Learning Fine-Grained Vision-Language Alignment from Discriminative Part Descriptions

Sibo Yin[1] **and Yuxin Peng**[1,†]
[1]Wangxuan Institute of Computer Technology, Peking University
yinsibo@stu.pku.edu.cn pengyuxin@pku.edu.cn

Abstract

Vision-Language Pre-trained (VLP) models such as CLIP learn strong representations from large-scale image–text pairs and demonstrate impressive zero-shot transfer. However, by learning to align visual and textual features only at a global level, they often suffer from limited interpretability and weak fine-grained perception. This issue stems from pre-training data that overlooks key visual details of object parts, which are often important for distinguishing between subordinate categories (e.g., species of birds or models of cars). Multimodal Large Language Models (MLLMs) built on CLIP-style vision encoders inherit this weakness, limiting both accuracy and trustworthiness. To address this, we propose Part-Aware CLIP (*PA-CLIP*), a framework that improves fine-grained perception while enhancing interpretability. First, we leverage MLLMs to construct a new dataset, *FG-Part*, containing about 850K image-text pairs that explicitly describe discriminative components (e.g., beak shape or wing patterns). Second, we introduce a part-aware training strategy that encourages explicit grounding of fine-grained textual descriptions to corresponding image regions, strengthening part-level cross-modal alignment. Extensive experiments show that PA-CLIP achieves state-of-the-art results on multiple fine-grained visual recognition benchmarks, validating the benefit of part-level captions for capturing subtle details. Moreover, evaluations on general tasks such as cross-modal retrieval indicate that these improvements do not compromise the model's core generalist capabilities.

1 Introduction

The advent of large-scale pre-training has marked a paradigm shift in artificial intelligence, with Vision-Language Pre-trained (VLP) models leading the charge in multimodal understanding. Models such as CLIP (Contrastive Language-Image Pre-training) [Radford et al., 2021], SigLIP [Zhai et al., 2023] have demonstrated remarkable success by learning rich, transferable visual representations from hundreds of millions of image-text pairs from the web. Demonstrating robust zero-shot capabilities in visual understanding, these models serve as a powerful foundation for a wide range of downstream tasks, such as image classification [Radford et al., 2021, Hu et al., 2025], cross-modal retrieval [Csizmadia et al., 2025]. Notably, in recent years, influential MLLMs such as the LLaVA [Liu et al., 2023, 2024] and MiniGPT [Zhu et al., 2023, Chen et al., 2023] series have widely adopted pre-trained CLIP models as their vision encoder to extract visual representations.

Despite their impressive capabilities, CLIP series models exhibit a critical weakness when dealing with tasks that demand a fine-grained understanding of visual information [Xie et al., 2025a, Asokan et al., 2025]. By learning to align images and text at a coarse, global level, CLIP and its successors often fail at fine-grained visual recognition (FGVR)–the task of distinguishing between subordinate categories, such as species of birds, models of cars, or types of aircraft [Welinder et al., 2010,

†: Corresponding author.

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

Maji et al., 2013]. This deficiency arises as the pre-training process overlooks the subtle, part-level visual details (e.g., the shape of a bird's beak, the curvature of a car's headlights) that are often the sole differentiators between fine-grained classes. Recent works on improving the fine-grained understanding abilities of CLIP models have primarily concentrated on two main directions. The first approach [Zhang et al., 2024a, Xie et al., 2025a] focuses on extending the text token sequence beyond the 77-token limit of the original CLIP models. This allows the model to process longer and more detailed captions during the training phase, which contains more textual information. A second, more direct strategy involves fine-tuning the model on curated, fine-grained image-text datasets [Xie et al., 2025a, Xiao et al., 2024]. This method provides explicit fine-grained supervision, teaching the model to associate specific visual features with their corresponding textual descriptions.

However, while these approaches have incrementally improved the fine-grained capabilities of CLIP, they still suffer from two significant drawbacks: a persistent lack of interpretability [Rasekh et al., 2024] and insufficient fine-grained understanding. On challenging FGVR benchmarks, their accuracy still lags considerably behind models designed specifically for these tasks. The primary reason for this persistent gap is that even the fine-grained image-text data remains fundamentally "coarse-grained." Although previous methods provide longer, more detailed text, these descriptions often focus on the holistic scene rather than the discriminative, part-level attributes. For instance, the critical visual distinction between two visually similar bird species, the 'Great Crested Flycatcher' and the 'White-browed Chat-tyrant', may lie solely in a subtle pattern on the wing feathers—a minute detail that is decisive for correct classification. Existing image-text datasets, even those curated for fine-grained tasks, are unable to consistently capture such subtle, part-level visual differences in their textual descriptions. Consequently, models trained on this data are not equipped with the necessary supervision to learn these critical features, leading to an inability to reliably discern between visually similar fine-grained categories.

To overcome these fundamental limitations, we introduce Part-Aware CLIP (PA-CLIP), a novel framework that systematically enhances both the fine-grained perception and interpretability of VLP models. First, we propose a scalable data generation pipeline which utilizes powerful MLLMs to generate fine-grained image-text data. This pipeline allows us to construct a large-scale dataset named FG-Part, which contains approximately 850K image-text pairs. Unlike previous long textual captions, our generated descriptions function as explicit textual pointers to discriminative, part-level visual evidence (e.g., "The throat of the bird exhibits a pale bluish-gray color, with a relatively smooth texture"), capturing the kind of subtle features that are essential for expert-level recognition but absent in existing datasets.

Second, building upon this curated dataset, we design a part-aware fine-tuning strategy. This strategy explicitly compels the model to ground these nuanced textual descriptions in their corresponding local image regions, forcing an alignment between specific words and the visual features they describe. This process improves the model's understanding of part-level semantics, improving both its fine-grained recognition accuracy and its inherent interpretability, as the model's attention is now guided by discriminative object characteristics.

On several challenging FGVR benchmarks, our method consistently surpasses existing state-of-the-art approaches. Concurrently, experiments on general vision-language benchmarks, such as cross-modal retrieval, demonstrate that PA-CLIP maintains highly competitive performance with existing methods. Experimental results prove that our approach enhances fine-grained perception while preserving the versatile, generalist capabilities of the foundational CLIP architecture.

2 Related Work

2.1 Contrastive Vision-Language Pre-training

Learning visual representations from natural language supervision has become a dominant paradigm in computer vision. This approach was popularized by models like CLIP [Radford et al., 2021] and ALIGN [Jia et al., 2021], which pioneered the use of contrastive learning on a large scale. The resulting models learn robust and generalizable representations that enable impressive zero-shot performance on a wide array of classification tasks without direct supervision. Following this initial success, research has largely advanced in two main directions. One dominant trend has been to scale up the training data. The creation of massive public datasets, most notably the LAION series [Schuhmann et al., 2022], provided billions of image-text pairs. Building upon this larger data

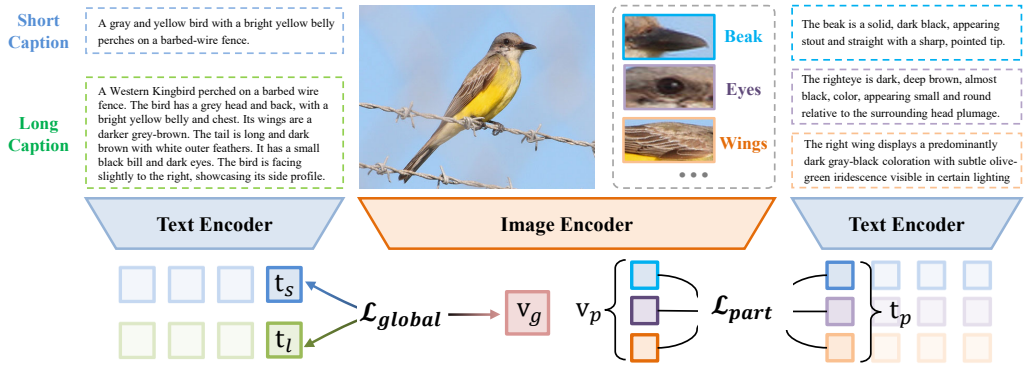


Figure 1: An overview of our proposed PA-CLIP framework. The global loss ($\mathcal{L}_{\text{global}}$) aligns the global image representation ($\mathbf{v}_g$) with embeddings from both short ($\mathbf{t}_s$) and long ($\mathbf{t}_l$) captions. Concurrently, the part-level loss ($\mathcal{L}_{\text{part}}$) aligns dedicated part-level visual features ($\mathbf{v}_p$) with their corresponding textual embeddings ($\mathbf{t}_p$).

enabled the development of more powerful models like OpenCLIP [Cherti et al., 2023]. A second direction has focused on improving the quality of the textual supervision. Recognizing that noisy web captions are often short and lack detail, some works [Zhang et al., 2024a, Asokan et al., 2025] have sought to improve CLIP's fine-grained understanding by using longer and more detailed text. This is often done by employing sophisticated captioning models to generate longer, more detailed sentences for the entire image or regions within images, with the goal of providing a richer and more textual information.

2.2 Fine-grained Visual Recognition

While recognizing general objects like "car" or "bird" is a largely solved problem, the next frontier is fine-grained understanding: the ability to perceive and reason about subtle, detailed visual information. This capability is essential for real-world applications, from identifying specific plant diseases in agriculture to describing nuanced product details in e-commerce [Zhang et al., 2024b]. As discussed, the VLP models that form the basis of today's MLLMs [Peng et al., 2026] lack this crucial skill due to their global-level pre-training. Fine-grained visual recognition (FGVR) is essential for fine-grained understanding, as the model needs to accurately identify the fine-grained categories of objects before performing fine-grained analysis (e.g., visual question answering). Historically, research in FGVR has followed two main paths. The first involved specialized models that were explicitly designed to perform fine-grained image classification [Zheng et al., 2019, Ji et al., 2020, Sun et al., 2023a]. Based on CNN or transformers architecture, these methods show excellent performance on specific datasets. The second path involves using large Vision-Language Models for classification. The standard approach is to perform zero-shot classification by crafting a set of text prompts, such as "a photo of a {class name}" [Radford et al., 2021]. The model then classifies an image by comparing its embedding to the embeddings of the different prompts.

3 Methodology

In this section, we introduce our proposed Part-Aware CLIP (PA-CLIP), a novel framework designed to inject fine-grained perception and interpretability into existing VLP models. Our approach tackles the inherent limitations of models like CLIP by explicitly grounding textual descriptions of object parts to their corresponding visual regions. This is achieved through a multi-level contrastive learning objective powered by a meticulously constructed dataset of part-level image-text pairs. The overall architecture of PA-CLIP is depicted in Figure 1.

3.1 PA-CLIP

Following previous work [Xie et al., 2025a, Xiao et al., 2024], we first extend the limitation of text token sequence in CLIP models from 77 to 248, enabling models to capture more detailed textual information. Our PA-CLIP framework adopts the same architecture as the original CLIP, comprising

a text encoder and an image encoder, and introduces a specialized fine-tuning stage. This stage is orchestrated by a multi-level contrastive learning objective, which operates on both global image-text pairs and fine-grained, part-level correspondences. This dual-level supervision ensures that the model not only retains its robust general-purpose knowledge but also develops a deep understanding of subtle visual details critical for fine-grained tasks.

3.1.1 Image and Text Encoders

We adopt the standard dual-encoder architecture for our PA-CLIP, comprising an image encoder $f_I(\cdot)$ and a text encoder $f_T(\cdot)$. For the image encoder, we primarily utilize a Vision Transformer (ViT) [Dosovitskiy et al., 2020] architecture. When processing a full image x, the encoder $f_I(x)$ produces a global embedding $\mathbf{v}_g \in \mathbb{R}^D$ from its output [CLS] token, capturing the holistic content of the scene. The text encoder $f_T(\cdot)$ processes textual inputs (short captions, long captions, part descriptions) to produce corresponding D-dimensional text embeddings $\mathbf{t} \in \mathbb{R}^D$. All embeddings are L2-normalized before computing similarities.

3.1.2 Multi-Level Contrastive Loss

Our training objective for PA-CLIP is designed to simultaneously learn global image-text alignments and fine-grained, part-level correspondences. This is achieved by combining distinct contrastive loss components. For a given mini-batch of N image-text sets, the total loss $\mathcal{L}$ is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{global}} + \mathcal{L}_{\text{part}} \tag{1}$$

where $\mathcal{L}_{\text{global}}$ is the global image-text contrastive loss and $\mathcal{L}_{\text{part}}$ is the part-level contrastive loss.

Global Contrastive Loss. To ensure PA-CLIP retains its broad zero-shot capabilities, we incorporate a global contrastive loss that considers two levels of global textual descriptions: short captions and long captions. For each image x_i in the batch, we construct a short caption $y_{i,s}$ and a long caption $y_{i,l}$. Then, the corresponding text embeddings are $\mathbf{t}_{i,s} = f_T(y_{i,s})$ and $\mathbf{t}_{i,l} = f_T(y_{i,l})$. The global image embedding is $\mathbf{v}_{i,g} = f_I(x_i)_{\text{CLS}}$.

