT-L-14-336 [45] and DINOv2 ViT-L/14 with registers [9] as the default VLM and vision-centric backbone, respectively. Notably, our framework is designed to be model-agnostic and resolution-flexible, capable of aligning arbitrary VLMs with various vision models.

We train exclusively on the ImageNet-1K [10] training set. In contrast to the massive datasets required for CLIP's pre-training or the distillation processes of aggregation models, our data regime is exceptionally lightweight and purely image-based, independent of text. For inference, the auxiliary CGA is removed, leaving only the finetuned CLIP for downstream tasks to ensure zero extra cost.

We train for a total of four epochs using a three-stage curriculum strategy to progressively optimize structural alignment and fine-grained semantic reconstruction. Specifically, we begin with a one-epoch structural warm-up, where alignment is driven by DMA with the reconstruction loss disabled. During the subsequent two-epoch transition, we gradually decrease the DMA weight while increasing the reconstruction weight, shifting the training emphasis from relational structure to feature reconstruction. In the final epoch, reconstruction becomes the dominant alignment objective, while DMA remains

Table 1: Zero-shot object recognition performance on various benchmarks. We report Top-1 accuracy (%) on ImageNet-1K and 11 other datasets. The best results are highlighted in **bold**. † indicates the model reproduced by us on ImageNet-1K.

Method	ImageNet	Zero-Shot											
		CIFAR-10	CIFAR-100	Pets	Caltech-101	RESISC45	PCam	DTD	EuroSAT	ImageNet-O	FER2013	ImageNetV2	**Average**
CLIP	76.55	94.92	74.33	93.68	83.43	63.78	60.72	55.69	61.50	32.75	49.11	70.89	67.35
RADIOv2.5†	75.02	90.97	64.94	86.70	83.78	37.35	50.02	43.62	26.69	**61.05**	34.68	67.82	58.87
un^{2}CLIP	71.25	91.84	69.95	91.74	**85.45**	57.33	61.78	52.93	**62.46**	36.15	**51.49**	64.91	66.00
KUEA	76.97	95.86	76.95	93.79	83.43	64.43	59.60	**56.33**	61.89	35.65	48.08	71.40	67.95
SALM-Self (Ours)	*77.13*	*95.58*	*77.37*	*93.87*	*83.79*	*63.95*	*64.74*	*55.96*	*61.72*	*36.80*	*49.60*	*71.27*	*68.60*
SALM (Ours)	**77.21**	**96.38**	**77.87**	**94.09**	84.04	**64.63**	**67.54**	56.27	62.19	37.05	50.11	**71.53**	**69.25**

Table 2: Quantitative results on zero-shot fine-grained understanding and dense prediction. We report Top-1 accuracy (%) for zero-shot fine-grained understanding tasks, alongside mean IoU (mIoU) for semantic segmentation. † indicates the model reproduced by us on ImageNet-1K.

Method	ZS. Fine-grained Tasks			Linear Probing Segmentation				
	SVHN	CLEVR Distance	CLEVR Counts	ADE20K	Cityscapes	VOC2012	COCO-Stuff	Context
CLIP	55.97	15.81	20.01	36.96	48.06	69.81	31.34	42.80
RADIOv2.5†	41.25	**19.45**	15.60	42.35	51.82	75.31	35.11	48.73
un^{2}CLIP	54.77	15.93	22.73	37.46	48.11	71.64	32.74	44.44
KUEA	57.74	15.95	20.81	36.38	46.74	69.02	31.46	42.45
SALM-Self (Ours)	*57.92*	*15.83*	22.78	*39.23*	*49.43*	*72.33*	*34.60*	*45.61*
SALM (Ours)	**57.99**	16.21	**22.81**	**42.83**	**53.42**	**77.15**	**37.00**	**49.93**

an auxiliary structural constraint. For SALM-Self, which omits the reconstruction loss, we instead use fixed loss weights throughout training. We use the 6th CLIP layer by default as its self-reference feature, which provides a balance between local structural details and semantic consistency.

Detailed training configurations and additional experiments evaluating generalization across architectures and resolutions are provided in the **Appendix**.

Compared Methods. We benchmark our method against several SOTA approaches for improving CLIP, including: (1) RADIOv2.5 [19], an aggregation method that distills diverse vision models into a single one. For a fair comparison, we reproduced it on ImageNet-1K following its protocol: using CLIP and DINOv2 as teachers and a student model that matches the CLIP vision encoder in both architecture and initialization; (2) un^{2}CLIP [35], a generative-based approach that employs unCLIP [46] to inject fine-grained semantics into CLIP; and (3) KUEA [13], a kernel-based alignment method which is also trained on ImageNet-1K that employs kernel matrix alignment to mitigate the disruption of the original space. All compared methods maintain identical model sizes.

Evaluated Tasks and Benchmarks. We conduct evaluations across multiple benchmarks:

- Object Recognition: (1) General Objects (ImageNet [10], CIFAR-10/100 [29], Caltech-101 [12]); (2) Fine-grained Domains (Pets [44], DTD [7], EuroSAT [20], FER2013 [14]); (3) Specialized Domains (RESISC45 [6], PCam [57], GTSRB [53]); and (4) Distribution Shifts (ImageNet-O [21], ImageNetV2 [49]).
- Fine-grained Understanding: (1) Object Counting and Spatial Reasoning (CLEVR [26]), and (2) Complex Number Recognition (SVHN [42]).
- Dense Prediction: We perform linear probing for semantic segmentation on ADE20K [71], Cityscapes [8], PASCAL VOC 2012 [11], COCO-Stuff [3], and PASCAL Context [40].
- MLLM Integration: We perform our experiments using LLaVA-1.5-7B [37]. For fair comparison, following [13], all methods skip the feature alignment step and utilize the original SFT dataset for visual instruction tuning via LoRA [22]. The evaluation benchmarks include: (1) Open-ended VQA: VQA v2 [15]; (2) Diagram Understanding: AI2D [28];

Table 3: Linear probing classification performance on various benchmarks. We report Top-1 accuracy (%) on 14 datasets. The best results are highlighted in **bold**. † indicates the model reproduced by us on ImageNet-1K.

