

Grounding Everything in Tokens for Multimodal Large Language Models

Xiangxuan Ren¹ Zhongdao Wang² Liping Hou² Pin Tang¹ Guoqing Wang¹ Chao Ma^{1*}

¹ MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

² Huawei Noah’s Ark Lab

{bunny.renxiangxuan, pin.tang, guoqing.wang, chaoma}@sjtu.edu.cn

wcdl7@mails.tsinghua.edu.cn, houliping1@huawei.com

Project page: <https://getokpage.github.io>

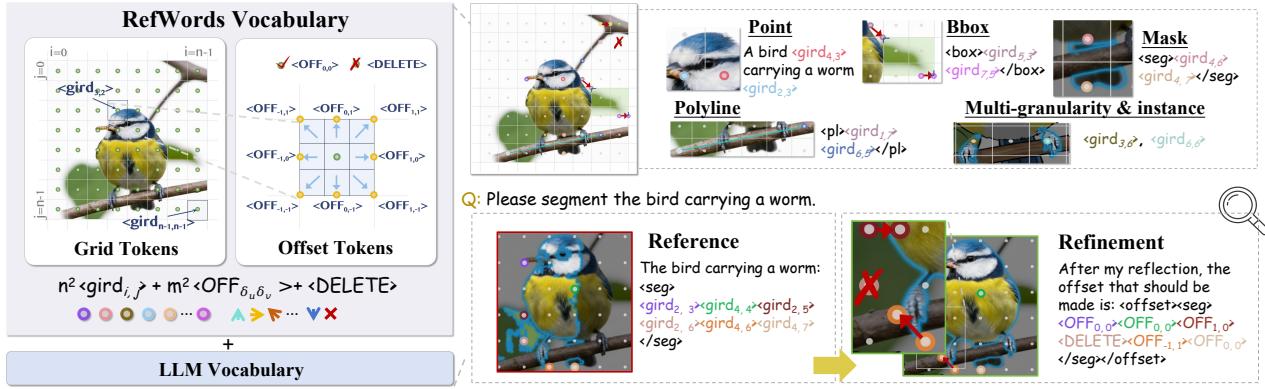


Figure 1. Overview of GETok. GETok equips MLLMs with pre-defined, learnable discrete tokens tied to uniformly distributed anchor points on the image plane, enabling unified grounding from diverse inputs such as text, points, bounding boxes, and segmentation masks. A localization refinement scheme further supports coarse-to-fine correction and iterative recovery from initial grounding errors.

Abstract

Multimodal large language models (MLLMs) have made significant advancements in vision understanding and reasoning. However, the autoregressive Transformer architecture used by MLLMs requires tokenization on input images, which limits their ability to accurately ground objects within the 2D image space. This raises an important question: how can sequential language tokens be improved to better ground objects in 2D spatial space for MLLMs? To address this, we present a spatial representation method for grounding objects, namely GETok, that integrates a specialized vocabulary of learnable tokens into MLLMs. GETok first uses grid tokens to partition the image plane into structured spatial anchors, and then exploits offset tokens to enable precise and iterative refinement of localization predictions. By embedding spatial relationships directly into tokens, GETok significantly advances MLLMs in native 2D space reasoning without modifying the autoregressive architecture. Extensive experiments demonstrate that GETok achieves superior performance over the state-of-the-art methods across various referring tasks in both supervised fine-tuning and reinforcement learning settings.

1. Introduction

Recent years have witnessed significant advancements in multimodal large language models (MLLMs) [2, 26, 27, 29, 31, 46, 61, 93] concerning vision understanding, reasoning, and interaction. The impressive successes of autoregressive Transformers in language modeling [4, 12, 14, 21, 47] have established them as the foundational architecture for MLLMs. Autoregressive Transformers typically require tokenization of input images, similar to that used for text. However, this image tokenization often leads to a substantial loss of spatial information [18, 42, 60]. As a result, current MLLMs encounter a notable limitation in their ability to reason accurately about precise spatial localization.

Numerous efforts have been made to address this issue. One straightforward solution is to use text to describe object locations [8, 11, 78], as demonstrated by Qwen-VL [3]. However, this text-based approach struggles to preserve spatial topology [57], resulting in large syntactic overhead and tokenization bias, as illustrated in Fig. 2(a). More recently, a number of methods directly project image patches into visual tokens via linear projection [20, 42, 60, 67], as shown in Fig. 2(b). However, the patch size is fixed by the image encoder, and the linear projection is tied to this

patch partition, entangling texture with geometry and often confusing texture-similar objects at different spatial locations. Alternatively, bin-based methods [9, 64, 70] use one-dimensional bins to describe bounding boxes for grounding objects (Fig. 2(c)). While promising, slight changes in one-dimensional indices do not accurately reflect smooth changes in 2D topology, so bin tokens benefit less from recent reinforcement learning schemes such as GRPO[55], where small action changes can unexpectedly cause large reward fluctuations.

In this work, we identify the core challenge in advancing MLLMs toward precise 2D reasoning as establishing a reliable mapping between discrete sequential tokens and continuous 2D space. As such, we propose *GETok* to Ground Every object in *Tokens* via a set of learnable *spatial vocabulary* terms. As shown in Fig. 1, our GETok comprises two core types of tokens: i) **Grid tokens** first establish a structured spatial topology by discretizing the image plane into an $n \times n$ uniform grid. Each grid cell is associated with a learnable token added to the model’s vocabulary, yielding a set of *spatial anchors*, each of which is responsible for referring to objects within its local region. While this 2D lattice provides native spatial awareness, it introduces a vocabulary bottleneck, as the number of tokens grows quadratically with increasing resolution. ii) **Offset tokens** overcome this issue and refine spatial reasoning via a set of discrete displacement vectors together with a <DELETE> token. Building upon the structural regularity of grid tokens, offset tokens enable high-precision spatial refinement at a minimal vocabulary cost. For example, a 32^2 anchor grid can be upgraded to 64^2 effective precision by adding ten offset tokens, instead of introducing $64^2 - 32^2 = 3072$ new grid tokens. Furthermore, the use of offset tokens yields an emergent benefit of progressive localization refinement. Because the <DELETE> token can recursively reject errors during localization, transforming the process from a one-shot prediction into an iterative reasoning approach.

GETok presents three significant advantages over state-of-the-art methods: First, GETok provides a unified representation for various tasks, ranging from points to masks, all within a standard autoregressive framework. This integration eliminates the need for task-specific modules, simplifying the architecture while ensuring generalizability and precision. Second, the integrated offset mechanism facilitates self-correction through iterative refinement. This feature allows the model to adjust its spatial predictions, addressing a common limitation in existing methods where initial errors often go uncorrected. Third, the geometric foundation that correlates token shifts with smooth spatial changes creates a low-entropy action space. This results in stable reward landscapes and more efficient exploration, significantly enhancing policy optimization compared to unstructured representations. Building on these advantages, we introduce a novel

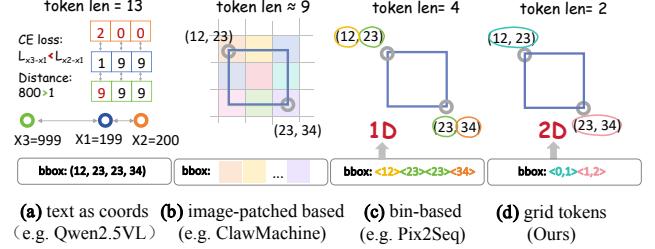


Figure 2. Comparison of token-based representations for grounding objects in MLLMs. Note that 2D grid tokens preserves spatial topology with shorter sequences than coordinate-, patch-, or 1D bin-based formulations.

self-improving reinforcement learning framework that explicitly models spatial dynamics and employs GRPO-style preference optimization to refine locations through iterative self-correction. Comprehensive experiments across various referring benchmarks show that our GETok achieves superior performance under both the supervised fine-tuning and reinforcement learning settings.

In summary, the main contributions of this work are:

- We propose a lexical spatial representation that embeds a vocabulary of learnable tokens to empower MLLMs to accurately reason over native 2D spatial space without modifying autoregressive frameworks.
- We develop a localization refinement scheme on top of offset tokens that provides coarse-to-fine correction and iterative recovery from initial grounding errors.
- We propose a geometry-aware policy optimization framework to perform self-improving reinforcement learning to facilitate spatial reasoning.

2. Related Work

MLLMs for Vision Perception. Enabling MLLMs to understand, manipulate, and output image regions is a central goal for vision-language intelligence [32–35]. Current methods primarily follow two training paradigms: Under the SFT paradigm, methods have explored diverse referring representations including points [13], bounding boxes [8], and masks [25]. Early approaches demonstrated that text-based coordinate representations [8, 15, 18, 19, 41, 66, 78, 85, 93] can enable basic referring dialogue, but these methods suffered from ambiguous region-text alignment. This limitation prompted methods like GPT4RoI [88] to introduce specialized region modules [7, 41, 63] for improved alignment at the cost of architectural complexity. For complex shape representation, mask-based referring has emerged as a promising direction. Ferret [78] and Osprey [83] designed specialized pooling mechanisms for irregular mask inputs, while LISA [25] pioneered an embedding-as-mask paradigm using dedicated segmentation tokens that trigger external decoders, a design that has inspired a series of subsequent works [51, 62, 71]. Nev-

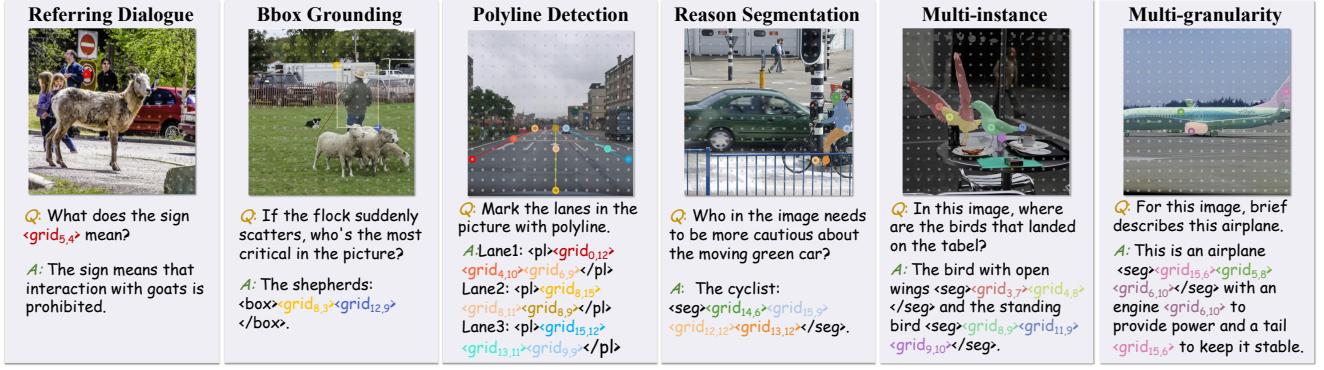


Figure 3. GETok supports both input and output references with multiple format conversions, including boxes, polylines, and masks. It is seamlessly compatible with multi-instance and multi-granularity capabilities. Best viewed in color.

ertheless, SFT-based approaches remain limited when handling complex spatial reasoning tasks. The emergence of RL-based methods has demonstrated significant potential for overcoming these limitations [1, 45, 72, 79, 84], such as VLM-R1 [56] and Visual-RFT [38]. Approaches like Seg-Zero [36] and VisionReasoner [37] further demonstrate the effectiveness of decoupled architectures where reasoning chains generate prompts for external segmenters, establishing RL as a promising direction for mask output. In this work, we revisit both paradigms through the lens of a lexical spatial vocabulary and a self-improving RL framework to endow MLLMs with native 2D spatial reasoning.

