

Pixel-SAIL: Single Transformer for Pixel-Grounded Understanding

Tao Zhang^{1,2}, Xiangtai Li¹, Zilong Huang¹, Yanwei Li¹, Weixian Lei¹
Xueqing Deng¹, Shihao Chen², Shunping Ji², Jiashi Feng¹

¹ByteDance Seed

²Wuhan University

xiangtai94@gmail.com; zhang_tao@whu.edu.cn Corresponding: Xiangtai Li

Abstract

Multimodal large language models (MLLMs) achieve remarkable performance on fine-grained, pixel-level understanding tasks. However, state-of-the-art methods for pixel-grounded understanding rely heavily on additional components, such as a vision encoder (e.g., CLIP), segmentation experts, a mask tokenizer, and a detokenizer. This dependence increases system complexity and limits model scaling. In this work, we investigate whether a highly simplified MLLM can achieve state-of-the-art performance on pixel-grounded tasks without external vision encoders or segmentation experts at inference time. Our work is motivated by recent Single trAnsformer as a unified vIsion-Language Model (SAIL) designs, which jointly learn vision and text tokens within a transformer. We find that a single-transformer architecture can align visual and language signals while retaining the fine-grained features required for pixel-level understanding. Based on this finding, we present Pixel-SAIL, a single-transformer model for pixel-grounded MLLM tasks. Pixel-SAIL combines visual prompt injection, a learnable upsampling module, and vision-expert distillation to enhance pixel-level perception and visual prompt understanding. In addition, we construct a comprehensive pixel-grounded understanding benchmark (**PerBench**) with manual verification. It includes three tasks: detailed object description, visual prompt-based question answering, and visual-text referring segmentation. Extensive experiments on referring segmentation, visual prompt understanding, and PerBench show that Pixel-SAIL matches or outperforms more complex architectures. The code will be released.

1 Introduction

Multimodal large language models (MLLMs) have garnered significant research attention, driven by advances in large language models (LLMs) (Touvron et al., 2023; Hugo Touvron et al., 2023; Yang et al., 2024a). While most studies focus on open-ended visual question-answering tasks, there is growing interest (Rasheed et al., 2024; Zhang et al., 2024b) in fine-grained, pixel-level understanding. Such capabilities support applications including region-level editing and generation, as well as detailed understanding of designated image regions.

Recent pixel-grounded MLLMs (Lai et al., 2024; Ren et al., 2024; Xia et al., 2024; Zhang et al., 2024d; Rasheed et al., 2024; Zhang et al., 2024b; Yuan et al., 2025a; Yan et al., 2024) mainly adopt vision-language fusion frameworks, following design patterns (Liu et al., 2023c; Ding et al., 2021; Yang et al., 2022b) established before the LLM era. For example, LAVT (Yang et al., 2022b) adopts an encoder-fusion approach that injects language embeddings generated by BERT (Devlin et al., 2019) into a vision transformer. With the advent of LLMs (Hugo Touvron et al., 2023; Yang et al., 2024a;b), recent works (Lai et al., 2024; Ren et al., 2024; Zhang et al., 2024b; Yuan et al., 2025a) integrate state-of-the-art segmentation

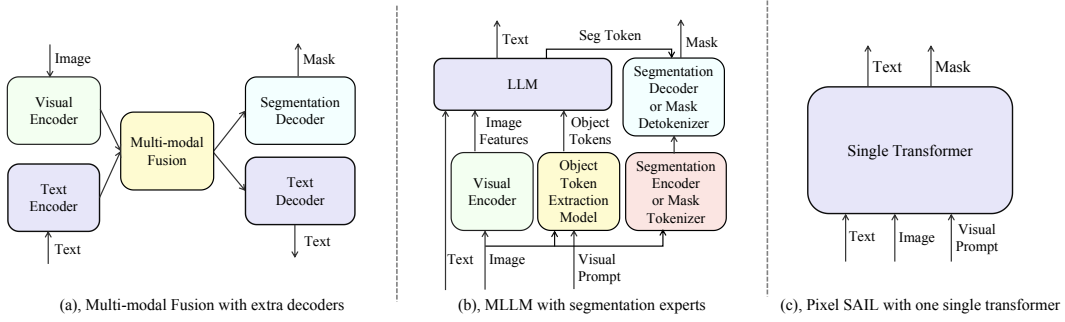


Figure 1: **Comparison between current methods for pixel-grounded understanding and our method.** (a) Pixel-grounded models without LLMs cannot process complex text instructions or support flexible dialogue. (b) Current pixel-grounded MLLMs use complex architectures that may include an LLM, a CLIP-like vision backbone, an object-token extraction model, a segmentation backbone or mask tokenizer, and a SAM-like decoder or mask detokenizer. (c) Pixel-SAIL uses a single transformer without an external vision encoder or segmentation expert at inference time.

models (Kirillov et al., 2023; Ravi et al., 2024; Li et al., 2024d; Shu et al., 2025) for pixel-level understanding, either by attaching them to LLM outputs or by embedding an LLM within a segmentation pipeline. While effective, the overall architectures are complex, requiring specialized components such as vision-language fusion modules and additional decoders. Moreover, their final performance often depends heavily on either MLLMs or segmentation models, which may lead to suboptimal results due to limitations within individual submodules. Although some works (Lan et al., 2025; Wang et al., 2023b; Pramanick et al., 2024) model segmentation differently, for example by outputting contour coordinates (Wang et al., 2023b; Pramanick et al., 2024; Xie et al., 2020; Zhang et al., 2022) or RLE-like textual representations (Lan et al., 2025), a substantial performance gap remains relative to state-of-the-art methods (Yuan et al., 2025a; Wei et al., 2024b).

In this work, we explore a simple pixel-grounded MLLM design inspired by recent advances in SAIL architectures (Diao et al., 2025b; Lei et al., 2025; Diao et al., 2024; Luo et al., 2025), also known as *encoder-free* MLLMs. These methods remove the external vision encoder and jointly train vision and language tokens on large-scale datasets, resulting in a simpler design. Moreover, they show competitive performance on image-level VQA tasks compared with LLaVA (Liu et al., 2023b; 2024a). Motivated by this success, we extend the framework to pixel-level understanding tasks, aiming to reduce the complexity of existing approaches. To the best of our knowledge, this is the first study to investigate a minimal encoder-free architecture for pixel-grounded MLLM tasks, including referring segmentation and visual prompt understanding.

We first build Pixel-SAIL on the SAIL architecture and generate segmentation masks by multiplying the final-layer vision-token representations by the predicted segmentation-token representation. This baseline learns strong text-driven segmentation while retaining the conversational capabilities of the base SAIL model and improving performance on some vision-centric QA tasks, although its segmentation boundaries remain imprecise. We therefore introduce a learnable upsampling module and a vision-expert distillation loss to improve segmentation quality.

For visual prompt understanding, we design a visual prompt injection method that represents visual prompts with special token embeddings without introducing an additional visual prompt encoder. Adding visual prompt embeddings to vision tokens enables early fusion and eliminates the need for a separate visual prompt fusion module.

To evaluate Pixel-SAIL and support research on pixel-grounded MLLMs, we introduce a challenging benchmark, **PerBench**. Compared with previous pixel-grounded MLLM benchmarks, PerBench has three distinctive features. *First*, it includes detailed object captions, whereas most existing benchmarks contain only short, coarse-grained captions. *Second*, it evaluates visual-prompt understanding through multiple-choice VQA, following MME (Fu et al., 2023) and MMBench (Liu et al., 2024b), to provide a more objective evalu-

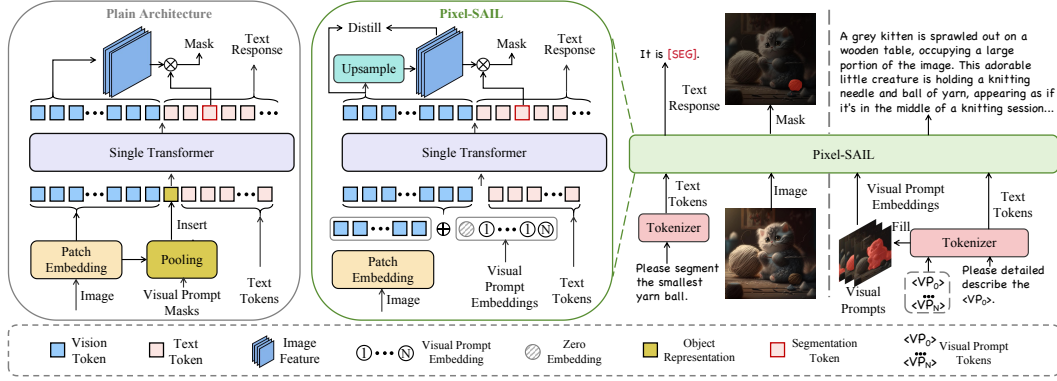


Figure 2: **Architecture of Pixel-SAIL.** The examples on the right illustrate its capabilities in general conversation and pixel-grounded understanding.

ation than free-form region captioning. *Third*, it introduces a task that requires segmenting objects jointly referenced by visual prompts and text. Our benchmark reveals limitations of current state-of-the-art pixel-grounded MLLMs on fine-grained understanding and multi-modal referring tasks.

Pixel-SAIL is jointly trained on a mixture of referring segmentation, VQA, and visual prompt datasets. Experimental results show that our method performs strongly on four public referring segmentation benchmarks and PerBench. In particular, on the RefCOCO+ and RefCOCOg validation splits, Pixel-SAIL 3B reaches 75.7 and 78.7 cIoU, outperforming Sa2VA 4B by 1.4 and 2.0 cIoU points, respectively, while using a simpler inference pipeline. On PerBench, Pixel-SAIL 3B achieves 24.2 METEOR, 74% accuracy, 33.4 cIoU, and an overall score of 42.2, surpassing GLaMM 7B and Sa2VA 4B, which obtain overall scores of 15.3 and 39.0, respectively.

2 Method

2.1 Pixel-SAIL Architecture

Recently, several encoder-free MLLMs (Chen et al., 2024d; Luo et al., 2025; Diao et al., 2025b;a; Lei et al., 2025) have achieved performance comparable to models with external vision encoders. These models jointly learn vision and text features in a single transformer, yielding a much simpler architecture. In particular, SAIL (Lei et al., 2025) uses a projection layer to map images to visual tokens, combines them with language tokens, and feeds the resulting sequence to the transformer. However, this architecture has not been thoroughly explored for fine-grained vision-language tasks such as region captioning and referring segmentation.