Method	ImageNet	CIFAR-10	CIFAR-100	Pets	SVHN	Caltech-101	RESISC45	PCam	DTD	EuroSAT	GTSRB	CLEVR D.	FER2013	CLEVR C.	Average
CLIP	80.09	97.42	85.49	94.73	78.5	95.97	96.01	84.46	80.58	97.14	92.76	55.88	71.92	75.62	84.76
RADIOv2.5†	80.43	97.30	86.36	94.71	80.56	96.31	96.00	84.48	81.34	97.05	91.14	56.08	72.11	76.86	85.05
un^{2}CLIP	80.02	97.02	84.51	93.76	**81.63**	**96.53**	95.82	84.14	80.37	97.70	92.5	**60.59**	72.21	77.43	85.30
KUEA	80.72	98.13	87.83	94.93	81.48	96.49	96.25	84.99	**81.86**	97.40	93.63	57.07	72.12	77.83	85.77
SALM-Self (Ours)	***81.57***	*98.16*	*88.23*	*95.11*	*81.40*	*96.42*	*96.06*	*84.59*	*81.55*	*97.02*	*93.39*	*57.01*	*72.39*	*77.85*	*85.77*
SALM (Ours)	80.94	**98.17**	**88.28**	**95.20**	81.61	96.45	**96.90**	**85.49**	**81.86**	**97.73**	**93.66**	57.88	**72.50**	**79.75**	**86.17**

Table 4: Performance evaluation on MLLM benchmarks. We compare the performance of different visual encoders when integrated into the LLaVA. "+FT" indicates that the model was fine-tuned using the same SFT data.

Method	AI2D	POPE	TallyQA	VSR	RefCOCO	RefCOCO+	RefCOCOg	VQA v2	Average
LLaVA-7B	54.08	86.89	62.14	51.47	54.86	49.55	51.04	76.54	60.82
+ FT	53.95	86.90	62.45	52.70	66.42	58.64	58.93	77.23	64.65
un^{2}CLIP	52.92	86.93	62.75	51.47	65.42	60.66	59.82	77.40	64.68
KUEA	53.33	**87.30**	61.21	52.04	66.94	60.74	60.48	77.27	64.92
SALM-Self (Ours)	*54.19*	*86.92*	*62.53*	*53.44*	*66.99*	*60.74*	*61.37*	*76.99*	*65.40*
SALM (Ours)	**54.31**	87.06	**63.12**	**54.58**	**69.63**	**63.68**	**63.48**	**77.68**	**66.69**

(3) Spatial Awareness: RefCOCO, RefCOCO+, RefCOCOg [27, 68], and VSR [36]; (4) Counting Ability: TallyQA [1]; and (5) Hallucination Evaluation: POPE [34].

4.2 Evaluation on Vision-centric Tasks

Zero-shot Object Recognition. As presented in Table 1, we evaluate zero-shot classification performance on ImageNet-1K and 11 downstream datasets. Our method outperforms all baselines, achieving an average accuracy of 69.25% across the 11 downstream tasks. First, the distillation strategy used by RADIOv2.5, particularly with limited training data, yields only marginal gains and frequently fails to retain zero-shot capabilities. This confirms that simple multi-model aggregation is often insufficient. Second, compared to KUEA, we not only achieve higher accuracy on ImageNet-1K but also surpass it in average accuracy. This indicates that the kernel-based method can only achieve sub-optimal alignment. In contrast, our methods capture the fine-grained geometric structures of vision-centric models more effectively. Finally, we notice that un^{2}CLIP exhibits alignment degradation across multiple datasets. This suggests that its generative strategy disrupts the distribution stability of the original feature space. Conversely, SALM effectively enhances visual representation while preserving the integrity of the original distribution and CLIP's zero-shot generalization.

Fine-grained Understanding. We further evaluate fine-grained understanding capabilities in Table 2. Our method achieves the best performance on the geometry-sensitive CLEVR benchmarks and the SVHN recognition task. Specifically, it reaches 22.81% on CLEVR Counts, significantly surpassing all baselines. This directly demonstrates that the geometric constraints introduced via DMA successfully endow the model with the core capability to handle local spatial relationships and object counting, effectively bridging the perceptual gap inherent in CLIP. In contrast, un^{2}CLIP suffers from severe performance degradation on SVHN, further corroborating that its generative injection strategy disrupts the stability of the original feature distribution. Conversely, our approach maintains superior robustness while significantly enhancing structural understanding.

Dense Prediction. To validate the quality of local dense features, we conducted linear probing semantic segmentation experiments on multiple benchmark datasets (see Table 2, right). For instance, it achieves substantial improvements on COCO-Stuff and VOC2012, reaching 37.00 and 77.15 mIoU, respectively. Notably, the kernel-based alignment method, KUEA, performs slightly worse than the original CLIP on this task. This aligns with our expectations: while kernel methods effectively align global semantic distributions, they neglect the intrinsic local structural information within samples.

Furthermore, although other methods attempt to enhance representations through various strategies, none match our level of dense prediction capability. SALM-Self consistently improves segmentation performance without an external teacher, demonstrating the value of CLIP's internal structural cues. SALM further improves performance by incorporating vision-centric priors.

Linear Probing. To further validate the robustness and linear separability of the learned visual representations under frozen parameters, we evaluated the model's classification performance without reliance on textual priors. We followed the linear probing protocol, where the backbone is frozen and only a linear head is trained. As shown in Table 3, our method achieved an average accuracy of 86.17%, outperforming existing methods designed for enhancing VLM visual representations. In particular, the improvement on CLEVR Counts suggests that counting-related information is more readily accessible through a linear readout, complementing the gains on general object recognition. This further corroborates that our enhanced visual representations capture richer semantic structural information, yielding high-quality features that are both more discriminative and fine-grained.

4.3 Multimodal Large Language Model Evaluation

To evaluate the efficacy of our aligned encoder on MLLMs, we integrated the SALM-aligned encoder into the LLaVA framework, benchmarking it against the vanilla encoder and the kernel-based KUEA method. For a fair comparison, the fine-tuned LLaVA baseline (+FT) and all enhanced-encoder variants use the same visual instruction-tuning dataset and finetuning protocol. As reported in Table 4, our approach delivers an SOTA average score of 66.69%, surpassing both the fine-tuned LLaVA baseline and KUEA by a significant margin. It is worth noting that, enabled by the explicit constraints on local geometric information via our dual-matrix alignment strategy, our model demonstrates superior performance in tasks requiring spatial reasoning. Specifically, compared to KUEA, our method improves by 2.69%, 2.94%, and 3.00% on RefCOCO, RefCOCO+, and RefCOCOg, respectively. Moreover, the performance boosts on VSR and TallyQA further confirm that SALM effectively addresses CLIP's limitations in fine-grained perception and local structure capture, successfully elevating overall multimodal understanding.

4.4 Ablation Studies

Effectiveness of Proposed Components. As shown in Table 5, our method outperforms both the original CLIP baseline and the naive fine-tuning strategy. Here, "+FT" denotes finetuning CLIP on ImageNet-1K using the class-name prompt "This is a photo of {class name}". Its limited gains suggest that additional finetuning alone is insufficient to achieve comparable improvements, highlighting the importance of the proposed alignment objectives. Specifically, DMA improves local structural representations, increasing ADE20K segmentation performance from 36.96 to 42.07 mIoU. Meanwhile, CGA improves zero-shot classification performance through implicit semantic alignment, with particularly strong gains on PCam. Combining both components yields the best results across all evaluated tasks in the table, supporting their complementary roles in enhancing fine-grained understanding and dense prediction. More comprehensive comparisons and ablations are provided in the **Appendix**, including direct feature distillation, the individual contributions of the two DMA matrices, and the effects of the training curriculum and self-reference layer selection.