Token-based Referring Representation. A complementary research direction focuses on unifying visual and linguistic representations through discrete tokenization. One line of work employs special tokens to aggregate positional information [25, 71, 78, 86, 87, 89], which encode image regions through dedicated tokens but typically require additional architectural modifications to process this spatial information. Another line leverages image patches as visual tokens [6, 20, 42, 58–60], though this remains tightly coupled with the visual encoder [49], limiting transferability. Bin-based methods [9, 70, 76] discretize coordinates into tokens selected from a fixed vocabulary. While these methods introduce learnable spatial tokens, they remain limited to 1D indexing, which does not adequately capture explicit 2D structural relationships. The closest related work, Kosmos-2 [48], also envisions 2D spatial tokens but is restricted to basic bounding box grounding, lacking the ability to support more complex spatial tasks. In contrast, our method aims to establish a general-purpose spatial lexicon in which grid tokens create an explicit 2D lattice. This lattice enables referencing across various formats, including points, boxes, polylines, and masks (See Fig. 3). Moreover, offset tokens facilitate coarse-to-fine corrections with minimal vocabulary expansion. This effectively improves localization precision while preserving architectural simplicity.

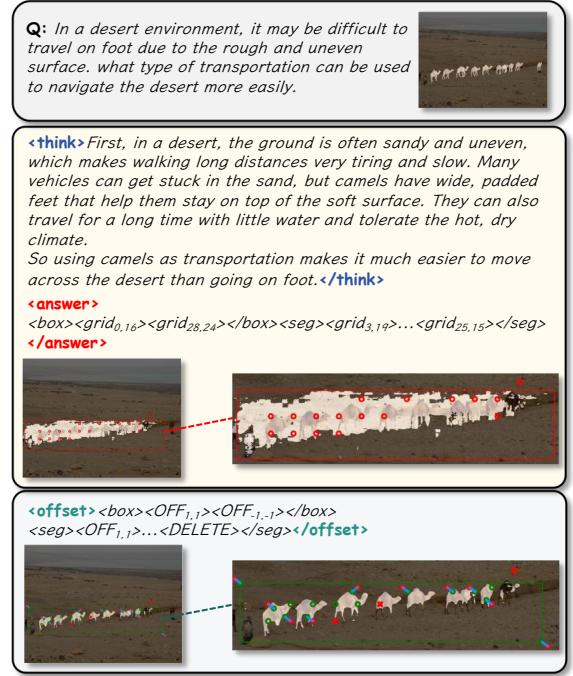


Figure 4. Example of the propose-and-refine mechanism in GETok. Grid tokens provide coarse localization, while offset tokens enable precise adjustment.

3. GETok: Grounding Everything in Tokens

3.1. Overview

To endow MLLMs with the ability to interpret and generate spatial references in a native token-based manner, we propose to augment their vocabulary with a set of learnable spatial tokens. Specifically, we first use grid tokens to discretize the image into $n \times n$ anchors: $\mathcal{T}_{\text{grid}} = \{\langle \text{grid}_{i,j} \rangle \mid i, j \in \{0, \dots, n-1\}\}$. Then we use offset tokens to refine local positions: $\mathcal{T}_{\text{offset}} = \{\langle \text{OFF}_{\delta_u, \delta_v} \rangle\} \cup \{\langle \text{DELETE} \rangle\}$, where $\delta_u, \delta_v \in \{-1, 0, 1\}$. The complete vocabulary

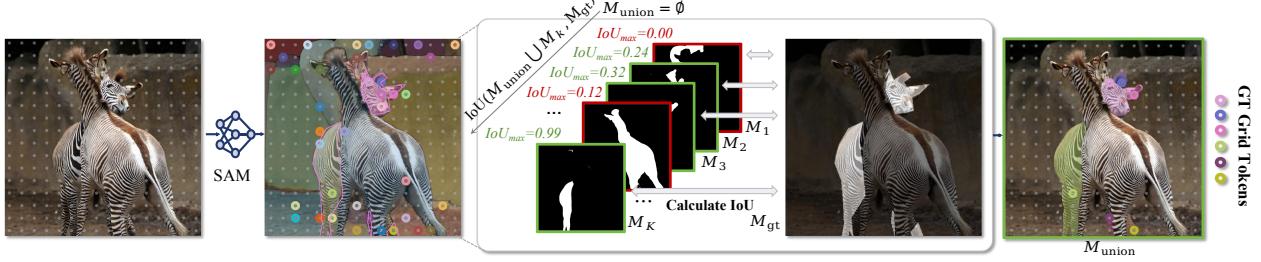


Figure 5. We use a greedy algorithm to generate the ground-truth grid tokens referring to the ground-truth mask. This conversion automatically transforms continuous masks into discrete tokens, enabling scalable data expansion.

$\mathcal{V} = \mathcal{V}_{LLM} \cup \mathcal{T}_{\text{grid}} \cup \mathcal{T}_{\text{offset}}$ facilitates spatial reasoning as precise spatial pronouns. These two types of tokens collectively reason about localization through a *propose-and-refine* chain. Fig. 4 illustrates an example where a complex mask cannot be adequately represented by bounding boxes. When using grid tokens to initiate coarse proposals and offset tokens to iteratively refine them, we succeed in constructing precise representations for intricate masks. The following will detail how to realize GETok under the supervised fine-tuning and reinforcement learning settings.

3.2. Supervised Fine-Tuning

Since GETok does not require modifying the architecture of base MLLMs, the key to applying supervised fine-tuning (SFT) lies in constructing training data. As such, SFT can properly utilize the GETok vocabulary through automated annotation conversion and sequence simulation. While grid tokens provide a unified representation for points, bounding boxes, and polylines through straightforward mappings as shown in Fig. 3, two core challenges remain: (1) how to construct discrete point representations for dense masks, and (2) how to create effective training data for offset tokens to perform localization refinement.

3.2.1. Greedy Mask-to-Token Conversion

When converting dense masks into discrete points, current token-based approaches generally use single points, bounding boxes, combinations of bounding boxes with one or two fixed points, or randomly sampled points within a mask [13, 36, 37]. However, we have noticed that these formats often exhibit significant redundancy and ambiguity, especially when dealing with multiply-connected mask regions. To address this issue, we have developed a greedy algorithm to facilitate the transformation from masks to grid tokens. Importantly, this conversion process is training-free and does not incur any additional computational costs or require changes to the network design.

As illustrated in Fig. 5, we initially input the image along with n^2 grid points as prompts into the SAM¹. This process generates K masks, denoted as $\mathcal{M} = \{M_1, \dots, M_K\}$.

Each mask corresponds uniquely to an input grid, defined by a mapping $\theta : \{i\}_{i=1}^{n^2} \rightarrow \{k\}_{k=1}^K$. Typically, $K < n^2$ because of mask deduplication during post-processing. Given a ground-truth mask M_{gt} , our goal is to identify a minimal set of grid points such that the union of their corresponding masks approximates M_{gt} . Formally, this objective is written as:

$$\begin{aligned} \pi^* = \arg \min_{\pi \in \{0,1\}^{n^2}} \|\pi\|_0 \\ \text{s.t. } \text{IoU}\left(M_{\text{gt}}, \bigcup_{k: \pi_k=1} M_{\theta(k)}\right) \geq \tau. \end{aligned} \quad (1)$$

Here, π is a binary selection vector over the grid tokens, and τ is a quality threshold that ensures a minimum Intersection-over-Union (IoU). Eq. 1 denotes a constrained multi-objective optimization problem. To solve it efficiently, we develop a simple yet effective greedy algorithm starting with $\pi = \mathbf{0}$, $M_{\text{union}} = \mathbf{0}$, and $\text{IoU}_{\text{max}} = 0$. First, we compute the IoUs between M_{gt} and all K mask proposals, sorting them in descending order. Then, we iterate through all masks. For the k -th iteration, we calculate $\text{IoU}^* = \text{IoU}(M_{\text{union}} \cup M_k, M_{\text{gt}})$. If $\text{IoU}^* > \text{IoU}_{\text{max}}$, we update $\pi_k \leftarrow 1$, $M_{\text{union}} \leftarrow M_{\text{union}} \cup M_k$, and set $\text{IoU}_{\text{max}} \leftarrow \text{IoU}^*$. At the end of the iterative process, we obtain an approximately optimal π^* that identifies the grid points corresponding to the ground truth mask. Finally, a mask can be represented as an unordered sequence of matched grid tokens, for example, $\langle \text{seg} \rangle \langle \text{grid}_{i_1, j_1} \rangle \dots \langle \text{grid}_{i_n, j_n} \rangle \langle \backslash \text{seg} \rangle$.

3.2.2. Offset-Aware Dataset Construction

To generate high-quality training data for offset tokens, we develop a systematic approach that categorizes grid points based on their spatial relationship to mask boundaries. Using morphological operations scaled to the offset step size, we define four distinct regions around each mask boundary: i) **Inside**: Stable interior points mapped to zero offset ($\langle \text{OFF}_{0,0} \rangle$). ii) **Ring**: Boundary-proximal exterior points requiring non-zero offsets; iii) **Far**: Distant negatives mapped to deletion ($\langle \text{DELETE} \rangle$). iv) **Hard-Delete**: Challenging edge cases also mapped to $\langle \text{DELETE} \rangle$. Each

¹We utilize its *segment anything* mode [22].

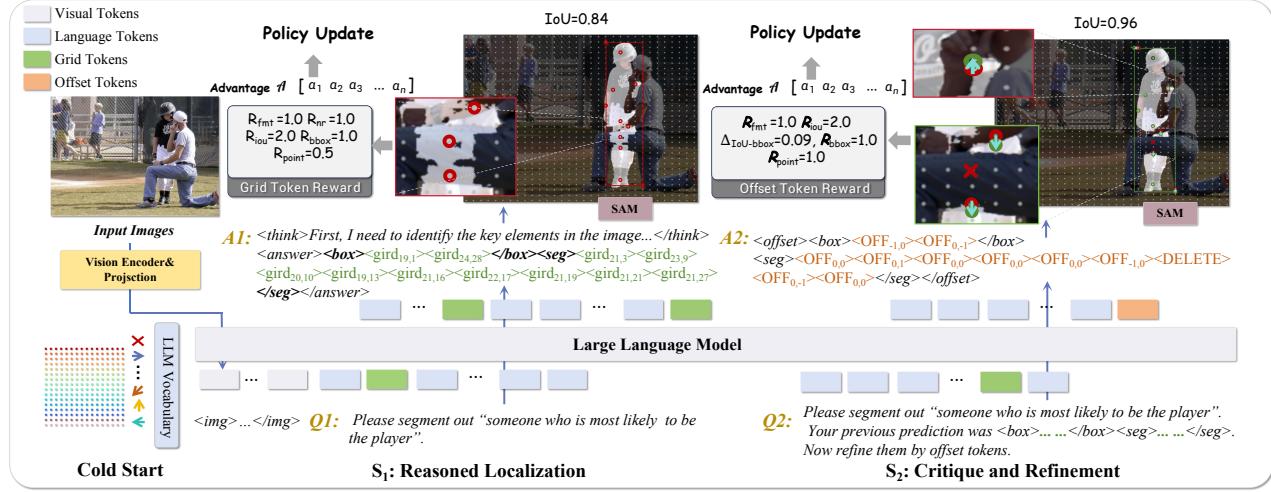


Figure 6. **Overview of the Self-Improving RL Framework.** Our framework models 2D spatial localization as a two-step generative task. First, grid tokens are generated to propose anchor regions in the image. Second, offset tokens refine the region proposals to precise points.

grid point is assigned to exactly one region through an ordered decision rule that prioritizes educationally valuable cases. Training pairs are sampled with bias toward INSIDE and RING regions where offset corrections provide the most learning value.