Plain Architecture. To address this gap, we first construct a plain single-transformer architecture, motivated by previous ViT-based MLLMs (Lai et al., 2024; Yuan et al., 2025a). We start with a pretrained encoder-free MLLM. For segmentation tasks, we adapt previous mask-generation methods to operate on features produced by the single transformer. First, we reshape the hidden states of the last transformer layer of vision tokens $\mathcal{V} \in \mathbb{R}^{N \times C}$ into image features $\mathcal{F} \in \mathbb{R}^{\frac{H}{S} \times \frac{W}{S} \times C}$. N represents the number of vision tokens, C denotes the channel dimension, H and W denote the image height and width, and S is the downsampling stride. The image features are then multiplied by the hidden states of the predicted segmentation tokens $\mathcal{Q} \in \mathbb{R}^{K \times C}$ to generate segmentation masks $\mathcal{M} \in \mathbb{R}^{K \times \frac{H}{S} \times \frac{W}{S}}$. K denotes the number of predicted segmentation tokens, following previous works (Lai et al., 2024; Rasheed et al., 2024). For visual prompt understanding, we employ a pooling-based method (Yuan et al., 2024b) to derive object representations $\mathcal{O} \in \mathbb{R}^{M \times C}$ from image patch embeddings $\mathcal{P} \in \mathbb{R}^{\frac{H}{P} \times \frac{W}{P} \times C}$. These object embeddings are fed into the single transformer to represent the corresponding objects. M represents the number of visual prompts, and P

denotes the patch size. For segmentation tasks, we use an additional mask loss. For VQA and visual prompt understanding, we use the standard text-generation loss.

The plain architecture demonstrates effective pixel-text alignment because the segmentation and visual prompt tokens are jointly learned with the vision and language tokens. However, it has two limitations: 1) Segmentation masks have imprecise boundaries because of the large feature downsampling stride (16 or 32), even when pixel shuffle or bilinear interpolation is used for upsampling. 2) The single transformer struggles to identify the object referenced by a visual prompt because the object representation is pooled from patch embeddings with limited semantic information.

To address these limitations, we first integrate a learnable upsampling module to improve the spatial detail of the single-transformer features. Second, we develop a visual prompt injection mechanism that helps the model interpret visual prompts. Our method enables early-stage fusion between vision tokens and visual prompt embeddings. Finally, we introduce a dense feature distillation strategy to improve the model’s fine-grained visual features. These improvements collectively address the shortcomings of the plain baseline while maintaining its architectural simplicity.

Learnable Upsampling Module. Inspired by (Li et al., 2022b), we incorporate a learnable upsampling module $\mathcal{U}$ to generate high-resolution features $F_h \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$ for pixel-level grounding. The module comprises multiple upsampling blocks, each consisting of a transposed 2D convolution followed by a depthwise convolution. It effectively upscales the low-resolution features $F_l \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C}$, derived from resized vision tokens, to one-quarter of the original resolution.

Visual Prompt Injection. Previous works (Yuan et al., 2024b; Rasheed et al., 2024; Yuan et al., 2025a) represent referenced objects by pooling vision-token features from a ViT encoder. Encoder-free MLLMs lack the semantically enriched visual features produced by a separate pretrained vision encoder. Consequently, object representations pooled directly from patch embeddings are dominated by low-level cues, such as edges, and may not identify the referenced object precisely.

To overcome this limitation, we propose a visual prompt injection mechanism. Our approach adds multiple visual prompt special tokens $\{VP_i \mid i \in [1, M]\}$ to the language model’s vocabulary, where M is the number of visual prompts. The corresponding text embeddings $\mathcal{VP}^t \in \mathbb{R}^{M \times C}$ are assigned to the regions defined by mask prompts $\mathcal{M}^{vp} \in \mathbb{R}^{M \times \frac{H}{P} \times \frac{W}{P}}$, producing visual prompt tokens $\mathcal{VP} \in \mathbb{R}^{\frac{HW}{P^2} \times C}$. The visual prompt tokens $\mathcal{VP}$ are added to the vision tokens $\mathcal{V} \in \mathbb{R}^{\frac{HW}{P^2} \times C}$ before the sequence is processed by the single transformer. This design enables the model to identify referenced objects through the corresponding special tokens $\{VP_i \mid i \in [1, M]\}$ in the text instructions.

Dense Feature Distillation. Without large-scale, high-quality segmentation data such as SA-1B (Kirillov et al., 2023), the plain architecture produces imprecise masks, particularly at object boundaries. However, directly training on such datasets would be costly and may degrade the original instruction-following capabilities. To address both limitations, we distill knowledge from pretrained segmentation experts into the single transformer, improving boundary detail while preserving VQA capabilities. We use mask features from the Mask2Former (Cheng et al., 2022) pixel decoder to supervise the upsampled features $F_h \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$ and features from the SAM2 (Ravi et al., 2024) image encoder to supervise the low-resolution features $F_l \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C}$. This distillation strategy improves segmentation quality while increasing training time by approximately 3%, and the frozen teachers are removed at inference time.

2.2 Benchmark and Dataset Engine

Our Benchmark: PerBench. We construct a manually verified benchmark named **PerBench** (Pixel-grounded Understanding **Benchmark**). PerBench addresses three limitations of existing pixel-grounded understanding benchmarks.

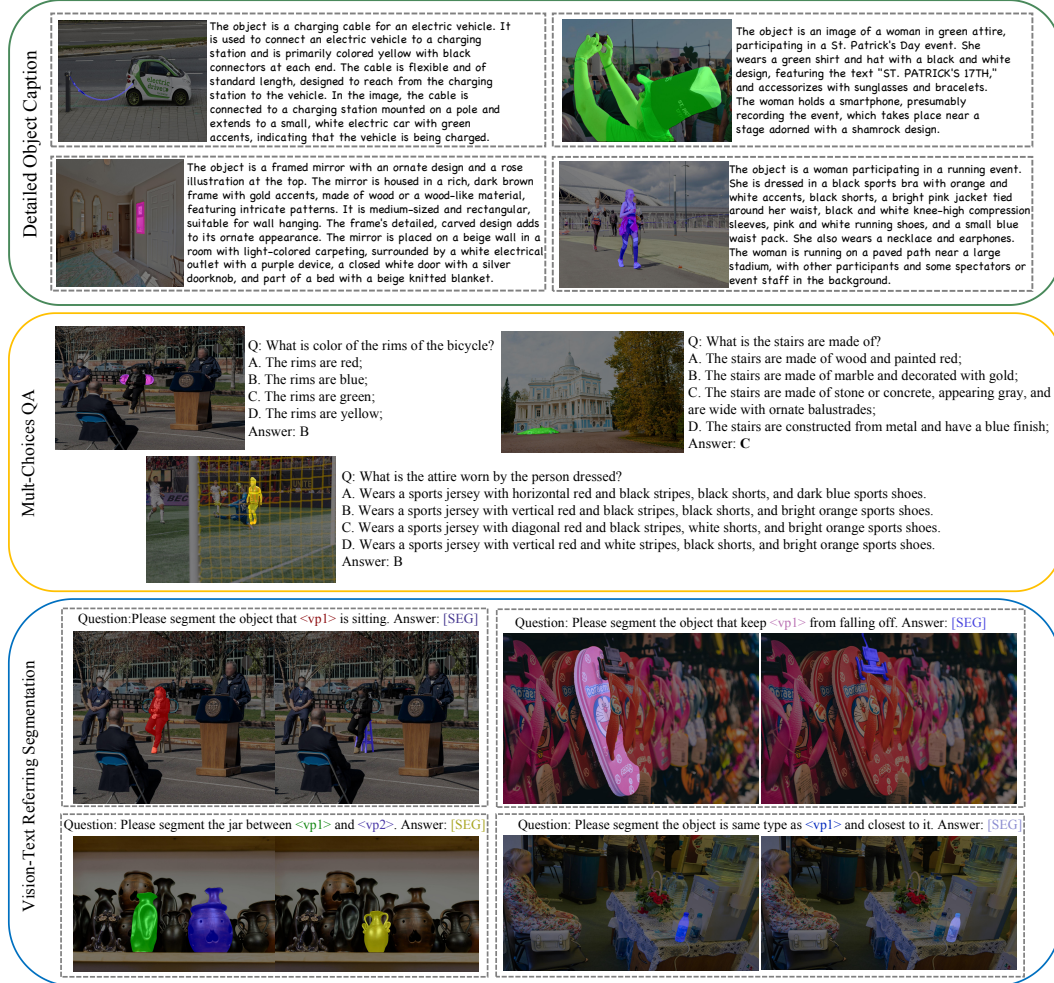


Figure 3: Examples from PerBench. Best viewed in color and at high magnification.

The first is the lack of detailed object captions. Previous works (Chen et al., 2023; Li et al., 2024e) have emphasized more detailed image captions, demonstrating that comprehensive captions significantly enhance model performance. However, current object-caption datasets such as Osprey-724k (Yuan et al., 2024b) and evaluation benchmarks such as Ref-COCOg provide only brief object captions. To address this limitation, we use InternVL2.5-78B (Chen et al., 2024g) and Qwen2.5-VL-72B (Bai et al., 2025) to generate detailed object captions. We then screen and refine them through **manual review**, yielding 500 object captions for evaluation. We use METEOR (Banerjee & Lavie, 2005) as the evaluation metric for this task.

The second is the evaluation of visual-prompt understanding in a multiple-choice format. Although captioning tasks can accurately reflect a model’s ability to understand visual prompts, precise and fair evaluation is difficult. Rule-based metrics such as CIDEr (Vedantam et al., 2015) and METEOR (Banerjee & Lavie, 2005) are affected by response length, format, and ground-truth quality. Using traditional models as evaluators, such as BERT (Devlin et al., 2019), makes it difficult to distinguish subtle differences, while employing large language models as judges inevitably introduces model bias. Therefore, a fair and quantitative visual-prompt understanding benchmark is necessary. Inspired by MMBench (Liu et al., 2024b) and MME (Fu et al., 2023), we manually annotate 500 multiple-choice questions based on the detailed object captions. The questions assess models’ understanding of each referenced object’s appearance, attributes, function, and relationship to surrounding

objects. Models must perceive the attributes of referenced objects and follow instructions to select the appropriate answer. We use accuracy as the evaluation metric.

The third task requires segmenting objects jointly referenced by visual prompts and text, which we abbreviate as V-T RES. It aims to test the model’s ability to understand objects indicated by user-input visual prompts and segment associated objects according to text instructions. This task evaluates visual prompt understanding, textual reasoning, and pixel-level grounding jointly. We manually annotate 500 V-T RES samples and have five expert annotators verify them. As in the RefCOCO series datasets, we select cloU and gloU as the evaluation metrics for the V-T RES task. The overall PerBench score is the average of the three task scores after each is normalized to the range 0–100.