The global contrastive loss can be then expressed as:

$$\mathcal{L}_{\text{global}} = \mathcal{L}_{\text{CLIP}}(\mathbf{v}_g, \mathbf{t}_s) + \mathcal{L}_{\text{CLIP}}(\mathbf{v}_g, \mathbf{t}_l) \tag{2}$$

where $\mathcal{L}_{\text{CLIP}}(\mathbf{V}, \mathbf{T})$ is a standard symmetric contrastive loss function. Given a batch of N image embeddings $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ and N text embeddings $\mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_N\}$, this loss is defined as:

$$\mathcal{L}_{\text{CLIP}}(\mathbf{V}, \mathbf{T}) = \frac{1}{2N} \sum_{i=1}^{N} \left[-\log \frac{\exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{v}_i, \mathbf{t}_j)/\tau)} - \log \frac{\exp(\text{sim}(\mathbf{t}_i, \mathbf{v}_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{t}_i, \mathbf{v}_j)/\tau)} \right] \tag{3}$$

Here, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, and τ is a learnable temperature parameter.

Part-Level Contrastive Loss. While the global loss ensures the model understands the overall scene, the part-level loss is specifically designed to focus on the fine-grained visual details within it. This component enforces a direct alignment between nuanced textual descriptions and their specific visual counterparts at the sub-object level. To achieve this, for each image x_i annotated with K_i object parts, we generate a set of dedicated part-specific visual embeddings.

Specifically, for each part k of image x_i, defined by its bounding box $\text{BBox}_{i,p}^k$, we first crop this region from the original high-resolution image, yielding a part-specific image patch $x_{i,p}^k$. This patch is then resized to the standard input resolution of the image encoder and processed in a separate forward pass. The CLS token embedding from this pass serves as the dedicated visual feature for that part:

$$\mathbf{v}_{i,p}^k = f_I(x_{i,p}^k)_{\text{CLS}} \tag{4}$$

This approach yields a feature vector $\mathbf{v}_{i,p}^k$ that represents the part in isolation, which is crucial for learning subtle, context-independent visual attributes. The corresponding part-level text description $y_{i,p}^k$ is then encoded by the text encoder to produce its embedding $\mathbf{t}_{i,p}^k = f_T(y_{i,p}^k)$.

The part-level contrastive loss $\mathcal{L}_{\text{part}}$ is then computed by encouraging each part's visual feature $\mathbf{v}_{i,p}^k$ to be close to its matching textual description $\mathbf{t}_{i,p}^k$, and distant from all other part descriptions in the

Table 1: Comparison of FGVR performance on 7 benchmarks. Top-1 accuracy is reported. **Best** is in bold and second best is underlined within each backbone group. Δ_{SOTA} denotes absolute gains of PA-CLIP over the best prior method for the same backbone.

Backbone	Method	Venue	CUB	Aircraft	Dogs	Pets	Nabirds	Flowers	Cars	Average
ViT-B/16	CLIP Radford et al. [2021]	ICML 2021	48.1	23.4	57.1	84.4	42.4	64.5	61.7	54.5
	EVA-CLIP Sun et al. [2023b]	arXiv 2023	51.7	20.9	63.4	86.8	49.1	70.4	79.5	60.3
	SigLIP Zhai et al. [2023]	ICCV 2023	50.2	36.7	66.4	87.5	50.5	83.9	**91.1**	66.6
	SigLIP2 Tschannen et al. [2025]	arXiv 2025	59.8	31.7	71.2	89.9	37.3	80.8	88.7	65.6
	Long-CLIP Zhang et al. [2024a]	ECCV 2024	46.4	21.6	54.3	83.1	40.1	62.8	61.0	52.8
	DreamLIP Zheng et al. [2024]	ECCV 2024	33.0	7.9	22.8	64.1	27.2	51.8	22.7	32.8
	FG-CLIP Xie et al. [2025a]	ICML 2025	55.0	22.6	61.8	88.0	48.3	68.0	85.9	61.4
	SmartCLIP Xie et al. [2025b]	CVPR 2025	52.0	21.0	58.1	83.7	40.0	61.2	62.8	54.1
	PA-CLIP (Ours)		**68.8**	**36.9**	**72.6**	**93.8**	**53.4**	**87.8**	87.4	**71.5**
	Δ_{SOTA}		+9.0	+0.2	+1.4	+3.9	+2.9	+3.9	-3.7	+4.9
ViT-L/14	CLIP Radford et al. [2021]	ICML 2021	56.1	32.0	66.4	89.8	51.5	72.0	73.0	63.0
	EVA-CLIP Sun et al. [2023b]	arXiv 2023	61.5	32.6	69.2	87.5	57.6	75.3	90.2	67.7
	SigLIP Zhai et al. [2023]	ICCV 2023	60.1	43.4	74.0	92.7	58.6	84.6	92.4	72.3
	SigLIP2 Tschannen et al. [2025]	arXiv 2025	65.8	39.0	77.9	91.3	46.0	85.0	**93.5**	71.2
	Long-CLIP Zhang et al. [2024a]	ECCV 2024	45.1	27.9	56.2	82.7	47.7	63.2	71.5	56.3
	FG-CLIP Xie et al. [2025a]	ICML 2025	64.2	29.8	69.9	90.4	59.9	70.7	90.0	67.8
	SmartCLIP Xie et al. [2025b]	CVPR 2025	57.0	30.4	61.5	83.5	49.7	64.4	71.5	59.7
	PA-CLIP (Ours)		**78.6**	**52.7**	**80.6**	**96.5**	**66.4**	**92.6**	92.7	**80.0**
	Δ_{SOTA}		+12.8	+9.3	+2.7	+3.8	+6.5	+7.6	-0.8	+7.7

batch. This is a symmetric contrastive loss operating on part-level visual and textual embeddings:

$$\mathcal{L}_{\text{part}} = \mathcal{L}_{\text{CLIP}}(\mathbf{V}_P, \mathbf{T}_P) \tag{5}$$

where $\mathbf{V}_P$ and $\mathbf{T}_P$ represent the concatenated sets of all part visual embeddings and part text embeddings, respectively, within the current mini-batch. By forcing this explicit alignment between cropped part images and their detailed descriptions, PA-CLIP learns to perceive and ground discriminative local features in a way that global-only training cannot achieve.

3.2 FG-Part Dataset

The core challenge in instilling part-level awareness in VLP models is the scarcity of datasets containing explicit, high-quality descriptions of object parts. However, manually annotating hundreds of thousands of image parts with detailed textual descriptions is infeasible. To circumvent this bottleneck and enable scalable training, we developed a sophisticated, automated pipeline for generating a large-scale dataset named FG-Part of part-level image-text pairs using MLLMs.

Part Localization and Extraction. First, we detect the discriminative regions of objects within an image. Specifically:

- **For datasets with existing part annotations:** For established fine-grained datasets such as CUB-200-2011 [Wah et al., 2011] and Birdsnap [Berg et al., 2014], which already provide high-quality human-annotated part-level keypoints (e.g., 'beak', 'wing', 'eyes' for birds), we directly utilize these ground-truth annotations to extract object parts.
- **For datasets without part annotations:** To vastly expand our data sources to datasets that lack part labels (e.g., StanfordCars [Krause et al., 2013] and FGVC-Aircraft [Maji et al., 2013]), we employ an open-vocabulary object detection model, OV-DINO [Wang et al., 2024a]. We first curate a domain-specific *part name list* containing key discriminative components (e.g., for cars: 'logo', 'wheel hub', 'headlight', 'grille'). We then use this list to query the OV-DINO model, which automatically localizes these predefined parts within the images and generates the corresponding bounding boxes.

For each obtained bounding box, we crop the corresponding image region, creating a collection of individual part images for the next stage.

Prompt Design and Description Generation. After extracting part regions, we use MLLMs to generate detailed descriptions for each localized object part. Rather than asking for generic image captions, we design domain-specific prompts that specify the object superclass, candidate part attributes, and several in-context examples, encouraging the model to focus on discriminative visual

Table 2: Comparison of cross-modal retrieval performance on five benchmarks. Recall@1 (R@1) is reported. I2T and T2I denote Image-to-Text and Text-to-Image retrieval, respectively.

	Method	Venue	MSCOCO		DCI		Flickr		DOCCI		Urban1k	
			I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I
ViT-B/16	CLIP Radford et al. [2021]	ICML 2021	51.8	32.7	45.5	43.0	82.2	62.1	–	–	–	–
	EVA-CLIP Sun et al. [2023b]	arXiv 2023	58.7	41.6	41.9	41.2	85.7	71.2	–	–	–	–
	Long-CLIP Zhang et al. [2024a]	ECCV 2024	57.6	40.4	51.7	57.3	85.9	70.7	–	–	78.9	79.5
	FineCLIP Jing et al. [2024]	NeurIPS 2024	54.5	40.2	35.5	34.4	82.5	67.9	78.1	80.0	91.2	90.0
	FG-CLIP Xie et al. [2025a]	ICML 2025	64.1	45.4	**61.8**	**60.6**	**90.7**	**76.4**	**83.9**	**84.4**	**94.0**	**93.4**
	SmartCLIP Xie et al. [2025b]	CVPR 2025	61.9	42.4	53.4	52.8	90.2	72.6	77.4	78.1	90.0	87.4
	PA-CLIP (Ours)		**65.3**	**46.3**	60.8	59.2	88.9	76.1	83.5	83.3	93.9	92.9
ViT-L/14	CLIP Radford et al. [2021]	ICML 2021	58.0	37.1	37.2	36.4	87.4	67.3	–	–	–	–
	EVA-CLIP Sun et al. [2023b]	arXiv 2023	64.2	47.9	47.2	47.8	89.2	77.9	–	–	–	–
	Long-CLIP Zhang et al. [2024a]	ECCV 2024	62.8	46.3	44.2	52.5	90.0	76.2	–	–	82.7	86.1
	FineCLIP Jing et al. [2024]	NeurIPS 2024	–	–	40.1	46.2	–	–	83.7	84.4	94.5	93.9
	FG-CLIP Xie et al. [2025a]	ICML 2025	68.9	50.9	**66.7**	**66.1**	93.7	81.5	**87.4**	**89.3**	96.4	96.5
	SmartCLIP Xie et al. [2025b]	CVPR 2025	66.0	48.5	57.6	58.5	92.1	77.8	81.7	82.6	93.0	90.1
	PA-CLIP (Ours)		**70.3**	**51.4**	65.6	65.6	**94.1**	**81.9**	87.3	87.8	**97.1**	**96.7**

details such as shape, color, texture, and local patterns. A representative prompt template is provided in Appendix F.

Specifically, we utilize Google's Gemma3 27B [Gemma Team, 2025] and Meta's Llama4 Scout [Meta AI, 2025] for part-level description generation. These models are deployed across 16 NVIDIA RTX 5090 and 20 NVIDIA L40 GPUs using the vLLM toolkit [Kwon et al., 2023]. Through this pipeline, we curate FG-Part, a dataset of approximately 850K image-text pairs.

4 Experiments

4.1 Experimental Settings

Table 3: Ablation studies on fine-grained visual recognition. We report the average Top-1 accuracy (%) across 7 FGVR benchmarks for different combinations of long global (#*Long*), short global (#*Short*), and part-level (#*Part*) captions.

#*Long*	#*Short*	#*Part*	Backbone	CUB	Aircraft	Dogs	Pets	Nabirds	Flowers	Cars	Average
✓			ViT-B-16	59.9	25.7	61.2	88.7	53.3	67.6	83.9	**62.9**
	✓		ViT-B-16	60.3	36.8	66.5	89.7	54.4	69.0	87.4	**66.3**
		✓	ViT-B-16	65.2	30.7	70.7	93.9	49.2	84.8	86.9	**68.8**
	✓	✓	ViT-B-16	68.3	34.3	71.4	93.6	53.7	87.1	87.5	**70.8**
✓		✓	ViT-B-16	68.4	29.8	70.6	93.9	53.3	86.5	86.1	**69.8**
✓	✓		ViT-B-16	60.2	36.9	66.0	89.0	54.0	69.1	87.2	**66.1**
✓	✓	✓	ViT-B-16	68.8	36.9	72.6	93.8	53.4	87.8	87.4	**71.5**
✓			ViT-L-14	70.5	35.7	70.0	88.7	63.8	71.6	90.1	**70.1**
	✓		ViT-L-14	70.9	50.8	73.4	89.6	66.3	73.3	91.9	**73.7**
		✓	ViT-L-14	77.6	43.2	78.9	95.8	65.3	89.4	92.1	**77.5**
	✓	✓	ViT-L-14	78.2	50.8	79.5	94.8	67.3	90.9	92.3	**79.1**
✓		✓	ViT-L-14	77.9	40.4	78.2	95.4	66.3	90.7	92.0	**77.3**
✓	✓		ViT-L-14	70.3	49.1	72.6	89.6	65.1	72.7	91.8	**73.0**
✓	✓	✓	ViT-L-14	78.6	52.7	80.6	96.5	66.4	92.6	92.7	**80.0**

Implementation Details. Following prior work [Xie et al., 2025a, Asokan et al., 2025], we initialize models with CLIP weights [Radford et al., 2021] and use ViT-B/16 and ViT-L/14. Stage 1 (FineHARD) uses global batch size 1280, AdamW with 1e-6 learning rate and cosine schedule, plus DeepSpeed ZeRO-2. Stage 2 (FG-Part) uses global batch size 512, AdamW with 1e-6 learning rate, cosine schedule with 300 warm-up steps, and 20 epochs. See Appendix L for full hyperparameters.

