ONESEG: IMPROVING GENERALIZED REFERRING SEGMENTATION WITH SEMANTIC-INSTANCE DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalized referring segmentation extends conventional referring segmentation to scenarios where a referring expression may correspond to zero, one, or multiple instances, requiring models to jointly reason about referential semantics, instance cardinality, and target existence. Existing MLLM-based methods either autoregressively generate multiple [SEG] tokens to represent individual targets, making subsequent segmentation vulnerable to missing or duplicated tokens, or employ mask queries for parallel prediction, underutilizing the MLLM's native autoregressive reasoning capability. In this work, we propose **OneSEG**, a decoupled framework that preserves autoregressive referential reasoning while deferring instance prediction to decoding. Specifically, OneSEG employs a **single [SEG] token** to capture the shared referential semantics of all potential targets. Our **Semantic-to-Instance Query Decomposition (SIQD)** then transforms this holistic representation into instance-aware queries conditioned on visual evidence, enabling flexible multi-instance prediction without explicit instance enumeration during multimodal reasoning. Furthermore, we introduce an **Instance Evidence Verifier (IEV)** that verifies decoded candidates using explicit spatial evidence. Extensive experiments on generalized referring segmentation benchmarks demonstrate that OneSEG achieves state-of-the-art performance, validating the effectiveness of decoupling semantic reasoning from instance prediction.

1 INTRODUCTION

Referring expression segmentation (RES) aims to segment image regions described by a natural-language expression (Hu et al., 2016; Wang et al., 2022; Yang et al., 2022). While conventional RES typically assumes a single valid referent, generalized referring expression segmentation (GRES) further allows an expression to correspond to zero, one, or multiple instances (Liu et al., 2023). This extension fundamentally changes the problem: a model must not only understand complex vision-language semantics, but also predict a variable-size set of instances and determine whether valid referents exist. While earlier approaches mainly rely on task-specific cross-modal architectures (Dai et al., 2025a; Chen et al., 2024a; Hu et al., 2023a), Multimodal Large Language Models (MLLMs) (Li et al., 2023; Wang et al., 2024; Chen et al., 2024c; Liu et al., 2024; Bai et al., 2025) have recently emerged as a powerful foundation for GRES with stronger semantic understanding, multimodal alignment and reasoning capabilities.

Existing MLLM-based GRES approaches mainly follow two paradigms: 1) Autoregressive instance enumeration, where MLLMs generate multiple [SEG] tokens as instance representations for mask decoding, while using a [REJ] token to indicate the absence of valid referents. The resulting [SEG] representations are subsequently fed into a learnable decoder to produce pixel-level masks (Xia et al., 2024; Wang et al., 2025a; Zhang et al., 2026b; 2024a); and 2) Query-based set prediction, where additional mask queries are introduced within the MLLM and subsequently decoded into masks by a set-based decoder (Zhang et al., 2024b; Wei et al., 2024; 2025). As shown in Figure 1(a), while the first paradigm effectively extends the reasoning capabilities of MLLMs to generalized scenarios, it introduces two inherent limitations: (i) the number of generated [SEG] tokens directly determines the mask candidates, making instance cardinality vulnerable to missing or duplicated [SEG], especially as the number of targets increases (Yuan et al., 2026); and (ii) [REJ] turns target existence

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

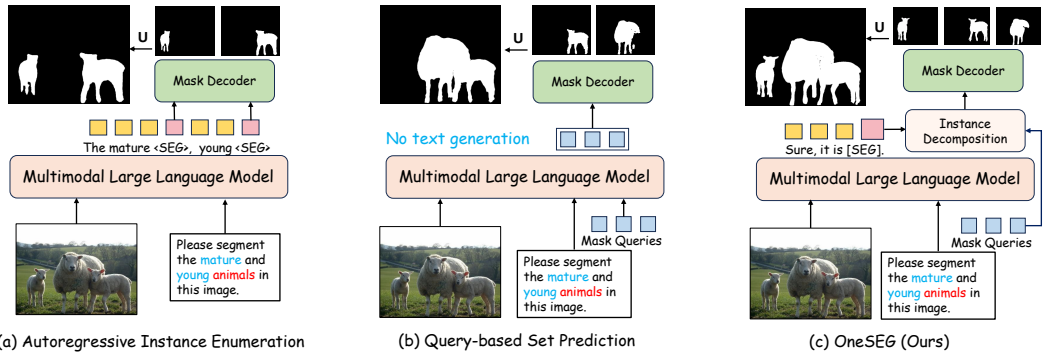

Figure 1: Overview of MLLM-based GRES paradigms. (a) Autoregressive Multiple [SEG]/[REJ] generation directly couples instance cardinality and target existence with discrete token generation. (b) Query-based mask decoding introduces multiple mask queries into the MLLM for parallel set prediction, at the cost of additional computation and reduced use of autoregressive generation. (c) Our **OneSEG** decouples semantic reasoning from instance enumeration: the MLLM generates a single referential representation, which is subsequently decomposed into multiple instance queries for parallel existence prediction and mask decoding.

into an early hard decision, preventing downstream pixel-level evidence from correcting false rejections. To address these issues, the query-based paradigm decouples instance cardinality from autoregressive generation through parallel set prediction, as illustrated in Figure 1(b). However, it still suffers from two limitations: (i) it underutilizes the native autoregressive generation capability of MLLMs, limiting flexibility in reasoning-intensive segmentation (Wei et al., 2024); and (ii) the additional mask queries incur substantial computational overhead when propagated through large MLLMs (Wen et al., 2026). These observations raise a more fundamental question: **is it necessary to explicitly maintain multiple instance representations during multimodal reasoning?**

We argue that instance enumeration can be deferred from multimodal reasoning to the subsequent spatial decoding stage. Based on this insight, we propose **OneSEG**, a decoupled framework that separates referential reasoning from instance prediction. Specifically, OneSEG is built upon three key designs: 1) **One-SEG semantic reasoning**. Instead of enumerating instances within the MLLM, we use a single [SEG] token to summarize the shared referential semantics distilled through multimodal reasoning. Since all valid instances are grounded by the same referring expression, this decoupled design preserves the autoregressive reasoning capability of the MLLM while avoiding both brittle instance enumeration and redundant query computation within the language model, leaving instance cardinality and localization to downstream decoding. 2) **Semantic-to-Instance Query Decomposition (SIQD)**. Although the [SEG] representation captures holistic vision-language semantics of the referred targets, it lacks the instance-level specificity required for multi-instance prediction. To bridge this gap, we introduce a **Semantic-to-Instance Query Decomposition** module, which decomposes the single [SEG] representation into a set of instance-aware queries conditioned on both referential semantics and visual evidence. 3) **Instance Evidence Verifier (IEV).** Rather than determining instance existence from an early [REJ] decision or query-level confidence alone, we verify each decoded instance using its predicted mask together with visual features and instance semantics, allowing explicit pixel-level evidence to suppress spurious candidates. Extensive experiments demonstrate that **OneSEG** establishes a new state-of-the-art on generalized referring segmentation benchmarks while reducing redundant MLLM computation, validating the effectiveness of single-token semantic reasoning, semantic-to-instance query decomposition, and evidence-aware instance verification. The contributions of this work are summarized as follows:

- We propose **OneSEG**, a decoupled MLLM-based GRES framework that separates referential reasoning from instance prediction, preserving the autoregressive reasoning capability of MLLMs while avoiding brittle instance enumeration and redundant in-MLLM query computation.
- We introduce **Semantic-to-Instance Query Decomposition** and **Instance Evidence Verifier** to derive instance-aware queries from holistic referential semantics and verify decoded candidates with pixel-level evidence, enabling robust multi-instance prediction and no-target handling.

- We conduct extensive experiments on widely used generalized referring segmentation benchmarks to demonstrate the performance superiority of our proposed method over previous SOTA approaches, validating the effectiveness of our strategy.

2 Related Work

2.1 Referring and Reasoning Segmentation

Referring Expression Segmentation (RES) aims to segment objects specified by natural-language expressions (Hu et al., 2016), with prior methods mainly relying on cross-modal vision-language fusion for target grounding (Wang et al., 2022; Yang et al., 2022; Deng et al., 2021; Zhu et al., 2022; Chen et al., 2024a; Dai et al., 2024; Xiao et al., 2024; Dai et al., 2025b;c; Yu et al., 2018; Luo et al., 2020). Generalized Referring Expression Segmentation (GRES) further extends RES to arbitrary numbers of targets, including multi-target and no-target cases (Liu et al., 2023; Hu et al., 2023a; Shah et al., 2024; Luo et al., 2025; Dai et al., 2025a). With the emergence of MLLMs, LISA (Lai et al., 2024) introduced ReasonSeg and connected language reasoning to mask prediction through the hidden embedding of a special [SEG] token, inspiring subsequent methods to exploit MLLMs for grounding complex and implicit referring intents (Yang et al., 2023; Zhang et al., 2024a; Wang et al., 2025c). While these advances enhance semantic reasoning, extending MLLM-based segmentation from determining *what to segment* to *how many instances to segment* introduces a distinct challenge.