This procedure yields a variable number K of grid–offset token pairs per image for supervised training. Empirically, this simulated supervision outperforms real-generated alternatives by focusing on boundary-proximal scenarios, creating a curated set of high-value training cases that foster effective refinement strategies. Detailed algorithms are provided in supplementary.

3.3. Self-Improving Reinforcement Learning

The structured nature of GETok offers an ideal framework for reinforcement learning due to its 2D lattice organization, which creates a geometrically grounded action space rich in spatial semantics. We introduce a novel self-improving reinforcement learning framework that utilizes the grid-offset hierarchy of GETok to enable iterative self-correction. Unlike traditional RL methods that optimize for one-time predictions, our method incorporates a multi-turn refinement process allowing the model to critique and adjust its spatial predictions using offset tokens and a `<DELETE>` command.

As illustrated in Figure 6, our pipeline starts with a cold-start model that is pre-trained through supervised fine-tuning on the GETok vocabulary. This process provides the policy π_θ with a foundation for generating spatially grounded responses in GETok. The training then follows a two-stage procedure utilizing GRPO [55]. The first stage focuses on generating grid tokens, rewarding spatial accuracy and structural validity, while the second stage introduces offset tokens in multi-turn dialogues, incentivizing precision improvement through iterative refinement. This

self-correcting mechanism significantly enhances localization precision while ensuring conversational coherence.

We design specific reward functions for grid and offset tokens. Grid token rewards promote accurate regional grounding, while offset token rewards aim to adjust misaligned grid anchors through local shifts. This dual-phase optimization distinctly separates token placement from token movement, achieving geometry-aware self-correction.

3.3.1. Reward for Grid Token Generation

The key distinction from existing methods is that we train the model to generate a variable-size set of semantic-critical points to handle complex scenes, rather than reducing each mask to only one or two points. Specifically, the reward for grid-token generation is designed as follows:

Format Reward. This reward encourages structured output with reasoning in `<think>` tags and spatial predictions in `<answer>` tags containing `<box>` and `<seg>` tokens.

Non-repeat Reward. This reward penalizes sentence-level repetition when multiple identical sentences appear.

Mask Reward. This reward favors large IoU scores between the generated masks and GT masks using a piecewise function. For automated quality assessment, we employ SAM [22] and convert predicted boxes and points into spatial prompts to generate masks.

Box Reward. This reward favors large IoU scores and small L1 corner distance between the predicted and ground-truth boxes.

Semantic-Critical Points Reward. This reward evaluates segmentation quality by combining hit ratio (points inside ground-truth masks) and spatial distribution. We balance point density using an exponential saturation term $(1 - e^{-m_p/5})$ to prevent sparse predictions and a linear penalty $(0.02m_p)$ to avoid excessive points. Additional details are provided in the supplementary material.

Table 1. **Referring Expression Segmentation** results on the ReasonSeg and RefCOCO (+/g) datasets.

Methods	Training Mask Dec.	ReasonSeg		refCOCO			refCOCO+			refCOCOg		Avg.	
		Val.	Test	Val.	T-A	T-B	Val.	T-A	T-B	Val.	Test		
<i>Supervised Fine-Tuning Models</i>													
LAVT [77]	✓	-	-	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1	-	
ReLA [28]	✓	-	-	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0	-	
CRIS [68]	✓	-	-	70.5	73.2	66.1	65.3	68.1	53.7	59.9	60.4	-	
PixelLM [54]	✓	-	-	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5	-	
LISA [25]	✓	44.4	36.8	76.0	78.8	72.9	65.0	70.2	58.1	69.5	70.5	64.2	
Qwen2.5-VL-7B [3]	✗	55.4	51.5	72.5	76.4	70.0	64.3	70.5	58.4	68.1	69.9	65.7	
GETok-SFT-grid	✗	58.1	54.4	74.3	77.9	72.3	65.6	71.9	58.8	68.0	70.9	67.2	
GETok-SFT	✗	59.2	55.8	76.1	79.2	73.2	66.4	72.3	59.9	69.4	70.9	68.2	
<i>Reinforcement Learning Models</i>													
Seg-Zero [36]	✗	62.6	57.5	-	80.3	-	-	-	76.2	-	-	72.6	69.8
SAM-R1 [17]	✗	64.0	60.2	-	79.2	-	-	-	74.7	-	-	73.1	70.2
VisionReasoner [37]	✗	66.3	63.6	-	79.3	-	-	-	72.2	-	-	72.2	70.7
GETok-R1-grid	✗	64.2	63.7	-	79.8	-	-	-	74.3	-	-	73.9	71.2
GETok-R1	✗	65.9	64.2	-	80.8	-	-	-	77.4	-	-	75.2	72.7

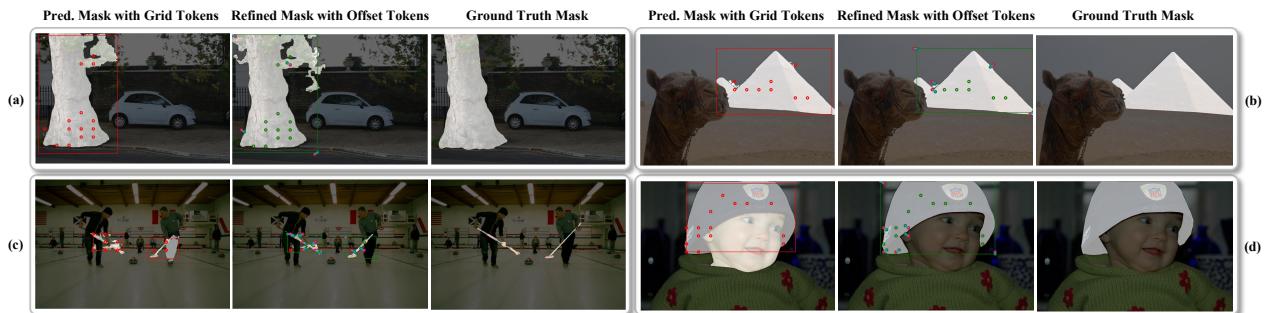


Figure 7. **GETok Qualitative Results on RES [80].** We visualize the two-step localization process: red dots are grid-token proposals, blue lines show the applied offset vectors, and green dots represent the final offset-refined points. Our method demonstrates adaptive corrections, achieving precise localization across diverse scenarios, including small objects and complex shapes.

3.3.2. Reward for Offset Token Refinement

In our experiments, inferior reward formulations do not predict any offset. To prevent this, we design more effective rewards to support a geometric update.

Format Reward. This reward enforces a minimal schema on per-instance `<offset>` tokens containing `<box>` and `<seg>` serializations, as we find that the `<think>` preamble provides negligible benefit for offsets.

Point Refinement Reward. This reward assesses point-level refinement with a ternary score $s_{k,p} \in \{-1, 0, 1\}$ per point: -1 for moves that leave the ground-truth mask, $+1$ for corrections that enter the mask, stay inside it, or perform a valid deletion, and 0 otherwise. A deletion is counted as valid only when `<DELETE>` is predicted and no point in the 3×3 neighborhood of the original position lies inside the ground-truth mask. The instance-level reward is obtained by averaging scores over all points of that instance.

Box Refinement Reward. This reward measures the IoU gain between the initial and refined bounding boxes. For each instance, we assign a positive reward when the refined box increases *bounding boxes* IoU over the initial prediction, and zero otherwise.

Mask IoU Gain Reward. This reward favors large IoU improvements by measuring the relative gain, defined as the IoU improvement from the initial proposal to the refined result normalized by the maximum possible improvement.

4. Experiments

4.1. Experimental Setup

Training Details. We use Qwen2.5-VL-7B [3], a powerful open-source VLM, as the base model for GETok. For GETok-SFT, we use the `ms_swift` framework [90] with LoRA [16] (rank=64), a batch size of 16, and a learning rate of 1×10^{-6} , training on publicly available corpora spanning image-level reasoning, referring grounding, and segmentation. For GETok-RL, we employ the GRPO algorithm [55] via the `easy_r1` framework [91], initializing from a cold-start model trained on referring segmentation data and open-source multimodal instruction data (e.g., LLaVA-CoT-100k [73]). GRPO training in stage 1 uses a 9K dataset containing LISA++ [75] and referring segmentation samples [43, 80], with a batch size of 16 (8 samples per step), learning rate of 1×10^{-6} , and weight decay of 0.01. Refinement training in stage 2 is limited to 200 steps

Table 2. Referring Expression Comprehension results on the RefCOCO (+/g) datasets.

Methods	refCOCO			refCOCO+			refCOCOg		Avg.
	Val.	Test-A	Test-B	Val.	Test-A	Test-B	Val.	Test	
—Supervised Fine-Tuning Models (Acc@0.5)—									
VisionLLM [65]	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2	82.9
UNINEXT-L [74]	91.4	93.7	88.9	83.1	87.9	76.2	86.9	87.5	87.0
Shikra [8]	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2	82.9
Ferret [78]	87.5	91.4	82.5	80.8	87.4	73.1	83.9	84.8	83.9
Groma [41]	89.5	92.1	86.3	83.9	88.9	78.1	86.4	87.0	86.5
ClawMachineX [42]	89.7	92.5	86.9	84.4	88.9	78.0	86.7	87.1	86.8
Qwen2.5-VL-7B [3]	90.0	92.5	85.4	84.2	89.1	76.9	87.2	87.2	86.6
GETok-SFT-grid	90.4	93.8	86.9	86.3	90.8	79.4	87.1	87.5	87.8
GETok-SFT	90.6	93.7	87.2	86.7	90.9	79.9	88.5	88.4	88.2
—Supervised Fine-Tuning Models (Acc@0.8)—									
Qwen2.5-VL-7B [3]	72.6	77.2	67.5	66.6	74.3	61.2	66.3	68.9	69.3
GETok-SFT-grid	73.8	78.9	68.1	67.9	75.1	63.1	68.8	71.1	70.9
GETok-SFT	74.9	79.9	69.6	69.1	77.9	66.3	70.1	72.9	72.6
—Reinforcement Learning Models (Acc@0.5)—									
VisionReasoner [†] [37]	89.6	91.1	-	85.4	89.0	-	88.2	89.0	88.7
GETok-R1-grid	90.2	92.9	-	86.7	89.9	-	89.2	88.7	89.6
GETok-R1	90.9	93.6	-	87.1	90.8	-	89.9	89.2	90.3
—Reinforcement Learning Models (Acc@0.8)—									
VisionReasoner [†] [37]	72.4	76.8	-	67.3	74.9	-	68.5	71.2	71.6
GETok-R1-grid	74.1	78.3	-	68.1	75.5	-	71.2	72.9	73.4
RefEdit-R1	75.1	79.2	-	68.9	76.9	-	72.9	73.2	74.4

to prevent overfitting, given the concise nature of offset tokens. All experiments are conducted on 8x NVIDIA A800 GPUs using the DeepSpeed engine [52], with a grid size of 32 and an offset size of 64. Detailed dataset composition is in the supplementary.