Evaluation Tasks and Metrics. We conduct extensive experiments across two tasks: Fine-Grained Visual Recognition and cross-modal retrieval. FGVR is a core challenge that directly measures

a model's ability to distinguish between highly similar subordinate categories. We use the Top-1 Accuracy as the evaluation metric, which measures the percentage of test images correctly classified into their ground-truth category. Cross-Modal Retrieval evaluates the model's general-purpose representation quality for aligning images and text [Wang et al., 2024b]. We report performance on both text-to-image (T2I) and image-to-text retrieval (I2T) setting. The key metric is Recall@1 (R@1), which measures the percentage of queries where the correct corresponding item is ranked as the top result.

Inference. The part-level branch, the open-vocabulary detector (OV-DINO), and the MLLM are used only for constructing FG-Part and computing the training objective; none of them is involved at test time. At inference, PA-CLIP is used in exactly the same way as a standard CLIP model. For zero-shot FGVR, the full image is encoded in a single forward pass and classified by cosine similarity to the text embeddings of class-name prompts; for cross-modal retrieval, each full image and caption is encoded in a single forward pass and ranked by cosine similarity. No part crops, part descriptions, or part annotations are required for any test sample.

4.2 Results

We conduct a comprehensive evaluation of PA-CLIP to demonstrate its capabilities in both specialized and general domains. The primary evaluation of our method's fine-grained perception is conducted on several challenging FGVR benchmarks. Furthermore, to ensure that our improvements in fine-grained perception do not come at the cost of generalist capabilities, we also evaluate our method on a diverse set of cross-modal retrieval tasks.

4.2.1 Fine-Grained Visual Recognition.

For FGVR task, we utilize a wide range of standard FGVR datasets, covering diverse domains from birds to cars. These include CUB-200-2011 (CUB) [Wah et al., 2011], FGVC Aircraft (Aircraft) [Maji et al., 2013], Stanford Dogs (Dogs) [Khosla et al., 2011], Oxford-IIIT Pets (Pets) [Parkhi et al., 2012], NABirds [Van Horn et al., 2015], Oxford Flowers (Flowers) [Nilsback and Zisserman, 2008], and Stanford Cars (Cars) [Krause et al., 2013].

The main results for fine-grained visual recognition are presented in Table 1. Our proposed PA-CLIP demonstrates a substantial and consistent improvement in average accuracy over all baseline and state-of-the-art methods across both ViT-B-16 and ViT-L-14 backbones. Specifically, with the ViT-L-14 backbone, PA-CLIP achieves the state-of-the-art average accuracy of 80.0% across the 7 FGVR datasets. This result surpasses the next best competing method, SigLIP, by 7.7%. With the ViT-B-16 backbone, PA-CLIP achieves an average accuracy of 71.5%, outperforming the strongest baseline (SigLIP at 66.6%) by 4.9%. This consistent superiority across different model scales highlights the effectiveness of our PA-CLIP and FG-Part dataset.

On individual datasets, PA-CLIP sets new SOTA results on six of the seven benchmarks with the ViT-L/14 backbone. The gains are particularly significant on datasets where localized features are critical for recognition. For instance, on Aircraft, PA-CLIP-L-14 achieves 52.7% accuracy, outperforming the CLIP-L-14 baseline by +20.7%. On CUB and Dogs, we observe improvements of 22.5% and 14.2%, respectively. This demonstrates that our method is highly effective at capturing the discriminative, part-level features required for challenging fine-grained tasks.

4.2.2 Cross-Modal Retrieval

To verify that PA-CLIP maintains robust general-purpose representations, we evaluate it on five standard cross-modal retrieval benchmarks covering different domains and caption lengths: MSCOCO 5K [Lin et al., 2014] and Flickr 1K [Young et al., 2014] (short captions, five per image), as well as DOCCI [Onoe et al., 2024], DCI [Urbanek et al., 2024], and Urban1K [Zhang et al., 2024a] (long captions). As shown in Table 2, PA-CLIP achieves competitive performance with state-of-the-art methods. With the ViT-L/14 backbone, it sets new state-of-the-art results on MSCOCO (70.3% I2T, 51.4% T2I) and Flickr, outperforming FG-CLIP. On long-caption benchmarks, PA-CLIP remains highly competitive: it achieves a state-of-the-art 97.1% on Urban1K I2T and is within 0.1% of the best on DOCCI I2T. The results indicate that our model does not lose its ability to understand long text while improving fine-grained perception capability. We further compare PA-CLIP with MLLM-based embedding models, GME [Zhang et al., 2025] and Qwen3-VL-Embedding [Li et al., 2026], in

Table 4: Ablation studies on cross-modal retrieval. We report the average Recall@1 (R@1) across 5 benchmarks for different combinations of long global (#*Long*), short global (#*Short*), and part-level (#*Part*) captions.

Long	*Short*	*Part*	Backbone	MSCOCO		DCI		Flickr		DOCCI		Urban1k		Average
				I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	
✓			ViT-B-16	66.2	46.3	62.2	60.1	91.0	75.9	84.7	85.2	93.2	93.2	**75.8**
	✓		ViT-B-16	64.9	47.6	60.9	61.3	89.5	76.9	84.0	85.2	92.9	92.3	**75.5**
		✓	ViT-B-16	62.2	41.9	57.3	53.2	89.2	73.5	78.0	77.4	90.7	89.0	**71.2**
	✓	✓	ViT-B-16	63.9	44.8	57.6	56.8	88.8	75.1	81.3	81.2	92.7	91.2	**73.3**
✓		✓	ViT-B-16	64.5	43.4	59.7	56.1	90.1	74.4	82.5	81.5	93.6	92.5	**73.8**
✓	✓		ViT-B-16	65.4	47.3	62.4	60.3	89.8	76.6	84.6	85.0	92.9	92.4	**75.7**
✓	✓	✓	ViT-B-16	65.3	46.3	60.8	59.2	88.9	76.1	83.5	83.3	93.9	92.9	**75.0**
✓			ViT-L-14	71.3	51.9	68.6	67.5	94.5	81.9	88.7	90.1	97.4	97.6	**81.0**
	✓		ViT-L-14	70.8	53.4	65.8	67.8	94.0	83.0	87.8	90.4	96.8	96.8	**80.7**
		✓	ViT-L-14	68.1	49.4	62.4	63.5	92.8	79.8	83.5	84.6	95.4	95.3	**77.5**
	✓	✓	ViT-L-14	69.6	51.0	64.3	65.4	94.0	81.5	86.0	87.1	96.6	96.1	**79.2**
✓		✓	ViT-L-14	70.4	50.1	66.0	64.4	94.7	80.7	86.4	86.9	96.6	96.7	**79.3**
✓	✓		ViT-L-14	71.1	53.3	68.1	67.7	94.2	82.8	88.8	90.3	97.3	97.9	**81.1**
✓	✓	✓	ViT-L-14	70.3	51.4	65.6	65.6	94.1	81.9	87.3	87.8	97.1	96.7	**79.8**

Table 5: Generalization to other CLIP-family models. We apply the same PA-CLIP training data, loss function, and training pipeline to different CLIP-family models. FGVR Avg. denotes the average Top-1 accuracy across 7 FGVR benchmarks, and CMR Avg. denotes the average Recall@1 across retrieval benchmarks and both retrieval directions.

Backbone	Base Model	PA-trained Variant	FGVR Avg.			CMR Avg.		
			Original	+ PA Training	Gain	Original	+ PA Training	Gain
ViT-B/16	SigLIP	PA-SigLIP	66.6	77.3	+10.7	62.8	67.3	+4.5
	SigLIP2	PA-SigLIP2	65.6	73.2	+7.6	50.6	57.6	+7.0
	LongCLIP	PA-LongCLIP	52.8	62.1	+9.3	65.3	65.6	+0.3
ViT-L/14	SigLIP	PA-SigLIP	72.3	83.1	+10.8	68.3	73.5	+5.2
	SigLIP2	PA-SigLIP2	71.2	81.8	+10.6	59.3	68.0	+8.7
	LongCLIP	PA-LongCLIP	56.3	68.0	+11.7	67.6	69.0	+1.4

Appendix C. With more than an order of magnitude fewer parameters, PA-CLIP ViT-L/14 achieves an average R@1 comparable to Qwen3-VL-Embedding-8B (79.8% vs. 80.6%) and outperforms GME-7B (72.6%).

4.3 Ablation Studies

We conduct a detailed ablation study to validate the effectiveness PA-CLIP. We report results on both FGVR (Table 3) and cross-modal retrieval (Table 4) tasks. In our experiments, we train the model with different caption types: long captions (#*Long*), short captions (#*Short*), and part-level captions (#*Part*).

Fine-Grained Visual Recognition. The results in Table 3 show that our part-level supervision is the most critical component for FGVR performance. Training with part-level supervision alone yields an average accuracy of 77.5% on the ViT-L-14 model, substantially outperforming models trained only on long (70.1%) or short (73.7%) global captions. Combining short and part supervision further improves performance to 79.1%, indicating a positive synergy between global context and local fine-grained details. Our full model, utilizing all three components, achieves the highest average accuracy of 80.0%, demonstrating that while part-level data is the primary driver, incorporating varied global supervision provides additional benefits.

Cross-Modal Retrieval. The results of ablation studies on cross-modal retrieval are summarized in Table 4. The average performance is highest when training only with global captions (#*Long*, #*Short*, or both). Conversely, using only part-level supervision (#*Part*) leads to a notable performance degradation. This is expected, as part-level descriptions lack the global scene context essential for cross-modal retrieval. For instance, with the ViT-B-16 backbone, the Part-only model's average R@1

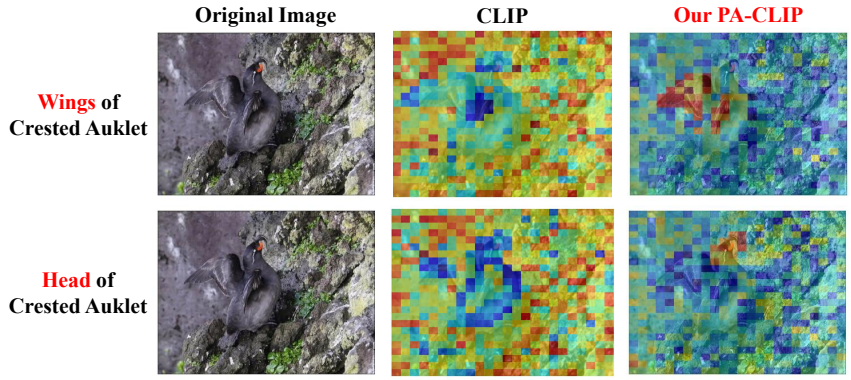


Figure 2: Part-aware grounding visualization based on image–text token cosine similarity. PA-CLIP localizes queried object parts more accurately than CLIP.

drops by 4.6% compared to the Long-only variant (71.2% vs. 75.8%). Crucially, our full model largely mitigates this gap, achieving a competitive performance of 75.0%.

Generalization to Other CLIP-family Models. We further study whether the proposed part-aware training strategy generalizes to different CLIP-family models. To this end, we apply the same training data, loss function, and training pipeline to several representative CLIP-family models, including SigLIP, SigLIP2, and LongCLIP. As shown in Table 5, part-aware training consistently improves FGVR performance across different models and model scales. For example, PA-SigLIP improves the FGVR average accuracy by +10.7% and +10.8% on ViT-B/16 and ViT-L/14, respectively, while PA-LongCLIP also brings +9.3% and +11.7% gains. Moreover, the retrieval results are generally maintained or improved, suggesting that part-aware training enhances fine-grained perception without substantially sacrificing general cross-modal alignment.