Table 5: Ablation study on the effectiveness of proposed components. We analyze the impact of the DMA and CGA modules. "+FT" represents finetuning CLIP solely on ImageNet-1K with the text "This is a photo of {class name}".

Method	ZS. Cls.				LP. Seg.
	RESISC	PCam	SVHN	CLEVR Distance	ADE20K
CLIP	63.78	60.72	55.97	15.81	36.96
+FT	63.76	60.80	55.99	15.85	36.84
w/ DMA	63.48	63.00	<u>57.51</u>	<u>16.13</u>	<u>42.07</u>
w/ CGA	<u>63.83</u>	<u>64.46</u>	56.77	15.99	39.80
Ours	**64.63**	**67.54**	**57.99**	**16.21**	**42.83**

Impact of Data Scale. We evaluate the scalability of our method by varying the data size. As illustrated in Figure 5, we observe consistent performance gains as the data volume increases. Notably, using only 10% of the training data already yields clear improvements over the original CLIP, indicating that useful structural knowledge can be transferred with limited data. The continued gains with additional training images further demonstrate the data efficiency and scalability of our method within the evaluated range.

Impact of Masking Ratio. We vary the masking ratio to analyze its impact. Results in Figure 5 demonstrate that SALM maintains strong performance across a broad range of partial masking ratios. Lower ratios may allow greater reliance on visible reference features, whereas complete masking removes the reference context that provides spatial anchors for reconstruction. The similar performance across intermediate ratios suggests that the improvements do not depend on a narrowly tuned masking configuration. A masking ratio of 75% achieves the best performance in our evaluation, balancing the reconstruction challenge with sufficient contextual guidance.

4.5 Visualization Analysis

We visualize the PCA-based feature maps of different encoders in Figure 1. The original CLIP features exhibit coarse and scattered activation patterns with blurred boundaries, confirming its limitation in capturing local geometric details. While DINOv2 demonstrates superior structural coherence, previous alignment methods like un^{2}CLIP and KUEA struggle to fully reconstruct this spatial precision, often retaining significant background noise or ambiguous edges. In contrast, our method generates feature maps with remarkably clear semantic layouts, effectively distinguishing foreground objects from the background.

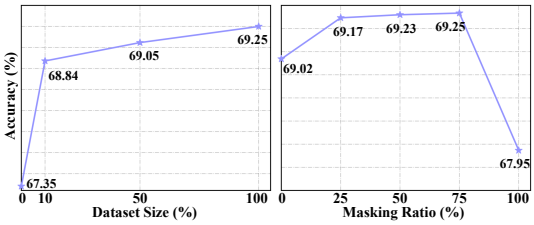

Figure 5: Ablation analysis of data scale and masking ratio evaluated on ZS. datasets. Left: Performance scales consistently with increasing data volume. Right: SALM exhibits robustness across masking ratios, with 75% striking an optimal balance for capturing fine-grained semantics while retaining necessary spatial anchors.

5 Conclusions

In this paper, we propose SALM, a novel unsupervised alignment framework that effectively reconciles the tension between the robust global semantics of CLIP and the fine-grained structural priors of vision-centric models. Unlike prior kernel-based or distillation approaches, SALM synergizes the explicit geometric constraints of Dual-Matrix Alignment with the implicit reconstruction of latent masked modeling. This dual-pathway mechanism successfully injects intricate geometric details into the CLIP embedding space without compromising its pre-trained manifold integrity. Extensive experiments validate that SALM yields consistent gains in zero-shot tasks while achieving superior performance on dense prediction benchmarks. Furthermore, our aligned encoder significantly empowers MLLMs with versatile fine-grained capabilities, ranging from precise object counting to complex spatial reasoning, in a cost-effective manner. We hope this work establishes an efficient paradigm for evolving foundational vision encoders towards more unified multimodal understanding. Future work will explore adaptive structural supervision and stronger patch-text alignment to better preserve global semantics and support open-vocabulary dense prediction.

Acknowledgments and Disclosure of Funding

This work was supported by National Natural Science Foundation of China (NSFC) 62672187 and GuangDong Basic and Applied Basic Research Foundation 2025B1515120037. The authors declare no competing interests.

References

[1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8076–8084, 2019.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang

Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] 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.

[4] Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. In *Forty-second International Conference on Machine Learning*, 2025.

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[6] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[7] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[9] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[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] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Pattern Recognition Workshop*, 2004.

[13] Shizhan Gong, Yankai Jiang, Qi Dou, and Farzan Farnia. Kernel-based unsupervised embedding alignment for enhanced visual representation in vision-language models. *arXiv preprint arXiv:2506.02557*, 2025.

[14] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *Neural information processing: 20th international conference, ICONIP 2013, daegu, korea, november 3-7, 2013. Proceedings, Part III 20*, pages 117–124. Springer, 2013.

[15] 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.

[16] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer, 2024.

[17] Boyu Han, Qianqian Xu, Shilong Bao, Zhiyong Yang, Ruochen Cui, Xilin Zhao, and Qingming Huang. Guiding diffusion-based reconstruction with contrastive signals for balanced visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[19] Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2. 5: Improved baselines for agglomerative vision foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22487–22497, 2025.

[20] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 204–207. IEEE, 2018.

[21] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021.

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[23] Yuhang Huang, Jiazhao Zhang, Shilong Zou, Xinwang Liu, Ruizhen Hu, and Kai Xu. Ladi-wm: A latent diffusion-based world model for predictive manipulation. In *CoRL*, 2025.

[24] Mohammad Jalali, Bahar Dibaei Nia, and Farzan Farnia. Towards an explainable comparison and alignment of feature embeddings. In *Forty-second International Conference on MachineLearning*, 2025.

[25] Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023.

[26] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[27] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[28] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016.

[29] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[30] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 70–88. Springer, 2024.

[31] Junho Lee, Jeongwoo Shin, Hyungwook Choi, and Joonseok Lee. Latent diffusion models with masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17422–17431, 2025.

[32] Martha Lewis, Nihal Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1487–1500, 2024.

[33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[34] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[35] Yinqi Li, Jiahe Zhao, Hong Chang, RuiBing Hou, Shiguang Shan, and Xilin Chen. un^{2}CLIP: Improving CLIP's visual detail capturing ability via inverting unCLIP. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[36] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

[37] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[38] Haoran Lou, Chunxiao Fan, Ziyan Liu, Yuexin Wu, and Xinliang Wang. Llava-sp: Enhancing visual representation with visual spatial tokens for mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22014–22024, 2025.

[39] Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, BINGYUE PENG, and XIAOJUAN QI. Unitok: a unified tokenizer for visual generation and understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[40] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014.

[41] Ilan Naiman, Emanuel Ben-Baruch, Oron Anschel, Alon Shoshan, Igor Kviatkovsky, Manoj Aggarwal, and Gerard Medioni. Lv-mae: Learning long video representations through masked-embedding autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[42] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.