2.2 MLLM-based Generalized Referring Segmentation

Existing MLLM-based generalized segmentation methods mainly follow two paradigms. Generation-based methods such as GSVA (Xia et al., 2024) autoregressively generate multiple [SEG] tokens for multiple targets and use [REJ] for absent referents. Although this preserves the reasoning capability of MLLMs, it couples semantic understanding with instance enumeration, making mask prediction sensitive to missing or duplicated tokens and increasingly costly as the number of targets grows (Wang et al., 2025a; Zhang et al., 2026b; 2024a; Rasheed et al., 2024; Yuan et al., 2025). Alternatively, query-based methods (Zhang et al., 2024b; Wei et al., 2024; 2025) employ multiple mask queries and parallel set decoding to avoid autoregressive instance enumeration. However, maintaining many instance-specific queries inside the MLLM introduces substantial redundant computation, while recent studies show that mask queries mainly interact with visual tokens in deeper layers (Wen et al., 2026). Recently, SELF1E (Zhang et al., 2026a) demonstrates that a single [SEG] token can serve as an effective vision-language representation for segmentation. In contrast, OneSEG focuses on a different problem: how such a holistic representation can support variable-cardinality instance prediction in GRES. Specifically, we treat the single [SEG] token as shared referential semantics and defer instance decomposition outside the MLLM, where it is transformed into instance-aware queries for separate mask prediction and existence verification.

3 Method

In this section, we first provide an overview of the proposed OneSEG framework in **Sec. 3.1**. We then introduce One-SEG Referential Reasoning in **Sec. 3.2**, followed by the Semantic-to-Instance Query Decomposition module in **Sec. 3.3**. **Sec. 3.4** details the Visual-Semantic Fusion strategy, while **Sec. 3.5** presents the Instance Mask Decoding and Verification process. Finally, **Sec. 3.6** describes the training objectives used to optimize the entire framework.

3.1 Overview

Given an image I and a referring expression T, OneSEG first performs multimodal reasoning to obtain a single holistic referential representation h_{SEG}. The proposed SIQD module then decomposes h_{SEG} into a fixed set of instance-aware queries $\mathbf{Q}_{\mathrm{ins}}$, which are subsequently decoded into candidate masks and verified using explicit pixel-level evidence. As illustrated in Fig. 2, OneSEG consists of three stages: One-SEG referential reasoning, semantic-to-instance query decomposition, and instance mask decoding with evidence verification.

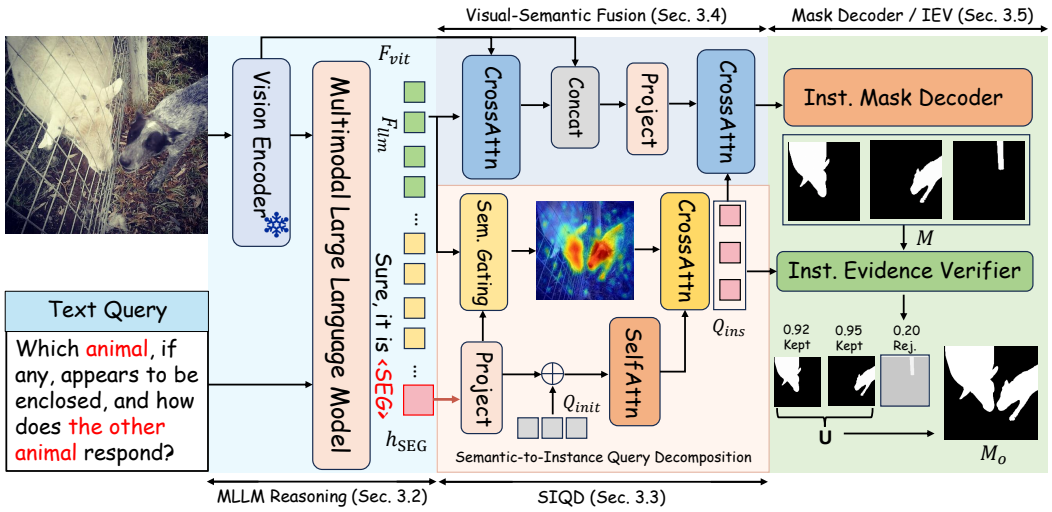

Figure 2: Overview of **OneSEG**. The MLLM produces a single holistic [SEG] representation for referential reasoning. Our **SIQD** module transforms it into multiple instance-aware queries by jointly leveraging referential semantics and visual evidence. The resulting queries are decoded into instance masks and further verified by the **IEV** using explicit pixel-level evidence.

3.2 One-SEG Referential Reasoning

Given an image I and an expression T, we first extract visual features using a vision encoder $\mathcal{E}_v$:

$$F_{\text{vit}} = \mathcal{E}_v(I), \tag{1}$$

which are projected into the embedding space of the MLLM and jointly processed with the textual tokens. The MLLM then performs autoregressive reasoning and generates a response containing a single [SEG] token. We denote the final-layer hidden state corresponding to this token as

$$h_{\text{SEG}} = H[p_{\text{SEG}}] \in \mathbb{R}^d, \tag{2}$$

where H denotes the output hidden states of the MLLM and p_{SEG} is the position of [SEG].

Different from prior multi-[SEG] approaches that associate each [SEG] token with an individual target instance, OneSEG uses h_{SEG} as a *holistic referential representation*. Since all candidate instances are governed by the same referring expression, h_{SEG} summarizes their shared vision-language semantics without explicitly encoding instance cardinality or localization. Accordingly, the MLLM always produces a single [SEG] representation, while instance discovery and target-existence determination are deferred to subsequent spatial decoding. We additionally retain the MLLM-enhanced visual representations F_{llm}, which provide language-conditioned visual semantics for the subsequent visual-semantic fusion and query decomposition stages.

3.3 Semantic-to-Instance Query Decomposition

Although h_{SEG} captures holistic vision-language semantics shared by the referred targets, it does not explicitly encode instance-level distinctions. We therefore introduce **Semantic-to-Instance Query Decomposition** to transform this shared referential representation into a set of instance-aware queries. Specifically, we maintain N_q learnable query embeddings $Q_{\text{init}} \in \mathbb{R}^{N_q \times d}$ and condition each query on the shared referential representation:

$$q_s^i = q_{init}^i + \phi_s(h_{\text{SEG}}), \qquad i = 1, \ldots, N_q, \tag{3}$$

where q_{init}^i denotes the i-th learnable query and $\phi_s(\cdot)$ projects the holistic [SEG] representation into the query embedding space. In this way, all queries inherit the shared referential semantics while preserving their individual embeddings. We further apply self-attention among Q_s to facilitate query interaction and encourage different queries to specialize toward distinct instance hypotheses:

$$\widetilde{Q}_s = \text{SelfAttn}(Q_s). \tag{4}$$

To provide spatial evidence for instance decomposition, we introduce a semantic-guided visual gating mechanism. We first project the visual features F_{vit} and the referential representation h_{SEG} into a shared embedding space, and compute a spatial relevance score for each visual token:

$$A = \sigma\left(\frac{\phi_v(F_{\text{vit}})\phi_g(h_{\text{SEG}})}{\sqrt{d}}\right), \tag{5}$$

where A denotes the semantic relevance weights and $\sigma(\cdot)$ is the sigmoid function. The visual features are then spatially reweighted as

$$F_g = A \odot \phi_v(F_{\text{vit}}), \tag{6}$$

such that regions consistent with the referential semantics are emphasized while irrelevant visual responses are suppressed. The semantically conditioned queries are then grounded into individual target regions through cross-attention over the gated visual features:

$$Q_{\text{ins}} = \text{CrossAttn}\left(\widetilde{Q}_s, F_g, F_g\right), \tag{7}$$

where $Q_{\text{ins}} \in \mathbb{R}^{N_q \times d}$ denotes the resulting instance-aware query set. By attending to referent-relevant visual regions, different queries can capture distinct spatial hypotheses while sharing the same referential semantics. In this way, SIQD decomposes a single holistic [SEG] representation into multiple spatially grounded instances without introducing specific queries into the MLLM.