Transfer to MLLMs. Since many MLLMs adopt CLIP-style vision encoders as their visual front-end, we further conduct a preliminary transfer study on LLaVA-1.5-7B. Specifically, we replace the original CLIP vision tower with our trained PA-CLIP vision encoder, followed by the standard pre-training and instruction fine-tuning stages. PA-CLIP-LLaVA improves the overall FGVC accuracy from 74.73% to 76.03%. This result suggests that the fine-grained visual representations learned by PA-CLIP can be transferred to MLLMs to further enhance their fine-grained perception capabilities. Full per-dataset results are provided in Appendix B.

Visualization. We visualize part-aware grounding by computing the cosine similarity between image tokens and part-level text tokens. As shown in Figure 2, PA-CLIP produces more localized activations on queried object parts, such as "wings" and "head", while CLIP tends to respond to broader or less relevant regions. More examples are provided in Appendix I. In addition, we visualize the open-vocabulary part detection results used during FG-Part construction in Appendix J.

5 Conclusion

In this paper, we addressed the critical limitations of standard Vision-Language Pre-trained (VLP) models: their poor fine-grained perception and lack of interpretability, which stem from a reliance on global-only image-text supervision. We introduced Part-Aware CLIP (PA-CLIP), a framework which introduces a multi-level contrastive loss to instill part-level awareness into these models. Extensive experiments demonstrate that PA-CLIP achieves new state-of-the-art performance on a comprehensive suite of fine-grained visual recognition benchmarks. In addition, we introduce FG-Part, a novel dataset of approximately 850K image-text pairs with part-level, fine-grained descriptions.

Limitations and Future Work. First, the part-level contrastive loss treats parts as a flat set of crop–description pairs and does not explicitly model hierarchical or compositional relations among parts (e.g., the eye within the head), which are only implicitly captured by global captions. Explicitly modeling part structures, e.g., via part–whole consistency constraints or structure-aware negative sampling, is a promising direction. Second, FG-Part is constructed automatically and inevitably contains residual noise from part localization and MLLM-generated descriptions, and its coverage

is currently limited to a few fine-grained domains. Third, the potential of PA-CLIP as a vision encoder for a wider range of MLLMs remains to be explored, and combining part-aware training with LLM-side alignment methods may further improve fine-grained multimodal reasoning.

Broader Impacts. Our work can benefit fine-grained visual applications such as ecological monitoring and medical imaging, but may also be misused in surveillance. We encourage responsible deployment with proper safeguards, dataset governance, and legal compliance.

Acknowledgments and Disclosure of Funding

This work was supported by the grants from the National Natural Science Foundation of China (62432001, 62525201, 62132001) and Beijing Natural Science Foundation (L247006, L257005). The authors declare no competing interests.

References

Mothilal Asokan, Kebin Wu, and Fatima Albreiki. Finelip: Extending clip's reach via fine-grained alignment with longer text inputs, 2025. URL `https://arxiv.org/abs/2504.01916`.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2011–2018, 2014.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, 2014.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023.

Daniel Csizmadia, Andrei Codreanu, Victor Sim, Vighnesh Prabhu, Michael Lu, Kevin Zhu, Sean O'Brien, and Vasu Sharma. Distill clip (dclip): Enhancing image-text retrieval via cross-modal transformer distillation, 2025. URL `https://arxiv.org/abs/2505.21549`.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2004.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Saihui Hou, Yushan Feng, and Zilei Wang. Vegfru: A domain-specific dataset for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Wenshuai Hu, Wei Li, Hengchao Li, Fenghua Huang, and Ran Tao. Language dynamic-guided spatial-spectral network for cross-domain few-shot hyperspectral image classification. *Chinese Journal of Electronics*, 34(6):1845–1860, 2025.

Ruyi Ji, Longyin Wen, Libo Zhang, Dawei Du, Yanjun Wu, Chen Zhao, Xianglong Liu, and Feiyue Huang. Attention convolutional binary neural tree for fine-grained visual categorization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10468–10477, 2020.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. URL https://arxiv.org/abs/2102.05918.

Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and Zhiwu Lu. Fineclip: Self-distilled region-based clip for better fine-grained understanding. *Advances in Neural Information Processing Systems*, 37:27896–27918, 2024.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Meta AI. Llama 4. https://ai.meta.com/, 2025.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, et al. DOCCI: Descriptions of connected and contrasting images. *arXiv preprint arXiv:2404.19753*, 2024.

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

Yuxin Peng, Zishuo Wang, Geng Li, Xiangtian Zheng, Sibo Yin, and Hulingxiao He. A survey on fine-grained multimodal large language models. *Chinese Journal of Electronics*, 35(2):771–803, 2026.

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.

Ali Rasekh, Sepehr Kazemi Ranjbar, Milad Heidari, and Wolfgang Nejdl. Ecor: Explainable clip for object recognition, 2024. URL https://arxiv.org/abs/2404.12839.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

Hongbo Sun, Xiangteng He, Jiahuan Zhou, and Yuxin Peng. Fine-grained visual prompt learning of vision-language models for image recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5828–5836, 2023a.

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*, 2023b.

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

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

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 595–604, 2015.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Hao Wang, Pengzhen Ren, Zequn Jie, Xiao Dong, Chengjian Feng, Yinlong Qian, Lin Ma, Dongmei Jiang, Yaowei Wang, Xiangyuan Lan, et al. Ov-dino: Unified open-vocabulary detection with language-aware selective fusion. *arXiv preprint arXiv:2407.07844*, 2024a.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: A systematic review of methods and future directions, 2024b. URL https://arxiv.org/abs/2308.14263.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD birds 200. 2010.

xAI. Grok-1.5 vision preview, 2024. URL https://x.ai/blog/grok-1.5v.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

Rui Xiao, Sanghwan Kim, Mariana-Iuliana Georgescu, Zeynep Akata, and Stephan Alaniz. FLAIR: VLM with Fine-grained Language-informed Image Representations, December 2024. URL http://arxiv.org/abs/2412.03561. arXiv:2412.03561 [cs].

Chunyu Xie, Bin Wang, Fanjing Kong, Jincheng Li, Dawei Liang, Gengshen Zhang, Dawei Leng, and Yuhui Yin. Fg-clip: Fine-grained visual and textual alignment, 2025a. URL https://arxiv.org/abs/2505.05071.

Shaoan Xie, Lingjing Kong, Yujia Zheng, Yu Yao, Zeyu Tang, Eric P Xing, Guangyi Chen, and Kun Zhang. Smartclip: Modular vision-language alignment with identification guarantees. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29780–29790, 2025b.

Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3973–3981, 2015.

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.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL `https://arxiv.org/abs/2303.15343`.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pages 310–325. Springer, 2024a.

Ruru Zhang, E Haihong, Lifei Yuan, Yanhui Wang, Lifei Wang, and Meina Song. Fgm-spcl: Open-set recognition network for medical images based on fine-grained data mixture and spatial position constraint loss. *Chinese Journal of Electronics*, 33(4):1023–1033, 2024b.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9274–9285, 2025.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024c.

Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, Jiebo Luo, and Tao Mei. Learning rich part hierarchies with progressive attention networks for fine-grained image recognition. *IEEE Transactions on Image Processing*, 29:476–488, 2019.

Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions, 2024. URL `https://arxiv.org/abs/2403.17007`.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Generalization to Other CLIP-family Models

In the main paper, we report a compact summary of the generalization results on other CLIP-family models. Here, we provide the full per-dataset results on both fine-grained visual recognition and cross-modal retrieval. For each CLIP-family model, we apply the same PA-CLIP training data, loss function, and training pipeline, and compare the resulting PA-trained variant with its original counterpart. The goal of this analysis is not to introduce new baselines, but to verify whether the proposed part-aware training strategy consistently generalizes to different CLIP-family models.

Table 6 reports the detailed FGVR results. Across SigLIP, SigLIP2, and LongCLIP, part-aware training consistently improves the average accuracy on both ViT-B/16 and ViT-L/14. The improvements are especially clear on fine-grained object-centric datasets such as CUB, Aircraft, Dogs, and Flowers, indicating that the proposed part-level supervision is broadly useful for enhancing fine-grained perception.

Table 7 further reports the detailed cross-modal retrieval results. Compared with the original CLIP-family models, the PA-trained variants generally maintain or improve retrieval performance. This suggests that the part-aware training strategy improves fine-grained recognition ability without substantially damaging the general image-text alignment learned by the original models. For LongCLIP, we follow the reported settings of the original paper and compute the retrieval average over the available benchmarks excluding DOCCI.

B Transfer to Multimodal Large Language Models

To provide more details for the preliminary MLLM transfer study in the main paper, we report the per-dataset FGVC results in Table 8. We replace the original CLIP vision tower of LLaVA-1.5-7B with our trained PA-CLIP vision encoder, and then perform the standard pre-training and instruction fine-tuning stages. The evaluation is conducted on the complete official test split of each dataset (60,199 images in total over the seven FGVR benchmarks).

General Multimodal Benchmarks. To examine whether replacing the vision encoder affects general multimodal capabilities, we further evaluate both models on MMMU [Yue et al., 2024] (the full 10,500-sample test split) and RealWorldQA [xAI, 2024] (all 765 samples). As shown in Table 9, PA-CLIP-LLaVA performs on par with the original LLaVA (MMMU +0.36, RealWorldQA -0.13). This is expected: MMMU mainly stresses subject knowledge and multi-step reasoning, where the bottleneck lies in the LLM rather than the vision encoder, while RealWorldQA mainly tests scene-level spatial understanding rather than the part-level fine-grained discrimination that PA-CLIP targets. These results show that PA-CLIP improves the fine-grained recognition of downstream MLLMs while preserving their general reasoning and scene understanding.

Discussion on the Magnitude of Gains. The FGVR gain of PA-CLIP-LLaVA (+1.30 on average) is obtained under a controlled setting in which the vision encoder is the only difference, while the LLM, training data, and training recipe are identical. Prior studies have shown that the FGVR bottleneck of MLLMs lies mainly on the LLM side rather than in the vision encoder: Zhang et al. [2024c] find that visually-grounded language models underperform their own CLIP encoders on image classification and attribute this to insufficient decoding of visual information during LLM training, and He et al. [2025] identify the misalignment between object information and category names within the LLM representation space as a root cause. Given this downstream bottleneck, the gain obtained by replacing only the vision encoder suggests that the encoder-level improvement is genuine, and combining PA-CLIP with LLM-side alignment methods is a promising orthogonal direction.

C Comparison with MLLM-based Embedding Models

Recent universal multimodal retrieval models build image and text embeddings on top of MLLMs. We compare PA-CLIP with two representative models of this kind, GME-Qwen2-VL-7B-Instruct [Zhang et al., 2025] and Qwen3-VL-Embedding-8B [Li et al., 2026], using their publicly released checkpoints. Both models are evaluated on the same five cross-modal retrieval benchmarks and with the same protocol as in Table 2. The results are reported in Table 10.

Table 6: Detailed FGVR results of generalization to other CLIP-family models. Top-1 accuracy is reported on 7 FGVR benchmarks. "PA" denotes the model obtained by applying the same PA-CLIP training data, loss function, and training pipeline to the corresponding CLIP-family model.

Backbone	Base Model	Method	CUB	Aircraft	Dogs	Pets	NABirds	Flowers	Cars	Avg.
ViT-B/16	SigLIP	Original	50.2	36.7	66.4	87.5	50.5	83.9	91.1	66.6
		PA-SigLIP	70.2	53.6	77.6	95.8	55.2	96.7	91.7	77.3
		Δ	+20.0	+16.9	+11.2	+8.3	+4.7	+12.8	+0.6	+10.7
	SigLIP2	Original	59.8	31.7	71.2	89.9	37.3	80.8	88.7	65.6
		PA-SigLIP2	67.0	47.1	77.4	95.4	40.5	93.7	91.5	73.2
		Δ	+7.2	+15.4	+6.2	+5.5	+3.2	+12.9	+2.8	+7.6
	LongCLIP	Original	46.4	21.6	54.3	83.1	40.1	62.8	61.0	52.8
		PA-LongCLIP	58.9	28.1	68.7	91.9	44.1	75.2	67.8	62.1
		Δ	+12.5	+6.5	+14.4	+8.8	+4.0	+12.4	+6.8	+9.3
ViT-L/14	SigLIP	Original	60.1	43.4	74.0	92.7	58.6	84.6	92.4	72.3
		PA-SigLIP	78.7	61.8	84.2	96.5	68.7	97.1	94.8	83.1
		Δ	+18.6	+18.4	+10.2	+3.8	+10.1	+12.5	+2.4	+10.8
	SigLIP2	Original	65.8	39.0	77.9	91.3	46.0	85.0	93.5	71.2
		PA-SigLIP2	71.1	71.6	84.4	96.4	57.1	97.0	95.1	81.8
		Δ	+5.3	+32.6	+6.5	+5.1	+11.1	+12.0	+1.6	+10.6
	LongCLIP	Original	45.1	27.9	56.2	82.7	47.7	63.2	71.5	56.3
		PA-LongCLIP	64.4	34.7	73.3	93.8	53.3	79.5	76.8	68.0
		Δ	+19.3	+6.8	+17.1	+11.1	+5.6	+16.3	+5.3	+11.7

PA-CLIP ViT-L/14 outperforms GME-7B on all ten retrieval settings, and achieves an average R@1 comparable to Qwen3-VL-Embedding-8B (79.8% vs. 80.6%) with more than an order of magnitude fewer parameters (0.43B vs. 8B). It also surpasses Qwen3-VL-Embedding-8B on MSCOCO and Flickr in both retrieval directions. Even the ViT-B/16 variant (0.15B) exceeds GME-7B in average R@1 (75.0% vs. 72.6%). Qwen3-VL-Embedding-8B performs better on the long-caption benchmarks (DCI, DOCCI, and Urban1k), likely benefiting from the stronger language modeling capacity of its LLM backbone compared with the CLIP text encoder. Overall, these results show that PA-CLIP remains competitive with much larger MLLM-based embedding models on general cross-modal retrieval, while retaining the efficiency of a lightweight dual-encoder architecture.