Benchmark Settings. GETok addresses a broad spectrum of visual referring tasks. We conduct quantitative evaluations on six benchmarks: (i) Referring Expression Comprehension (REC), (ii) Referring Expression Segmentation (RES), (iii) Reasoning Segmentation, (iv) Referring Captioning, (v) Generalized Referring Expression Segmentation (gRES), and (vi) Lane Polyline Detection. We also build (vii) a driving case study that mixes polylines (lanes), polygons (drivable area), and boxes (dynamic objects), demonstrating unified supervision in complex scenes.

For GETok-SFT, we perform exhaustive validation across all seven settings (i)–(vii), establishing strong and consistent SFT baselines under a shared training setting and decoding budget. For GETok-RL, we focus on (i)–(iii), which reflect mainstream benchmarks for R1 paradigm referring models. Due to space limitations, we put the complete results and ablation studies in the supplementary.

4.2. Overall Performance

Referring Expression Segmentation. As shown in Tab. 1, GETok-SFT demonstrates competitive performance compared to specialized methods while maintaining architectural simplicity. When trained with our reinforcement learning framework, GETok-RL achieves state-of-the-art performance, fully realizing the potential of our token design with a significant gain of +4.5% over supervised fine-tuning. This highlights the substantial capability of our regularized

2D token representation in RL paradigms, where the structured action space facilitates stable policy optimization and efficient exploration.

The offset mechanism proves essential in both training paradigms, providing consistent gains in resolution enhancement of +1.0% in SFT and +1.5% in RL over grid-only configurations. This improvement is particularly critical for mask generation tasks, where even minor localization errors can be enlarged during the decoding process, highlighting the importance of precise spatial refinement.

Fig. 7(a) shows that using off-the-shelf SAM allows us to preserve its generalization capability, resulting in high-quality masks with fine-grained edge details. We note that this can sometimes lead to discrepancies when compared to lower-quality ground truth annotations. Figs. 7(b) and (d) demonstrate the adaptability of our refinement mechanism, which applies small corrections to accurate proposals (b) and larger corrections to less precise ones (d). Fig. 7(c) specifically showcases the effectiveness of our *propose-then-refine* approach for small targets, where precise localization is particularly challenging.

Referring Expression Comprehension. As indicated in Tab. 2, GETok-SFT demonstrates solid performance under the conventional accuracy metric (Acc@0.5), with a gain of +1.6% over the Qwen2.5-VL-7B baseline. To better evaluate localization accuracy, we report results using the more demanding Acc@0.8 metric. Under this stricter evaluation, the combination of grid and offset tokens shows significant improvements in spatial reasoning. The visualizations reveal particularly pronounced gains for small objects under both the SFT and RL settings.

Unlike the ReasonSeg dataset [25] in segmentation

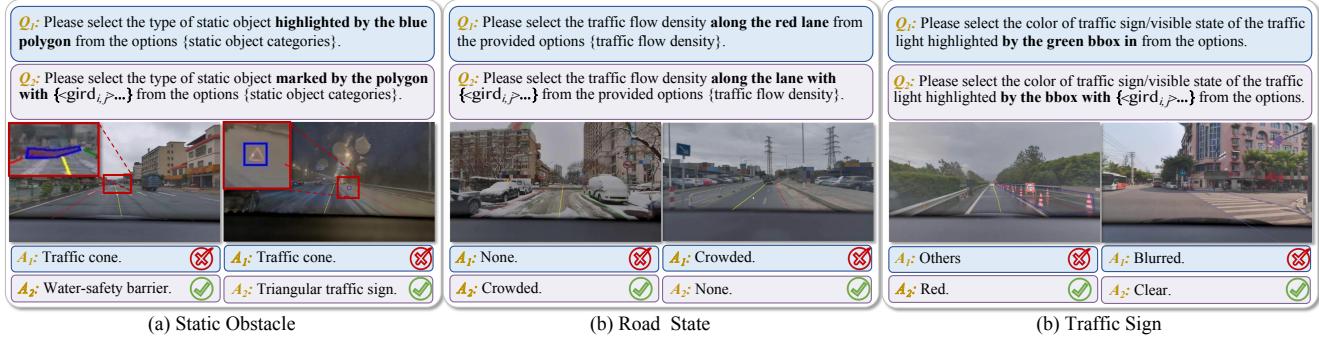


Figure 8. **Qualitative results of the proposed grid tokens in the driving scene.** Challenging examples from three referring categories demonstrate that the proposed GETok offers superior region-referencing ability compared to conventional visual referring prompts.

Table 3. Performance comparison of different grid resolutions on REC (Acc@0.8) and RES (gIoU).

Grid Size	REC	RES	Avg. Token Len. per Mask
16 × 16	68.9	66.2	5.2
64 × 64	71.2	67.1	14.6
32 × 32	70.9	67.2	8.7
w/ offset	72.6	68.2	9.2

benchmarks, which comprises complex reasoning chains, RefCOCO expressions are relatively straightforward, limiting the potential of RL. This contrast highlights that our GETok-RL achieves the greatest advantages when tackling complex reasoning tasks that benefit from iterative refinement and chain-of-thought processing.

4.3. Grid Resolution

The grid size n is a crucial parameter for GETok, governing the trade-off between spatial precision and vocabulary expansion. As shown in Tab. 3, we identify two key observations: First, the 32×32 configuration achieves comparable performance to 64×64 while maintaining significantly lower token length and vocabulary overhead. Second, offset tokens demonstrate remarkable efficiency, outperforming the costly doubling of grid resolution with only 10 additional tokens. This minimal expansion delivers superior performance to the 64×64 configuration.

4.4. Real-World Driving Case Study

We further evaluate grid tokens using a proprietary driving dataset that features diverse urban scenarios, annotated in three ways: lanes (polylines), static obstacles (bounding boxes), and traffic signs (key points). More details can be found in the supplementary materials. For general scene understanding, GETok consistently outperforms traditional visual prompts across all tasks, achieving significant improvements in challenging scenarios: a +12.24% increase in traffic sign color recognition and a +7.95% increase in static obstacle classification, as shown in Tab. 5. Fig. 8 illustrates the success of GETok in complex driving scenarios,

Table 4. Performance comparison of using GETok in the driving scene with the visual referring prompt.

Category	Task	Baseline	GETok
Static obstacle	Classification	81.69	89.64
	Visible State	90.60	93.49
Road	Blockage Status	86.07	87.25
	Surface Condition	95.46	95.68
	Traffic Density	84.31	86.39
Traffic Sign	Color	71.43	83.67
	Visible State	63.27	67.35

Table 5. Comparative results for lane polyline detection.

Methods	Lane Polyline		
	Precision	Recall	F1
Coords-based	0.49	0.47	0.48
GETok	0.52	0.65	0.58

demonstrating its ability to handle diverse reference types through a unified representation without requiring architectural modifications. Additionally, we report lane detection results for GETok, highlighting its particular strength in handling curved lanes. For lane detection, GETok transforms continuous coordinate regression into discrete point selection, resulting in a +3% increase in precision, +18% increase in recall, and a +10% increase in F1-score compared to coordinate-based methods, as shown in Tab. 5.

5. Conclusion

We presented GETok, a novel spatial representation that addresses the fundamental challenge of 2D spatial reasoning in MLLMs. By introducing learnable grid and offset tokens, GETok provides a unified framework for precise spatial localization while maintaining architectural simplicity. The offset mechanism yields the emergent benefit of progressive localization refinement, enabling iterative self-correction. Extensive experiments demonstrate competitive performance across diverse referring tasks under both the supervised and reinforcement learning settings.

Grounding Everything in Tokens for Multimodal Large Language Models

Supplementary Material

We provide supplementary material for further study and analysis related to the main paper, arranged as follows:

- Additional experimental results extending the main findings (Sec. A)
- Real-world driving dataset Curation (Sec. B)
- More implementation details, including training setup, offset-aware dataset construction, and reward design (Sec. C)
- Additional qualitative results and visual analysis (Sec. D)

A. Additional Experiment Results

A.1. More Benchmarks

Referring Captioning evaluates region understanding given referring inputs (e.g., bbox, mask). We evaluate region-based caption generation on refCOCOg [43] and Visual Genome [23]. As shown in Table 6, GETok achieves superior or comparable performance to models using specialized region feature extractors (✓), confirming the efficacy of GETok in enhancing region-aware comprehension. GETok excels at handling scenarios with overlapping objects, where traditional bounding boxes often fail to precisely capture targeted regions.

Table 6. **Region-Level Captioning** results on the refCOCOg and visual genome datasets.

Methods	Region Feat. Extractor	refCOCOg		Visual Genome	
		METEOR	CIDEr	METEOR	CIDEr
GRIT [69]	✓	15.2	71.6	17.1	142.0
SLR [82]	✓	15.9	66.2	-	-
GPT4RoI [88]	✓	-	-	17.4	145.2
GLaMM [51]	✓	16.2	106.0	19.7	180.5
Groma [41]	✓	16.8	107.3	19.0	158.4
Kosmos-2 [48]	✗	14.1	62.3	-	-
Shikra-7B [8]	✗	15.2	72.7	-	-
GETok-SFT	✗	16.9	110.5	19.0	165.9

Generalized RES validates multi-instance resolution through grid token sequences, demonstrating simultaneous referencing capability for multiple objects within a single spatial representation. GETok naturally supports multi-instance expressions. We validate the effectiveness of our method for multi-instance segmentation on the gRefCOCO dataset. As shown in Tab. 7, the results on the gRefCOCO demonstrate the effectiveness of GETok in multi-instance segmentation, achieving competitive performance compared to specialized methods while maintaining architectural simplicity.

Object Pointing evaluates precise coordinate localization, while GETok offers flexible point annotations by marking representative object positions, yielding more adapt-

Table 7. **Generalized Referring Expression Segmentation** results (cIoU) on the RefCOCO (+/g) datasets.

Methods	Training M-Dec.	Validation	Test-A	Test-B	Average
LAVT [77]	✓	58.4	65.9	55.8	60.0
ReLA [28]	✓	63.6	70.0	61.0	64.9
LISA [25]	✓	63.5	68.2	61.8	64.5
GSVA [71]	✓	68.0	71.8	63.8	67.9
GETok-SFT	✗	66.9	72.3	64.1	67.8
GETok-RL	✗	67.4	74.1	65.6	69.0

able localization than rigid bounding boxes. As shown in Tab. 8, GETok achieves competitive performance across all datasets compared to methods trained with substantially more data. The advantage is particularly pronounced in dense object scenarios, where grid tokens reduce coordinate representation from multiple sequential tokens (e.g., `[('(', '124', ',', '143', ')')]`) to a single spatial token (e.g., `<grid12,14>`), eliminating the formatting errors that accumulate with longer text-based coordinate sequences.