3.4 Visual-Semantic Fusion

Following prior MLLM-based segmentation methods (Zhang et al., 2024b; Wang et al., 2025a), we exploit the complementary representations from the vision encoder and the MLLM. Specifically, F_{vit} preserves fine-grained spatial details, whereas F_{llm} contains higher-level visual semantics enriched by multimodal reasoning. We therefore integrate them to construct a representation that is both spatially precise and semantically informative.

We first employ cross-attention to inject the semantic information from F_{llm} into visual features:

$$F_{\text{sem}} = \text{CrossAttn}\left(F_{\text{vit}}, F_{\text{llm}}, F_{\text{llm}}\right), \tag{8}$$

we then concatenate it with the original visual features and project them into a unified feature space:

$$F_{\text{fuse}} = \phi_f\left(\text{Concat}\left(F_{\text{sem}}, F_{\text{vit}}\right)\right), \tag{9}$$

where $\phi_f(\cdot)$ denotes a learnable projection and Concat denotes the concatenation operation.

Finally, we allow each instance query to retrieve its visual context from the fused visual representation. Specifically, each instance query serves as the query, while F_{fuse} provides the keys and values:

$$\begin{aligned} q^i_{\text{ins}} &= \text{CrossAttn}\left(q^i_{\text{ins}}, F_{\text{fuse}}, F_{\text{fuse}}\right), \\ F^i_m &= F_{\text{fuse}} + \text{Broadcast}_L\left(q^i_{\text{ins}}\right), \qquad i = 1, \ldots, N_q, \end{aligned} \tag{10}$$

where $q^i_{\text{ins}} \in \mathbb{R}^d$ denotes the query-specific visual context aggregated from all fused visual tokens, and $\text{Broadcast}_L(\cdot)$ replicates it over the L spatial positions. This query-to-visual interaction enables each instance hypothesis to retrieve relevant visual evidence, while the residual fusion preserves the spatial structure of F_{fuse}. Consequently, each F^i_m retains fine-grained spatial information while being conditioned on the instance query, and is subsequently decoded into its candidate mask M_i.

3.5 Instance Mask Decoding and Verification

Instance Mask Decoder. Given the query-specific spatial features F_m, we employ a lightweight instance mask decoder consisting of a sub-pixel coarse prediction head and an RGB-guided residual refinement branch. The coarse head directly projects each query-specific feature into a full-resolution mask through PixelShuffle rearrangement (Shi et al., 2016), while the refinement branch progressively incorporates high-resolution image evidence and coarse-mask priors to recover fine spatial details. The resulting residual correction is added to the coarse logits to obtain the final candidate masks $\{M_i\}_{i=1}^{N_q}$. Detailed architectural configurations are provided in the Appendix.

Instance Evidence Verifier. To determine whether each mask matches a valid referent, we introduce an **Instance Evidence Verifier** that evaluates candidate using mask-conditioned visual evidence.

Existing approaches typically determine target existence either through an early discrete decision, such as the [REJ] token, or through confidence scores derived primarily from instance queries. Such decisions, however, do not explicitly examine whether the region predicted by a candidate mask contains visual-semantic evidence consistent with the referred target. In contrast, IEV performs verification *after* mask decoding, allowing the predicted spatial support itself to provide evidence for candidate validity.

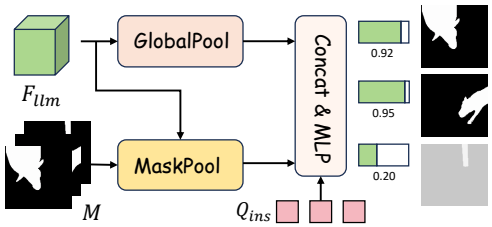

Figure 3: Overview of the **Instance Evidence Verifier**. IEV verifies each decoded mask by jointly leveraging mask-conditioned visual-semantic evidence, global context, and the corresponding instance query.

Specifically, for each predicted mask M_i, we aggregate the corresponding MLLM-enhanced visual features F_{llm} through mask pooling:

$$e_i^{\text{mask}} = \text{MaskPool}\left(F_{\text{llm}}, M_i\right), \tag{11}$$

where e_i^{mask} captures the region-level visual semantics within the predicted mask. We further obtain a global visual-semantic representation

$$e^{\text{global}} = \text{GlobalPool}\left(F_{\text{llm}}\right). \tag{12}$$

The mask-conditioned evidence, global context, and the corresponding instance query are then jointly used to predict the validity of each candidate:

$$p_i^{\text{ver}} = \sigma\left(\text{MLP}\left(\text{Concat}\left(q_{\text{ins}}^i, e_i^{\text{mask}}, e^{\text{global}}\right)\right)\right), \tag{13}$$

where MLP denotes a 3-layer multilayer perceptron. By conditioning verification on the visual-semantic content covered by each decoded mask, IEV evaluates candidate validity using explicit region-level evidence, enabling more reliable suppression of spurious predictions.

3.6 Training Objectives

We employ Hungarian matching with the same Focal-Dice cost used for mask supervision to build one-to-one assignments between masks and ground truth. The matched masks are supervised with

$$\mathcal{L}_{\text{mask}} = \lambda_{\text{focal}}\mathcal{L}_{\text{focal}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}. \tag{14}$$

For instance verification, matched queries are treated as positives and unmatched queries as negatives. We optimize the verification scores with binary cross-entropy:

$$\mathcal{L}_{\text{ver}} = -\frac{1}{N_q}\sum_{i=1}^{N_q}\left[y_i \log p_i^{\text{ver}} + (1-y_i)\log(1-p_i^{\text{ver}})\right], \tag{15}$$

where $y_i \in \{0,1\}$ denotes the existence label. The overall objective is

$$\mathcal{L} = \lambda_{\text{lm}}\mathcal{L}_{\text{lm}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{ver}}\mathcal{L}_{\text{ver}}. \tag{16}$$

4 Experiments

4.1 Experiment Settings

Implementation Details. We adopt Qwen3-VL-4B-Instruct (Bai et al., 2025) as our base MLLM and fine-tune it with LoRA (Hu et al., 2021). The input resolution is set to 448×448. We use AdamW with a learning rate of 4×10^{-5} and a cosine schedule, with a batch size of 32 and gradient accumulation of 2. We set $N_q = 6$ and jointly train OneSEG on RefCOCO/+/g, gRefCOCO, and MUSE for 50K steps. All loss coefficients are set to 1. All experiments are conducted on 4 NVIDIA RTX 4090 GPUs (24 GB), and full training takes approximately 48 hours.

Table 1: **Comparison with prior methods on the gRefCOCO benchmark.** We report generalized IoU (gIoU), cumulative IoU (cIoU), and null-target accuracy (N-acc) on the val, testA, and testB splits. "–" indicates that the corresponding result is not reported. * denotes zero-shot evaluation.

Method	val			testA			testB		
	gIoU	cIoU	N-acc	gIoU	cIoU	N-acc	gIoU	cIoU	N-acc
Specialized Methods									
CRIS (Wang et al., 2022)	56.3	55.3	–	63.4	63.8	–	51.8	51.0	–
LAVT (Yang et al., 2022)	58.4	57.6	49.3	65.9	65.3	49.3	55.8	55.0	48.5
ReLA (Liu et al., 2023)	63.6	62.4	56.4	70.0	69.3	59.0	61.0	59.9	58.4
DeRIS-L (Dai et al., 2025c)	77.7	72.0	82.2	75.3	73.7	78.3	68.0	67.4	66.8
MLLM-based Methods									
LISA-13B (Lai et al., 2024)	65.2	64.0	57.5	70.0	71.0	55.4	62.1	62.3	56.3
GSVA-13B (Xia et al., 2024)	70.0	66.4	66.0	73.3	72.8	64.7	65.5	63.2	62.5
SAM4MLLM-8B (Chen et al., 2024b)	71.9	67.8	66.1	74.2	72.2	63.9	65.3	63.4	60.0
UniRES++-7B (Liu et al., 2025a)	74.4	69.9	–	76.0	74.5	–	69.8	66.6	–
PSALM-1.3B* (Zhang et al., 2024b)	43.3	42.0	–	54.5	52.4	–	52.5	50.6	–
HyperSeg-3B* (Wei et al., 2024)	–	47.5	–	–	57.3	–	–	52.5	–
Text4Seg-13B (Lan et al., 2025)	74.8	69.8	–	75.1	74.3	–	68.0	67.1	–
MLLMSeg-8B (Wang et al., 2025a)	75.1	71.6	73.2	77.0	76.9	72.4	69.7	68.5	65.5
HiMTok-8B (Wang et al., 2025b)	72.1	70.4	–	73.5	74.9	–	71.7	72.0	–
RAS-13B (Cao et al., 2025)	74.6	70.5	69.1	77.5	77.0	64.6	69.4	67.9	62.9
UGround-7B (Qian et al., 2025)	72.5	65.6	74.5	74.3	70.9	73.9	66.9	61.8	71.2
SetCon-8B (Zhang et al., 2026b)	78.4	72.0	–	78.5	78.1	–	73.1	**72.4**	–
SegCompass-13B (Lu et al., 2026)	76.8	72.2	–	77.7	77.3	–	71.9	72.4	–
SAMTok-RL-3B (Zhou et al., 2026)	79.4	73.7	81.5	79.4	77.7	77.7	71.3	69.8	72.0
AnchorSeg-7B (Qian et al., 2026)	74.8	68.7	76.0	75.8	73.4	71.8	68.3	64.6	68.3
ARGenSeg-8B (Wang et al., 2026b)	74.7	72.2	–	73.7	73.6	–	72.4	70.4	–
OneSEG-4B (Ours)	**82.5**	**76.2**	**88.5**	**80.4**	**80.6**	**80.2**	**74.8**	72.0	**74.9**