[43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification.

[44] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[45] 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.

[46] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[47] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12490–12500, June 2024.

[48] Mike Ranzinger, Greg Heinrich, Collin McCarthy, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. C-radiov4 (tech report). *arXiv preprint arXiv:2601.17237*, 2026.

[49] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[50] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *ICML*, 2024.

[51] Cheng Shi, Yizhou Yu, and Sibei Yang. Vision transformers need more than registers. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2026.

[52] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.

[53] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012.

[54] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

[55] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13019–13029, 2024.

[56] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024.

[57] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. June 2018.

[58] Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. Declip: Decoupled learning for open-vocabulary dense perception. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14824–14834. IEEE, 2025.

[59] Wenxuan Wang, Quan Sun, Fan Zhang, Yepeng Tang, Jing Liu, and Xinlong Wang. Diffusion feedback helps CLIP see better. In *The Thirteenth International Conference on Learning Representations*, 2025.

[60] XuDong Wang, Xingyi Zhou, Alireza Fathi, Trevor Darrell, and Cordelia Schmid. Visual lexicon: Rich image features in language space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19736–19747, 2025.

[61] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[62] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. VILA-u: a unified foundation model integrating visual understanding and generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[63] Youqi WU, Jingwei Zhang, and Farzan Farnia. When kernels multiply, clusters unify: Fusing embeddings with the kronecker product. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[64] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pages 320–337. Springer, 2024.

[65] Yin Xie, Kaicheng Yang, Xiang An, Kun Wu, Yongle Zhao, Weimo Deng, Zimin Ran, Yumeng Wang, Ziyong Feng, Roy Miles, et al. Region-based cluster discrimination for visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1793–1803, 2025.

[66] Zipeng Yan, Yinjie Chen, Chong Zhou, Bo Dai, and Andrew Luo. Vision transformers with self-distilled registers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[67] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[68] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016.

[69] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[70] Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 35:27127–27139, 2022.

[71] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

[72] Hao Zhou, Zhanning Gao, Zhili Chen, Maosheng Ye, Qifeng Chen, Tongyi Cao, and Honggang Qi. Hints of prompt: Enhancing visual representation for multimodal llms in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6165–6175, 2025.

[73] Chenchen Zhu, Saksham Suri, Cijo Jose, Maxime Oquab, Marc Szafraniec, Wei Wen, Yunyang Xiong, Patrick Labatut, Piotr Bojanowski, Raghuraman Krishnamoorthi, et al. Efficient universal perception encoder. *arXiv preprint arXiv:2603.22387*, 2026.

[74] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Unlocking Fine-Grained Perception in CLIP via Structurally-Aware Latent Masked Modeling (Appendix)

Contents

A More Implementation Details **15**

A.1 Regularization 15

A.2 Three-Stage Curriculum Strategy 15

A.3 Experimental Setup 16

B More Experimental Results **16**

B.1 Generalization across Architectures and Resolutions 16

B.2 Compare with Simple Feature Distillation 17

B.3 Impact of Three-Stage Curriculum Strategy 18

B.4 Decoupled Ablation Analysis of Dual-Matrix Alignment 19

B.5 Reference Layer Selection for SALM-Self 19

B.6 Evaluation on Open-Vocabulary Segmentation 19

B.7 Evaluation on MMVP VLM Benchmark 20

B.8 Evaluation on Zero-shot Image-Text Retrieval 21

B.9 Visualization of Global Semantic Features. 21

B.10 More PCA Visualizations 21

C Limitations **21**

D Broader Impact **22**

A More Implementation Details

A.1 Regularization

To prevent catastrophic forgetting and maintain a healthy feature space distribution, we introduce two regularization terms:

- **Reference Regularization [50, 13]:** We freeze the weights of the original CLIP model to serve as a reference. The global features of the current model are constrained using MSE losses to ensure semantic consistency.
- **Uniformity Loss [70]:** Applied to the CLS token, this loss prevents feature collapse and maintains the uniformity of the feature distribution.

A.2 Three-Stage Curriculum Strategy

To effectively balance the difficulty between structural learning and feature reconstruction, we design a dynamic three-stage curriculum strategy that adjusts the weights of $\mathcal{L}_{\text{dma}}$ and $\mathcal{L}_{\text{recon}}$ (denoted as γ and β) over the course of the training. In the first stage, Structural Warm-up (1 epoch), we exclusively enable DMA to prioritize learning DINO's macroscopic spatial geometric structures without interference from specific feature values. Subsequently, in the Progressive Transition stage (2 epochs), we employ linear interpolation to gradually decay γ from 1.0 to 0.1 while simultaneously increasing β from 0.0 to 1.0. This phase smoothly introduces the more challenging reconstruction task while gradually relaxing the strong global structural constraints. Finally, during the Fine-grained Alignment stage (1 epoch), the training focuses on precise feature reconstruction dominated by MRA ($\beta = 1.0$), with DMA kept at a minimal weight ($\gamma = 0.1$) as an auxiliary constraint to achieve

high-fidelity alignment with DINO features. **Notably, we emphasize that our core alignment framework is inherently robust**; even without this curriculum strategy, the model achieves significant performance gains over the baseline CLIP across all evaluated tasks. As demonstrated in our ablation studies (see Section B.3), while the three-stage schedule is not a prerequisite for convergence, it serves as a highly effective optimization heuristic that further stabilizes the training trajectory and unlocks the model's peak fine-grained potential. It is worth noting that for the SALM-Self variant, since the reconstruction loss is removed, we simply apply constant loss weights throughout the entire training process.

A.3 Experimental Setup

We trained our model on the ImageNet-1K dataset using 4 NVIDIA RTX 4090 24GB GPUs. The detailed hyperparameter configurations are listed in Table 6.

Table 6: Detailed Hyperparameter Configurations.

Hyperparameter	**Value / Description**
Dataset	ImageNet-1K (Train split)
Default Input Resolution	336×336
Global Batch Size	16
Cross-Guided Adapter Layers	4
Optimizer	AdamW
Learning Rate	1×10^{-5}
Weight Decay	0.05
LR Scheduler	Cosine Annealing
Total Training Epochs	4
Reg Loss Weight	5.0
Uni Loss Weight	0.01

B More Experimental Results

B.1 Generalization across Architectures and Resolutions

To comprehensively evaluate the generalization and flexibility of our proposed framework, we conducted extensive experiments on various combinations of VLMs and Vision-Centric Models. We aim to verify whether the framework can effectively adapt to diverse model architectures, as well as varying input resolutions and feature map sizes. Specifically, we introduced several representative model variants in our experimental setup: regarding VLMs, in addition to the standard OpenAI CLIP (ViT-L-14 and ViT-L-14-336) [45], we also tested the architecturally distinct and more powerful Google SigLIP (so400m-Patch14-224) [69]. For Vision-Centric Models, we utilized MAE (ViT-MAE-Large) [18], DINOv2 (ViT-L/14 with Registers) [9] and the latest DINOv3 (VIT-L/16-Pretrain-lvd1689m) [52].