D Data Splits and Evaluation Setting

FG-Part is constructed exclusively from the official training splits of its source datasets; no test images, annotations, or part crops of any evaluation benchmark are included. At test time, FGVR is still performed by zero-shot class-name prompting.

To assess the contribution of in-domain images, we conduct two controls with ViT-B/16. First, after removing all CUB-derived samples from FG-Part and retraining, PA-CLIP still achieves 59.6% on CUB, above CLIP (48.1%) and FG-CLIP (55.0%). Second, on CompCars [Yang et al., 2015], which is not used in FG-Part construction, PA-CLIP achieves 32.8% versus 18.6% for CLIP and 29.9% for FG-CLIP.

E Compute-Matched Baseline

To disentangle the effect of the part-aware objective from that of longer training on in-domain images, we extend the Long+Short variant to the same 11,640 iterations, keeping the images, backbone (ViT-B/16), optimizer, and schedule unchanged. As shown in Table 11, four times more iterations improve the average FGVR accuracy by only 0.7 points (66.1% $\rightarrow$ 66.8%) and slightly reduce the average retrieval R@1 (75.7% $\rightarrow$ 74.7%), suggesting that training on global captions alone saturates. In contrast, PA-CLIP reaches 71.5% under the same iteration budget, i.e., +4.7 points over the compute-matched baseline, indicating that the part-aware objective, rather than additional compute, is the primary source of the gains.

Table 7: Detailed cross-modal retrieval results of generalization to other CLIP-family models. Recall@1 is reported. CMR Avg. denotes the average over all reported retrieval scores. For LongCLIP, DOCCI results are not available in the original paper, so the average is computed over MSCOCO, DCI, Flickr, and Urban1k only.

Backbone	Base Model	Method	MSCOCO		DCI		Flickr		DOCCI		Urban1k		Avg.
			I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	
ViT-B/16	SigLIP	Original	65.3	47.2	44.2	41.7	89.0	74.6	70.2	70.6	63.1	62.4	62.8
		PA-SigLIP	68.0	48.4	48.7	46.4	89.2	74.7	76.5	74.6	76.0	70.6	67.3
		Δ	+2.7	+1.2	+4.5	+4.7	+0.2	+0.1	+6.3	+4.0	+12.9	+8.2	+4.5
	SigLIP2	Original	64.8	48.1	29.3	28.8	89.7	73.7	48.4	49.3	38.5	35.3	50.6
		PA-SigLIP2	66.1	49.7	36.1	37.1	90.3	74.9	55.7	58.5	53.0	54.7	57.6
		Δ	+1.3	+1.6	+6.8	+8.3	+0.6	+1.2	+7.3	+9.2	+14.5	+19.4	+7.0
	LongCLIP	Original	57.6	40.4	51.7	57.3	85.9	70.7	–	–	78.9	79.5	65.3†
		PA-LongCLIP	57.7	40.3	52.8	54.8	86.0	69.4	–	–	84.7	78.9	65.6†
		Δ	+0.1	-0.1	+1.1	-2.5	+0.1	-1.3	–	–	+5.8	-0.6	+0.3†
ViT-L/14	SigLIP	Original	68.9	51.2	48.9	47.1	91.5	77.4	76.2	77.0	72.9	71.9	68.3
		PA-SigLIP	72.8	54.0	56.1	53.3	92.7	81.0	82.0	82.8	83.6	76.9	73.5
		Δ	+3.9	+2.8	+7.2	+6.2	+1.2	+3.6	+5.8	+5.8	+10.7	+5.0	+5.2
	SigLIP2	Original	68.2	50.0	39.4	38.7	91.0	76.9	62.1	63.8	54.1	49.1	59.3
		PA-SigLIP2	71.2	54.2	48.3	50.4	92.7	80.7	70.0	72.7	69.6	69.9	68.0
		Δ	+3.1	+4.1	+8.9	+11.7	+1.7	+3.8	+7.9	+8.9	+15.5	+20.8	+8.6
	LongCLIP	Original	62.8	46.3	44.2	52.5	90.0	76.2	–	–	82.7	86.1	67.6†
		PA-LongCLIP	63.8	45.9	45.2	55.7	90.4	75.2	–	–	90.7	84.9	69.0†
		Δ	+1.0	-0.4	+1.0	+3.2	+0.4	-1.0	–	–	+8.0	-1.2	+1.4†

† For LongCLIP, the average is computed over the available retrieval scores excluding DOCCI. Gains are computed before rounding.

Table 8: Per-dataset FGVC results of transferring PA-CLIP to LLaVA-1.5-7B.

Model	CUB	Aircraft	Dogs	Pets	NABirds	Flowers	Cars	Avg.
Ori-CLIP-LLaVA	69.83	65.41	70.45	78.50	73.67	70.74	94.52	74.73
PA-CLIP-LLaVA	70.19	66.70	70.41	84.19	75.21	70.26	95.25	76.03
Δ	+0.36	+1.29	-0.04	+5.69	+1.54	-0.48	+0.73	+1.30

F Prompt Template for Part-level Description Generation

We provide the prompt template used for generating part-level descriptions in FG-Part. For each object category, we specify the object superclass, a list of candidate part attributes, and several in-context examples. The goal is to guide the MLLM to produce descriptions that focus on localized and discriminative visual details rather than holistic image-level captions. A representative prompt for aircraft images is shown below.

Object SuperClass: Aircraft

Object Attributes: Wings, Engines, Tail

Examples:

- **#1** The aircraft's engines are mounted under the wings, with large circular intakes and smooth metallic nacelles...

- **#2** This aircraft's vertical tail is tall and swept back, featuring a painted livery with a sharp trailing edge...

- **#3** The wings are long and slightly swept, with upturned winglets at the tips...

Table 9: Results of transferring PA-CLIP to LLaVA-1.5-7B on general multimodal benchmarks. We report per-discipline and overall accuracy (%) on MMMU and overall accuracy (%) on RealWorldQA.

Model	MMMU							RealWorldQA
	Art & Design	Business	Science	Health & Medicine	Human. & Social Sci.	Tech & Eng.	Overall	
Ori-CLIP-LLaVA	44.11	26.68	25.56	33.79	45.83	28.48	31.74	54.38
PA-CLIP-LLaVA	45.14	24.86	25.56	33.05	46.36	30.60	32.10	54.25
Δ	+1.03	-1.82	0.00	-0.74	+0.53	+2.12	+0.36	-0.13

Table 10: Comparison with MLLM-based embedding models, GME [Zhang et al., 2025] and Qwen3-VL-Embedding [Li et al., 2026], on cross-modal retrieval. Recall@1 is reported. Avg. denotes the average over all ten retrieval scores. The best result in each column is shown in bold.

Method	Backbone	#Params	MSCOCO		DCI		Flickr		DOCCI		Urban1k		Avg.
			I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	I2T	T2I	
GME-Qwen2-VL-7B-Instruct	MLLM	7B	64.9	49.1	54.7	57.4	88.5	75.1	79.9	82.0	87.3	87.1	72.6
Qwen3-VL-Embedding-8B	MLLM	8B	69.1	42.8	**75.0**	**75.8**	92.2	73.2	**91.5**	**91.1**	**97.2**	**97.6**	**80.6**
PA-CLIP (Ours)	ViT-B/16	0.15B	65.3	46.3	60.8	59.2	88.9	76.1	83.5	83.3	93.9	92.9	75.0
PA-CLIP (Ours)	ViT-L/14	0.43B	**70.3**	**51.4**	65.6	65.6	**94.1**	**81.9**	87.3	87.8	97.1	96.7	79.8

> **Instruction**: You are a professional aviation expert and computer vision expert. Please generate detailed textual descriptions for the given aircraft image regions, focusing on visible part-level attributes such as shape, color, texture, and local patterns.

Similar templates are used for other fine-grained domains by replacing the object superclass and candidate part attributes. For example, for birds, the candidate attributes include parts such as the beak, wings, and belly; for cars, they include components such as the logo, wheel hub, headlight, grille, and taillight.

G Dataset Documentation for FG-Part

FG-Part is a newly constructed dataset containing approximately 850K image–text pairs for improving fine-grained vision-language alignment. Each sample consists of an image or a localized object-part region and a corresponding textual description that focuses on discriminative local attributes such as shape, color, texture, and visual patterns. The dataset is built from existing fine-grained visual recognition datasets and automatically generated descriptions.

Dataset Construction. For datasets with human-annotated part locations, such as CUB-200-2011 and Birdsnap, we directly use the available part annotations to crop object regions. For datasets without part annotations, such as Stanford Cars and FGVC-Aircraft, we use an open-vocabulary detector to localize predefined discriminative parts, such as headlights, grilles, wings, engines, tails, and fuselages. After extracting part regions, we use MLLMs to generate detailed part-level descriptions based on domain-specific prompts. The prompt template is provided in Appendix F.

Statistics. Table 12 summarizes the composition of FG-Part. It contains 851,017 image–text pairs from 14 source datasets, including 308,898 image-level pairs and 542,119 part-level pairs over 322,622 images.

Localization Audit. For sources without human-annotated part locations, part boxes are produced by OV-DINO with a confidence threshold of 0.5, and boxes below the threshold are discarded. To quantify localization quality, we randomly sample 1,000 retained boxes (200 from each of five OV-DINO-based sources) and manually label each as *Correct*, *Partially shifted*, or *Wrong*. As shown in Table 13(a), 82.1% of the boxes are correct, 12.0% are partially shifted, and only 5.9% are clearly wrong. Accuracy is lower on Stanford Cars and Oxford 102 Flowers, where parts are less visually delimited (e.g., densely clustered floral structures).

Table 11: Compute-matched comparison on ViT-B/16. Top-1 accuracy is reported on 7 FGVR benchmarks. Ret. Avg. denotes the average Recall@1 over the five cross-modal retrieval benchmarks in both directions.

Setting	Objective	Iters	CUB	Aircraft	Dogs	Pets	NABirds	Flowers	Cars	FGVR Avg.	Ret. Avg.
Global	Long+Short	2,900	60.2	36.9	66.0	89.0	54.0	69.1	87.2	66.1	75.7
Compute-matched	Long+Short	11,640	60.4	37.8	67.6	89.9	54.4	69.6	88.1	66.8	74.7
PA-CLIP (Ours)	Long+Short+Part	11,640	68.8	36.9	72.6	93.8	53.4	87.8	87.4	71.5	75.0

Table 12: Statistics of FG-Part by source dataset.

Source	Images	Image-level	Part-level	Total
CUB-200-2011	5,994	5,994	71,861	77,855
NABirds	23,929	23,929	205,403	229,332
Birdsnap	14,529	6,761	141,080	147,841
Stanford Dogs	12,000	12,000	26,871	38,871
Oxford-IIIT Pets	3,680	3,680	21,573	25,253
FGVC-Aircraft	6,667	6,667	19,896	26,563
Stanford Cars	8,144	8,144	14,018	22,162
Oxford 102 Flowers	2,040	2,040	8,143	10,183
Caltech-101	4,346	4,346	0	4,346
Food-101	58,952	52,996	26,965	79,961
Fru92	13,800	13,800	2,801	16,601
Veg200	30,000	30,000	3,508	33,508
SUN397	19,850	19,850	0	19,850
DeepFashion	118,691	118,691	0	118,691
Total	**322,622**	**308,898**	**542,119**	**851,017**

Caption Audit. We further assess the accuracy of the generated captions with two strong multimodal judges, Gemini 3.6 Flash and GPT-5.6-Sol, both independent of the generation models. We randomly sample 1,000 captions of each type (short, long, and part-level; 3,000 in total). Each judge receives the image (or the part crop for part-level captions) together with the caption and outputs one of three verdicts: *Yes*, if the described content and all meaningful visible attributes are consistent with the image; *Partially*, if the subject matches but at least one meaningful attribute is incorrect or unsupported; and *No*, if the described part or core content clearly conflicts with the image. The two verdicts are merged conservatively: a caption is labeled *Yes* only if both judges answer *Yes*, *No* if either judge answers *No*, and *Partially* otherwise. As shown in Table 13(b), 87.8% of the sampled captions are judged fully accurate and only 3.2% clearly inaccurate; part-level captions, the core supervision of PA-CLIP, reach 86.1%. By source dataset (Table 13(c)), accuracy ranges from 78.0% on FGVC-Aircraft to 92.8% on Oxford 102 Flowers.