4.2 Quantitative Results

Results on GRES benchmarks. To validate the efficacy of our method, we compare it against the SOTA approaches on gRefCOCO and MUSE datasets. As shown in Table 1, OneSEG achieves SOTA performance on gRefCOCO across most metrics, outperforming SAMTok-RL-3B, the strongest prior method in average gIoU and N-acc, by 2.5 and 4.1 points across the three splits. Notably, the substantial gains in N-acc indicate that our decoupled design provides a more reliable basis for target-existence estimation by separating shared referential understanding from downstream instance-level prediction. OneSEG-4B consistently outperforms several larger counterparts, further validating the effectiveness of decoupling referential reasoning from instance prediction.

On the MUSE benchmark, as shown in Table 2, OneSEG achieves SOTA performance. Compared with SetCon-8B, the strongest prior method on both splits, OneSEG achieves consistent improvements on both sets. In particular, OneSEG delivers substantial gains of 13.8 gIoU and 8.3 cIoU on the test set, demonstrating strong generalization to challenging generalized referring scenarios.

Results on RES benchmarks. As shown in Table 3, OneSEG demonstrates strong performance on conventional RES benchmarks, achieving the best results on seven of eight splits and the highest average score of 84.4, outperforming the previous best X-SAM by 2.3 points. These results demonstrate that the proposed decoupled paradigm generalizes effectively to standard single-target referring segmentation without compromising segmentation accuracy.

Component Ablation. Table 4 evaluates each component on the gRefCOCO validation split using an LISA-style baseline with the same Qwen3-VL-4B backbone and a lightweight decoder. VSF improves all metrics, while SIQD yields the largest gain, increasing N-acc from 74.6 to 81.9. The mask decoder and IEV further improve performance, resulting in 82.5 gIoU, 76.2 cIoU, and 88.5 N-acc. Overall, OneSEG outperforms the baseline by 9.3, 6.2, and 17.4 points, respectively, without relying on a high-resolution decoder such as SAM2 (Ravi et al., 2025).

Table 2: **Comparison with prior methods on the MUSE benchmark.** We report gIoU and cIoU on the validation and test sets. † denotes results reproduced by subsequent work.

Method	val		test	
	gIoU	cIoU	gIoU	cIoU
LISA-13B (Lai et al., 2024)	43.6	50.2	41.9	50.5
GSVA-13B† (Xia et al., 2024)	41.7	50.3	40.7	48.8
GLaMM-7B† (Rasheed et al., 2024)	41.5	48.0	38.1	44.5
PixelLM-13B (Ren et al., 2024)	44.8	54.1	42.3	51.0
POPEN-13B (Zhu et al., 2025)	48.0	59.1	46.0	56.9
UGround-7B (Qian et al., 2025)	52.4	56.4	40.1	50.4
VisionReasoner-7B† (Liu et al., 2025b)	42.4	41.6	36.1	36.8
StAR-8B + MV (Yun et al., 2026)	58.2	56.7	–	–
SetCon-8B (Zhang et al., 2026b)	56.9	59.5	53.2	60.0
OneSEG-4B (Ours)	**59.7**	**61.5**	**67.0**	**68.3**

Table 3: **Comparison with prior methods on RefCOCO/+/g benchmarks.**

Model	RefCOCO			RefCOCO+			RefCOCOg		Avg.
	val	testA	testB	val	testA	testB	val	test	
LISA-7B (Lai et al., 2024)	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	69.9
PixelLM-7B (Ren et al., 2024)	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5	69.2
GSVA-13B (Xia et al., 2024)	79.2	81.7	77.1	70.3	73.8	63.6	75.7	77.0	74.8
GLaMM-7B (Rasheed et al., 2024)	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	75.6
SAM4MLLM-8B (Chen et al., 2024b)	79.8	82.7	74.7	74.6	80.0	67.2	75.5	76.4	76.4
PSALM-1.3B (Zhang et al., 2024b)	83.6	84.7	81.6	72.9	75.5	70.1	73.8	74.4	77.1
Text4Seg-13B (Lan et al., 2025)	79.2	81.7	75.6	72.8	77.9	66.5	74.0	75.3	75.4
MLLMSeg-8B (Wang et al., 2025a)	81.0	82.4	78.7	76.4	79.1	72.5	79.9	80.8	78.9
HyperSeg-3B (Wei et al., 2024)	84.8	85.7	83.4	79.0	83.5	75.2	79.4	78.9	81.2
VisionReasoner-7B (Liu et al., 2025b)	76.8	78.9	72.5	70.9	74.9	64.6	72.9	71.3	72.9
RICE-7B (Xie et al., 2025)	83.5	85.3	81.7	79.4	82.8	75.4	79.8	80.4	81.0
X-SAM-3.8B (Wang et al., 2026a)	85.1	**87.1**	83.4	78.0	81.0	74.4	83.8	83.9	82.1
InstructSeg-3B (Wei et al., 2025)	85.8	86.6	84.0	80.1	83.8	75.6	79.3	80.3	81.9
SELF1E-8B (Zhang et al., 2026a)	82.5	83.8	80.1	77.6	81.6	73.7	79.1	79.8	79.8
SetCon-8B (Zhang et al., 2026b)	83.7	84.4	81.6	79.4	82.3	75.6	80.9	80.0	81.0
OneSEG-4B (Ours)	**85.9**	86.6	**84.4**	**82.2**	**84.7**	**79.6**	**85.6**	**86.0**	**84.4**

Ablation on SIQD. Table 5 shows that directly decoding the [SEG] representation into a global mask provides a strong baseline, indicating that it captures effective holistic referential semantics. Conditioning learnable queries on [SEG] improves gIoU and N-acc, while self-attention and semantic gating further specialize and ground the queries, yielding the best overall performance.

Ablation on IEV. We first use the instance query alone with an MLP for existence prediction, which serves as the query-level baseline. Introducing mask-pooled visual evidence improves verification performance, and using F_{llm} consistently outperforms F_{vit}, indicating that instance existence is more effectively determined from semantically enriched features than from fine-grained visual representations alone. Finally, incorporating global pooled context further improves all metrics.

Effect of Query Number. We further investigate the effect of the number of instance queries N_q. As shown in Fig. 4, both gIoU and N-acc improve substantially as N_q increases, indicating that sufficient query capacity is important for capturing multiple instance hypotheses. The performance saturates around $N_q = 6$, while further increasing the query number brings limited gains or slight degradation, suggesting that excessive queries may introduce redundant instance hypotheses.

4.3 Qualitative Results

As shown in Fig. 5, OneSEG exhibits effective query specialization across different target cardinalities. For multi-target cases, distinct queries capture different referred instances while invalid

Table 4: **Component ablation on the gRefCOCO validation split.**

Setting	gIoU	cIoU	N-acc
Baseline	73.2	70.0	71.1
+ VSF	74.2	71.9	74.6
+ SIQD	79.4	74.7	81.9
+ Mask Decoder	80.3	75.6	83.5
+ IEV	**82.5**	**76.2**	**88.5**

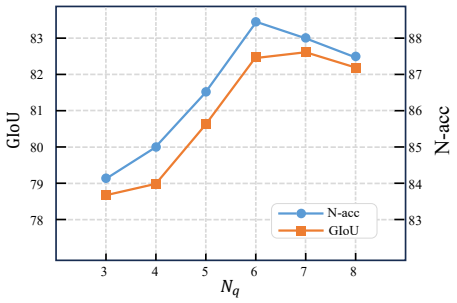

Figure 4: **Effect of Query Number N_q.**

Table 5: **Ablation of the SIQD Module.**

Setting	gIoU	cIoU	N-acc
Direct [SEG]	78.3	73.7	80.4
Query + [SEG]	80.1	74.2	83.6
+ Self-Attention	81.1	74.8	85.5
+ Semantic Gating	**82.5**	**76.2**	**88.5**

Table 6: **Ablation of the IEV Module.**

Setting	gIoU	cIoU	N-acc
Query-only MLP	80.7	74.5	83.0
+ Mask Pooling (F_{vit})	81.1	74.2	84.0
+ Mask Pooling (F_{llm})	81.8	75.4	86.7
+ Global Pooling	**82.5**	**76.2**	**88.5**

candidates are suppressed by IEV. In the no-target example, IEV successfully rejects a plausible but incorrect mask prediction. When the number of targets increases, the query slots maintain clear instance-wise specialization. For more visualization, please refer to Appendix.