As shown in Table 7, different model combinations result in significant discrepancies in input resolutions and feature map sizes. For instance, the input resolution for CLIP is 336^2 (feature map is 24^2), whereas for SigLIP it is 224^2 (feature map size is 16^2); in the CLIP + DINOv3 combination, the feature map sizes are 24^2 and 21^2, respectively, and similarly, in the CLIP + MAE combination, the feature map sizes are 16^2 and 14^2. To address the inconsistency between the feature map resolutions of the source and target models, our framework demonstrates exceptional adaptability: we resize the features of the source model via interpolation to match the feature dimensions of the target model, thereby achieving effective feature alignment and fusion. Furthermore, to address discrepancies in feature embedding dimensions between the two encoders, we project the source model's features onto the target dimension via a linear layer.

Experimental results indicate that our method significantly enhances the zero-shot performance of VLMs, regardless of variations in base architecture or resolution. By incorporating MAE as the reference model, the average accuracy of the baseline CLIP (ViT-L-14) across 11 datasets improves from 66.64% to 67.17%. After introducing DINOv2 as the reference model, the average accuracy

Table 7: Generalization and flexibility analysis. We report the performance across various model combinations. The columns *Input Res.* and *Feat. Map* denotes the input resolution and the extracted feature map size, respectively.

Method	Settings		ImageNet	Zero-Shot Performance											
	Input Res.	Feat. Map		CIFAR-10	CIFAR-100	Pets	Caltech-101	RESISC45	PCam	DTD	EuroSAT	ImageNet-O	FER2013	ImageNetV2	Average
CLIP (ViT-L-14)	224^2	16^2	75.53	95.59	75.77	93.18	83.27	63.36	51.98	55.42	62.53	32.25	49.86	69.84	66.64
CLIP (ViT-L-14-336)	336^2	24^2	76.55	94.92	74.33	93.68	83.43	63.78	60.72	55.69	61.50	32.75	49.11	70.89	67.35
SigLIP	224^2	16^2	82.02	96.74	84.12	95.33	86.05	69.68	50.58	71.01	62.98	29.55	49.83	76.06	70.18
CLIP (L-14) + MAE	224^2 / 224^2	$\mathbf{16^2}$/$\mathbf{14^2}$	76.13	95.63	77.29	93.51	84.29	63.48	50.4	54.95	61.85	36.6	50.68	70.23	**67.17**
CLIP (L-14) + DINOv2	224^2 / 224^2	16^2 / 16^2	76.00	96.42	77.19	93.62	84.38	64.40	53.13	56.06	62.59	36.75	50.15	69.86	**67.69**
CLIP + DINOv2	336^2 / 336^2	24^2 / 24^2	77.21	96.38	77.87	94.09	84.04	64.63	67.54	56.27	62.19	37.05	50.11	71.53	**69.25**
SigLIP + DINOv2	224^2 / 224^2	16^2 / 16^2	82.38	97.63	84.55	95.67	86.26	69.13	51.04	71.48	63.68	36.00	50.84	76.74	**71.18**
CLIP + DINOv3	336^2 / 336^2	$\mathbf{24^2}$/$\mathbf{21^2}$	77.23	96.16	77.75	93.79	83.89	64.47	67.07	55.90	60.22	37.55	52.79	71.49	**69.19**

Table 8: Ablation analysis of simple feature distillation strategy and three-stage curriculum strategy on zero-shot benchmarks. We report Top-1 accuracy (%) on ImageNet-1K and 11 other datasets. The best results are highlighted in **bold**.

Method		Zero-Shot											
	ImageNet	CIFAR-10	CIFAR-100	Pets	Caltech-101	RESISC45	PCam	DTD	EuroSAT	ImageNet-O	FER2013	ImageNetV2	Average
CLIP	76.55	94.92	74.33	93.68	83.43	63.78	60.72	55.69	61.50	32.75	49.11	70.89	67.35
Simple Distillation	77.06	95.94	77.40	93.13	83.09	64.39	58.26	55.85	60.92	**37.90**	49.27	71.39	67.96
Ours w/o 3-stage	77.02	96.30	77.24	93.97	**84.07**	64.10	62.89	55.95	62.16	36.65	49.81	71.50	68.60
Ours w/ 3-stage	**77.21**	**96.38**	**77.87**	**94.09**	84.04	**64.63**	**67.54**	**56.27**	**62.19**	37.05	**50.11**	**71.53**	**69.25**

of CLIP across 11 datasets improved from 67.35% to 69.25%; the more advanced SigLIP also saw an improvement from 70.18% to 71.18%, demonstrating the universality of our framework across different VLMs. Even when replacing the model with DINOv3 and facing the challenge of mismatched feature map sizes, the CLIP + DINOv3 combination still achieved an outstanding performance of 69.19%, surpassing the original CLIP. Notably, despite the spatial interpolation required by this structural mismatch, DINOv3 exhibits distinct advantages in handling distribution shifts and specialized fine-grained domains. Specifically, it outperforms the DINOv2 counterpart on ImageNet-O (37.55% *v.s.* 37.05%) and FER2013 (52.79% *v.s.* 50.11%). These results compellingly demonstrate the high flexibility and generalization capability of our framework. It not only allows for the seamless replacement of various VLMs and Vision-Centric Models but also effectively resolves resolution discrepancies between models through interpolation strategies, consistently maintaining stable performance gains and leveraging the unique strengths of different vision encoders.

B.2 Compare with Simple Feature Distillation

To further verify the necessity of our proposed dual-matrix alignment and masked reconstruction alignment modules, we compared SALM with a naive feature distillation baseline. In this baseline, the source CLIP features are first mapped via a linear projection layer to match the target dimension. We then directly minimize the Mean Squared Error (MSE) between the projected CLIP features and the target DINOv2 features at both the CLS token and patch token levels, while incorporating a regularization term $\mathcal{L}_{\text{reg}}$ to preserve the original semantic space. The objective is defined as:

$$\mathcal{L} = \|h_\omega(\mathbf{Z}^{\text{src}}) - \mathbf{Z}^{\text{ref}}\|_2^2 + \alpha\mathcal{L}_{\text{reg}},$$

where h_ω denotes the linear projection layer.

As shown in Table 8 and 9, experimental results reveal the limitations of relying solely on point-wise feature alignment. While simple distillation achieves high performance on the ADE20K semantic segmentation task, indicating that strong constraints can inject DINO information, this comes at the cost of the model's general semantic representation capability. This trade-off is particularly evident in the PCam and SVHN task, where the accuracy of simple distillation drops significantly from the original CLIP's 55.97% to 51.87%, suggesting that overly rigid feature approximation disrupts CLIP's

Table 9: Ablation analysis of simple feature distillation strategy and three-stage curriculum strategy on zero-shot fine-grained understanding and dense prediction. We report Top-1 accuracy (%) for zero-shot fine-grained understanding tasks, alongside mean IoU (mIoU) for semantic segmentation.