Table 8. **Object pointing** results on HumanRef and RefCOCOg datasets (F1-scores).

Methods	HumanRef	refCOCOg val	refCOCOg test
OVIS2.5-9B [40]	62.3	85.0	84.5
Molmo-7B-D [13]	70.0	83.7	83.6
Qwen2.5-VL-7B [3]	65.1	78.9	79.4
GETok-SFT	70.7	84.1	82.9

A.2. More Discussions

How Should Points be Represented? We analyze three representation formats that operate purely through *vocabulary-level modification*: text coordinates, bin tokens, and grid tokens, all of which require no architectural changes. Among them, bin tokens and text coordinates share the same 1D numerical nature, with bin tokens merely quantizing coordinates into discrete indices, and empirical evidence shows that bin-based methods can even underperform text coordinates [8]. The key difference, therefore, lies between these 1D schemes and the native 2D spatial encoding of grid tokens, which addresses three fundamental limitations:

1) *ID-2D Representation Gap*: A single 1D token cannot directly represent a 2D location; instead, multiple tokens must be combined to denote a coordinate. This composition hinders the implicit semantic features of the 2D space from being effectively mapped into the token embeddings.

2) *Format Brittleness*: Syntactic elements introduce ex-

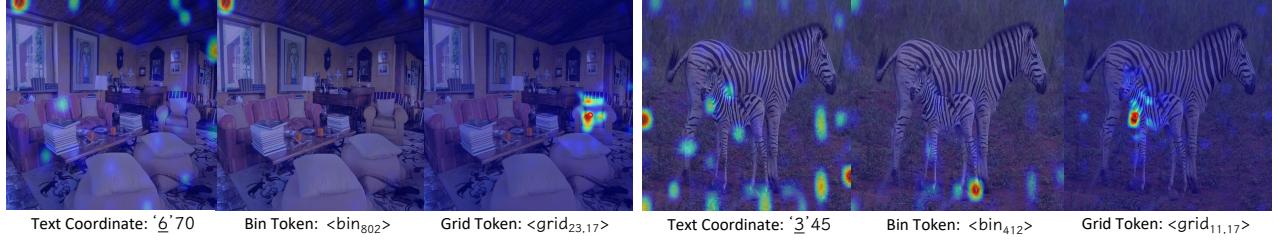


Figure 9. **Visualization of spatial responses for different localization vocabularies.** We aggregate attention maps between location tokens and image patches to obtain heatmaps for text coordinates, 1D bin tokens, and grid tokens. Grid tokens produce smooth, topology-aware activations that align with object extents.

ponential failure rates that are particularly problematic in multi-object scenarios. For example, with 98% per-token accuracy, a 12-token box sequence has 78% validity probability, dropping to 48% for three boxes (36 tokens).

3) *Metric–Objective Mismatch*: Token cross-entropy on digit sequences correlates poorly with geometric error. Small changes in token indices can correspond to large jumps in image space.

Using Qwen2.5-VL-7B with identical RefCOCO/+g instruction-tuning data, we compare text, bin, and grid formats in Tab. 9, and observe a clear advantage for grid tokens. Furthermore, as shown in Fig. 9, grid tokens produce smooth, locally coherent activations that closely follow object extents because each token is tied to a fixed 2D region in the image plane. In contrast, text and bin tokens yield fragmented, geometry-agnostic responses without a stable 2D correspondence.

Table 9. Ablation on **point representation formats** for REC on the RefCOCO/+g datasets.

Methods	refCOCO Test-A	refCOCO+ Test-A	refCOCOg Test
Text Coordinates	92.9	89.9	87.4
Bin token	92.3	89.9	87.1
Grid token	93.0	90.6	87.6

Why GRPO Works with GETok? GETok’s structured representation creates an ideal action space for GRPO optimization. As shown in Fig. 10, GETok achieves accelerated convergence and consistently higher reward levels at equivalent training steps compared to text coordinates, validating its structured action space advantage for GRPO optimization. We attribute this advantage to two key factors: (1) The 2D grid structure provides a stable foundation for policy learning, unlike text coordinates, where minor token changes yield discontinuous spatial shifts. (2) The finite $n \times n$ token format is easier to learn than text coordinates. This compact set allows the model to focus on spatial layout rather than complex text patterns, leading to faster convergence.

How to Represent Masks with Sparse Geometry? We analyze existing sparse geometric representations, such as

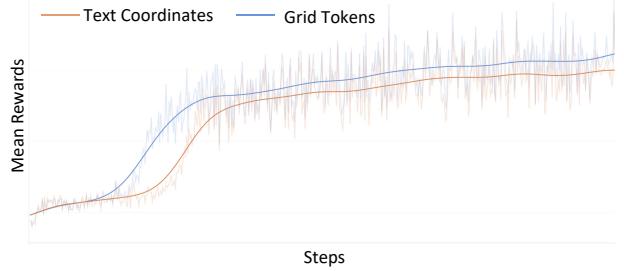


Figure 10. **Reward curve comparison between grid tokens and text coordinates.** GETok achieves faster convergence and higher rewards than text coordinates.

single points, bounding boxes, fixed sets of one or two points, or randomly sampled points, all of which suffer from redundancy and an inability to unambiguously capture complex mask semantics as shown in Fig. 11. We introduce a novel greedy algorithm that automatically extracts an appropriate set of such tokens from a target mask. Compared to methods that require training a dedicated mask decoder [25, 51, 71], this design offers several advantages:

1) *At training time*, our method avoids any mask-specific loss, decoder, or supervision, offering a much simpler alternative compared to methods that rely on task-specific decoders.

2) *At inference time*, our method offers strong flexibility as our decoder is purely plug-and-play and can be seamlessly updated without retraining the referring VLM. For example, replacing SAM [22] with advanced SAM2 [53], our method achieves a performance gain of 0.8% cIoU on refCOCO val at no cost. In contrast, LISA has to retrain the full model for this replacement, which is particularly costly.

A.3. Ablation Studies

Image Preprocessing. We investigate the impact of different image preprocessing strategies on localization performance as shown in Tab. 10. Padding gives the worst results, because the added gray borders effectively downscale the informative region and distract the model from relevant content. Center cropping risks semantic distortion by removing

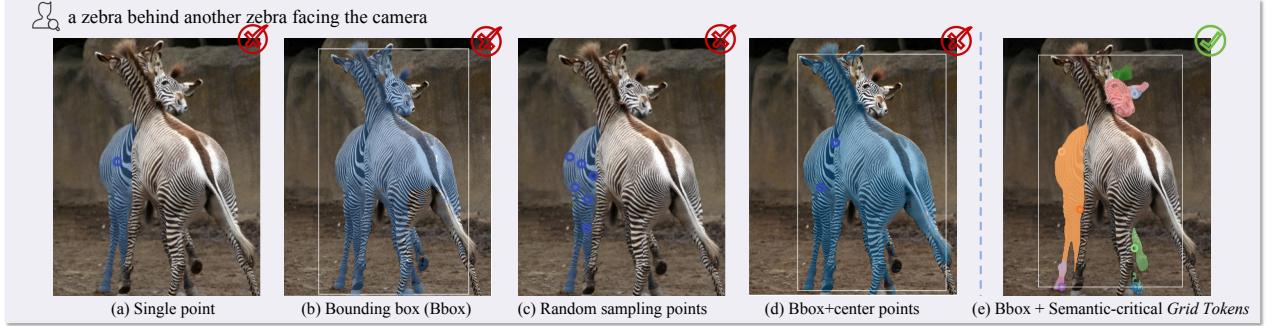


Figure 11. **Comparison of mask representation strategies.** We convert continuous masks into discrete, segment-critical grid tokens to achieve precise region referencing.

peripheral image areas. For example, in a referring expression such as “the person on the far left,” cropping may exclude the target entirely, leading to ground-truth mismatch. In contrast, resizing and dynamic resolution achieve comparable performance in our experiments. We therefore adopt simple resizing as our default preprocessing strategy.

Table 10. Ablation on **image preprocessing** strategies for REC on RefCOCOg.

Methods	RefCOCOg
Padding	85.9
Center Crop	86.2
Dynamic	87.1
Resize	87.4

Reward Function. For grid token generation, removing the semantic-critical points reward causes the model either to collapse to one or two high-confidence points or to overpopulate a small region with redundant points as shown in Tab. 11. Removing the box reward yields the largest drop, and visual inspection shows that points become scattered in the absence of a stable coarse prior. By contrast, the mask reward mainly provides fine-grained geometric supervision, especially for thin structures and concave regions that are not well constrained by box and point-level signals alone.

For offset token refinement, we focus on whether offsets perform genuine geometric corrections. The mask IoU gain and box refinement rewards provide instance-level guidance that promotes updates with improved mask and box IoU. The point refinement reward further stabilizes behavior by reducing large mask changes caused by a few erroneous point adjustments.

Reasoning vs. No Reasoning for Offset Refinement. The `<think>` process has been shown to be beneficial for multimodal understanding, especially in cases that require complex semantic reasoning [36, 39, 56]. We further examine its role in the refine stage. Empirically, the performance gap between using and omitting `<think>` during refinement

Table 11. Ablation on **reward design** for grid-token generation and offset-token refinement.

Variant	Reward for Grid Token Generation			
	Mask	Box	Sem. points	ReasonSeg
w/o Sem. points	✓	✓		58.6
w/o Mask reward		✓	✓	59.1
w/o Box reward	✓		✓	57.2
Full (ours)	✓	✓	✓	60.1
Variant	Reward for Offset Token Refinement			
	Point gain	Box gain	Mask IoU gain	ReasonSeg
w/o Mask IoU gain	✓	✓		61.8
w/o Box ref.	✓		✓	61.2
w/o Point ref.		✓	✓	60.5
Full (ours)	✓	✓	✓	62.8

is negligible (0.1% gIoU), suggesting that offset refinement does not substantially benefit from additional verbal reasoning. We observe that the model rarely produces meaningful explanations for point-level updates and instead repeats almost the same `<think>` content as in the propose step, so we do not enforce `<think>` generation in this stage.

B. Real-World Driving Dataset

We constructed a proprietary autonomous driving dataset to validate our method in complex scenarios in a fair comparison with state-of-the-art approaches. This dataset contains 1,988 training samples (29,825 annotations) and 980 test samples (14,524 annotations), covering diverse urban scenarios like intersections, highways, and pedestrian zones.

As illustrated in Fig. 12(a), the dataset categorizes driving targets into three classes: Traffic Lanes, Static Obstacles, and Traffic Signs with hierarchical annotations for multi-granular reasoning. We then design a series of classification tasks to evaluate the model’s ability to understand and refer to these specific regions.