5 Conclusion

In this work, we present **OneSEG**, a decoupled framework for generalized referring segmentation that separates referential reasoning from instance prediction. Instead of explicitly enumerating instances within the MLLM, OneSEG uses a single [SEG] representation to capture shared referential semantics and delegates instance discovery to downstream spatial decoding. SIQD further decomposes the holistic semantics into visually grounded instance queries, while IEV verifies decoded candidates with explicit visual-semantic evidence. Experiments on GRES and conventional RES benchmarks demonstrate that OneSEG achieves state-of-the-art performance, showing that instance enumeration can be deferred to spatial decoding rather than performed within the MLLM.

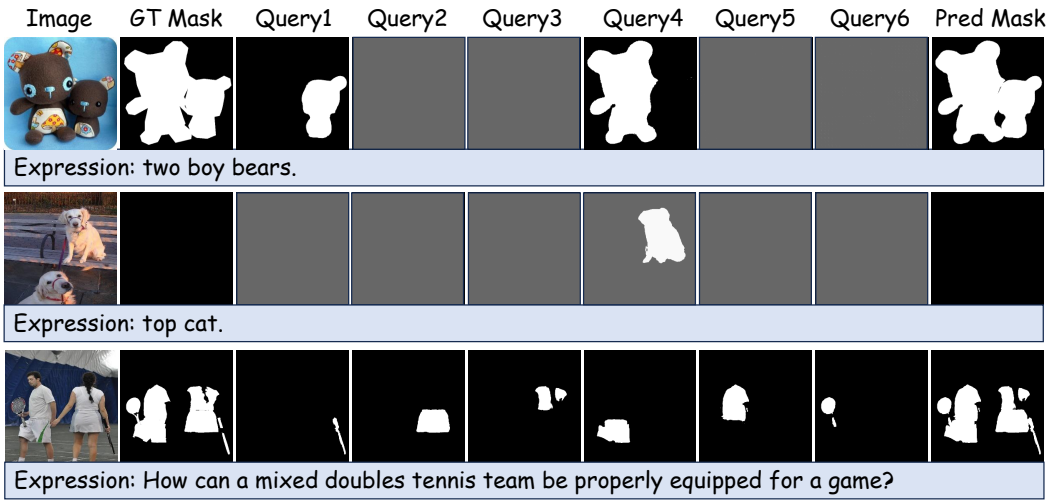

Figure 5: **Qualitative visualization of instance query decomposition.** We show the input image, ground-truth mask, per-query predictions, and final mask. Different queries capture distinct target instances, while redundant queries are suppressed.

AI USE STATEMENT

In this work, we used generative AI tools solely to aid and polish the writing of the manuscript, including improving grammar, clarity, and phrasing. We did not use generative AI tools to generate research ideas, formulate hypotheses, design experiments, develop or debug code, analyze data, produce experimental results, or draw scientific conclusions. Other disclosure categories are not applicable to this work. All AI-assisted text was carefully reviewed and edited by the authors to ensure its accuracy, consistency with the underlying research, and faithful representation of our intended claims. The authors take full responsibility for the final content of this work, including all text and claims produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or the collection of sensitive data. We are not aware of any specific ethical concerns arising from the methodology or experimental setup presented in this work. All datasets and resources used in our experiments are employed in accordance with their respective licenses and intended research purposes.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of our results. The main paper describes the proposed method and experimental protocol, while the appendix provides additional implementation details, including training configurations, hyperparameters, dataset processing, and evaluation settings. We also report the necessary experimental details and evaluation procedures to enable independent reproduction of the results presented in this work.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Shengcao Cao, Zijun Wei, Jason Kuen, Kangning Liu, Lingzhi Zhang, Jiuxiang Gu, HyunJoon Jung, Liang-Yan Gui, and Yu-Xiong Wang. Refer to any segmentation mask group with vision-language prompts. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21853–21863. IEEE, 2025.

Wei Chen, Long Chen, and Yu Wu. An efficient and effective transformer decoder-based framework for multi-task visual grounding. In *European Conference on Computer Vision*, pp. 125–141. Springer, 2024a.

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. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024b.

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*, 2024c.

Ming Dai, Lingfeng Yang, Yihao Xu, Zhenhua Feng, and Wankou Yang. Simvg: A simple framework for visual grounding with decoupled multi-modal fusion. *Advances in neural information processing systems*, 37:121670–121698, 2024.

Ming Dai, Kai Chen, Wenxuan Cheng, Jiedong Zhuang, Zhenhua Feng, Pengfei Zhu, and Wankou Yang. Gc 3 vg: Generalized multi-task visual grounding with coarse-to-fine consistency constraints. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025a.

Ming Dai, Wenxuan Cheng, Jiang-Jiang Liu, Lingfeng Yang, Zhenhua Feng, Wankou Yang, and Jingdong Wang. Improving generalized visual grounding with instance-aware joint learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Ming Dai, Wenxuan Cheng, Jiang-jiang Liu, Sen Yang, Wenxiao Cai, Yanpeng Sun, and Wankou Yang. Deris: Decoupling perception and cognition for enhanced referring image segmentation through loopback synergy. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19936–19946. IEEE, 2025c.

Ming Dai, Wenxuan Cheng, Jiedong Zhuang, Jiang-jiang Liu, Hongshen Zhao, Zhenhua Feng, and Wankou Yang. Propvg: End-to-end proposal-driven visual grounding with multi-granularity discrimination. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7058–7068. IEEE, 2025d.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1749–1759. IEEE, 2021.

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 preprint arXiv:2106.09685*, 2021.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European conference on computer vision*, pp. 108–124. Springer, 2016.

Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4044–4054. IEEE, 2023a.

Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4044–4054. IEEE, 2023b.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589. IEEE, 2024.

Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wei Zhang. Text4seg: Reimagining image segmentation as text generation. In *International Conference on Learning Representations*, volume 2025, pp. 1634–1661, 2025.

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

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23592–23601. IEEE, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

Jing Liu, Wenxuan Wang, Yisi Zhang, Yepeng Tang, Xingjian He, Longteng Guo, Tongtian Yue, and Xinlong Wang. Towards unified referring expression segmentation across omni-level visual target granularities. *arXiv preprint arXiv:2504.01954*, 2025a.

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 e-prints*, pp. arXiv–2505, 2025b.

Zhenyu Lu, Liupeng Li, Jinpeng Wang, Haoqian Kang, Yan Feng, Ke Chen, and Yaowei Wang. Segcompass: Exploring interpretable alignment with sparse autoencoders for enhanced reasoning segmentation. *arXiv preprint arXiv:2605.22658*, 2026.

Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10031–10040. IEEE, 2020.

Zhuoyan Luo, Yinghao Wu, Tianheng Cheng, Yong Liu, Yicheng Xiao, Hongfa Wang, Xiao-Ping Zhang, and Yujiu Yang. Cohd: A counting-aware hierarchical decoding framework for generalized referring expression segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22685–22694. IEEE, 2025.

E-Ro Nguyen, Hieu Le, Dimitris Samaras, and Michael S Ryoo. Phrase-instance alignment for generalized referring segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8486–8495, 2026.

Rui Qian, Xin Yin, Chuanhang Deng, Zhiyuan Peng, Jian Xiong, Wei Zhai, and Dejing Dou. Uground: Towards unified visual grounding with unrolled transformers. *arXiv preprint arXiv:2510.03853*, 2025.

Rui Qian, Chuanhang Deng, Qiang Huang, Jian Xiong, Mingxuan Li, Yingbo Zhou, Wei Zhai, Jintao Chen, and Dejing Dou. Anchorseg: Language grounded query banks for reasoning segmentation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20490–20505, 2026.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13009–13018. IEEE, 2024.

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. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26364–26373. IEEE, 2024.

Nisarg A Shah, Vibashan VS, and Vishal M Patel. Lqmformer: Language-aware query mask transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12903–12913, 2024.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Hao Wang, Limeng Qiao, Zequn Jie, Zhijian Huang, Chengjian Feng, Qingfang Zheng, Lin Ma, Xiangyuan Lan, and Xiaodan Liang. X-sam: From segment anything to any segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26187–26196, 2026a.