Method	**ZS. Fine-grained Tasks**			**LP. Seg.**
	SVHN	CLEVR Distance	CLEVR Counts	ADE20K
CLIP	55.97	15.81	20.01	36.96
Simple Distillation	51.87	15.92	22.63	**43.09**
Ours w/o 3-stage	<u>57.95</u>	<u>15.93</u>	<u>22.78</u>	42.29
Ours w/ 3-stage	**57.99**	**16.21**	**22.81**	<u>42.83</u>

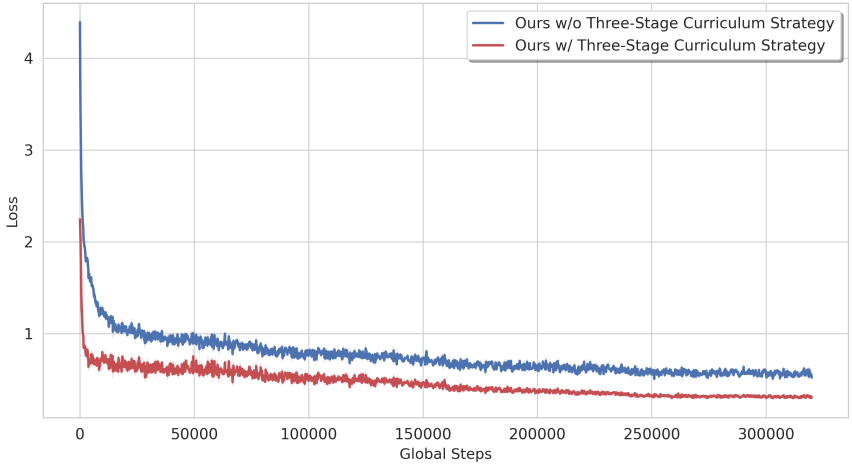

Figure 6: Ablation analysis of the training schedule. We compare the weighted total objective loss curves of models trained with and without the three-stage curriculum strategy. Both settings successfully converge, showing that the model can still train and converge stably when the curriculum strategy is removed.

pre-trained semantic manifold and leads to catastrophic forgetting. In contrast, SALM maintains a high level of performance on dense prediction tasks comparable to simple distillation, while significantly outperforming simple distillation in fine-grained understanding and overall zero-shot classification. This demonstrates that SALM's dual-matrix alignment and masked reconstruction mechanisms are not redundant; rather than simply replicating the teacher model's values, they learn its structure and contextual relationships to inject fine-grained details while effectively protecting and enhancing the original multimodal semantic space, thereby achieving an optimal balance between accuracy and robustness.

B.3 Impact of Three-Stage Curriculum Strategy

Our default training pipeline employs a three-stage curriculum strategy (Structural Warm-up $\rightarrow$ Progressive Transition $\rightarrow$ Fine-grained Alignment) to stabilize the optimization process. To assess the sensitivity of our framework to this training schedule, we conducted an ablation study where the model was trained without the curriculum strategy, utilizing constant loss weights (α, β, γ) from the beginning. Experimental results in Figure 6 demonstrate that SALM is robust to the removal of the curriculum strategy, successfully converging without divergence. As shown in Table 8 and 9, however, we observe a slight performance degradation in accuracy compared to the full three-stage training. This indicates that while the curriculum strategy is not strictly mandatory for convergence, it serves as an effective optimization heuristic. Prioritizing coarse-grained structural constraints before introducing fine-grained reconstruction objectives helps the model navigate the optimization landscape more smoothly, thereby unlocking the model's peak potential in capturing intricate visual details.

Table 10: Ablation analysis of the Dual-Matrix Alignment (DMA) strategy on various downstream tasks. We decouple the effects of the Spatial Relation Matrix (K) and the Energy Difference Matrix (D).

Method	ZS. Fine-grained Tasks			LP. Seg.
	SVHN	CLEVR Distance	CLEVR Counts	ADE20K
w/o DMA	56.77	15.99	20.99	39.80
w/ K	56.73	<u>16.05</u>	22.17	<u>40.98</u>
w/ D	<u>56.87</u>	16.01	<u>22.39</u>	39.97
w/ DMA	**57.99**	**16.21**	**22.81**	**42.83**

B.4 Decoupled Ablation Analysis of Dual-Matrix Alignment

To further investigate the individual contributions of the Spatial Relation Matrix (K) and the Energy Difference Matrix (D) within the Dual-Matrix Alignment (DMA) strategy, we conducted additional decoupled ablation experiments across various downstream tasks. As shown in Table 10, the experimental results thoroughly validate the complementarity of these two matrices in capturing different dimensions of geometric priors.

In dense prediction tasks that impose stringent requirements on local coherence (semantic segmentation on ADE20K), the model incorporating solely the Spatial Relation Matrix (w/ K) achieved an mIoU of 40.98, significantly outperforming the variant relying exclusively on the Energy Difference Matrix (w/ D), which obtained 39.97. This indicates that by constraining the cosine similarity between features, the K matrix successfully injects geometric priors essential for understanding local topological structures and delineating precise object boundaries. The advantage of the D matrix becomes more pronounced in the counting task, scoring 22.39 compared to 22.17 for the K matrix. The core challenge in counting tasks lies in isolating spatially connected or visually similar objects into distinct individual entities. By characterizing the variance in feature norms, the D matrix accurately captures the relative saliency ranking of local regions, thereby effectively facilitating the separation of discrete instances.

Finally, a striking synergistic effect is observed in the complex number recognition task (SVHN), which is characterized by extremely noisy backgrounds. Relying solely on either the K matrix or the D matrix resulted in a performance bottleneck, achieving results comparable to or slightly lower than the baseline model with only implicit reconstruction (w/ CGA, 56.77). However, the combination of both matrices (w/ DMA) yields a substantial performance leap to 57.99. This comprehensively demonstrates that in highly distractive environments, the model intrinsically requires the D matrix to suppress background noise and focus on salient targets, while simultaneously depending on the K matrix to discern fine-grained geometric topologies.

B.5 Reference Layer Selection for SALM-Self

SALM-Self utilizes self-generated shallow features as the reference to provide local structural guidance without relying on external vision models. The choice of reference layer is important, as shallow layers contain richer local details but may lack semantic abstraction, while deeper layers gradually aggregate global semantics and lose fine-grained spatial information. Therefore, we investigate different CLIP layers as the self-reference feature to analyze the trade-off between local structure preservation and semantic consistency.

As shown in Table 11, we evaluate different layers of CLIP as the reference feature. The 6th layer achieves the best overall performance, obtaining 77.13% Top-1 accuracy on ImageNet-1K and 68.60% average performance across evaluation benchmarks. Meanwhile, other nearby layers achieve comparable results, indicating that SALM-Self is not highly sensitive to the exact layer selection. These results suggest that intermediate layers provide an effective balance between retaining local visual structures and maintaining semantic consistency.