PerBench evaluates pixel-grounded MLLMs and highlights challenges in detailed object understanding and joint visual-text reference resolution.

Training Dataset Engine. Following previous works (Diao et al., 2025b; Luo et al., 2025), we collect diverse pixel-grounded data, including segmentation and visual-prompt understanding datasets.

For segmentation-related data, we first use RefCOCO/+g (Kazemzadeh et al., 2014; Yu et al., 2016) and COCO (Lin et al., 2014) semantic segmentation data used in LISA (Lai et al., 2024), the Grandf dataset (214k samples) used in GLaMM (Rasheed et al., 2024), and MUSE data (246k samples) used in PixelLM (Ren et al., 2024). We also use the recent Pixel2Cap (You et al., 2025) dataset, which comprises 20k images, and convert it to a referring-segmentation format. Finally, we add COCO (Lin et al., 2014) panoptic segmentation data and structure it as follows: “Question: Please segment the {class name} in instance mode. Answer: {class name}-1 [SEG], . . . , {class name}-n [SEG].”

For visual prompt understanding, we employ two public datasets: Osprey-724k (Yuan et al., 2024b) and Pixel2Cap (You et al., 2025). Additionally, we reformat the COCO dataset into a question-answer structure specifically designed to query object categories. To enhance fine-grained object description, we prompt InternVL2.5-78B (Chen et al., 2024g) to generate approximately 300k detailed object captions for objects in 10k SA-1B (Kirillov et al., 2023) images. Lastly, to maintain the ability to follow instructions, we also integrate the LLaVA-1.5 (Liu et al., 2023b) 665k dataset into our training data.

Training. We combine all the aforementioned data for joint training. The loss function consists of the next-token prediction loss $\mathcal{L}_{ntp}$, the segmentation loss $\mathcal{L}_{seg}$, and the distillation loss $\mathcal{L}_{distill}$:

$$\mathcal{L} = \mathcal{L}_{ntp} + \mathcal{L}_{seg} + \alpha \mathcal{L}_{distill}, \quad \mathcal{L}_{seg} = \lambda \mathcal{L}_{ce} + \beta \mathcal{L}_{dice}, \quad (1)$$

where α is set to 0.5, λ to 2.0, and β to 0.5.

3 Experiments

Implementation Details. We evaluate our architecture using two open-source encoder-free multimodal large language models: SAIL (Lei et al., 2025) and EVEv2 (Diao et al., 2025b). For EVEv2, we retain its original architecture and weights without any modifications. We build the 0.5B and 3B Pixel-SAIL models on SAIL and the 7B model on EVEv2. When training Pixel-SAIL based on SAIL, we maintain the original resolution of input images. For images with a long side exceeding 1024, we preserve the aspect ratio and resize the long side to 1024. When training Pixel-SAIL on EVEv2, we resize images to approximately 800×800 pixels to reduce training cost, rather than using the original 1600×1600 setting. We train on 32 A100 (80 GB) GPUs using the AdamW (Loshchilov & Hutter, 2017) optimizer with a cosine-decay learning-rate schedule. We set the initial learning rate to 4×10^{-5} , the warm-up ratio to 0.03, and the batch size to 256. The training duration for the 0.5B and 3B models is 12 hours and 24 hours, respectively.

Evaluation Setup. For visual prompt understanding and general image QA, we use the same settings as the corresponding base MLLM. For segmentation tasks, we use greedy decoding. If the model does not predict a [SEG] token, we output an all-zero mask and count the example as a failure; this case does not occur on the RefCOCO validation set.

Table 1: Performance on referring segmentation benchmarks. The evaluation metric is cloU. “ft” denotes fine-tuning on the specific dataset.

Method	MLLM Size	RefCOCO+			RefCOCOg		RefCOCO			GRES(gRefCOCO)		
		val	testA	testB	val(U)	test(U)	val	testA	testB	val	testA	testB
Referring Segmentation Specialists Without MLLMs												
VL2 (Ding et al., 2021)	-	56.3	61.0	50.1	55.0	57.7	67.5	70.5	65.2	52.5	62.2	50.5
CRIS (Wang et al., 2022)	-	62.3	68.1	53.7	59.9	60.4	70.5	73.2	66.1	55.3	63.8	51.0
LAVT (Yang et al., 2022b)	-	62.1	68.4	55.1	61.2	62.1	72.7	75.8	68.8	57.6	65.3	55.0
PolyFormer-L (Liu et al., 2023c)	-	69.3	74.6	61.9	69.2	70.2	76.0	78.3	73.3	-	-	-
ReLA (Liu et al., 2023a)	-	66.0	71.0	57.7	65.0	66.0	73.8	76.5	70.2	56.4	59.0	58.4
MLLMs With Vision Experts												
LISA (ft) (Lai et al., 2024)	7B	65.1	70.8	58.1	67.9	70.6	74.9	79.1	72.3	-	-	-
PixelLM (Ren et al., 2024)	7B	66.3	71.7	58.3	69.3	70.5	73.0	76.5	68.2	-	-	-
GSVA (ft) (Xia et al., 2024)	7B	64.5	67.7	58.6	71.1	72.0	76.4	77.4	72.8	61.7	69.2	60.3
GroundHog (Zhang et al., 2024d)	7B	70.5	75.0	64.9	74.1	74.6	78.5	79.9	75.7	66.7	-	-
GLaMM (ft) (Rasheed et al., 2024)	7B	72.6	78.7	64.6	74.2	74.9	79.5	83.2	76.9	-	-	-
SAM4MLLM (Chen et al., 2024e)	7B	73.5	77.8	65.8	74.5	75.6	79.6	82.8	76.1	66.3	70.1	63.2
LaSagnA (Wei et al., 2024a)	7B	66.4	70.6	60.1	70.6	71.9	76.8	78.7	73.8	38.1	50.4	42.1
OMG-LLaVA (ft) (Zhang et al., 2024b)	7B	69.1	73.1	63.0	72.9	72.9	78.0	80.3	74.1	-	-	-
F-LLM (Wu et al., 2025)	7B	65.8	75.2	58.5	70.1	71.7	75.8	79.5	72.4	-	-	-
UFO (ft) (Tang et al., 2025)	8B	72.3	76.8	66.6	73.7	74.3	78.0	79.7	75.6	-	-	-
Sa2VA (Yuan et al., 2025a)	4B	74.3	-	-	76.7	-	80.4	-	-	-	-	-
MLLMs Without Vision Experts												
Pixel-SAIL	0.5B	70.8	75.8	65.4	75.4	76.7	77.9	80.5	75.9	63.9	71.5	63.6
Pixel-SAIL (ft)	0.5B	73.0	77.0	68.0	75.6	76.1	79.1	81.7	77.0	68.0	74.0	66.8
Pixel-SAIL	3B	75.7	79.7	72.0	78.7	80.4	80.8	82.6	79.0	67.7	74.6	67.1
Pixel-SAIL (ft)	3B	76.2	79.7	71.2	78.5	79.4	81.8	83.4	78.8	72.1	77.1	70.4

Table 2: Performance on PerBench. Because LISA cannot accept visual prompts, it scores 0 on all tasks.

Model	Size	Detailed Caption METEOR	MCQ Acc	V-TRES cloU	Overall Score
LISA (Lai et al., 2024)	7B	0	0	0	0
Osprey (Yuan et al., 2024b)	7B	13.4	0.12	0	8.5
GLaMM (Rasheed et al., 2024)	7B	12.6	0.14	24.3	15.3
Sa2VA (Yuan et al., 2025a)	4B	19.2	0.71	31.9	39.0
Pixel-SAIL	0.5B	21.4	0.69	29.7	38.4
Pixel-SAIL	3B	24.2	0.74	33.4	42.2

Table 3: Performance on the VQA benchmarks. ★ refers to the use of an 800² resolution, which differs from the 1600² resolution in the pre-trained model.

Model	LLM Size	MME	MMBench	SEED	MMStar
SOLO	0.5B	523.2/222.5	13.8	45.5	26.2
SOLO	3B	1155.7/257.5	53.4	65.4	40.3
EVEv2★	7B	1128.0/240.7	60.3	54.2	44.9
Pixel-SAIL	0.5B	564.1/150.7	31.8	52.2	26.3
Pixel-SAIL	3B	1187.3/242.9	56.3	66.1	40.1
Pixel-SAIL★	7B	1081.0/260.4	58.9	64.7	44.3

3.1 Main Results

Results on Referring Segmentation Benchmarks. We compare Pixel-SAIL with other pixel-grounded MLLMs and segmentation specialists on RefCOCO+ (Yu et al., 2016), RefCOCOg (Yu et al., 2016), RefCOCO (Kazemzadeh et al., 2014), and gRefCOCO (Liu et al., 2023a), with results shown in Tab. 1. Without dataset-specific fine-tuning, Pixel-SAIL 0.5B obtains 70.8, 75.4, and 77.9 cloU on the RefCOCO+, RefCOCOg, and RefCOCO validation splits, respectively, while retaining image-chat capabilities. After dataset-specific fine-tuning, Pixel-SAIL 0.5B outperforms the SAM-based LISA 7B (Lai et al., 2024) by 4.2, 7.9, and 7.7 cloU points on RefCOCO, RefCOCO+, and RefCOCOg, respectively. On gRefCOCO, the fine-tuned Pixel-SAIL 0.5B outperforms GSVA 7B (Xia et al., 2024) by 6.3, 4.8, and 6.5 cloU points on the validation, testA, and testB splits, respectively.

When scaled to 3B parameters, Pixel-SAIL obtains 75.7, 78.7, 80.8, and 67.7 cloU on the validation splits of RefCOCO+, RefCOCOg, RefCOCO, and gRefCOCO, respectively, outperforming the reported validation results of larger 7B MLLMs with vision experts. Pixel-SAIL 3B also outperforms Sa2VA 4B (Yuan et al., 2025a), which uses InternVL2-4B (Chen et al., 2024f) and SAM2-L (Ravi et al., 2024), by 1.4 and 2.0 cloU points on the RefCOCO+ and RefCOCOg validation splits.

Results on Visual Prompt Understanding Benchmarks. We evaluate region captioning on RefCOCOg, with results shown in Tab. 4. Because Pixel-SAIL is not trained on the RefCOCOg region-captioning data, we evaluate it in a zero-shot setting. Pixel-SAIL 0.5B achieves a METEOR score of 16.0, exceeding OMG-LLaVA 7B by 0.7 points. Scaling Pixel-SAIL to 3B raises the score to 17.6, exceeding Osprey 7B and GLaMM 7B by 1.0 and 1.4 points, respectively.