Jingchao Wang, Zhijian Wu, Dingjiang Huang, Yefeng Zheng, and Hong Wang. Unlocking the potential of mllms in referring expression segmentation via a light-weight mask decoder. *arXiv preprint arXiv:2508.04107*, 2025a.

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*, 2024.

Tao Wang, Changxu Cheng, Lingfeng Wang, Senda Chen, and Wuyue Zhao. Himtok: Learning hierarchical mask tokens for image segmentation with large multimodal model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23267–23278. IEEE, 2025b.

Xiaolong Wang, Lixiang Ru, Ziyuan Huang, Kaixiang Ji, Dandan Zheng, Jingdong Chen, and Jun Zhou. Argenseg: Image segmentation with autoregressive image generation model. *Advances in Neural Information Processing Systems*, 38:28312–28336, 2026b.

XuDong Wang, Shaolun Zhang, Shufan Li, Kehan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, et al. Segllm: Multi-round reasoning segmentation with large language models. In *International Conference on Learning Representations*, volume 2025, pp. 56526–56547, 2025c.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11676–11685. IEEE, 2022.

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*, 2024.

Cong Wei, Yujie Zhong, Haoxian Tan, Yingsen Zeng, Yong Liu, Hongfa Wang, and Yujiu Yang. Instructseg: Unifying instructed visual segmentation with multi-modal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20193–20203. IEEE, 2025.

Changsong Wen, Zelin Peng, Yu Huang, and Wei Shen. Efficient segmentation with multimodal large language model via token routing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10593–10602, 2026.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3858–3869. IEEE, 2024.

Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. Oneref: Unified one-tower expression grounding and segmentation with mask referring modeling. *Advances in Neural Information Processing Systems*, 37:139854–139885, 2024.

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

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.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18134–18144. IEEE, 2022.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pp. 69–85. Springer, 2016.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1307–1315. IEEE, 2018.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, et al. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025.

Yuqian Yuan, Wentong Li, Zhaocheng Li, Yutong Lin, Juncheng Li, Siliang Tang, Jun Xiao, Yueting Zhuang, and Wenqiao Zhang. Instructsam: Segment any instance with any instructions. *arXiv preprint arXiv:2605.26102*, 2026.

Seokju Yun, Dongheon Lee, Noori Bae, Jaesung Jun, Chanseul Cho, and Youngmin Ro. Star: Segment anything reasoner. *arXiv preprint arXiv:2603.14382*, 2026.

Anqi Zhang, Xiaokang Ji, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, and Yunchao Wei. Rethinking mllm itself as a segmenter with a single segmentation token. *arXiv preprint arXiv:2603.19026*, 2026a.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in neural information processing systems*, 37:71737–71767, 2024a.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024b.

Zhixiong Zhang, Yizhuo Li, Shuangrui Ding, Yuhang Zang, Shengyuan Ding, Long Xing, Yibin Wang, Qiaosheng Zhang, and Jiaqi Wang. Setcon: Towards open-ended referring segmentation via set-level concept prediction. *arXiv preprint arXiv:2605.20110*, 2026b.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37852–37863, 2026.

Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pp. 598–615. Springer, 2022.

Lanyun Zhu, Tianrun Chen, Qianxiong Xu, Xuanyi Liu, Deyi Ji, Haiyang Wu, De Wen Soh, and Jun Liu. Popen: Preference-based optimization and ensemble for lvlm-based reasoning segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30231–30240. IEEE, 2025.

A APPENDIX

A.1 DATASETS AND ADDITIONAL EXPERIMENTAL DETAILS

RefCOCO/+/g. RefCOCO, RefCOCO+, and RefCOCOg (Yu et al., 2016) are widely used referring expression segmentation benchmarks built upon MS COCO. RefCOCO and RefCOCO+ contain 142,210 and 141,564 referring expressions over approximately 20K images, respectively, where RefCOCO+ excludes absolute location descriptions and thus places greater emphasis on appearance-based reasoning. RefCOCOg contains 95,010 expressions over 25,799 images and features longer and more complex referring expressions. Following standard protocols, we evaluate on the val/testA/testB splits of RefCOCO and RefCOCO+, and the val/test split of RefCOCOg.

gRefCOCO. gRefCOCO (Liu et al., 2023) extends the conventional RefCOCO benchmark to generalized referring segmentation by incorporating single-target, multi-target, and no-target expressions. It contains 278,232 referring expressions over 19,994 images and 60,287 distinct instances, including 80,022 multi-target and 32,202 no-target expressions. Pixel-level masks and bounding boxes are provided for all target instances. The dataset follows the UNC partition of RefCOCO and is divided into train, validation, testA, and testB splits. Following prior work, we report gIoU, cIoU, and N-acc for evaluation.

MUSE. MUSE (Ren et al., 2024) is a multi-target reasoning segmentation benchmark constructed from LVIS with GPT-4V-assisted data generation. It contains 246K question-answer pairs covering approximately 0.9M annotated instances, with an average of 3.7 targets per answer. Compared with conventional referring segmentation datasets, MUSE emphasizes open-set concepts, complex reasoning queries, and instance-level multi-target mask annotations. The dataset is split into 239K training, 2.8K validation, and 4.3K test samples.

To further evaluate the generalization ability of OneSEG under diverse target cardinalities, we additionally evaluate our model on Ref-ZOM.

Ref-ZOM. Ref-ZOM (Hu et al., 2023b) is a generalized referring segmentation benchmark constructed from COCO, covering one-to-zero, one-to-one, and one-to-many referring scenarios. It contains 55,078 images with 74,942 annotated objects and provides diverse referring expressions beyond the conventional one-to-one setting. Following the standard protocol, we report mIoU and oIoU for target-present samples and accuracy for one-to-zero cases.

Additional Implementation Details. We apply LoRA (Hu et al., 2021) to the LLM and LM head with rank $r = 64$ and scaling factor $\alpha = 128$, while keeping the vision encoder frozen. The VSF, SIQD, mask decoder, and IEV are fully optimized during training. We use DeepSpeed ZeRO-1 with a per-GPU batch size of 4 and enable gradient checkpointing for memory efficiency. Training on each 24 GB RTX 4090 reaches a peak memory usage of approximately 22 GB. Unless otherwise specified, the IEV threshold is set to 0.8 during inference.

A.2 ADDITIONAL RESULTS ON REF-ZOM

Results on Ref-ZOM. We further evaluate OneSEG on Ref-ZOM under both zero-shot and fine-tuning settings. As shown in Table 7, without training on Ref-ZOM, OneSEG already achieves 72.4 mIoU, outperforming previous methods in segmentation quality. After fine-tuning, OneSEG achieves 76.9 mIoU, 75.5 oIoU, and 99.6 Acc., surpassing the previous best PropVG (Dai et al., 2025d) by 5.7, 3.5, and 1.5 points, respectively. These results further demonstrate the strong generalization and adaptation capability of OneSEG across generalized referring scenarios.

A.3 MASK DECODER ARCHITECTURE

Given the query-specific spatial features $F_m \in \mathbb{R}^{N_q \times L \times C}$, where N_q is the number of instance queries, C is the feature dimension, and $L = H_d W_d$ denotes the number of spatial tokens, we employ a lightweight coarse-to-fine mask decoder for instance mask prediction. The decoder consists of a sub-pixel coarse prediction head and an RGB-guided residual refinement branch. The former directly decodes each query-specific representation into a full-resolution coarse mask, while the latter progressively refines its spatial details using image evidence and coarse-mask cues.

Table 7: **Comparison with prior methods on the Ref-ZOM test set.** "zs" and "ft" denote zero-shot evaluation and fine-tuning on Ref-ZOM, respectively.

Method	mIoU	oIoU	Acc.
DMMI (Hu et al., 2023b)	68.2	68.8	87.0
CoHD (Luo et al., 2025)	69.8	69.0	93.3
InstAlign (Nguyen et al., 2026)	70.8	71.1	94.2
PropVG (Dai et al., 2025d)	71.2	<u>72.0</u>	<u>98.1</u>
LISA-7B (Lai et al., 2024)	61.5	60.1	72.6
GSVA-7B (Xia et al., 2024)	68.0	67.1	82.7
LISA-7B (ft) (Lai et al., 2024)	65.4	66.4	93.4
GSVA-7B (ft) (Xia et al., 2024)	68.1	68.3	94.6
OneSEG-4B (zs)	<u>72.4</u>	69.9	84.0
OneSEG-4B (ft)	**76.9**	**75.5**	**99.6**

Coarse Mask Prediction. For the i-th query, we first reshape its spatial representation as

$$D_i = \mathcal{R}(F_m^i) \in \mathbb{R}^{C \times H_d \times W_d}, \tag{17}$$

where $\mathcal{R}(\cdot)$ denotes the operation that restores the flattened spatial tokens to a two-dimensional feature map. We then apply a three-layer convolutional network $\mathrm{Conv}_c(\cdot)$, followed by sub-pixel rearrangement:

$$Z_i^c = \mathrm{PixelShuffle}_r\left(\mathrm{Conv}_c(D_i)\right), \tag{18}$$

where r denotes the spatial upsampling factor determined by the ratio between the decoder feature resolution and the input image resolution. The resulting coarse mask logits satisfy

$$Z_i^c \in \mathbb{R}^{1 \times H_I \times W_I}, \tag{19}$$

where $H_I \times W_I$ denotes the input image resolution.