Fig. 12(b) shows an example from our dataset, where each sample is annotated with object categories selected from the options illustrated in Fig. 12(a). Overall, driving scenes provide a realistic setting that demands understand-

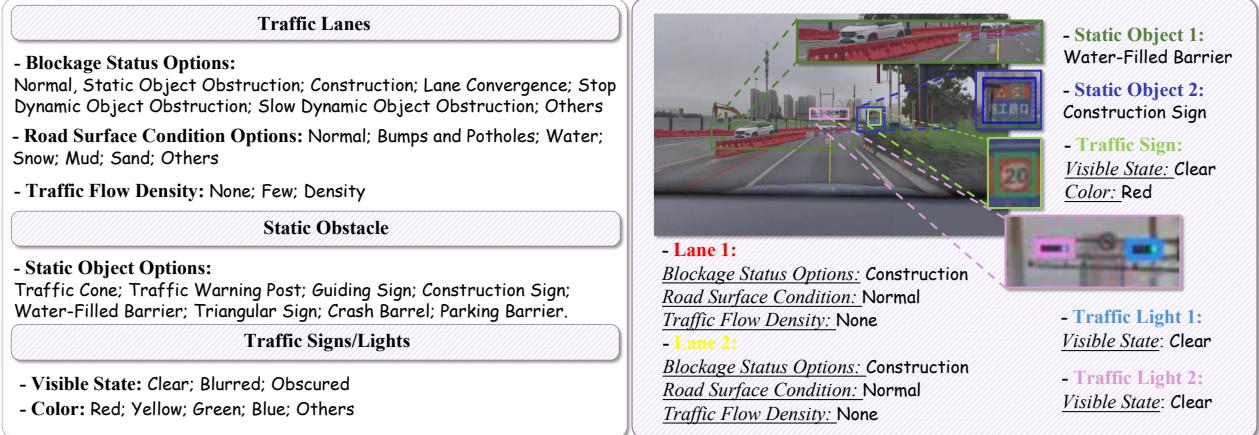


Figure 12. **Overview of driving dataset annotations information.** (a) Summarizes the taxonomy of annotated driving targets (lanes, static obstacles, and traffic signs) with hierarchical labels. (b) Illustrates an example scene annotated with points, polygons, polylines, bounding boxes, and masks for referring and safety-related queries.

ing and referring to regions in multiple formats, including points, polygons, polylines, bounding boxes, and masks, highlighting the application potential of a unified and robust localization framework.

C. More Implementation Details

C.1. Training Setup

Supervised Fine-Tuning. The model is fine-tuned on the mixed-task corpus summarized in Tab. 12. All location-related annotations (points, boxes, masks) are converted into GETok’s grid tokens. The offset-aware dataset is constructed on top of RefCOCO/+g and a more systematic description of the construction pipeline is provided in Sec. C.2. We use a per-device batch size of 2 with 8 gradient accumulation steps, yielding an effective batch size of 16 per device. All input images are resized to 840×840 , and training is conducted with bfloating16 mixed precision.

Reinforcement Learning. We first perform a cold-start stage to adapt the model to the newly introduced tokens while mixing in CoT-style instruction data, thereby preserving its original multimodal capabilities. Building on this checkpoint, we further optimize the policy with GRPO on both grid-token placement and offset-token refinement. Each update is regularized by a KL-divergence penalty to the SFT policy with coefficient 1×10^{-2} . For each prompt, we sample 8 candidate responses to estimate group-wise advantages. For offset tokens, we empirically find that about 200 steps are sufficient to obtain satisfactory refinement performance, which corresponds to roughly 5 hours of training on our setup.

Table 12. Summary of training data composition.

Stage	Datasets	Task
SFT	LLaVA-665K [30]	Image reasoning
	RefCOCO/+g [44, 81]	Referring grounding
	COCO-Stuff [5]; ADE20K [92]	Segmentation (seg.)
	Visual Genome [23]	Image captioning
	PACO-LVIS [50]; PASCAL-Part [10]	Part-level seg.
	gRefCOCO [28]	Multi-instance seg.
Cold Start	Pixmo-point [13]	Object pointing
	GETok-Offset	Referring refinement
GRPO	RefCOCO/+g [44, 81]	Referring seg.
	LLaVA-CoT-100K [73]	Instruction tuning
	GETok-Offset	Offset training
GRPO	RefCOCOg [44] subset (3.0K)	Single-target seg.
	LISA++ [75] (2.0K); gRefCOCO [28] (4.0K)	Multi-instance seg.

C.2. Offset-Aware Dataset Curation Details

Region Definitions. Let $\mathbf{M}_{gt} \in \{0, 1\}^{H \times W}$ be the binary foreground mask. We place an $n \times n$ grid and denote the pixel center of cell (i, j) by $\mathbf{c}_{i,j} = (x_{i,j}, y_{i,j})^\top$. To construct pools of candidate grid tokens, we employ morphology-based bands scaled according to the offset step size. Let $\mathcal{K}_k \in \{0, 1\}^{k \times k}$ represent a square structuring element with side length k pixels, $\mathbf{M}_{gt} \in \{0, 1\}^{H \times W}$ be the binary foreground mask. We define:

$$\begin{aligned} k_e &= \lfloor s_y \rfloor + 1, & \mathbf{E} &= \mathbf{M}_{gt} \ominus \mathcal{K}_{k_e}, \\ k_d &= 2\lfloor s_y \rfloor + 1, & \mathbf{D} &= \mathbf{M}_{gt} \oplus \mathcal{K}_{k_d}, \end{aligned} \quad (2)$$

where $\lfloor \cdot \rfloor$ denotes the floor operation, while \ominus and \oplus represent morphological erosion and dilation respectively. A thin boundary band is additionally defined as:

$$\mathbf{B} = \text{dilate}(\text{grad}(\mathbf{M}_{gt}), \mathcal{K}_b), \quad (3)$$

where $\text{grad}(\mathbf{M}_{\text{gt}})$ is the morphological gradient and b is a small width parameter. By construction, $\mathbf{E} \subset \mathbf{M}_{\text{gt}} \subset \mathbf{D}$: \mathbf{E} forms a step-sized interior buffer, \mathbf{D} creates a step-sized exterior halo, and \mathbf{B} captures edge uncertainty as a narrow boundary ribbon.

Grid Point Categorization and Sampling. We define a one-step hit test to determine reachability:

$$\text{Hit}(i, j) \triangleq \exists \delta \in \{-1, 0, 1\}^2 : \mathbf{M}_{\text{gt}}(\mathbf{c}_{i,j} + \mathbf{S}\delta) = 1. \quad (4)$$

Each grid center is assigned to exactly one category via the hierarchical decision rule:

$$\text{pool}(i, j) = \begin{cases} \text{Hard-Delete}, & \mathbf{B}(y_{i,j}, x_{i,j}) = 1 \\ & \wedge \mathbf{M}_{\text{gt}}(y_{i,j}, x_{i,j}) = 0 \\ & \wedge \neg \text{Hit}(i, j), \\ \text{Inside}, & \mathbf{E}(y_{i,j}, x_{i,j}) = 1, \\ \text{Ring}, & \mathbf{D}(y_{i,j}, x_{i,j}) = 1 \\ & \wedge \mathbf{M}_{\text{gt}}(y_{i,j}, x_{i,j}) = 0, \\ \text{Far}, & \text{otherwise}. \end{cases} \quad (5)$$

Following pool formation $\mathcal{P}_{\text{hard}} \rightarrow \mathcal{P}_{\text{inside}} \rightarrow \mathcal{P}_{\text{ring}} \rightarrow \mathcal{P}_{\text{far}}$, we sample $K \sim \pi_K$ grids per image with preferential selection from $\mathcal{P}_{\text{inside}}$ and $\mathcal{P}_{\text{ring}}$, while maintaining representation from all categories for robustness. Then, the complete construction process, detailed in Algorithm 1, processes each image-mask-query triple to automatically produce conversational data containing grid tokens and their corresponding offset targets.

C.3. Reward Details

Multi-object Matching. From each line in `<answer>` we extract p predicted instance consisting of an optional box $\hat{\mathbf{b}}_p \in \mathbb{R}^4$ and a point set $\mathcal{P}_p = \{\mathbf{q}\} \subset \mathbb{R}^2$. Let there be P predictions and G ground-truth (GT) instances with binary masks $\{\mathbf{M}_g\}_{g=1}^G$ and tight boxes $\{\mathbf{b}_g\}_{g=1}^G$. We define pairwise similarities between predicted p and GT g :

i) Box IoU:

$$\text{IoU}_{p,g} \in [0, 1]. \quad (6)$$

ii) Point-hit ratio: the fraction of predicted points that land inside \mathbf{M}_{gt} ,

$$H_{p,g} = \frac{1}{\max(1, |\mathcal{P}_p|)} \sum_{\mathbf{q} \in \mathcal{P}_p} \mathbb{1}\{\mathbf{q} \in \mathbf{M}_{\text{gt}}\} \in [0, 1]. \quad (7)$$

iii) Normalized L_1 box score:

$$S_{p,g}^{\ell_1} = \text{clip}\left(1 - \frac{\|\hat{\mathbf{b}}_p - \mathbf{b}_g\|_1/4}{\tau_{\ell_1}}, 0, 1\right). \quad (8)$$

These are combined into a similarity used only for the assignment:

$$\text{Sim}_{p,g} = \text{IoU}_{p,g} + H_{p,g} + S_{p,g}^{\ell_1}, \quad (9)$$

Algorithm 1: Offset-Supervised Data Construction

```

Input: Referring dataset  $\mathcal{D}$ ; grid size  $n$ ; offset granularity  $m$ ; IoU threshold  $\tau$ 
Output: JSONL conversations containing grids and offset targets
foreach  $(I, \mathbf{M}_{\text{gt}}, q) \in \mathcal{D}$  do
    Resize  $I, \mathbf{M}_{\text{gt}}$  to  $H \times W$ ; compute  $s_x = W/m$ ,
     $s_y = H/m$ ,  $\mathbf{S} = \text{diag}(s_x, s_y)$ ;
    // grid pools via morphology (cf. (2)–(3))
    Compute  $\mathbf{E}, \mathbf{D}, \mathbf{B}$ ; assign each grid cell  $(i, j)$  to one of INSIDE/RING/FAR/HARD-DELETE by rule (5);
    // Segmentation grids and offsets
    Sample  $K$  grids  $\{(i_k, j_k)\}_{k=1}^K$  from the pools;
    for  $k = 1$  to  $K$  do
        Set  $\mathbf{c}_k \leftarrow \mathbf{c}_{i_k, j_k}$ ;
        if  $\mathbf{M}_{\text{gt}}(y_{i_k}, x_{i_k}) = 1$  then
            | emit [OFF_0_0]
        else if Hit3×3( $i_k, j_k$ ) then
            | pick  $(\delta_u, \delta_v) \in \{-1, 0, 1\}^2$  with
            |  $\mathbf{M}_{\text{gt}}(\mathbf{c}_k + \mathbf{S}\delta) = 1$ , and emit
            | [OFF_δu-δv]
        else
            | emit <DELETE>
        // Bounding-box corner offsets
        Let  $B^* \leftarrow \text{BBox}(\mathbf{M}_{\text{gt}})$ ; jitter its TL/BR to grid corners  $(i_{\text{tl}}, j_{\text{tl}}), (i_{\text{br}}, j_{\text{br}})$ ;
        Evaluate all offset pairs for the two corners (apply  $\mathbf{S}$ -scaled displacements), obtain  $\text{IoU}_{\text{max}}$ ;
        if  $\text{IoU}_{\text{max}} \geq \tau$  then
            | emit the two corner offsets
        else
            | emit <DELETE> for both corners
        // Serialization
        Write a JSONL sample with image tag, user prompt  $q$  and grids (user turn), and the offsets (assistant turn);

```

We solve a Hungarian assignment [24] with costs $C_{p,g} = 3 - \text{Sim}_{p,g}$, yielding matched pairs $\mathcal{M} \subseteq \{1..P\} \times \{1..G\}$. We use $\tau_{\ell_1}=18$ px.