Table 4: Region caption performance on RefCOCOg dataset.

Method	Size	METEOR
Sa2VA	4B	17.3
OMG-LLaVA	7B	15.3
Osprey	7B	16.6
GLaMM	7B	16.2
Pixel-SAIL	0.5B	16.0
Pixel-SAIL	3B	17.6

Table 5: Ablation study on the components of Pixel-SAIL. "RC" denotes region caption on the RefCOCOg dataset.

Model	RefCOCO/+ /g	RC
Plain Architecture	74.1/67.3/71.9	1.2
+ Upsampling	76.2/69.6/73.8	1.4
+ VP Injection	77.4/70.4/75.2	16.1
+ Distillation	77.9/70.8/75.4	16.0

Table 6: Ablation study on the base MLLM. The training data include only LLaVA-665k and RefCOCO/+ /g.

MLLM	Size	RefCOCO/+ /g
SAIL	0.5B	69.7/62.5/65.3
SAIL	3B	73.2/66.4/69.1
EVEv2	7B	74.9/68.7/71.3

Table 7: Ablation study on the distillation strategy.

Teacher	RefCOCO/+ /g
w/o Distill	77.5/70.5/75.5
M2F	77.7/71.0/75.8
SAM2	77.8/70.9/75.9
Both	78.1/70.8/76.1

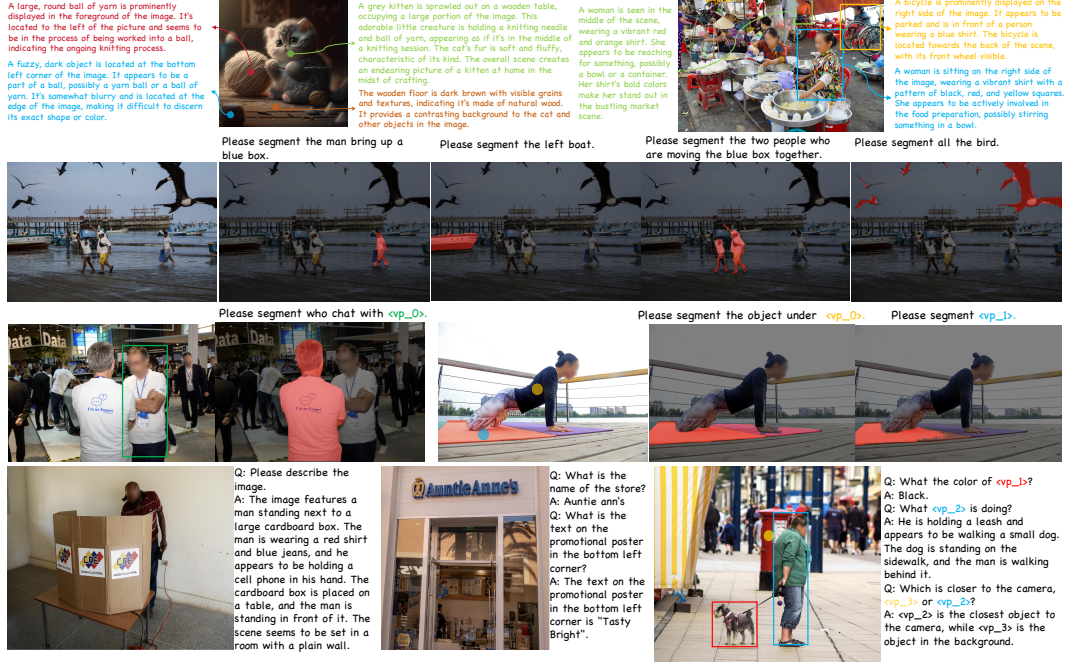


Figure 4: Qualitative results of Pixel-SAIL across diverse tasks. Best viewed on screen. From top to bottom, the examples show visual prompt-based object captioning, single- and multi-object referring segmentation, visual-text referring segmentation, image captioning and QA, and visual prompt-based conversation. Point and box prompts are converted into mask prompts using SAM (Kirillov et al., 2023). Additional results and comparisons with other MLLMs are provided in the appendix.

Results on PerBench. We evaluate several pixel-grounded MLLMs on PerBench, with results shown in Tab. 2. LISA (Lai et al., 2024) scores 0 on all tasks because it cannot accept visual prompts. Osprey (Yuan et al., 2024b) obtains 13.4 METEOR on detailed captioning and 12.0% accuracy on the multiple-choice task, possibly because its training data contain short object captions and provide limited instruction-following supervision. GLaMM (Rasheed et al., 2024) and Sa2VA (Yuan et al., 2025a) support both prompt understanding and segmentation, although GLaMM obtains only 14.0% accuracy on the multiple-choice task. Pixel-SAIL 0.5B achieves an overall score of 38.4, comparable to Sa2VA 4B despite Sa2VA using a larger base MLLM and an external segmentation expert. Pixel-SAIL 3B achieves an overall score of 42.2 and outperforms Sa2VA 4B on all three tasks.

Results on VQA Benchmarks. We compare the visual question-answering performance of Pixel-SAIL with the corresponding base MLLMs on MME (Fu et al., 2023), MMBench (Liu et al., 2024b), SEED (Li et al., 2024b), and MMStar (Chen et al., 2024b), with results shown in Tab. 3. At 0.5B parameters, Pixel-SAIL outperforms the base MLLM on all four benchmarks, particularly on MMBench, where the score increases from 13.8 to 31.8. At 3B and 7B parameters, Pixel-SAIL remains comparable to the base MLLMs. Further improvements may require a larger and more diverse collection of visual-prompt and segmentation data; the current training set contains fewer than 2M such examples.

3.2 Ablation Studies

Effectiveness of Each Component. We conduct comprehensive ablation studies on the proposed components, with results presented in Tab. 5. Our plain architecture achieves 74.1, 67.3, and 71.9 cIoU on the RefCOCO, RefCOCO+, and RefCOCOg datasets, respectively. However, this plain architecture performs poorly on visual prompt understanding, attaining only 1.2 METEOR on region captioning. Upon incorporating the learnable up-sampling module, segmentation quality improves dramatically, with the model achieving cIoU scores of 76.2, 69.6, and 73.8 on RefCOCO, RefCOCO+, and RefCOCOg, respectively. However, the model still cannot effectively interpret user-input visual prompts due to insufficient semantic information in the object representation. After adding visual prompt injection, the model reaches 16.1 METEOR on region captioning. The same component also improves referring segmentation, suggesting that stronger visual prompt understanding benefits pixel-text alignment. Finally, the distillation strategy further improves segmentation detail.

Ablation on Various MLLMs. To evaluate Pixel-SAIL across different base architectures and sizes, we train using only LLaVA-665k and RefCOCO+/g and report referring-segmentation results in Tab. 6. With SAIL 0.5B as the base MLLM, Pixel-SAIL obtains 69.7, 62.5, and 65.3 cIoU on RefCOCO+/g. Scaling SAIL to 3B improves these scores by 3.5, 3.9, and 3.8 points, respectively. With EVEv2 7B, which uses causal rather than full attention between vision tokens and adopts a mixture-of-experts architecture, Pixel-SAIL obtains 74.9, 68.7, and 71.3 cIoU. These results show that the method generalizes across base MLLM architectures and sizes.

Ablation on Distillation Strategy. We evaluate feature distillation in Tab. 7, using average cIoU across all splits as the metric. When only Mask2Former (Cheng et al., 2022) is employed to distill high-resolution mask features, Pixel-SAIL achieves performance gains of 0.2, 0.5, and 0.3 on RefCOCO+/g. When SAM2 (Ravi et al., 2024) is used to distill low-resolution image features, Pixel-SAIL obtains performance improvements of 0.3, 0.4, and 0.4 on RefCOCO+/g. When both teacher models are used, performance improves by 0.6, 0.3, and 0.6 points. For Pixel-SAIL 3B, the distillation strategy adds 22 A100 GPU hours, corresponding to approximately 3% additional training time, and incurs no inference-time cost (Tab. 12).

3.3 Visualization Analysis

Visual Comparison. Figure 4 presents qualitative results from Pixel-SAIL across diverse tasks. Pixel-SAIL interprets both visual prompts and textual instructions and responds with text and segmentation masks. Feature analysis, additional qualitative examples, and challenging cases are provided in Appendix Secs. C and E.

4 Conclusion

We investigate a minimal architecture for pixel-grounded understanding. Pixel-SAIL extends a SAIL-like MLLM with visual prompt injection, learnable upsampling, and training-time feature distillation. Our results show that a single-transformer MLLM can achieve strong performance on four public referring-segmentation benchmarks without an external vision encoder or segmentation expert at inference time. We also introduce PerBench, a benchmark for detailed object captioning, visual-prompt question answering, and visual-text referring segmentation. We hope these findings encourage further research on simple architectures for fine-grained multimodal understanding.

Limitations and Future Work. Pixel-SAIL is trained on only 1.7M examples. Future work will investigate larger training sets, including datasets with billions of masks and visual prompts Kirillov et al. (2023).

References

- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *ACL*, 2005.
- Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. Forest-of-thought: Scaling test-time compute for enhancing llm reasoning. *arXiv preprint arXiv:2412.09078*, 2024.
- Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Making large multimodal models understand arbitrary visual prompts. In *CVPR*, 2024a.
- Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024b.
- Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. *arXiv preprint arXiv:2406.16866*, 2024a.
- Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024b.
- Tao Chen, Enwei Zhang, Yuting Gao, Ke Li, Xing Sun, Yan Zhang, Hui Li, and Rongrong Ji. Mmict: Boosting multi-modal fine-tuning with in-context examples. *ACM TOMM*, 2024c.
- Xinghao Chen, Siwei Li, Yijing Yang, and Yunhe Wang. Deco: Unleashing the potential of convnets for query-based detection and segmentation. In *ICLR*, 2025.
- Yangyi Chen, Xingyao Wang, Hao Peng, and Heng Ji. A single transformer for scalable vision-language modeling. *TMLR*, 2024d.
- Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. *ECCV*, 2024e.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024f.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024g.
- Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2019.
- Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *arXiv preprint arXiv:2406.11832*, 2024.
- Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *NeurIPS*, 2025a.
- Haiwen Diao, Xiaotong Li, Yufeng Cui, Yueze Wang, Haoge Deng, Ting Pan, Wenxuan Wang, Huchuan Lu, and Xinlong Wang. Evev2: Improved baselines for encoder-free vision-language models. *arXiv preprint arXiv:2502.06788*, 2025b.

- Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 2021.
- Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *ACMMM*, 2024.
- Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *CVPR*, 2021.
- Qianhan Feng, Wenshuo Li, Tong Lin, and Xinghao Chen. Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language model. In *CVPR*, 2025.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiauwu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025.
- Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022.
- Kai Han, Jianyuan Guo, Yehui Tang, Wei He, Enhua Wu, and Yunhe Wang. Free video-llm: Prompt-guided visual perception for efficient training-free video llms. *arXiv preprint arXiv:2410.10441*, 2024.
- Louis Martin Hugo Touvron, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.
- Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *ECCV*, 2020.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.
- Ding Jia, Jianyuan Guo, Kai Han, Han Wu, Chao Zhang, Chang Xu, and Xinghao Chen. Geminifusion: Efficient pixel-wise multimodal fusion for vision transformer. *arXiv preprint arXiv:2406.01210*, 2024.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *ICCV*, 2023.
- Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, 2024.
- Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaying Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Text4seg: Reimagining image segmentation as text generation. In *ICLR*, 2025.
- Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. *arXiv preprint arXiv:2504.10462*, 2025.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.
- Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *CVPR*, 2024b.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022a.

- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023a.
- Wenshuo Li, Xinghao Chen, Han Shu, Yehui Tang, and Yunhe Wang. Excp: Extreme llm checkpoint compression via weight-momentum joint shrinking. *arXiv preprint arXiv:2406.11257*, 2024c.
- Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *CVPR*, 2024d.
- Xiaotong Li, Fan Zhang, Haiwen Diao, Yuezhe Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *arXiv preprint arXiv:2407.08303*, 2024e.
- Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *ECCV*, 2022b.
- Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. Rain: Your language models can align themselves without finetuning. *arXiv preprint arXiv:2309.07124*, 2023b.
- Chen Liang, Wenguan Wang, Tianfei Zhou, Jiaxu Miao, Yawei Luo, and Yi Yang. Local-global context aware transformer for language-guided video segmentation. *arXiv preprint arXiv:2203.09773*, 2022.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2403.20271*, 2024.
- Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *CVPR*, 2023a.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023b.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024a.
- Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. *CVPR*, 2023c.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2024b.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint*, 2017.
- Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *CVPR*, 2022.
- Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. *NeurIPS*, 2023.
- Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiwu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024a.
- Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-intervl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. *CVPR*, 2025.
- Yongdong Luo, Xiwu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *arXiv preprint arXiv:2411.13093*, 2024b.
- Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *ECCV*, 2024.

- Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Jiayi Ji, Jie Lou, Debing Zhang, and Rongrong Ji. Mllm-selector: Necessity and diversity-driven high-value data selection for enhanced visual instruction tuning. *arXiv preprint arXiv:2503.20502*, 2025.
- Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. *arXiv preprint arXiv:2411.04923*, 2024.
- Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks master of many: Designing general-purpose coarse-to-fine vision-language model. In *CVPR*, 2024.
- Lu Qi, Yi-Wen Chen, Lehan Yang, Tiancheng Shen, Xiangtai Li, Weidong Guo, Yu Xu, and Ming-Hsuan Yang. Generalizable entity grounding via assistance of large language model. *arXiv preprint arXiv:2402.02555*, 2024.
- 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 *ICML*, 2021.
- Miao Rang, Zhenni Bi, Chuanjian Liu, Yehui Tang, Kai Han, and Yunhe Wang. Eve: Efficient multimodal vision language models with elastic visual experts. *arXiv preprint arXiv:2501.04322*, 2025.
- Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *SIGKDD*, 2020.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *CVPR*, 2024.
- Yunhang Shen, Chaoyou Fu, Shaoqi Dong, Xiong Wang, Peixian Chen, Mengdan Zhang, Haoyu Cao, Ke Li, Xiaowu Zheng, Yan Zhang, et al. Long-vita: Scaling large multi-modal models to 1 million tokens with leading short-context accuracy. *arXiv preprint arXiv:2502.05177*, 2025.
- Han Shu, Wenshuo Li, Yehui Tang, Yiman Zhang, Yihao Chen, Houqiang Li, Yunhe Wang, and Xinghao Chen. Tinsam: Pushing the envelope for efficient segment anything model. In *AAAI*, 2025.
- Hao Tang, Chenwei Xie, Haiyang Wang, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Ufo: A unified approach to fine-grained visual perception via open-ended language interface. *arXiv preprint arXiv:2503.01342*, 2025.
- Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CVPR*, 2015.
- Haiyang Wang, Hao Tang, Li Jiang, Shaoshuai Shi, Muhammad Ferjad Naeem, Hongsheng Li, Bernt Schiele, and Liwei Wang. Git: Towards generalist vision transformer through universal language interface. In *ECCV*, 2024a.
- Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *ICCV*, October 2023a.

- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.
- Tao Wang, Changxu Cheng, Lingfeng Wang, Senda Chen, and Wuyue Zhao. Himtok: Learning hierarchical mask tokens for image segmentation with large multimodal model. In *ICCV*, 2025.
- Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and Jifeng Dai. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023b.
- Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, 2022.
- Cong Wei, Haoxian Tan, Yujie Zhong, Yujiu Yang, and Lin Ma. LaSagnA: Language-based segmentation assistant for complex queries. *arXiv preprint arXiv:2404.08506*, 2024a.
- Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Zheng Zhao, Jie Hu, and Yujiu Yang. Hyperseg: Towards universal visual segmentation with large language model. *arXiv preprint arXiv:2411.17606*, 2024b.
- Fangyun Wei, Jinjing Zhao, Kun Yan, Hongyang Zhang, and Chang Xu. A large-scale human-centric benchmark for referring expression comprehension in the Imm era. In *NeurIPS*, 2024c.
- Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. Towards open vocabulary learning: A survey. *arXiv pre-print*, 2023.
- Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. Controlmlm: Training-free visual prompt learning for multimodal large language models. *NeurIPS*, 2024.
- Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-Imm: Grounding frozen large multimodal models. *CVPR*, 2025.
- Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *CVPR*, 2024.
- Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *CVPR*, 2020.
- Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023.
- Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstathios Gavves. Visa: Reasoning video object segmentation via large language models. In *ECCV*, 2024.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yaqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024b.
- Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022a.
- Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022b.

- Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *ICLR*, 2024.
- Zuyao You, Junke Wang, Lingyu Kong, Bo He, and Zuxuan Wu. Pix2cap-coco: Advancing visual comprehension via pixel-level captioning. *arXiv preprint arXiv:2501.13893*, 2025.
- Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016.
- Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. *arXiv preprint*, 2024a.
- Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025a.
- Haobo Yuan, Tao Zhang, Xiangtai Li, Lu Qi, Zilong Huang, Shilin Xu, Jiashi Feng, and Ming-Hsuan Yang. 4th pvvw mevis 3rd place report: Sa2va. *arXiv preprint arXiv:2504.00476*, 2025b.
- Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *CVPR*, 2024b.
- Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *IJCV*, 2025.
- Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023.
- Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Chunyuan Li, Jainwei Yang, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *ECCV*, 2024a.
- Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023a.
- Tao Zhang, Shiqing Wei, and Shunping Ji. E2ec: An end-to-end contour-based method for high-quality high-speed instance segmentation. In *CVPR*, 2022.
- Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. DVIS: Decoupled video instance segmentation framework. In *ICCV*, 2023b.
- Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. Dvis++: Improved decoupled framework for universal video segmentation. *arXiv preprint arXiv:2312.13305*, 2023c.
- Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Change Loy Chen, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. In *NeurIPS*, 2024b.
- Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in vlms. *arXiv preprint arXiv:2411.09968*, 2024c.
- Xiaofeng Zhang, Fanshuo Zeng, Yihao Quan, Zheng Hui, and Jiawei Yao. Enhancing multimodal large language models complex reason via similarity computation. In *AAAI*, 2025.
- Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *CVPR*, 2024d.
- Dehua Zheng, Wenhui Dong, Hailin Hu, Xinghao Chen, and Yunhe Wang. Less is more: Focus attention for efficient detr. In *ICCV*, 2023.

- Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.
- Yikang Zhou, Tao Zhang, Shilin Xu, Shihao Chen, Qianyu Zhou, Yunhai Tong, Shunping Ji, Jiangning Zhang, Xiangtai Li, and Lu Qi. Are they the same? exploring visual correspondence shortcomings of multimodal llms. *arXiv preprint arXiv:2501.04670*, 2025.
- Yikang Zhou, Tao Zhang, Dengxian Gong, Yuanzheng Wu, Ye Tian, Haochen Wang, Haobo Yuan, Jiacong Wang, Lu Qi, Hao Fei, et al. Samtok: Representing any mask with two words. *CVPR*, 2026.
- Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *NeurIPS*, 2023.

Appendix

A Overview

We first provide additional training and testing details for Pixel-SAIL in Sec. B. We then present visual affinity map analysis in Sec. C, the detailed benchmark construction process in Sec. D, and challenging PerBench examples in Sec. E. Next, we provide additional comparisons with state-of-the-art pixel-grounded MLLMs in Sec. F. Additional related-work discussion appears in Sec. G. Finally, we provide additional analyses, including scaling and computational efficiency in Sec. H, robustness and alternative prompt representations in Sec. I, generalization across different base MLLMs and further distillation analysis in Sec. J, and a discussion of scope and limitations in Sec. K.

B More Detailed Training and Testing

Training. We provide further training details, including dataset sampling and the distillation methodology. For the RefCOCO-series datasets (Kazemzadeh et al., 2014; Yu et al., 2016), we randomly sample five referring expressions per image, organize them into a multi-round dialogue as one training example, and process all images for four epochs. For COCO (Lin et al., 2014), we sample five categories per image and randomly select either instance or semantic mode when constructing the responses. In instance mode, objects are ordered from left to right by their center points. We process COCO for one epoch. For Pixel2Cap (You et al., 2025), our generated detailed object captions, and the Osprey (Yuan et al., 2024b) object-description data, we randomly sample one to five visual prompts per image. We also include questions about nonexistent visual prompts, for which the response states that the prompt does not exist. We process these object-caption datasets for five epochs and all other segmentation or visual-prompt datasets for one epoch. During joint training, we sample LLaVA-665k examples at a 1:1 ratio with the remaining data to preserve the base MLLM’s instruction-following capability. When the input sequence, including vision tokens, exceeds 8192 tokens, we truncate the excess. We train the 0.5B model with DeepSpeed ZeRO-1 (Rasley et al., 2020) and the 3B and 7B models with DeepSpeed ZeRO-2 (Rasley et al., 2020).