RGB-Guided Residual Refinement. Although the coarse head provides direct full-resolution predictions, fine object boundaries can still be difficult to recover solely from low-resolution semantic features. We therefore introduce a residual refinement branch that revisits higher-resolution image evidence while preserving the coarse prediction.

We first project the query-specific feature using a convolutional transformation $\phi_{\mathrm{dec}}(\cdot)$ and resize it to an intermediate spatial resolution $H_1 \times W_1$:

$$S_i^{(1)} = \mathrm{Resize}_{H_1, W_1}\left(\phi_{\mathrm{dec}}(D_i)\right). \tag{20}$$

Here, $S_i^{(1)}$ retains the query-conditioned semantic information used for subsequent refinement.

In parallel, we derive explicit refinement cues from the coarse prediction. Let

$$P_i = \sigma(Z_i^c) \tag{21}$$

denote the foreground probability map, where $\sigma(\cdot)$ is the sigmoid function. We further construct an uncertainty map

$$U_i = 1 - |2P_i - 1|, \tag{22}$$

which emphasizes ambiguous regions, and a boundary response

$$B_i = \mathrm{Dilate}(P_i) - \mathrm{Erode}(P_i), \tag{23}$$

where $\mathrm{Dilate}(\cdot)$ and $\mathrm{Erode}(\cdot)$ denote morphological dilation and erosion operations, respectively, and B_i highlights the boundary regions of the coarse prediction. The three coarse-mask cues are concatenated and encoded by a convolutional projection $\phi_p(\cdot)$:

$$E_i^{(1)} = \phi_p\left(\mathrm{Resize}_{H_1, W_1}\left(\mathrm{Concat}[P_i, U_i, B_i]\right)\right), \tag{24}$$

where $\mathrm{Concat}[\cdot]$ denotes channel-wise concatenation.

We additionally extract image-detail features directly from the input image I at the same spatial scale:

$$R^{(1)} = \phi_r^{(1)}\left(\mathrm{Resize}_{H_1,W_1}(I)\right), \tag{25}$$

where $\phi_r^{(1)}(\cdot)$ denotes a lightweight convolutional encoder. The semantic feature, coarse-mask evidence, and RGB feature are then jointly fused:

$$D_i^{(1)} = \phi_f^{(1)}\left(\mathrm{Concat}\left[S_i^{(1)}, E_i^{(1)}, R^{(1)}\right]\right), \tag{26}$$

where $\phi_f^{(1)}(\cdot)$ denotes the convolutional fusion layer.

We further refine the spatial representation at a higher resolution $H_2 \times W_2$, where $H_2 = 2H_1$ and $W_2 = 2W_1$. Specifically,

$$\widetilde{D}_i^{(2)} = \phi_u\left(\mathrm{Resize}_{H_2,W_2}(D_i^{(1)})\right), \tag{27}$$

where $\phi_u(\cdot)$ denotes a convolutional projection applied after upsampling. At this stage, finer RGB features are extracted as

$$R^{(2)} = \phi_r^{(2)}\left(\mathrm{Resize}_{H_2,W_2}(I)\right), \tag{28}$$

and fused with the upsampled representation:

$$D_i^{(2)} = \phi_f^{(2)}\left(\mathrm{Concat}\left[\widetilde{D}_i^{(2)}, R^{(2)}\right]\right). \tag{29}$$

Finally, the refinement branch predicts a residual correction

$$\Delta Z_i = \mathrm{Resize}_{H_I,W_I}\left(\phi_o(D_i^{(2)})\right), \tag{30}$$

where $\phi_o(\cdot)$ denotes the output projection layer. The refined mask logits are obtained by residual prediction:

$$Z_i = Z_i^c + \alpha \Delta Z_i, \tag{31}$$

where α is a learnable residual scaling factor. The final instance mask is then given by

$$M_i = \sigma(Z_i). \tag{32}$$

The output layer of the refinement branch is zero-initialized such that the initial prediction is identical to the coarse mask. During training, the branch progressively learns residual corrections from high-resolution RGB evidence and explicit coarse-mask cues, improving spatial details while preserving the query-conditioned semantics of the original prediction.

A.4 Details of Attention Modules

For clarity, we specify the implementations of the attention modules used throughout OneSEG. Unless otherwise stated, both self-attention and cross-attention follow standard Transformer (Vaswani et al., 2017) architectures with multi-head attention, residual connections, normalization, and feed-forward transformation.

Self-Attention Layer. Given an input sequence $X \in \mathbb{R}^{N\times d}$, we first project it into query, key, and value representations:

$$Q = XW_Q, \qquad K = XW_K, \qquad V = XW_V, \tag{33}$$

where W_Q, W_K, and W_V are learnable projection matrices. Multi-head self-attention is then applied as

$$\widetilde{X} = X + \mathrm{Proj}\left(\mathrm{MHA}(Q,K,V)\right), \tag{34}$$

followed by a feed-forward network and normalization:

$$X' = \mathrm{Norm}\left(\widetilde{X} + \mathrm{FFN}(\widetilde{X})\right). \tag{35}$$

We use this standard self-attention layer for the $\mathrm{SelfAttn}(\cdot)$ operations described in the main paper.

Cross-Attention Layer. Given a query sequence $X_q \in \mathbb{R}^{N_q \times d}$ and a context sequence $X_k \in \mathbb{R}^{N_k \times d}$, we first project them into the corresponding query, key, and value representations:

$$Q = X_q W_Q, \qquad K = X_k W_K, \qquad V = X_k W_V. \tag{36}$$

The query sequence is then updated by attending to the context sequence:

$$\widetilde{X}_q = X_q + \mathrm{Proj}\left(\mathrm{MHA}(Q, K, V)\right), \tag{37}$$

followed by

$$X'_q = \mathrm{Norm}\left(\widetilde{X}_q + \mathrm{FFN}(\widetilde{X}_q)\right). \tag{38}$$

This formulation is used for all $\mathrm{CrossAttn}(\cdot)$ operations in OneSEG, where the first input provides the query sequence and the remaining inputs provide the key and value sequences.

A.5 Additional Ablation

Controlled Comparison of Segmentation Paradigms. To directly examine the motivation behind OneSEG, we construct controlled baselines that differ primarily in how instance representations are formed, while keeping the MLLM backbone and mask prediction components otherwise comparable. All variants adopt Qwen3-VL-4B as the MLLM backbone.

The first variant is a LISA-style single-[SEG] baseline, as illustrated in Fig. 6. The MLLM generates a single [SEG] representation that is directly decoded into one global mask. To isolate the effect of the segmentation paradigm, this baseline employs only the lightweight coarse prediction head, consisting of a convolutional projection followed by PixelShuffle, without SIQD or the residual refinement branch.

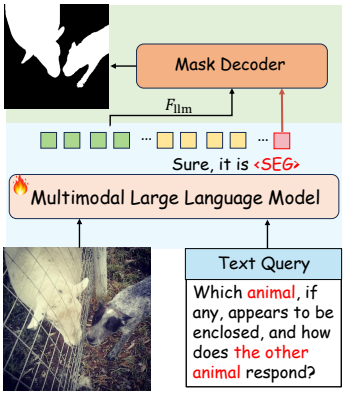

Figure 6: **Architecture of the single-[SEG] baseline.** The baseline follows a LISA-style design with Qwen3-VL-4B as the MLLM backbone. A single [SEG] representation is directly decoded into a global mask using the lightweight coarse mask prediction head.

The second variant follows a GSVA-style multi-[SEG] paradigm. Instead of generating a single holistic [SEG] representation, the MLLM autoregressively produces multiple [SEG] tokens, each serving as an instance-level representation for mask prediction. Apart from the supervision and generation of multiple [SEG] tokens, the backbone and mask prediction components are kept unchanged. This controlled setting allows us to directly examine whether explicit instance enumeration within the MLLM is necessary for effective multi-instance segmentation.