Intended Use. FG-Part is intended for academic research on fine-grained vision-language learning, part-aware representation learning, fine-grained visual recognition, and interpretable vision-language alignment. It is not intended for high-stakes decision-making, biometric identification, surveillance, or other privacy-sensitive applications.

Data Quality and Limitations. Since FG-Part is constructed automatically, it may contain noise from part localization errors and MLLM-generated descriptions. For example, a detected part region may be imprecise, or a generated description may contain inaccurate or overly generic visual attributes. Nevertheless, our experiments and ablation studies show that the dataset provides effective supervision for improving fine-grained visual recognition.

Licenses and Release. FG-Part is derived from existing datasets and models, and its release will follow the licenses and terms of the corresponding source datasets and base models. We are currently organizing the dataset, code, preprocessing scripts, and documentation. The dataset and code will be publicly released, subject to applicable license constraints.

Table 13: Quality audits of FG-Part. (a) Manual audit of 1,000 OV-DINO-detected part boxes. (b) Caption accuracy by caption type, judged by two independent MLLMs. (c) Caption accuracy by source dataset (all caption types combined).

(a) Localization

Source	Correct	Shifted	Wrong
Cars	75.0	16.0	9.0
Aircraft	89.5	7.5	3.0
Dogs	86.5	9.5	4.0
Pets	88.0	8.5	3.5
Flowers	71.5	18.5	10.0
Total	82.1	12.0	5.9

(b) Caption (by type)

Type	Yes	Partially	No
Short	91.7	7.9	0.4
Long	85.6	10.3	4.1
Part	86.1	8.9	5.0
Overall	87.8	9.0	3.2

(c) Caption (by source)

Source	Yes	Partially	No
Cars	87.2	9.5	3.3
Aircraft	78.0	16.8	5.2
Dogs	90.3	7.0	2.7
Pets	90.7	6.3	3.0
Flowers	92.8	5.5	1.7

Table 14: Pointing-based part grounding on CUB-200-2011 (ViT-L/14). Top-1 and Top-4 accuracy (%) is reported; Random denotes uniformly sampled patches.

Method	Metric	Beak	Belly	Breast	Crown	Forehead	Wing	Eye	Leg	Nape	Tail	Throat	Avg.
Random	Top-1	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.4	1.3	1.4	1.4
	Top-4	5.3	5.6	5.5	5.5	5.4	5.6	5.5	5.5	5.5	5.0	5.5	5.5
CLIP	Top-1	35.0	10.7	13.3	5.5	7.4	4.4	14.0	26.5	2.0	15.1	10.3	13.1
	Top-4	59.3	20.4	35.6	14.8	19.7	19.3	30.3	55.7	8.2	38.8	27.8	30.0
PA-CLIP (Ours)	Top-1	84.7	14.6	19.0	22.1	55.6	8.9	60.5	43.0	13.8	48.5	45.7	**37.9**
	Top-4	96.8	32.8	39.4	46.0	85.6	25.0	92.5	83.4	42.4	81.2	82.7	**64.3**

Maintenance. After release, we will maintain the dataset repository by providing data construction scripts, usage instructions, and updates for known issues. We also plan to further expand FG-Part to cover more object categories and visual domains in future work.

H Quantitative Grounding Evaluation

PA-CLIP is trained on cropped part regions and is never directly supervised for region–token grounding in the full image. We therefore evaluate part grounding with a pointing-based protocol on the CUB-200-2011 test set: for each part-level text query, we rank all patch tokens of the full image by cosine similarity and count a hit if the top-ranked patch (Top-1) or any of the top-4 patches (Top-4) falls within a tolerance box around the ground-truth part keypoint. As shown in Table 14, PA-CLIP improves the average Top-1 accuracy from 13.1% to 37.9% and the Top-4 accuracy from 30.0% to 64.3%, indicating that the part-level alignment learned from crops transfers to grounding in the full image.

I Additional Qualitative Visualizations

We provide additional qualitative visualization examples in Figures 3–11. These examples further show that PA-CLIP tends to produce more localized responses on discriminative object parts under different part-level textual queries, while the original CLIP often attends to broader or background regions.

J Visualization of FG-Part Dataset Construction

To further illustrate the construction process of FG-Part, we provide visualizations of the open-vocabulary part detection results used for extracting localized object parts. As shown in Figures 12–16, the detector can localize discriminative components such as wings, engines, fuselages, headlights, grilles, and other category-specific parts. These localized regions are then cropped and used for part-level description generation, forming the part-level image-text pairs in FG-Part.

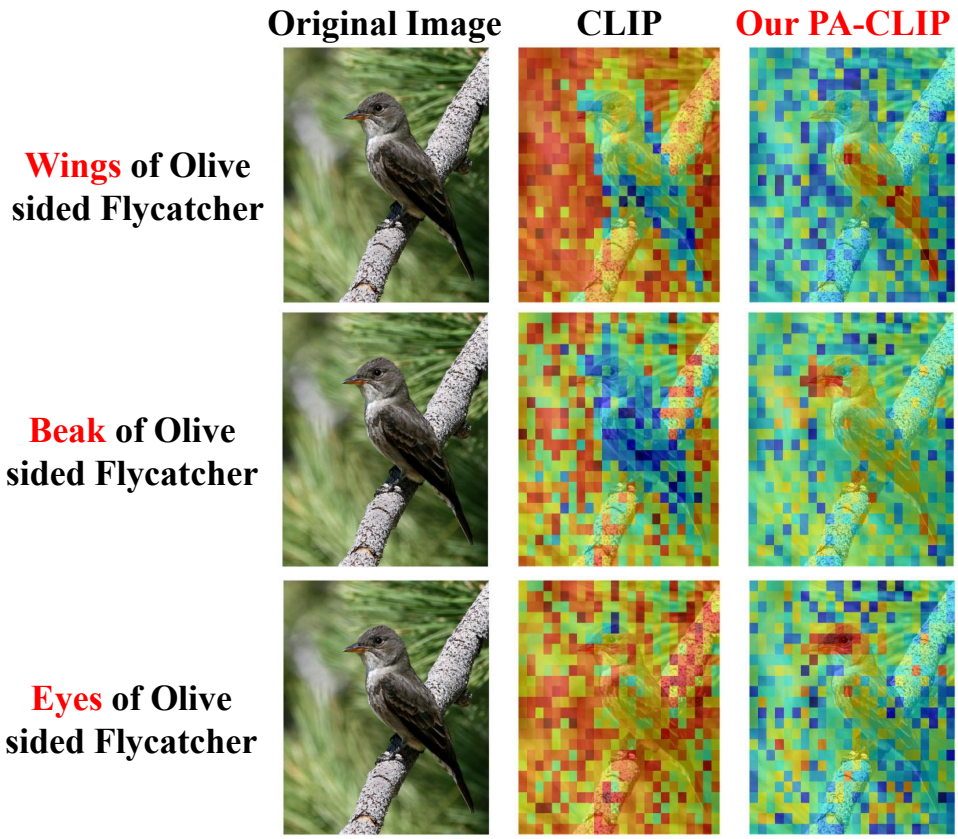


Figure 3: Additional visualization on a bird image. PA-CLIP produces more localized responses for fine-grained part queries such as wings, beak, and eyes.

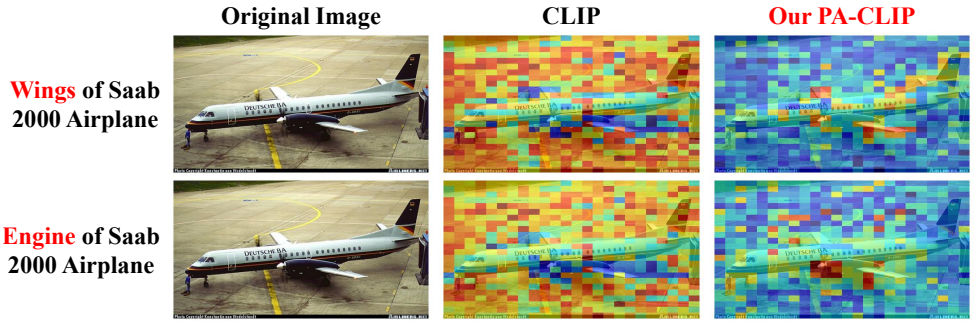


Figure 4: Additional visualization on an aircraft image. PA-CLIP more accurately highlights queried aircraft parts such as wings and engine.

K Licenses and Terms of Existing Assets

We use existing datasets, models, and software packages only for academic research, training, and evaluation. All assets are credited by citing their original papers or official project pages. We do not claim ownership of any existing image, annotation, model checkpoint, or codebase used in this work. For assets with non-commercial, attribution, redistribution, or acceptable-use restrictions, we use them only under the corresponding terms. For assets whose licenses are not explicitly specified in the official sources, we adopt a conservative policy: we use them only for academic evaluation or internal dataset construction, do not redistribute the original images, and will release only metadata, scripts, generated descriptions, or derived data when such release is permitted by the corresponding source terms.

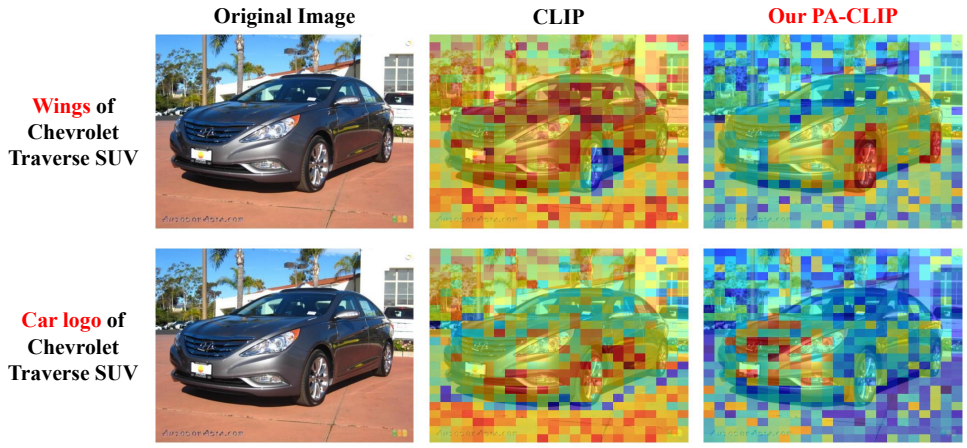


Figure 5: Additional visualization on a car image. PA-CLIP shows more concentrated activations on car-specific parts such as wheels and logo.

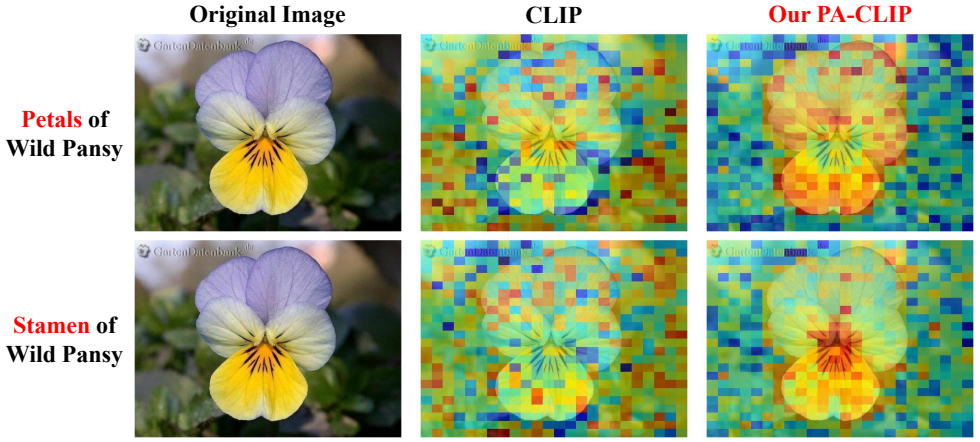


Figure 6: Additional visualization on a flower image. PA-CLIP better localizes fine-grained floral structures such as petals and stamen.