Semantic-Critical Points Reward. For each $(p, g) \in \mathcal{M}$, we compute a key points quality:

$$F_{p,g} \triangleq S(m_p) \left(w_H H_{p,g} + w_{\text{spr}} \text{Spread}_{p,g} \right) - \lambda_m m_p. \quad (10)$$

where $H_{p,g}$ is the hit ratio, and $\text{Spread}_{p,g}$ rewards larger nearest-neighbor spacing normalized by object scale:

$$\bar{d}_p = \frac{1}{m_p} \sum_{i=1}^{m_p} \min_{j \neq i} \|\mathbf{q}_i - \mathbf{q}_j\|_2, \quad (11)$$

$$\text{Spread}_{p,g} = \text{clip}(\bar{d}_p / (\rho_s r_g), 0, 1).$$

The multiplicative saturation $S(m) = 1 - \exp(-m/m_0)$ discourages degenerate few-point outputs, and the linear

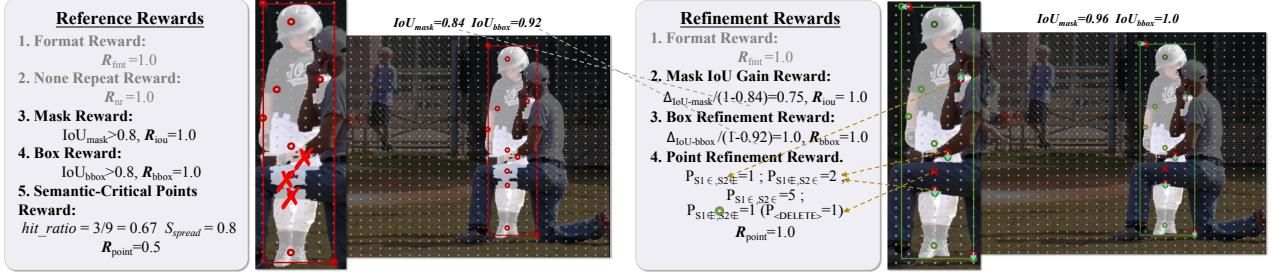


Figure 13. **Illustration of reward computation for grid token generation and refinement.** The diagram demonstrates how different reward components are calculated based on predicted outputs and ground-truth annotations.

term $\lambda_m m_p$ penalizes overly long point lists. We aggregate across matches with point-count weighting:

$$T = \text{clip}\left(\frac{\sum_{(p,g) \in \mathcal{M}} m_p F_{p,g}}{\sum_{p=1}^P \max(1, m_p)}, 0, 1\right). \quad (12)$$

We set $w_H=0.6$, $w_{\text{spr}}=0.4$, $\lambda_m=0.02$, $\rho_s=0.30$.

Point Refinement Reward. Let $\mathbf{M}_{\text{gt}}^{(k)} \subset \mathbb{Z}^2$ be the ground-truth mask of the k -th instance. The coarse point set is $\mathcal{C}_k = \{\mathbf{c}_{k,p}\}_{p=1}^{P_k}$ and the refined set is $\mathcal{C}_k^{\text{off}} = \{\mathbf{c}_{k,p}^{\text{off}}\}_{p=1}^{P_k}$, with a one-to-one correspondence over p (if a point is deleted, we keep its index p and mark a delete flag). Define the inclusion indicators $I_{k,p} = \mathbb{I}[\mathbf{c}_{k,p} \in \mathbf{M}_{\text{gt}}^{(k)}]$, $I_{k,p}^{\text{off}} = \mathbb{I}[\mathbf{c}_{k,p}^{\text{off}} \in \mathbf{M}_{\text{gt}}^{(k)}]$. The point-wise reward $s_{k,p} \in \{-1, 0, 1\}$ is

$$\begin{cases} -1, & I_{k,p} = 1 \wedge I_{k,p}^{\text{off}} = 0 \\ +1, & I_{k,p} = 0 \wedge I_{k,p}^{\text{off}} = 1 \\ +1, & I_{k,p} = 1 \wedge I_{k,p}^{\text{off}} = 1 \\ +1, & I_{k,p} = 0 \wedge \langle \text{DELETE} \rangle \wedge \mathcal{N}_{3 \times 3}(\mathbf{c}_{k,p}) \cap \mathbf{M} = \emptyset \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

where $\mathcal{N}_{3 \times 3}(\mathbf{c}_{k,p})$ is the 3×3 neighborhood centered at $\mathbf{c}_{k,p}$. The instance-level reward averages over its points. Fig. 13 provides a concrete example illustrating the reward computation process for better understanding.

D. Additional Visualization Results

Grid Tokens for Mask Representation. Fig. 14 presents additional qualitative results comparing predicted grid tokens, output masks, and GT annotations. The results are organized from top to bottom, ranging from predictions that are more precise than the GT mask to some failure cases. These visualizations highlight the following key observations:

(1) *High-Quality Predictions:* The model is capable of generating highly accurate grid tokens, which align well with the GT masks. These results demonstrate the effectiveness of grid tokens in precisely localizing and referring to objects in complex scenes.

(2) *Failure Cases:* In some cases, accurate grid-token predictions still yield imperfect masks due to discrepancies in SAM’s mask decoding. Nonetheless, as discussed in Sec. A.2, this training-free decoding remains advantageous compared to training task-specific mask decoders. Introducing offset tokens further mitigates these errors by refining point locations and aligning the decoded masks more closely with object boundaries.

The qualitative results underscore the robustness of grid tokens as a referring representation, even in cases where segmentation performance is suboptimal.

SFT Benchmarks Qualitative Results. Fig. 15 demonstrates the unified representation capability of GETok across diverse vision-language tasks. Our approach establishes a cohesive framework that processes various query types through a consistent token vocabulary, spanning image-, point-, box-, and mask-level formats while eliminating the need for task-specific output heads.

Self-Improving Mechanism. Fig. 16 presents additional qualitative examples demonstrating the propose-and-refine workflow of GETok for fine-grained mask prediction. The left panel shows that for interior points unambiguously inside the mask, the model correctly maintains their positions without unnecessary adjustments, focusing refinement efforts exclusively on boundary regions. The right panel illustrates a failure case primarily caused by erroneous refinement decisions resulting from initial tokens placed near misleading edge features. These examples collectively highlight the method’s capacity to maintain accurate localization through coordinated grid and offset token operations, even in challenging scenarios.

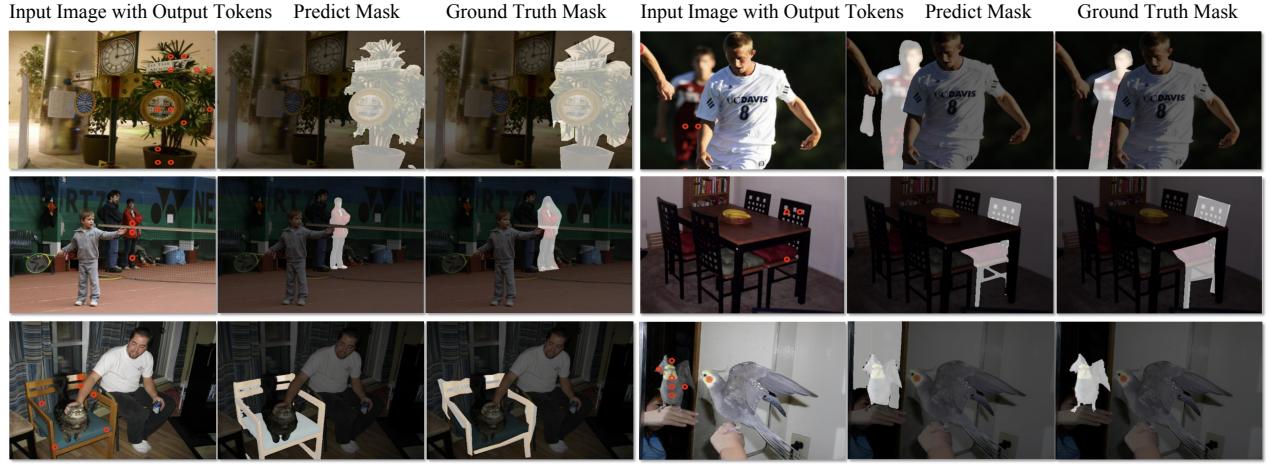


Figure 14. **More qualitative results of the segmentation task.** From top to bottom, the predictions are ordered by decreasing Intersection-over-Union (IoU) scores relative to the ground truth masks.

Image-Level	 (a) Image Understanding	Point-Level	 (b) Lane Detection
Bbox-Level	 (c) Object Grounding	 (d) Grounding Caption	
Pixel-Level	 (e) Region-Level Caption	 (f) Detailed Descriptions	 (g) Part-Level Segmentation
		 (h) Single Target Segmentation	 (i) Multiple Targets Segmentation

(a) Image Understanding
Q: Aside from the clock tower, what other architectural features can be observed in the building?
A: The building incorporates a distinctive architectural feature, with a section extending over the road to link different parts of the structure. Moreover, another building positioned directly behind the clock tower enhances the depth of the overall composition.

(b) Lane Detection
Q: Analyze the lane data in the image and number each lane according to the order from left to right and give the key grid tokens for each lane.

(c) Object Grounding
Q: Find a peacock walking on grass with grounding.

(d) Grounding Caption
Q: Describe in detail the visual characteristics of the region `<bbox><grid10,6><grid26,30></bbox>`.
A: A brown pony with black hooves is visible, it appears to be standing near the other horses and is looking towards our left which indicates its active engagement in its surroundings.

(e) Region-Level Caption
Q: What is in the region defined by region `<seg><grid5,13><grid14,15></seg>` in the image?
A: Scarf of the dog.

(f) Detailed Descriptions
Q: Describe the visual characteristics of the region `<seg><grid20,10><grid12,14><grid16,14><grid18,18><grid18,26></seg>` in the image.
A: Leading the row of food trucks is a striking all-white vehicle, likely serving as the first stop for attendees of the street food event. The open rooftop indicates that it is currently in operation.

(g) Part-Level Segmentation
Q: What object in the picture could be used to provide support and protection for the injured part?

(h) Single Target Segmentation
Q: Where is a dark car in front of a red car in the image?

(i) Multiple Targets Segmentation
Q: Where is the elephant butt showing to left and the elephant on the far right?

Figure 15. **Unified GETok representations across diverse vision-language tasks.** GETok provides a unified representation framework that handles diverse visual concepts without task-specific architectural modifications.