We distill the mask features produced by the Mask2Former (Cheng et al., 2022) pixel decoder and the lowest-resolution features produced by the SAM2 (Ravi et al., 2024) image encoder into Pixel-SAIL’s upsampled mask features and reshaped vision-token features, respectively. We use bilinear interpolation to align the spatial dimensions and a learnable linear layer to align the channel dimensions. The distillation objective uses an MSE loss with a weight of 0.5.

For dense feature distillation, we use Mask2Former decoder features to supervise the upsampled mask features and SAM2 encoder features to supervise the reshaped vision tokens. The reason is that, in SAM2, decoder image features depend on prompt embeddings and therefore vary across segmented objects, making them unsuitable as stable distillation targets. We also observe that simultaneously distilling the vision tokens with image features from both Mask2Former and SAM2 may occasionally cause training instability. In visual prompt injection, when multiple prompt masks overlap, we sort them by area from large to small and sequentially fill the prompt embeddings, so smaller masks have higher priority and overwrite the overlapping regions. We set the maximum number of simultaneously injected visual prompts to 15.

Architectural details. In the reported configuration, final-layer vision-token features form a stride-16 spatial grid. The learnable upsampler contains two blocks, each comprising a transposed 2D convolution followed by a depthwise convolution and increasing the spatial resolution by $2\times$; the resulting mask features therefore have one-quarter of the input resolution. We linearly project the hidden state of each generated [SEG] token and take its dot product with the upsampled pixel features to obtain the mask logits.

Testing. The main text describes evaluation on pixel-grounded benchmarks. For general image question-answering benchmarks, we follow the prompt settings of the base MLLMs

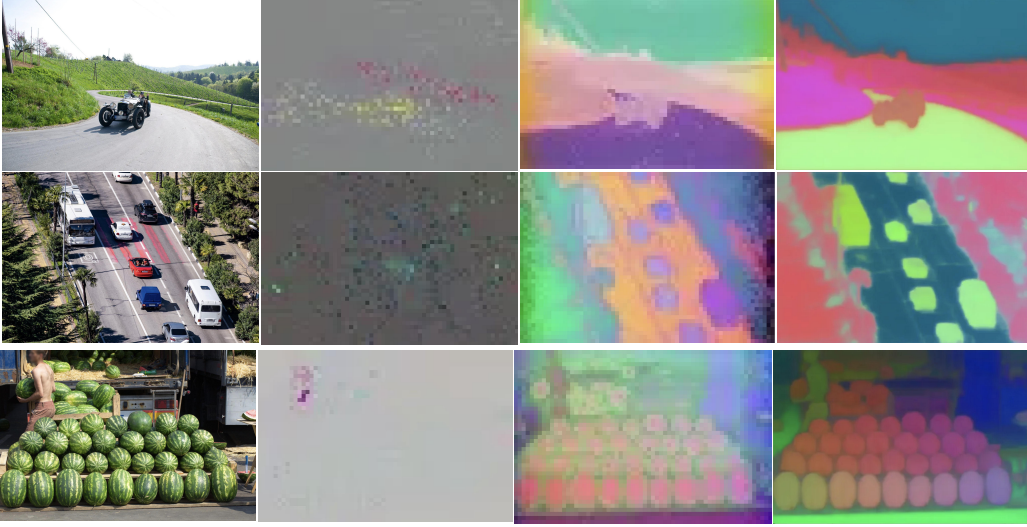


Figure 5: **Image-feature visualizations.** From left to right, the columns show image features from the base MLLM, image features from Pixel-SAIL, and mask features from Pixel-SAIL.

and use VLMEvalKit (Duan et al., 2024) without additional LLM assistance for answer extraction. For referring segmentation, we use greedy decoding and convert each generated [SEG] token into a mask. If no [SEG] token is generated, we return an all-zero mask and count the sample as a failure. The missing-[SEG] rate is 0% on the RefCOCO validation set.

C Visual Affinity Map Analysis

We use principal component analysis (PCA) to visualize the features in Fig. 5. Pixel-SAIL’s image features (third column) are denser and more diverse than those of the base MLLM (second column), while its upsampled mask features produce clearer object boundaries. The image features encode semantic attributes such as category, color, and position, whereas the mask features emphasize category and instance-level perception. For example, in the second row, the left and right cars have distinct image-feature representations but similar mask-feature representations.

D Detailed PerBench Construction Process

The construction of PerBench combines an automated model-generated pipeline with manual screening, correction, and annotation. The process is divided into three stages.

The **first stage** involves annotating detailed object captions. We crop objects and draw visual prompts on the original images, then prompt InternVL2.5-78B (Chen et al., 2024f) and Qwen2.5-VL-72B (Bai et al., 2025) to generate detailed captions. These captions are then cross-validated using Qwen2.5 72B (Yang et al., 2024b). If the captions are consistent, an LLM consolidates them into a single caption; otherwise, we discard the example. We manually select and correct 500 generated captions to form the detailed object-caption subset of the benchmark.

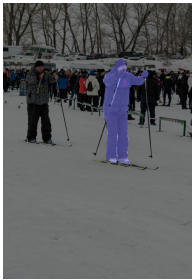
The **second stage** creates visual-prompt question-answering data in a multiple-choice question (MCQ) format. In this phase, we manually generate a multiple-choice question for each object caption obtained from the first stage. After completing the annotations, two quality control specialists perform cross-verification to identify and rectify any potential errors.



The object is a person standing behind a wooden podium covered with a blue cloth, addressing an audience outdoors. The podium has an emblem, and the person is dressed in a dark blue jacket with a logo, a scarf, and a white shirt. The audience includes people seated on a stool and a folding chair, with trees, parked cars, and a building with large windows visible in the background.



The object is a bronze statue of a young boy with exaggerated features. The statue is dressed in a formal outfit including a suit jacket, vest, and shorts, and it wears flip-flops. The boy holds a small object resembling a ball in his right hand. The statue's whimsical and playful appearance, characterized by exaggerated proportions and a sense of movement, is set on a rock-like structure in an urban street scene with a pedestrian crossing visible in the background.



The object is a person dressed in black and blue attire, standing on skis in a snowy outdoor setting, preparing for or participating in a skiing event. This person is wearing a black beanie, a black puffy jacket, black gloves, black pants, and a blue bib with the number "1" in red. They also have ski poles and ski boots. In the background, there are other people skiing or standing around, along with parked cars and trees, indicating a recreational or competitive skiing area.



The object is a Smart Fortwo, a compact city car known for its small size and maneuverability. This car features predominantly white with green accents, including green wheels and trim. Designed for urban environments with limited parking space, it is parked on a sidewalk next to a charging station, indicating its electric vehicle status. The charging cable is connected, suggesting it is currently being charged. The surrounding area includes a paved road with multiple lanes and a grassy area separated by a curb.



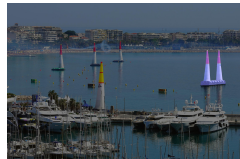
The object is a stack of white bangles adorned with small, colorful stones. These bangles, part of traditional jewelry, are worn on the arm of a person dressed in vibrant traditional attire, including a yellow scarf with colorful patterns and a floral-patterned garment, complementing the overall colorful and cultural appearance.



The object is a waterfall cascading down a steep, rocky cliff. The waterfall serves as the central focus, with water flowing smoothly from top to bottom. The area around the waterfall features rugged, uneven rock surfaces and patches of green vegetation. The cliff is part of a larger mountainous or hilly terrain, with dense foliage at the top. The waterfall's flow and movement create a striking contrast against the dark, textured rocks, highlighting the natural beauty of the scene.



The object is a black shoulder bag with a flap closure. It is made of a synthetic or leather-like material, has a smooth texture and sheen, and is carried over the shoulder of a person walking in a grassy area near a wooden fence, dressed in a black top and pink shorts.



The object is a navigational buoy used in maritime environments. It is a cylindrical structure with a pointed top, primarily white in color, featuring a pink top and a distinctive band. Mounted on a black base, it is situated in a body of water such as a harbor or marina, surrounded by other boats and buoys. Its function is to guide vessels through waterways, aiding navigation with its unique color pattern. This buoy is part of a larger maritime setting, often near populated areas or popular boating destinations.

Figure 6: More visualization examples of detailed object captions from our PerBench.

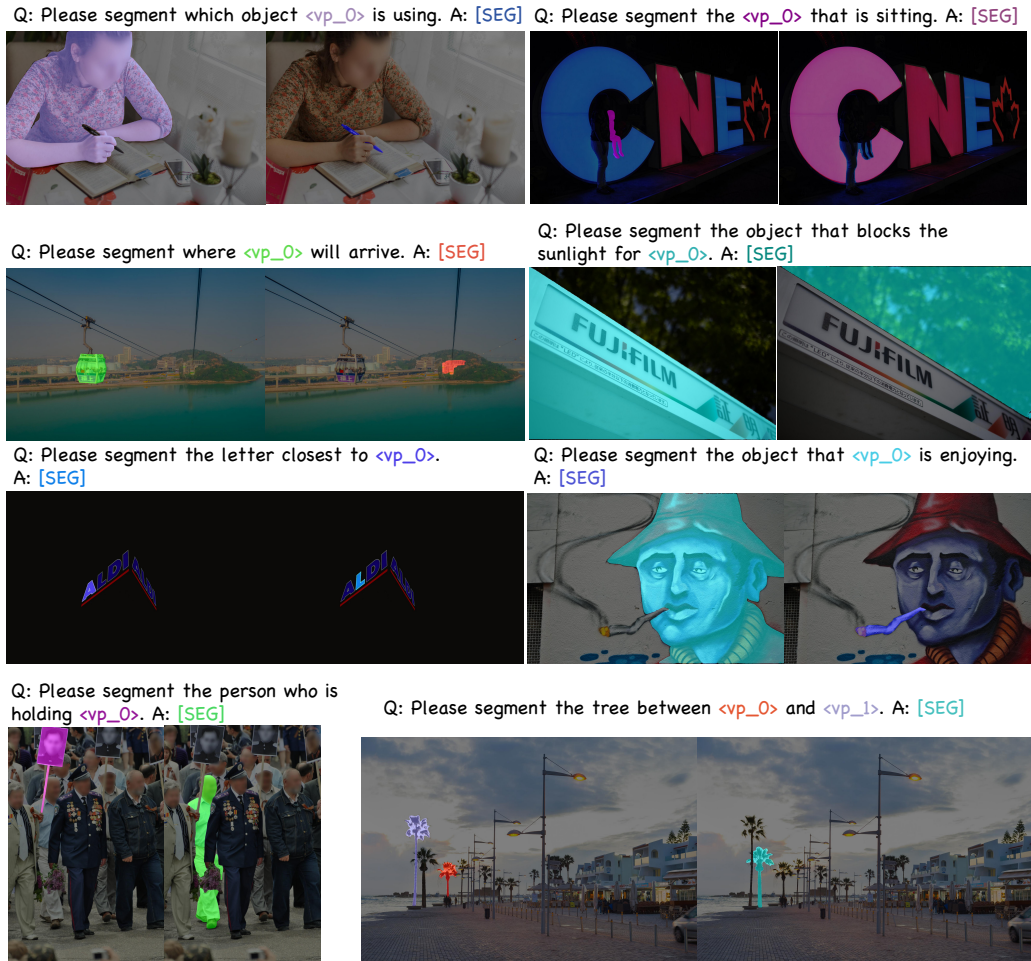


Figure 7: Additional examples of visual-text referring segmentation from PerBench.