B.6 Evaluation on Open-Vocabulary Segmentation

Although SALM is not specifically designed for open-vocabulary segmentation, we evaluate it in this setting to examine whether the enhanced local representations remain compatible with CLIP's original text space. Specifically, we compare SALM with CLIP and DeCLIP under the open-vocabulary

Table 11: Ablation study on the reference layer selection for SALM-Self.

Reference Layer	ImageNet-1K	Avg.
2	76.90	68.57
4	77.06	68.35
6	**77.13**	**68.60**
8	77.07	68.43
10	76.95	68.58

Table 12: Open-vocabulary segmentation results. We report mIoU (%) on five segmentation benchmarks.

Method	VOC2012	Context	COCO-Stuff	Cityscapes	ADE20K
CLIP	14.79	4.06	2.04	1.28	1.43
Simple Distillation	2.07	0.29	0.05	0.16	0.02
DeCLIP	**85.16**	**39.37**	**28.75**	**33.13**	**21.92**
SALM (Ours)	17.03	7.75	4.22	3.08	3.34

Table 13: Performance of CLIP on MMVP-VLM benchmark. Symbols for visual patterns are inherited: 🧭: Orientation and Direction, 🔍: Presence of Specific Features, 🔄: State and Condition, 🔢: Quantity and Count, 📍: Positional and Relational Context, 🎨: Color and Appearance, ⚙️: Structural and Physical Characteristics, 🅰️: Texts, 📷: Viewpoint and Perspective.

Method	🧭	🔍	🔄	🔢	📍	🎨	⚙️	🅰️	📷	Average
CLIP	0.0	20.0	40.0	**20.0**	6.7	20.0	**33.3**	6.7	**33.3**	20.0
KUEA	6.6	**26.7**	40.0	13.3	6.7	40.0	26.6	**13.3**	20.0	21.5
Ours	**13.3**	20.0	**46.7**	13.3	**13.3**	**53.3**	**33.3**	**13.3**	26.7	**25.9**

segmentation protocol. As shown in Table 12, SALM consistently improves over the original CLIP across all evaluated datasets, indicating that enhancing local visual structures does not compromise the compatibility between CLIP visual features and text embeddings. However, SALM remains inferior to DeCLIP, which achieves substantially better performance by explicitly modeling content/context features and optimizing dense alignment for open-vocabulary perception. We emphasize that this result does not indicate that SALM explicitly improves patch-text alignment, as SALM does not introduce region-level or patch-level text supervision. Instead, the improvement mainly comes from enhancing the intrinsic spatial structure and semantic quality of CLIP representations.

B.7 Evaluation on MMVP VLM Benchmark

To further evaluate the model's capability in capturing fine-grained visual details and spatial relationships, we conducted evaluations on the MMVP(Multimodal Visual Patterns)-VLM benchmark [56], which is specifically optimized for VLMs by simplifying language descriptions to isolate visual discrimination capabilities. It is organized into 9 distinct visual patterns, with each pattern containing 15 pairs of zero-shot questions. Unlike standard classification metrics, MMVP-VLM specifically challenges the VLM's ability to distinguish between visually similar images that require precise structural understanding, effectively probing the "visual blind spots" often found in CLIP-like models.

Table 13 presents the quantitative comparison on the MMVP-VLM benchmark. Our method achieves an average accuracy of 25.9%, significantly outperforming both the CLIP baseline (20.0%) and the competitive method KUEA (21.5%). This substantial improvement (+5.9% over CLIP) validates that distilling fine-grained features effectively mitigates the "visual blind spots" inherent in CLIP. Specifically, we observe remarkable gains in categories requiring precise spatial and attribute discrimination. Most notably, in the *Orientation and Direction* category (🧭), where the baseline completely fails (0.0%), our method successfully recovers valid predictions, achieving an accuracy of 13.3%. Furthermore, our approach demonstrates superior capability in capturing *Color and Appearance* (🎨), improving accuracy from 20.0% to 53.3%, and doubles the performance in *Positional and Relational Context* (📍) from 6.7% to 13.3%. These results confirm that the structural knowledge learned through our proposed Dual Matrix Alignment and Masked Reconstruction tasks significantly enhances the model's fine-grained visual reasoning capabilities.

Table 14: Zero-shot image-text retrieval results on Flickr30K and MSCOCO benchmarks. We report Recall@K (K=1, 5, 10) for both Image-to-Text and Text-to-Image retrieval tasks.

Method	Image-to-Text Retrieval						Text-to-Image Retrieval					
	Flickr30K			MSCOCO			Flickr30K			MSCOCO		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	**87.30**	98.30	**99.30**	57.92	81.20	87.86	67.34	89.02	93.28	37.07	61.64	71.51
KUEA	**87.30**	98.10	**99.30**	58.26	81.24	88.28	68.70	**89.80**	93.84	38.05	62.64	72.60
Ours	86.70	**98.40**	99.10	**59.12**	**82.08**	**88.36**	**68.88**	**89.80**	**94.20**	**38.54**	**63.25**	**73.10**

B.8 Evaluation on Zero-shot Image-Text Retrieval

We further conducted zero-shot image-text retrieval experiments to verify the fine-grained semantic alignment between visual and textual representations. This task evaluates the quality of the feature space directly, ensuring that semantically similar images and texts are projected closely. We utilized two standard benchmarks for this assessment: Flickr30K [67] and MSCOCO [5]. As reported in Table 14, our method achieves superior or competitive performance compared to CLIP and KUEA across both Image-to-Text and Text-to-Image retrieval tasks. Specifically, our model demonstrates robust improvements on the MSCOCO dataset, confirming that the learned representations maintain high fidelity even without task-specific fine-tuning.

B.9 Visualization of Global Semantic Features.

To further evaluate the quality of the learned representation space, we visualize the distribution of global semantic features using t-SNE in Figure 7. As illustrated, the embedding clusters of the original CLIP and recent alignment methods appear relatively dispersed, with ambiguous boundaries between adjacent semantic categories. This looseness suggests a latent instability in distinguishing semantically similar samples. In sharp contrast, our method produces a significantly more structured feature space, characterized by high intra-class compactness and clear inter-class separability. This evidence confirms that SALM's alignment strategy effectively integrates fine-grained geometric priors, and it refines the global semantic distribution, resulting in more robust and discriminative representations.