Table 8: **Controlled comparison of different segmentation paradigms on the gRefCOCO validation split.** All variants use Qwen3-VL-4B as the MLLM backbone.

Method	Instance Representation	gIoU	cIoU	N-acc
Single-[SEG] Baseline	One [SEG] for global mask	73.2	70.0	71.1
Multi-[SEG]	Autoregressive multiple [SEG]	76.7	72.4	78.6
OneSEG (Ours)	One [SEG] + downstream decomposition	**82.5**	**76.2**	**88.5**

As shown in Table 8, the single-[SEG] baseline achieves 73.2 gIoU, 70.0 cIoU, and 71.1 N-acc. Introducing autoregressive multi-[SEG] generation improves the performance to 76.7 gIoU, 72.4 cIoU, and 78.6 N-acc, confirming the importance of instance-level modeling for generalized referring segmentation. In contrast, OneSEG achieves 82.5 gIoU, 76.2 cIoU, and 88.5 N-acc by retaining a single holistic [SEG] representation during multimodal reasoning and deferring instance decomposition to the downstream decoding stage. These results demonstrate that effective multi-instance prediction does not require explicit instance enumeration within the MLLM, supporting our decoupled formulation of referential reasoning and instance prediction.

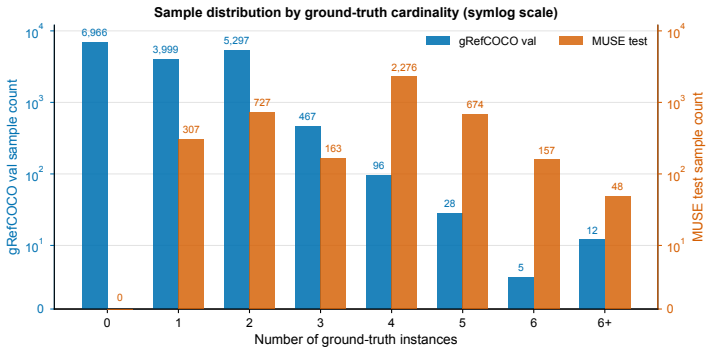

Figure 7: **Distribution of samples across ground-truth target cardinalities.** gRefCOCO val contains relatively few high-cardinality samples, whereas MUSE test provides substantially more examples with multiple target instances.

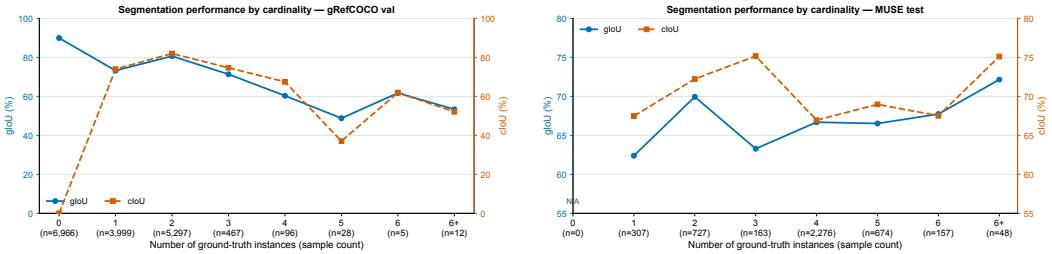

Figure 8: **Segmentation performance across different target cardinalities.** We report gIoU and cIoU on gRefCOCO val (left) and MUSE test (right). OneSEG maintains robust segmentation performance across a wide range of target cardinalities, particularly on MUSE where high-cardinality samples are more sufficiently represented.

A.6 Performance across Target Cardinalities

We further analyze the robustness of OneSEG under different numbers of ground-truth instances. Figure 7 shows that the two benchmarks exhibit substantially different cardinality distributions. In gRefCOCO val, samples with four or more targets are relatively scarce, whereas MUSE test contains substantially more high-cardinality examples.

As shown in Fig. 8, OneSEG maintains strong performance across different target cardinalities. On gRefCOCO, performance becomes more variable for high-cardinality groups, where only a limited number of samples are available. In contrast, on MUSE, where multi-target cases are considerably denser, OneSEG remains stable as the number of targets increases, achieving competitive gIoU and cIoU even for six or more instances. These results indicate that the proposed query decomposition generalizes effectively across varying target cardinalities rather than being restricted to low-cardinality cases.

A.7 Computation Efficiency

The majority of computation in MLLM-based segmentation originates from the MLLM backbone, while the remaining segmentation components contribute comparatively limited overhead. Moreover, directly comparing total FLOPs can obscure the computational difference introduced specifically by different instance-enumeration paradigms, since the MLLM backbone, VSF, mask decoder, and IEV are shared across the compared variants. We therefore focus on the additional computation required to expand a single semantic representation into multiple instance queries.

To quantify this property, we introduce **Query Expansion Efficiency (QEE)**, defined as

$$\mathrm{QEE}(K) = \frac{K-1}{C_{\mathrm{expand}}(K)}, \tag{39}$$

Table 9: **Query expansion efficiency under $K = 6$.** Expansion cost measures the additional computation required to increase the number of instance representations from one to six. Higher QEE indicates better efficiency.

Paradigm	Expansion Cost (GFLOPs) ↓	QEE ↑
Multi-[SEG]	83.2384	0.0601
Mask Token	83.3518	0.0600
OneSEG (Ours)	**0.0956**	**52.3013**

where K denotes the number of instance representations, $K-1$ is the number of additional queries relative to the single-query setting, and $C_{\text{expand}}(K)$ denotes the additional computation required for query expansion, measured in GFLOPs. Accordingly, QEE is measured in additional queries per GFLOP, with a larger value indicating more efficient query expansion.

Let L denote the MLLM sequence length under the single-query setting, and let $F_{\text{MLLM}}(\cdot)$ denote the corresponding MLLM computation. For the autoregressive Multi-[SEG] paradigm, expanding from one to K instance representations introduces $K-1$ additional [SEG] tokens, giving

$$C_{\text{MultiSEG}}(K) = F_{\text{MLLM}}(L+K-1) - F_{\text{MLLM}}(L). \tag{40}$$

For the mask-token paradigm, the additional queries are propagated through the MLLM and further interact with other representations. Its expansion cost is therefore defined as

$$C_{\text{MaskToken}}(K) = F_{\text{MLLM}}(L+K-1) - F_{\text{MLLM}}(L) + F_{\text{QueryInt}}(K), \tag{41}$$

where $F_{\text{QueryInt}}(K)$ denotes the additional computation introduced by query interaction.

In contrast, OneSEG maintains a single [SEG] representation throughout multimodal reasoning and performs query expansion only in the lightweight downstream query generator. Its expansion cost is thus

$$C_{\text{OneSEG}}(K) = F_{\text{QueryGen}}(K), \tag{42}$$

where $F_{\text{QueryGen}}(K)$ denotes the computation of the query generation module. Notably, we conservatively use the full query-generator FLOPs rather than subtracting its $K = 1$ computation, thereby slightly overestimating the expansion cost of OneSEG.

Table 9 reports the results under the default setting $K = 6$. OneSEG requires only 0.0956 GFLOPs for query expansion, compared with 83.2384 GFLOPs for Multi-[SEG] and 83.3518 GFLOPs for the mask-token paradigm. Consequently, OneSEG achieves a QEE of 52.3183 additional queries/GFLOP, substantially higher than the approximately 0.06 achieved by the two in-MLLM enumeration paradigms. These results demonstrate that deferring instance enumeration outside the MLLM enables substantially more efficient scaling with the number of instance queries.

A.8 Additional Qualitative Results

Figure 9 presents additional qualitative examples covering both attribute-based multi-object references and reasoning-intensive instructions involving spatial relations, object functions, and intended use. Across varying target cardinalities, different query slots specialize in complementary target instances, while unused slots remain inactive. The final predictions largely recover the ground-truth regions, demonstrating that SIQD adapts a fixed query set to diverse referents and that IEV effectively suppresses redundant candidates.

Limitations

Although OneSEG achieves strong overall segmentation performance, its lightweight instance mask decoder does not benefit from large-scale mask pre-training and may therefore be less effective than SAM2 (Ravi et al., 2025) at delineating fine boundaries and small objects. Since this work focuses on decoupling referential reasoning from instance prediction rather than maximizing mask fidelity, we leave mask decoder scaling to future work. The modular design of OneSEG allows a pre-trained decoder such as SAM2 to be readily incorporated for higher-quality mask prediction.



Figure 9: **Additional qualitative results.** Each row shows the input image, ground-truth mask, six per-query predictions, and final prediction. Active queries specialize in complementary target instances, while invalid or redundant queries are suppressed.