Table 8: Visual-prompt composition of the PerBench V-T RES subset.

Number of prompts	Samples	Proportion
1	364	72.8%
2	110	22.0%
≥ 3	26	5.2%
Total	500	100.0%

Table 9: Comparison with state-of-the-art pixel-grounded MLLMs. † denotes results obtained using the same segmentation data as Pixel-SAIL.

Method	MLLM Size	RefCOCO+	RefCOCOg	RefCOCO	GRES(gRefCOCO)
Sa2VA (Yuan et al., 2025a)	4B	74.3	76.7	80.4	-
Sa2VA† (Yuan et al., 2025a)	4B	74.6	76.8	80.2	65.1
Pixel-SAIL	3B	75.7	78.7	80.8	67.7

The **final stage** involves annotating visual-text referring segmentation data. We manually select and annotate object masks, visual prompts, and referring text from SAM images. The annotations cover position, interactions, appearance, and size, including cases with both single and multiple visual prompts. All 500 examples in the visual-text referring segmentation (V-T RES) subset are double-checked by five expert annotators. The relations include spatial relations (e.g., left or right) and dependency relations (e.g., an object worn by or otherwise associated with the visually prompted object).

Overlap audit and prompt composition. The source images are drawn from the SAM image collection and are manually reannotated for PerBench. An overlap audit against the Pixel-SAIL training data finds no shared examples. Table 8 reports the number of visual prompts in the 500 V-T RES samples. Most samples contain one prompt, while 27.2% require reasoning over multiple prompts.

E More Challenging Cases in PerBench

Figure 6 presents detailed object-caption examples from PerBench. The objects come from diverse scenes and categories, including people, human-made objects, and natural environments. The object captions include basic categories, attributes, purposes, and relationships with surrounding objects. These examples illustrate the breadth of the benchmark for evaluating visual prompt understanding.

More referring segmentation samples are illustrated in Fig. 7. Our manually annotated samples cover a variety of scenes, such as indoor and outdoor settings, and include objects of multiple granularities. The referring text describes positional and interaction relationships. This task is more challenging than conventional text-only referring segmentation because it requires joint interpretation of visual and textual references.

F Additional Comparison with State-of-the-Art Pixel-Grounded MLLMs

We further fine-tune Sa2VA using exactly the same segmentation data as Pixel-SAIL, with results shown in Tab. 9. Sa2VA improves only slightly on RefCOCO, RefCOCO+, and RefCOCOg after training with additional COCO and GRES data. One possible explanation is that SAM has already been trained on a substantially larger dataset. Even after training on the same data, Sa2VA 4B performs worse than Pixel-SAIL, particularly on multi-object segmentation in GRES. Pixel-SAIL also achieves these results with fewer model parameters than Sa2VA.

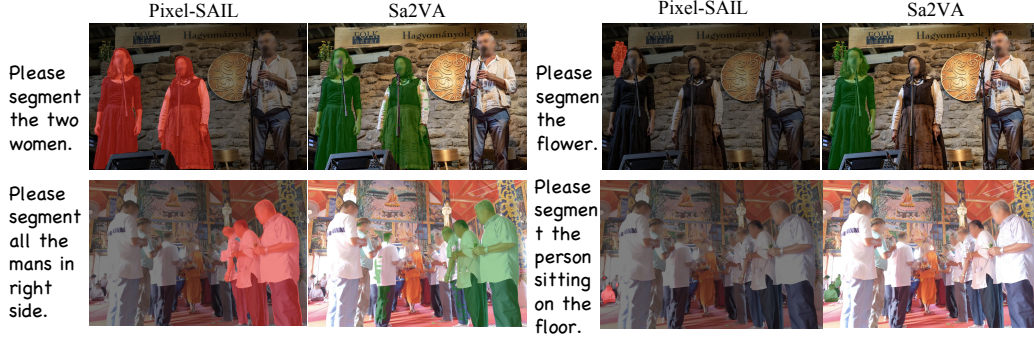


Figure 8: **Qualitative comparison between Sa2VA and Pixel-SAIL.**

Figure 8 qualitatively compares Pixel-SAIL with Sa2VA (Yuan et al., 2025a). Both methods perform well in many cases, but the examples reveal two recurring differences. In the examples on the left, Sa2VA misses more objects in multi-object segmentation. In the example on the right, it fails to segment an object near the image boundary, whereas Pixel-SAIL identifies it. These examples complement the quantitative comparison but do not by themselves establish a general causal explanation.

G More Related Work Discussion

Large Vision-Language Models. Starting with CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), modern vision-language models have used contrastive learning on large-scale image-text datasets to learn aligned representations. These models also perform well on open-vocabulary perception tasks such as segmentation (Yuan et al., 2024a; Lüddecke & Ecker, 2022; Zhang et al., 2023b;c; Shu et al., 2025) and detection (Gu et al., 2022; Wu et al., 2023; Wang et al., 2023a; Zareian et al., 2021; Jia et al., 2024; Chen et al., 2025; Zheng et al., 2023). Subsequent works (Li et al., 2022a; Zhai et al., 2023; Li et al., 2023a; Xu et al., 2023) retain similar architectures while exploring modified objectives, improved data quality, and data filtering. With the rise of large language models (Hugo Touvron et al., 2023; Yang et al., 2024a; Touvron et al., 2023; Cai et al., 2024b; Li et al., 2024c), recent works (Bai et al., 2023; Zhang et al., 2023a; Chen et al., 2024g; Liu et al., 2023b; Chen et al., 2024f; Tong et al., 2024; Zhang et al., 2024c; 2025; Rang et al., 2025; Bi et al., 2024; Han et al., 2024; Li et al., 2023b; Feng et al., 2025; Shen et al., 2025; Fu et al., 2025; Luo et al., 2024b; Ma et al., 2025; Zhou et al., 2024; Zhu et al., 2023; Wu et al., 2024; Luo et al., 2024a; Chen et al., 2024c; Luo et al., 2023) have focused on multimodal large language models for open-ended tasks such as visual question answering and OCR. One representative work, LLaVA (Liu et al., 2023b), uses CLIP to encode images into visual tokens and feeds them to an LLM. Subsequent works (Bai et al., 2023; Li et al., 2024a) improve this design through larger high-quality datasets and joint image-video training. Meanwhile, several recent works (Diao et al., 2025b; Chen et al., 2024d; Diao et al., 2024; Luo et al., 2025) also explore visual encoder-free designs that jointly learn image and text representations within a single transformer architecture. For example, SOLO (Chen et al., 2024d) trains one transformer on mixed language and vision datasets for VQA, while EVE (Diao et al., 2024) uses CLIP-based supervision to enhance visual-token learning. Our work follows the visual encoder-free design and goes a step further by exploring pixel-grounded understanding tasks, including grounding and visual prompt understanding. To our knowledge, we are the first to apply an encoder-free architecture for pixel-grounded understanding, conversation, and reasoning tasks.

Referring Expression Segmentation and Pixel-MLLMs. This task outputs specific masks driven by text description. Earlier works (Feng et al., 2021; Yang et al., 2022a; Hui et al., 2020; Liang et al., 2022; Liu et al., 2023a) explore various fusion architectures and modules to enhance the alignment of text and vision features. Equipped with LLMs, several recent advanced works (Lai et al., 2024; Zhang et al., 2024b; Xia et al., 2024; Qi et al., 2024; Yuan et al., 2025a;b; Munasinghe et al., 2024; Rasheed et al., 2024; Zhou et al., 2025; Wang

Table 10: Scaling Pixel-SAIL to a larger base model. We keep the training data and hyperparameters the same as Pixel-SAIL 3B. ‡ denotes zero-shot evaluation.

Model	Size	RefCOCO+	RefCOCOg	RefCOCO	GRES	RefCOCOg Caption
Pixel-SAIL (SAIL)	3B	75.7	78.7	80.8	67.7	17.6
Pixel-SAIL (SAIL)	7B	77.2	80.2	81.9	69.4	17.9
Sa2VA (combined approach)	8B	76.2	78.7	81.6	46.2‡	17.5

et al., 2025) propose more complex referring tasks, including reasoning-based referring, joint mask and caption generation. In particular, LISA (Lai et al., 2024) involves complex expression, while GLaMM (Rasheed et al., 2024) annotates a new dataset and proposes region-level caption and segmentation tasks. However, all these works contain complex designs: extra vision encoders, segmentation encoders, mask decoders, and prompt encoders. Some MLLMs (Zhang et al., 2024a; Zang et al., 2025; You et al., 2024; Bai et al., 2023; Wang et al., 2024b; Chen et al., 2024f;a; Wei et al., 2024c) focus on referring-expression comprehension, which can be integrated into textual output more easily. However, these methods output coarse bounding boxes that are insufficient for downstream tasks such as editing and matting. Additionally, GiT (Wang et al., 2024a) attempts to unify most vision tasks using a transformer structure; however, it lacks the powerful reasoning and conversational abilities of large language models. A contemporaneous work, UFO (Tang et al., 2025), also attempts to enable MLLMs to acquire pixel-grounded understanding capabilities more simply. However, even with careful design, there remains a considerable gap relative to state-of-the-art performance. Several works (Zhou et al., 2026; Wang et al., 2025) design new mask tokenizers for pixel-grounded MLLMs. However, these approaches still require **additional steps and components** for mask tokenization and detokenization. Our method, Pixel-SAIL, uses a single transformer to learn joint visual and language features. With the proposed data engine and architectural improvements, Pixel-SAIL achieves state-of-the-art results with a much simpler architecture.