B.10 More PCA Visualizations

As shown in Figure 8, we present additional PCA visualization results to further evaluate the performance of different models at the feature manifold level. It can be observed that the feature distribution of vanilla CLIP exhibits fragmented, blocky artifacts accompanied by substantial background noise, reflecting its inherent limitations in capturing fine-grained local structures. Although existing visual enhancement methods (un^{2}CLIP and KUEA) show marginal improvements, they still suffer from noticeable background interference and ambiguous semantic boundaries. In contrast, our proposed SALM-Self method exhibits remarkable noise reduction capabilities and clear semantic layouts under zero external model dependency. Furthermore, the complete SALM model, which incorporates external priors, achieves structural coherence most akin to DINOv2, accomplishing not only highly pure foreground-background separation but also exceptionally sharp local semantic boundaries, comprehensively validating the remarkable effectiveness of our proposed alignment framework in reshaping high-quality, fine-grained features.

C Limitations

While SALM effectively bridges the gap between global semantics and local structures, it possesses certain inherent limitations. First, the alignment performance is fundamentally capped by the representation quality and granularity of the vision-centric models used as references. Although SALM is model-agnostic and generalizes well across various teachers (*e.g.*, DINOv2/v3, MAE), any structural noise or lack of specific semantic details in the target model's latent space will inevitably be inherited during the distillation process. In scenarios where the target model fails to capture essential geometric priors, the enhancement of CLIP's fine-grained perception may reach a "representation ceiling." Second, the efficacy of our framework is constrained by the intrinsic architectural capacity and pre-trained stability of the VLM backbone. While we unlock fine-grained perception capabilities

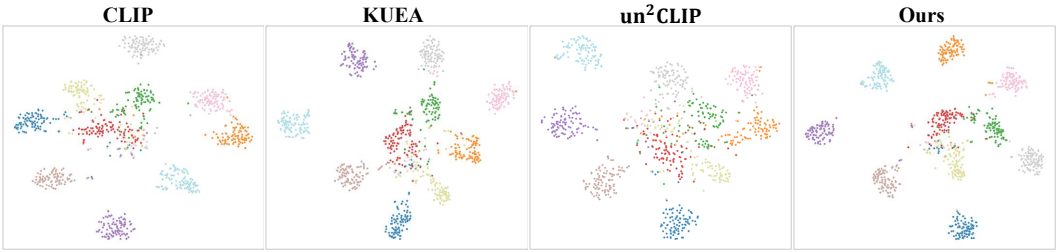

Figure 7: The t-SNE visualization of global semantic features. Compared to other methods, our approach generates more compact and well-separated clusters, demonstrating superior capability in learning discriminative global representations.

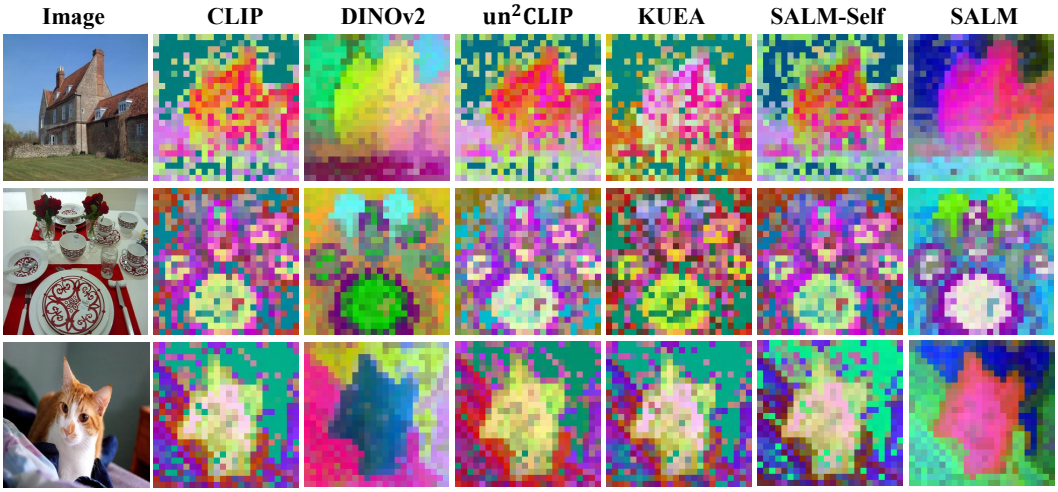

Figure 8: PCA-based visualizations. Unlike other enhancement methods, our approach excels at capturing fine-grained semantic nuances and preserving discriminative local details.

within CLIP's features, the fixed parameter count and the necessity of preserving the original multimodal semantic manifold impose a trade-off. Over-injecting dense structural information might eventually risk saturating the representation space, potentially causing a minor interference with the model's global reasoning in extreme edge cases. Finally, although our SALM-Self paradigm mitigates the dependency on external experts, it still operates within the boundaries of the VLM's existing pre-trained knowledge. Future research could investigate more advanced feature decoupling mechanisms, allowing the VLM to autonomously acquire out-of-distribution geometric priors without risking global semantic collapse.

D Broader Impact

This paper presents SALM, a framework designed to enhance the fine-grained visual perception of Vision-Language Models. By bridging the gap between global semantics and local structure, our approach has the potential to significantly advance real-world applications where precision is paramount. For instance, in assistive technologies, the improved capability to perform spatial reasoning and accurate object counting can lead to more reliable navigation aids for the visually impaired. Similarly, in embodied AI and autonomous systems, robust topological understanding is essential for safer interaction with complex environments, helping to mitigate perceptual hallucinations in safety-critical scenarios. While our method improves technical accuracy, it relies on pre-trained foundation models and may inherit their underlying biases. Furthermore, the heightened capacity for fine-grained visual recognition could potentially be misused for invasive surveillance, unauthorized tracking, or extracting sensitive personal attributes from images. We encourage the community to deploy these enhanced capabilities responsibly, particularly regarding user privacy and fairness.

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 presents the main idea, highlighting its innovative aspects that distinguish it from other works. The introduction clearly states the main contributions, which are supported by both method introductions (Section 3) and experimental results (Section 4). The claims align with the results presented in the paper, and the scope is well-defined.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Due to space constraints in the main text, the main body of the paper does not include a dedicated section on limitations and we have discussed limitations and several failure cases in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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: [N/A]

Justification: This paper does not involves theoretical result.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

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: We provide all the information needed about the main experiments in the experimental section (section 4) and appendix (section B). All necessary steps to reproduce the experiments are included either in the main text or in the supplementary material. The source code to be released later can reproduce the reported results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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: [Yes]

   Justification: The dataset benchmarks in the experiment, as well as the pre-trained vision model we used, are all open source. We will release the code and preprocessed data later.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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: The paper relies on public datasets which have provided the information of the data splits along with the datasets. We specify all the training and testing details in the experimental section (section 4) and appendix (section B).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

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 follow the convention in prior works and report the performance number.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: We provide sufficient information on the computer resources in the appendix (section B).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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 research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss both potential positive societal impacts and negative societal impacts in the and appendix.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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 paper does not describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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 creators or original owners of the assets (e.g., code, data, models) used in the paper, and explicitly mention and respect the corresponding licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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: [No]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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: The paper does not involve research with human subjects requiring IRB approvals.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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: This paper describe the usage of LLMs in the section of Experiment (4).

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.