Release policy for FG-Part. FG-Part is derived from existing fine-grained datasets, automatically localized part regions, and MLLM-generated descriptions. When releasing FG-Part, we will respect the licenses and terms of all source datasets and base models. For source datasets that prohibit redistribution or have unclear redistribution permissions, we will not redistribute the original images or restricted crops directly. Instead, we will provide data construction scripts, metadata, generated descriptions, and instructions for users to obtain the original data from the official sources when required. The released repository will include attribution, license notices, and usage restrictions for all included assets.

L Detailed Implementation Hyperparameters

Model Initialization and Architecture. We initialize all models using the original CLIP weights [Radford et al., 2021]. Experiments are conducted on two configurations: CLIP-ViT-Base/16 (ViT-B-16) and CLIP-ViT-Large/14 (ViT-L-14).

Stage 1: FineHARD Dataset. We train on the FineHARD dataset using a cluster of 10 NVIDIA A6000 GPUs. The per-GPU batch size is 128, resulting in a global batch size of 1280. We employ the AdamW optimizer with a learning rate of 1e-6 and a cosine annealing learning rate scheduler. The learnable temperature parameter τ in the contrastive loss is initialized to 0.07. DeepSpeed ZeRO stage 2 optimization is used to accelerate training.

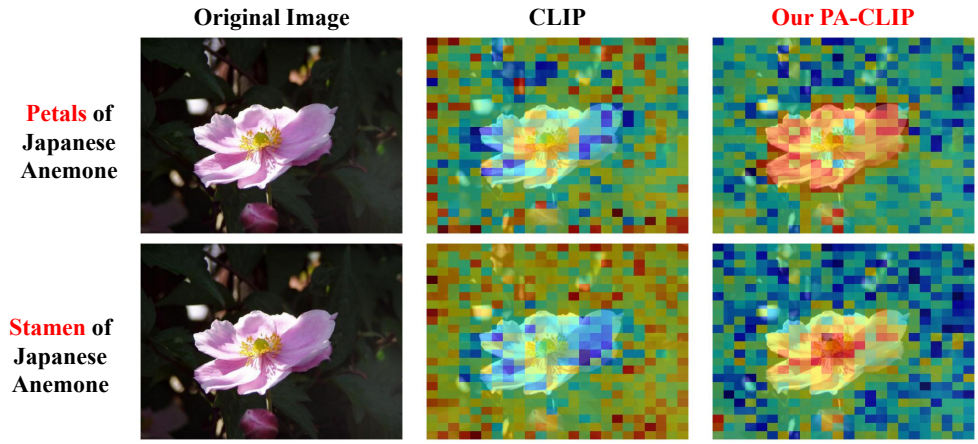


Figure 7: Additional visualization on a flower image. PA-CLIP produces more part-specific activations for petals and stamen.

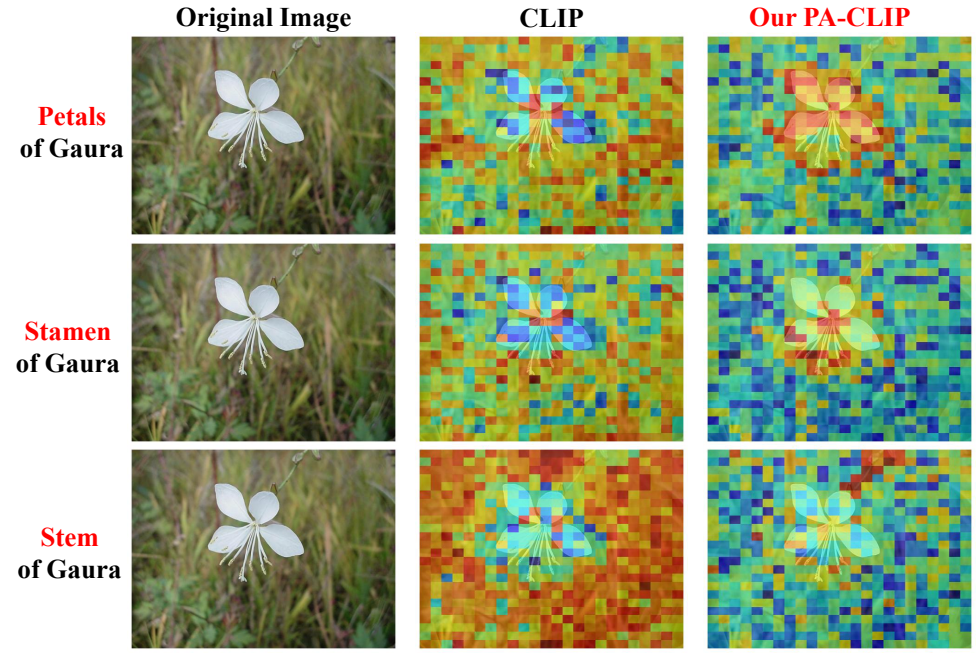


Figure 8: Additional visualization on a flower image with multiple queried parts. PA-CLIP localizes petals, stamen, and stem more clearly than CLIP.

Stage 2: FG-Part Dataset. We train on a subset of our FG-Part dataset using 8 NVIDIA RTX 5090 GPUs. The per-GPU batch size is 64, yielding a global batch size of 512. We use the AdamW optimizer with a learning rate of 1e-6 and a cosine annealing schedule. A warm-up period of 300 steps is applied, and the model is trained for a total of 20 epochs. DeepSpeed ZeRO stage 2 is again employed for optimization.

All other hyperparameters follow the default settings of Xie et al. [2025a], Asokan et al. [2025].

Part-level Description Generation. The MLLMs are served with vLLM using greedy decoding (temperature 0) and a maximum of 4,096 output tokens, producing one description per part region.

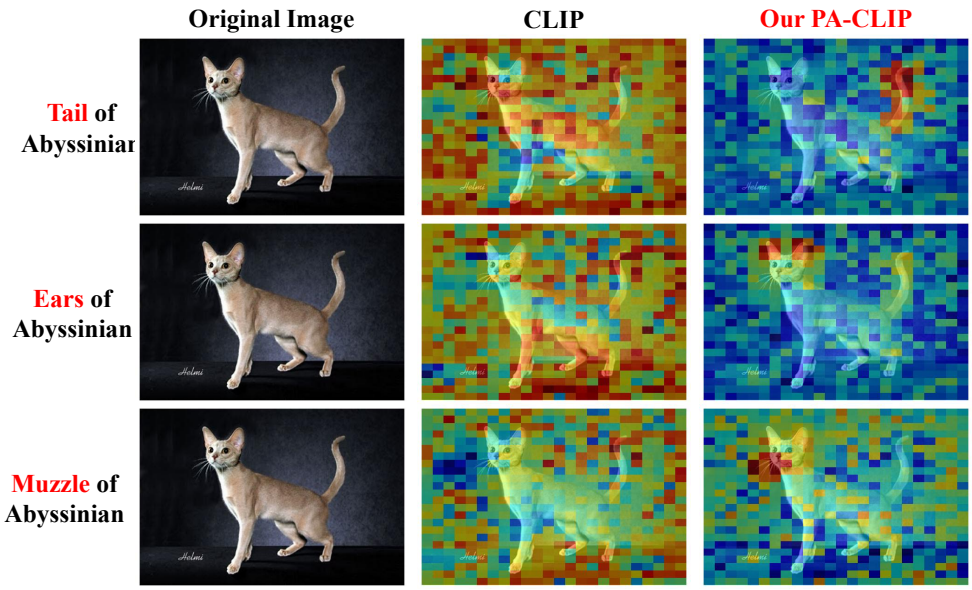


Figure 9: Additional visualization on a cat image. PA-CLIP focuses more accurately on queried parts such as tail, ears, and muzzle.

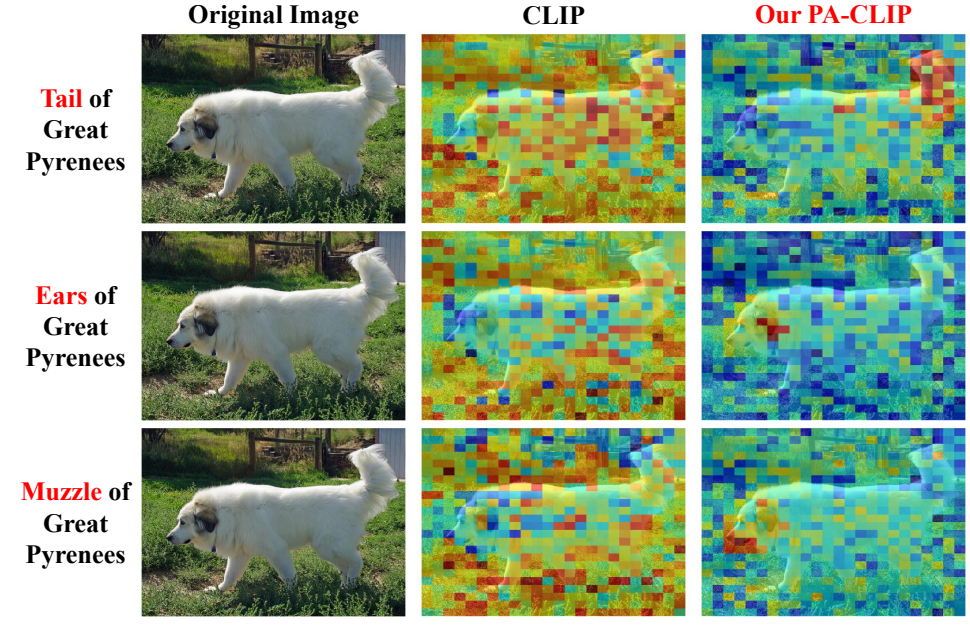


Figure 10: Additional visualization on a dog image. PA-CLIP provides more localized grounding for tail, ears, and muzzle queries.

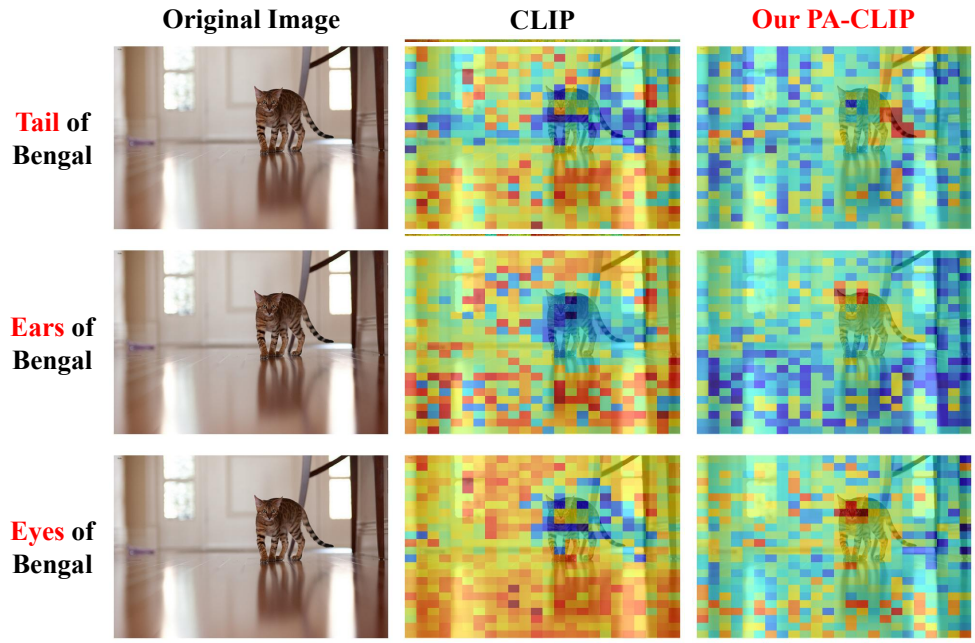
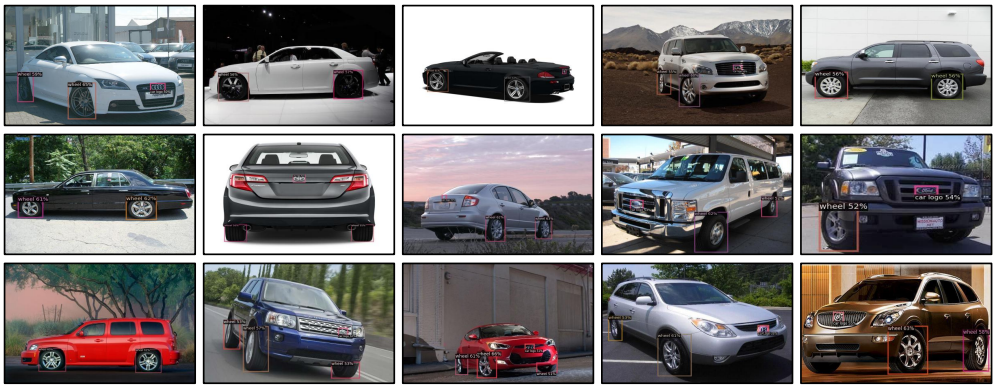


Figure 11: Additional part-aware grounding visualization on a cat image. PA-CLIP localizes queried parts such as tail more accurately than CLIP.