Visual Prompt Understanding. Visual prompt understanding is important for interaction between VLMs and users. Recent works (Cai et al., 2024a; Yuan et al., 2024b; Rasheed et al., 2024; Ma et al., 2024; Lin et al., 2024) introduce visual prompt datasets for region captioning and prompt-aware VQA. ViP-LLaVA (Cai et al., 2024a) overlays visual prompts directly on the image and fine-tunes LLaVA on a visual prompt dataset, while Osprey (Yuan et al., 2024b) encodes pixel-level mask regions within language instructions. Pixel-SAIL supports visual prompt understanding through the proposed prompt-token injection design.

H Additional Scaling and Efficiency Analysis

Scaling to a larger base MLLM. To isolate model-size scaling within one base-model family, we construct Pixel-SAIL 7B using SAIL-7B and keep the training data and hyperparameters identical to those of Pixel-SAIL 3B. This same-family experiment is distinct from the EVEv2-7B variant evaluated in the main paper, which uses a different base architecture and an approximately 800×800 input resolution. The results are summarized in Tab. 10. Pixel-SAIL 7B consistently improves over Pixel-SAIL 3B on all pixel-understanding and visual-prompt-understanding benchmarks and also outperforms Sa2VA 8B on several benchmarks. The gains are moderate rather than linear with parameter count: the four segmentation scores improve by 1.1–1.7 points, while RefCOCOg Caption improves by 0.3 METEOR. Thus, the result supports positive same-family scaling without implying a scaling law.

Training cost under comparable visual-token budgets. Pixel-SAIL requires more training time than Sa2VA mainly because current encoder-free MLLMs use a lower visual-token compression rate. To control the number of visual tokens, we downsample the input image by $2\times$ and retrain Pixel-SAIL 3B. As shown in Tab. 11, Pixel-SAIL (DS) reduces the training cost from 768 to 216 A100 GPU hours and remains competitive with or better than Sa2VA on the same benchmarks. This indicates that the higher training cost mainly results from

Table 11: Training cost under different visual-token budgets. DS denotes using $2\times$ downsampled images as input.

Model	Size	Training (A100 hours)	RefCOCO+	RefCOCOg	RefCOCO	RefCOCOg Caption
Sa2VA	4B	384	74.3	76.7	80.4	17.3
Pixel-SAIL	3B	768	75.7	78.7	80.8	17.6
Pixel-SAIL (DS)	3B	216	74.6	77.9	80.1	17.2

Table 12: Training-time overhead of dense feature distillation for Pixel-SAIL 3B.

Setting	Training (A100 hours)	Added cost
w/o distillation	746	–
w/ distillation	768	22 (2.9%)

Table 13: Inference speed comparison on the RefCOCO validation benchmark. DS denotes using $2\times$ downsampled images as input.

Model	FPS	RefCOCO+	RefCOCOg	RefCOCO	RefCOCOg Caption
Sa2VA	1.05	74.3	76.7	80.4	17.3
Pixel-SAIL	0.82	75.7	78.7	80.8	17.6
Pixel-SAIL (DS)	1.96	74.6	77.9	80.1	17.2

the number of visual tokens used by the base MLLM rather than from the proposed Pixel-SAIL components.

Isolating distillation overhead. Under the same Pixel-SAIL 3B configuration, training without dense feature distillation requires 746 A100 GPU hours, whereas enabling distillation requires 768 A100 GPU hours (Tab. 12). The additional cost is 22 A100 GPU hours, or 2.9%. Mask2Former and SAM2 are frozen teachers used only to produce training targets; they introduce no inference-time computation.

Inference speed. We further compare inference efficiency on the RefCOCO validation benchmark in Tab. 13. The full-resolution Pixel-SAIL 3B reaches 0.82 FPS, compared with 1.05 FPS for Sa2VA, because it processes more visual tokens. With $2\times$ downsampled inputs, Pixel-SAIL reaches 1.96 FPS, or $1.87\times$ the throughput of Sa2VA, while maintaining competitive performance. These results show that removing external experts does not by itself guarantee greater speed: throughput also depends strongly on the visual-token budget.

I Robustness and Alternative Representations of Visual Prompts

Robustness to noisy visual prompts. We evaluate the region captioning capability of Pixel-SAIL under noisy mask prompts on RefCOCOg. Specifically, we obtain an object’s bounding box from its mask, randomly scale the box by a factor sampled from a predefined range, and use SAM to generate a noisy mask from the scaled box. All noisy masks are generated offline so that Pixel-SAIL and Osprey are evaluated on exactly the same prompt inputs. Tab. 14 shows that Pixel-SAIL degrades gracefully as the noise increases and remains more robust than Osprey across all noise levels. These results suggest that Pixel-SAIL can tolerate moderate inaccuracies in interactive visual prompts.

Textualizing visual prompts with bounding boxes. We also test a simpler alternative that represents visual prompts as textual bounding-box coordinates instead of injecting visual prompt embeddings. The data and training hyperparameters are identical to those used for Pixel-SAIL. As shown in Tab. 15, this strategy performs substantially worse, especially on the V-T RES benchmark. This result supports our design choice of injecting visual prompt embeddings directly into the visual token stream instead of representing the prompt only as text.

Table 14: Robustness to noisy visual prompts on RefCOCOg caption. R- $x\%$ means that the bounding box is randomly scaled within $[1 - x\%, 1 + x\%]$ before converting it back to a noisy mask prompt with SAM.

Model	Size	w/o noise	R-10%	R-20%	R-25%	R-30%
Pixel-SAIL	3B	17.6	17.5	17.2	16.9	16.4
Osprey	7B	16.6	16.2	15.6	15.0	14.1

Table 15: Comparison between textualized visual prompts and visual prompt embedding injection. All settings, data, and training hyperparameters are kept the same.

Strategy	RefCOCOg Caption	V-T RES
Text bbox	13.5	18.9
VP Injection	17.6	33.4

Table 16: Generalizing Pixel-SAIL to a base MLLM with a vision encoder.

Model	Base MLLM	RefCOCO+	RefCOCOg	RefCOCO	RefCOCOg Caption
Sa2VA	InternVL2-4B	74.3	76.7	80.4	17.3
Pixel-SAIL	InternVL2-4B	76.2	79.0	80.7	17.8
Pixel-SAIL	SAIL-3B	75.7	78.7	80.8	17.6

Table 17: Additional ablation of expert distillation. Results on RefCOCO+/g are validation cIoU, and V-T RES is reported on PerBench.

Distill	V-T RES	RefCOCO	RefCOCO+	RefCOCOg
w/	33.4 (+3.2)	80.8 (+0.6)	75.7 (+0.6)	78.7 (+0.4)
w/o	30.2	80.2	75.1	78.3

J Generalization Across Base MLLMs and Further Distillation Analysis

Using a base MLLM with a vision encoder. Although our primary goal is to show that an encoder-free single transformer can simultaneously handle VQA, pixel understanding, and visual prompt understanding, the proposed Pixel-SAIL mechanism is not limited to encoder-free backbones. We further build Pixel-SAIL on top of InternVL2-4B. Similar to the encoder-free version, visual prompt embedding injection is applied after the patch embedding module, and segmentation masks are obtained by the dot product between the vision tokens output by the MLLM and the segmentation tokens. To accelerate training, we only use the thumbnail input, i.e., the maximum number of image splits is set to 1. The results in Tab. 16 show that Pixel-SAIL built on InternVL2-4B also achieves strong performance, demonstrating the generality of our design.

Additional analysis of expert distillation. We further analyze the effect of expert distillation on out-of-domain generalization using PerBench V-T RES, whose examples do not overlap with the training data. As shown in Tab. 17, distillation improves performance on all benchmarks and yields a larger gain on V-T RES (+3.2), indicating that the distillation strategy is particularly helpful for generalization beyond the training distribution. We note that some results in the main paper use the average cIoU over val/testA/testB, whereas Tab. 17 reports validation cIoU only on RefCOCO+/g; this difference explains the apparent inconsistency between tables. Overall, the teacher models guide early-stage learning and improve generalization rather than defining a strict upper bound for Pixel-SAIL.

K Discussion and Limitations

What enables pixel grounding without a vision encoder. Encoder-free pixel grounding presents two complementary challenges: preserving spatial detail after patch-level tok-

enization and aligning language-side reasoning tokens with dense visual features in a shared representation space. The lightweight upsampling path addresses the first challenge, while visual prompt injection addresses the second by assigning the same prompt identity to an image region and its textual reference. In the controlled component ablation of Tab. 5, all variants use the same data mixture, base model, and training schedule. Adding visual prompt injection raises RefCOCOg Caption from 1.4 to 16.1 METEOR and also improves referring segmentation, indicating that its effect is not explained by additional data. The no-distillation result in Tab. 17 further shows that the shared-transformer pathway learns pixel grounding from supervised data alone; distillation improves accuracy but is not required for the capability to emerge.

Scope of the “single transformer” claim. The claim refers to the shared inference-time backbone: image patches, text, visual-prompt identities, and [SEG] tokens are processed by one transformer, whose hidden states support both language generation and mask prediction. Pixel-SAIL still contains standard input embeddings, two lightweight learnable upsampling blocks, a linear projection of [SEG] states, and a dot product for mask logits. It does not use a separate pretrained CLIP/SAM vision encoder, Q-Former or resampler, visual-prompt encoder, segmentation expert, or pretrained mask decoder at inference time. Thus, “single transformer” describes backbone consolidation rather than the absence of all input and output layers. The individual operations are standard; the contribution is their reformulation into a shared encoder-free pixel-grounded system.

Attribution across systems and distillation scope. The controlled ablations isolate the effects of Pixel-SAIL’s modules, but comparisons with independently trained systems cannot fully separate architecture, pretraining, resolution, and data-mixture effects. The same-data Sa2VA comparison in Tab. 9 reduces the data confound but does not remove the remaining architectural differences. For distillation, we use a fixed MSE weight of 0.5 and do not claim that it is optimal; a broad weight sweep would require multiple full-scale training runs. Because the frozen teachers shape the learned representation, their biases may also be transferred. A useful future direction is staged training that uses expert guidance early and removes the objective later, together with broader teacher and loss-weight studies.

Data, domains, and reasoning coverage. Pixel-SAIL is trained on 1.7M examples, so the present results do not establish a data-scaling law or an optimal mixture. Changes in general VQA performance can reflect both the selected base MLLM and a training shift toward region-level and dense-prediction objectives, motivating more balanced mixtures. Our experiments focus on natural images; we have not evaluated medical imagery, grayscale or low-contrast inputs, or other specialized domains, where domain-specific data or adaptation may be necessary. PerBench includes spatial and dependency relations, but it does not replace dedicated spatial-reasoning benchmarks. Finally, compression below 0.5B parameters remains unstudied.