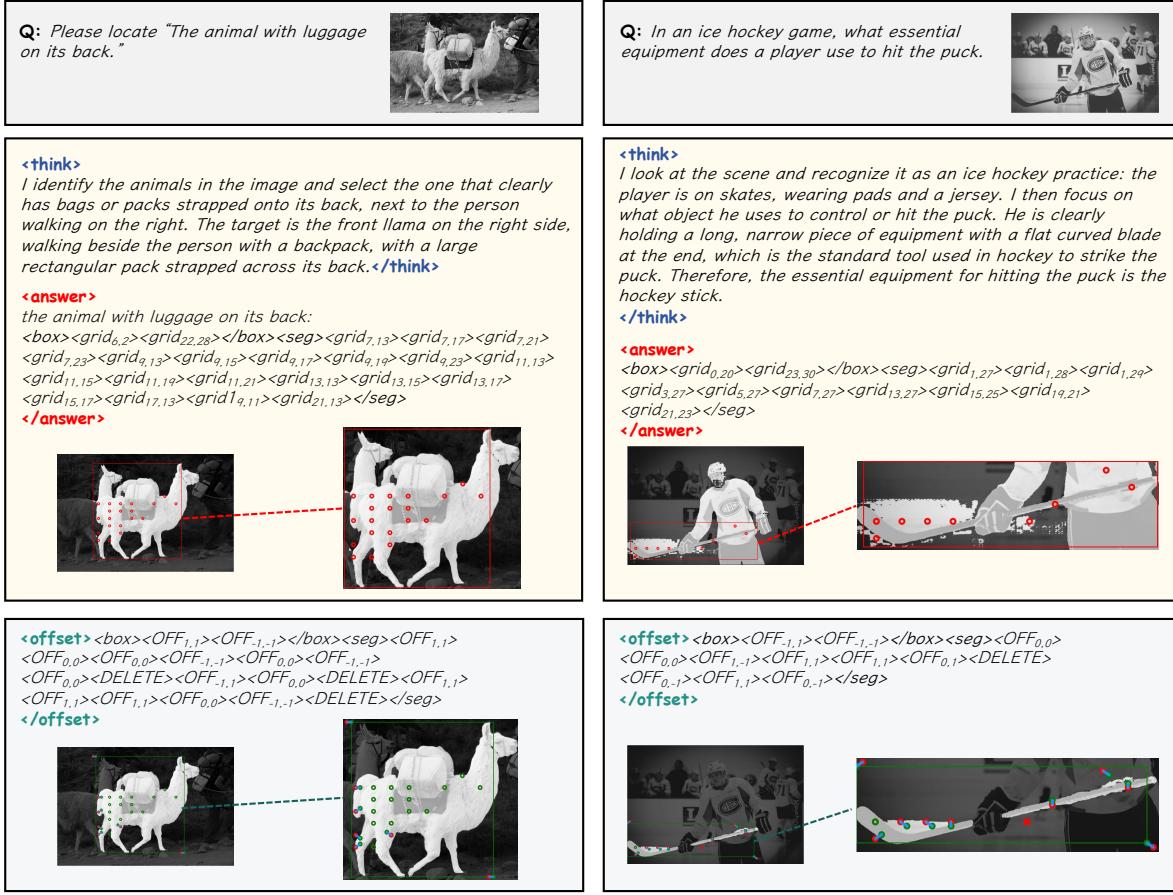


Figure 16. More qualitative results of the self-improving mechanism. Additional examples demonstrate how GETok establishes initial spatial proposals through grid tokens (red dots) and enables fine-grained adjustments via offset tokens (blue arrows), showing effective handling of objects across scales with enhanced precision on small targets.

References

- [1] David Acuna et al. Long grounded thoughts: Distilling compositional visual reasoning chains at scale. *arXiv preprint arXiv:2511.05705*, 2025. 3
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, pages 23716–23736, 2022. 1
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 6, 7
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, pages 1877–1901, 2020. 1
- [5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018. 4
- [6] ChameleonTeam. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 3
- [7] Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-Enhanced Visual Instruction Tuning for Multimodal Large Language Models. *arXiv preprint arXiv:2308.13437*, 2023. 2
- [8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1, 2, 7
- [9] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *ICLR*, 2022. 2, 3
- [10] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, pages 1971–1978, 2014. 4
- [11] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation mod-

- els and aligning for generic visual-linguistic tasks. *arXiv e-prints*, pages arXiv–2312, 2023. 1
- [12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2023. 1
- [13] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *CVPR*, pages 91–104, 2025. 2, 4, 1
- [14] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 1
- [15] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 2
- [16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv:2106.09685*, 2021. 6
- [17] Jiaqi Huang, Zunnan Xu, Jun Zhou, Ting Liu, Yicheng Xiao, Mingwen Ou, Bowen Ji, Xiu Li, and Kehong Yuan. Sam-r1: Leveraging sam for reward feedback in multimodal segmentation via reinforcement learning. *arXiv preprint arXiv:2505.22596*, 2025. 6
- [18] Qing Jiang, Gen Luo, Yuqin Yang, Yuda Xiong, Yihao Chen, Zhaoyang Zeng, Tianhe Ren, and Lei Zhang. ChatRex: Tam-ing Multimodal LLM for Joint Perception and Understanding. *arXiv preprint arXiv:2411.18363*, 2024. 1, 2
- [19] Qing Jiang, Lin Wu, Zhaoyang Zeng, Tianhe Ren, Yuda Xiong, Yihao Chen, Qin Liu, and Lei Zhang. Referring to any person. In *CVPR*, 2025. 2
- [20] Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Bin Chen, Chenyi Lei, An Liu, Chengru Song, Xiaojiaqiang Lei, et al. Unified language-vision pretraining with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023. 1, 3
- [21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 1
- [22] 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. In *ICCV*, pages 4015–4026, 2023. 4, 5, 2
- [23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017. 1, 4
- [24] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. 5
- [25] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, pages 9579–9589, 2024. 2, 3, 6, 7, 1
- [26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 1
- [27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*, 2023. 1
- [28] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, pages 23592–23601, 2023. 6, 1, 4
- [29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1
- [30] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 4
- [31] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1
- [32] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *MM*, pages 4235–4243, 2020. 2
- [33] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *CVPR*, pages 3042–3051, 2022.
- [34] Ye Liu, Jixuan He, Wanhua Li, Junsik Kim, Donglai Wei, Hanspeter Pfister, and Chang Wen Chen. r 2-tuning: Efficient image-to-video transfer learning for video temporal grounding. In *ECCV*, pages 421–438, 2024.
- [35] Ye Liu, Huifang Li, Chao Hu, Shuang Luo, Yan Luo, and Chang Wen Chen. Learning to aggregate multi-scale context for instance segmentation in remote sensing images. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):595–609, 2024. 2
- [36] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025. 3, 4, 6
- [37] Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. VisionReasoner: Unified Visual Perception and Reasoning via Reinforcement Learning. *arXiv preprint arXiv:2505.12081*, 2025. 3, 4, 6, 7
- [38] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025. 3
- [39] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-

- rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025. 3
- [40] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024. 1
- [41] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. *arXiv preprint arXiv:2404.13013*, 2024. 2, 7, 1
- [42] Tianren Ma, Lingxi Xie, Yunjie Tian, Boyu Yang, and Qixiang Ye. Clawmachine: Learning to fetch visual tokens for referential comprehension. *arXiv preprint arXiv:2406.11327*, 2024. 1, 3, 7
- [43] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 6, 1
- [44] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 4
- [45] OpenAI. Introducing openai o1-preview. <https://openai.com/index/introducing-openai-o1-preview/>, 2020. 3
- [46] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023. 1
- [47] OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [48] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3, 1
- [49] 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*, pages 8748–8763, 2021. 3
- [50] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *CVPR*, pages 7141–7151, 2023. 4
- [51] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abderrahman 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*, pages 13009–13018, 2024. 2, 1
- [52] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *SIGKDD*, pages 3505–3506, 2020. 7
- [53] 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. 2
- [54] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jia Shi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *CVPR*, pages 26374–26383, 2024. 6
- [55] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 5, 6
- [56] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model. *arXiv preprint arXiv:2504.07615*, 2025. 3
- [57] Yongyi Su, Haojie Zhang, Shijie Li, Nanqing Liu, Jingyi Liao, Junyi Pan, Yuan Liu, Xiaofen Xing, Chong Sun, Chen Li, et al. Patch-as-decodable-token: Towards unified multi-modal vision tasks in mllms. *arXiv preprint arXiv:2510.01954*, 2025. 1
- [58] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation. *arXiv preprint arXiv:2406.06525*, 2024. 3
- [59] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multi-modal models are in-context learners. *arXiv: 2312.13286*, 2023.
- [60] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *ICLR*, 2024. 1, 3
- [61] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Jo-han Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv: 2312.11805*, 2023. 1
- [62] UniPixel Team. Unipixel: A unified pixel-level multi-modal model for referring, segmentation and reasoning. In *NeurIPS*, 2025. 2
- [63] Yunjie Tian, Tianren Ma, Lingxi Xie, Jihao Qiu, Xi Tang, Yuan Zhang, Jianbin Jiao, Qi Tian, and Qixiang Ye. Chatterbox: Multi-round multimodal referring and grounding. *arXiv preprint arXiv:2401.13307*, 2024. 2
- [64] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhihang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, pages 23318–23340, 2022. 2
- [65] Wenhui Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In *NeurIPS*, pages 61501–61513, 2023. 7
- [66] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. In *NeurIPS*, pages 121475–121499, 2024. 2

- [67] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *arXiv*: 2409.18869, 2024. 1
- [68] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, pages 11686–11695, 2022. 6
- [69] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. In *ECCV*, pages 207–224, 2024. 1
- [70] Jiannan Wu, Muyan Zhong, Sen Xing, Ziqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhui Wang, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. In *NeurIPS*, pages 69925–69975, 2024. 2, 3
- [71] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *CVPR*, pages 3858–3869, 2024. 2, 3, 1
- [72] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. 3
- [73] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *CVPR*, pages 2087–2098, 2025. 6, 4
- [74] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Ze-huan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, pages 15325–15336, 2023. 7
- [75] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. Lisa++: An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023. 6, 4
- [76] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *ECCV*, pages 521–539, 2022. 3
- [77] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Heng-shuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, pages 18155–18165, 2022. 6, 1
- [78] 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. *arXiv preprint arXiv:2310.07704*, 2023. 1, 2, 3, 7
- [79] Zuyao You and Zuxuan Wu. Seg-r1: Segmentation can be surprisingly simple with reinforcement learning. *arXiv preprint arXiv:2506.22624*, 2025. 3
- [80] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 6
- [81] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 4
- [82] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *CVPR*, pages 7282–7290, 2017. 1
- [83] 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*, pages 28202–28211, 2024. 2
- [84] Simon Zhai. *Towards Vision-Language Foundation Models: Limitations, Improvements, and Generalization*. PhD thesis, University of California, Berkeley, 2025. Technical Report No. UCB/EECS-2025-9. 3
- [85] Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon: Spelling out all object locations at any granularity with large language models. In *ECCV*, pages 405–422, 2024. 2
- [86] Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. Next-chat: An Imm for chat, detection and segmentation. *arXiv:2311.04498*, 2023. 3
- [87] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024.
- [88] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 2, 1
- [89] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *CVPR*, pages 14227–14238, 2024. 3
- [90] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Hong Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, et al. Swift: A scalable lightweight infrastructure for fine-tuning, 2024. 6
- [91] Yaowei Zheng, Junting Lu, Shenzi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework, 2025. 6
- [92] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 4
- [93] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 1, 2