Visualization of open-vocabulary detection results on FGVC Aircraft dataset.

Figure 12: Visualization of open-vocabulary part detection results during FG-Part construction.



Visualization of open-vocabulary detection results on Stanford Cars dataset.

Figure 13: Visualization of open-vocabulary part detection results during FG-Part construction.



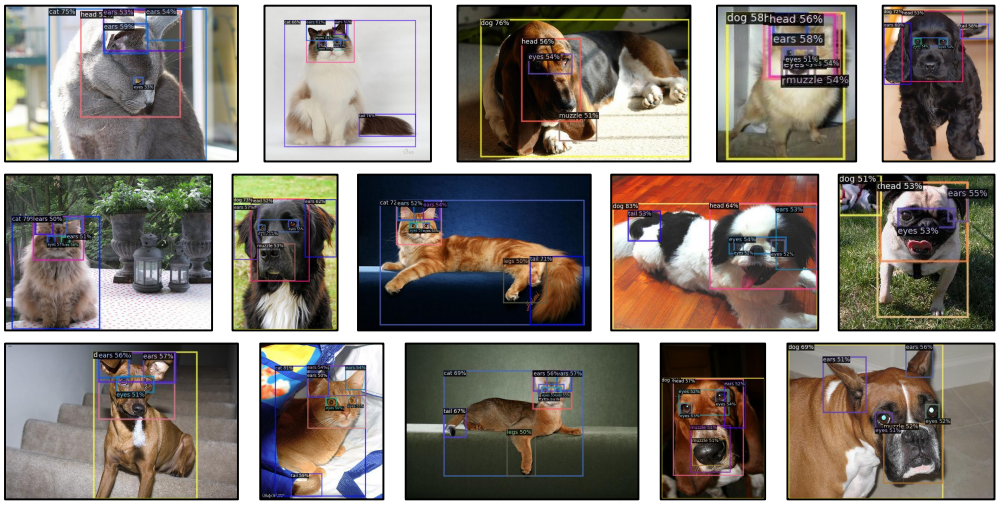
Visualization of open-vocabulary detection results on Oxford Flowers dataset.

Figure 14: Visualization of open-vocabulary part detection results during FG-Part construction.



Visualization of open-vocabulary detection results on Stanford Dog dataset.

Figure 15: Visualization of open-vocabulary part detection results during FG-Part construction.



Visualization of open-vocabulary detection results on Oxford-IIIT Pets dataset.

Figure 16: Visualization of open-vocabulary part detection results during FG-Part construction.

Table 15: Licenses and terms of existing datasets used in this paper.

Asset	Usage in this work	Citation	License / terms and compliance
CUB-200-2011	FGVR evaluation and FG-Part construction	Wah et al. [2011]	The official dataset page states that the images are restricted to non-commercial research and educational purposes, and that the dataset maintainers do not own the image copyrights. We use the dataset only for academic research and will not redistribute original images unless permitted by the source terms.
FGVC-Aircraft	FGVR evaluation and FG-Part construction	Maji et al. [2013]	The official dataset page states that the images are provided exclusively for non-commercial research purposes, and that the original authors retain image copyrights. We use the dataset only for academic research and do not redistribute original images.
Stanford Dogs	FGVR evaluation	Khosla et al. [2011]	We did not identify an explicit standalone license statement from the official dataset page. Since the dataset is based on ImageNet images and annotations, we use it only for academic evaluation and follow the corresponding ImageNet/source-image restrictions. We do not redistribute original images.
Oxford-IIIT Pets	FGVR evaluation	Parkhi et al. [2012]	The official dataset page states that the dataset is released under the Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0), while image copyright remains with the original owners. We cite the dataset and comply with the attribution and share-alike requirements.
SUN397	FGVR evaluation	Xiao et al. [2010]	We did not identify an explicit dataset license statement from the official dataset page. The asset is used only for academic evaluation, and we do not redistribute original images.
NABirds	FGVR evaluation	Van Horn et al. [2015]	The official terms state that the dataset is provided for non-commercial research use and restrict reproduction or distribution without permission. We use the dataset only for academic evaluation and do not redistribute the original data.
Oxford 102 Flowers	FGVR evaluation	Nilsback and Zisserman [2008]	We did not identify an explicit standalone license statement from the official dataset page. We use the dataset only for academic evaluation, cite the original work, and do not redistribute original images.
Stanford Cars	FGVR evaluation and FG-Part construction	Krause et al. [2013]	We did not identify an explicit standalone license statement from the currently accessible official source. We use the dataset only for academic evaluation and dataset construction, cite the original work, and do not redistribute original images unless permitted by the source terms.
Birdsnap	FG-Part construction	Berg et al. [2014]	We did not identify an explicit standalone license statement from the official public source. We use the dataset only for academic research and FG-Part construction, and will not redistribute original images or restricted derived crops unless redistribution permission is verified.
Food-101	FG-Part construction	Bossard et al. [2014]	The license agreement distributed with the dataset states that the images are collected from Foodspotting and are not the property of ETH Zurich, and that any use beyond scientific fair use must be negotiated with the respective image owners. We use the dataset only for academic research and FG-Part construction, and do not redistribute original images.
VegFru (Veg200 and Fru92)	FG-Part construction	Hou et al. [2017]	We did not identify an explicit dataset license statement from the official project page or repository. We use the dataset only for academic research and FG-Part construction, cite the original work, and do not redistribute original images.
Caltech-101	FG-Part construction	Fei-Fei et al. [2004]	The official CaltechDATA record releases the dataset under the Creative Commons Attribution 4.0 International License (CC BY 4.0). We cite the dataset and comply with the attribution requirement.
DeepFashion	FG-Part construction	Liu et al. [2016]	The official agreement states that the dataset is available for non-commercial research purposes only, that the images are not the property of the dataset maintainers, and that no portion of the dataset may be further copied, published, or distributed. We use the dataset only for academic research and FG-Part construction, and do not redistribute any original images or annotations.

Table 16: Licenses and terms of existing datasets used in this paper (continued).

Asset	Usage in this work	Citation	License / terms and compliance
MSCOCO	Cross-modal retrieval evaluation	Lin et al. [2014]	COCO annotations are commonly provided under the Creative Commons Attribution 4.0 License (CC BY 4.0), while the images are subject to Flickr Terms of Use and their original copyright holders. We use MSCOCO only for evaluation and follow the corresponding image and annotation terms.
Flickr30K	Cross-modal retrieval evaluation	Young et al. [2014]	The dataset documentation states that image use must abide by Flickr Terms of Use, and that images are provided only for non-commercial research and educational purposes. We use the dataset only for academic evaluation and do not redistribute images.
DCI	Cross-modal retrieval evaluation	Urbanek et al. [2024]	The DCI repository states that its code and data are released under CC-BY-NC, and that the images come from SA-1B and require acceptance of the SA-1B license. We use DCI only for non-commercial academic evaluation and comply with the SA-1B/source-data terms.
DOCCI	Cross-modal retrieval evaluation	Onoe et al. [2024]	The official DOCCI page states that the annotations and images are licensed under CC BY 4.0. We cite the dataset and comply with the attribution requirements.
Urban1k	Cross-modal retrieval evaluation	Zhang et al. [2024a]	The official dataset card lists the license as CC BY-NC 4.0. We use the dataset only for non-commercial academic evaluation and comply with the attribution and non-commercial restrictions.

Table 17: Licenses and terms of existing models, datasets, and software assets used in this paper.

Asset	Usage in this work	Citation	License / terms and compliance
CLIP	Model initialization and baseline	Radford et al. [2021]	The OpenAI CLIP code repository is released under the MIT License. We cite the original CLIP paper and use the released model weights/code in accordance with the corresponding license.
FG-CLIP	Baseline and related model	Xie et al. [2025a]	The official model card lists the FG-CLIP model/project content under Apache-2.0, while also noting that underlying datasets and checkpoints may be subject to their original licenses. We cite the original work and comply with the respective source-asset terms.
FineHARD	First-stage training dataset	Xie et al. [2025a]	The official dataset card lists FineHARD under Apache-2.0. Since FineHARD may include data derived from external image sources, we will respect the original source-data terms when releasing any derived assets.
OV-DINO	Open-vocabulary part localization	Wang et al. [2024a]	The official model card lists OV-DINO under Apache-2.0. We use it for part localization and cite the original work.
Gemma3 27B	MLLM-based part-level description generation	Gemma Team [2025]	Gemma models are governed by the Google Gemma Terms of Use and prohibited-use policy. We use Gemma3 only for generating part-level descriptions for academic research and comply with the applicable usage and redistribution terms.
Llama4 Scout	MLLM-based part-level description generation	Meta AI [2025]	Llama4 Scout is governed by the Meta Llama 4 Community License Agreement and Acceptable Use Policy. We use it only for part-level description generation in academic research and comply with the applicable attribution, redistribution, and acceptable-use requirements.
vLLM	Efficient MLLM inference	Kwon et al. [2023]	vLLM is released under the Apache License 2.0. We use it for efficient MLLM inference and cite the original work.

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 main claims in the abstract and Introduction are supported by the proposed PA-CLIP framework, the FG-Part dataset construction, and the experimental results on fine-grained visual recognition and cross-modal retrieval in Sections 3 and 4. The paper claims improved fine-grained perception and competitive general retrieval performance, which are evaluated in the corresponding experiments.

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper includes limitations paragraph in the conclusion section.

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: The paper does not present new theoretical results, theorems, or formal proofs. The contribution is empirical and methodological.

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: The paper describes the model architecture, multi-level contrastive objective, part-level data construction pipeline, training datasets, evaluation benchmarks, and main hyperparameters in Sections 3, 3.2, and 4.1. Code and data will be released. These details provide a reasonable path for reproducing or verifying the main experimental 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: We will publicly release the code and the FG-Part dataset.

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 specifies the training data, model backbones, optimizer, learning rate, batch size, temperature initialization, warm-up steps, number of epochs, DeepSpeed configuration, GPU types, evaluation tasks, metrics, and benchmark datasets in Section 4.1. The main experimental settings needed to understand the results are provided.

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: [Yes]

   Justification: Main results are averaged over three runs. This reduces the influence of random variation in training and evaluation.

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: The paper reports the main computational resources used for data generation and training. Specifically, we specify the GPU types and numbers used for MLLM-based data generation, the first-stage training on FineHARD, and the second-stage training on FG-Part. We also provide batch sizes, optimizer settings, training epochs, warm-up steps, and DeepSpeed configuration in the implementation details.

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: We have read the NeurIPS Code of Ethics and believe that the research conforms to it. The work uses existing public datasets and automatically generated part-level descriptions, and does not involve human subjects, private user data, or intentionally harmful applications.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The paper discusses potential broader impacts, including beneficial applications in fine-grained visual understanding and possible risks in privacy-sensitive or surveillance-related scenarios. We also emphasize responsible deployment with appropriate safeguards and compliance with relevant ethical and legal requirements.

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: The paper does not release a high-risk generative model or a system intended for autonomous decision-making. The proposed model is a vision-language representation model focused on fine-grained recognition and retrieval. Therefore, dedicated release safeguards for high-risk generative systems are not applicable.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We properly credit the original creators of the datasets, pretrained models, and software packages used in this paper through citations in the main text. We also provide license and terms-of-use summaries in Tables 15, 16, and 17, specifying how each asset is used and how we comply with its corresponding license or usage restrictions. For datasets with restricted or unclear redistribution terms, we use them only for academic research/evaluation and do not redistribute original images. The release of FG-Part, code, and checkpoints will follow the licenses and terms of the corresponding source datasets, models, and software packages.

13. **New assets**

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

Answer: [Yes]

Justification: The paper introduces FG-Part, a new dataset of approximately 850K image–text pairs. We describe its construction pipeline in Section 3.2, provide the prompt template in Appendix F, and include dataset documentation in Appendix G. The documentation describes the dataset composition, intended use, limitations, license considerations, release plan, and maintenance plan. We are currently organizing the dataset and code, and will release them subject to the licenses and terms of the source datasets and base models.

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

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

    Answer: [N/A]

    Justification: The paper does not use crowdsourcing, paid annotators, user studies, or human-subject experiments. The dataset is constructed automatically using existing datasets, open-vocabulary detection, and MLLM-based description generation.

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 human subjects research, private personal data, user studies, or interventions with participants. Therefore, IRB approval is not applicable.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: We use MLLMs as part of the FG-Part dataset construction pipeline. Specifically, Gemma3 27B and Llama4 Scout are used to generate part-level descriptions for localized object regions, as described in Section 3.2. We also provide the prompt template in Appendix F.