# SetCon: Towards Open-Ended Referring Segmentation via Set-Level Concept Prediction

**Zhixiong Zhang**[1,2,*] **Yizhuo Li**[1,2,*] **Shuangrui Ding**[3,‡] **Yuhang Zang**[4,†]
**Shengyuan Ding**[5,2] **Long Xing**[6] **Yibin Wang**[5,2] **Qiaosheng Zhang**[2,4] **Jiaqi Wang**[2,7,†]
[1]Shanghai Jiao Tong University [2]Shanghai Innovation Institute
[3]The Chinese University of Hong Kong [4]Shanghai Artificial Intelligence Laboratory
[5]Fudan University [6]University of Science and Technology of China [7]JD.com

## Abstract

Referring segmentation grounds natural-language queries to pixel-level masks, but extending it to complex scenarios with multiple instances, cross-category groups, or open-ended target sets remains challenging. Previous Large Vision Language Model (LVLM)-based methods represent referred targets with one or more special tokens sequentially, treating multiple targets as separate outputs rather than a coherent set and offering little incentive to capture set-level properties such as completeness and mutual exclusivity. We *reformulate* open-ended referring segmentation as explicit set-level concept prediction and propose **Set-Con**cept Segmentation (**SETCON**), which uses LVLM-generated natural-language concepts, instead of segmentation-specific tokens, as semantic conditions for joint mask-set decoding. A hierarchical semantic decomposition first predicts a shared set-level concept defining the target scope and then refines it into fine-grained concept groups aligned with target subsets. To support this, a two-stage annotation pipeline augments existing reasoning segmentation datasets with hierarchical semantic supervision (236k samples, 784k concept phrases). SETCON achieves state-of-the-art results on image benchmarks (**+3.3 gIoU** on gRefCOCO, **+12.1 gIoU** on MUSE), with margins that grow as the number of referred targets increases. The concept interface also transfers to video under a detect-and-track setting, yielding new state-of-the-art results on seven referring video benchmarks, including **+10.9** $\mathcal{J}\&\mathcal{F}$ on MeViS and **+12.4** $\mathcal{J}\&\mathcal{F}$ on Ref-SeCVOS. The code, model checkpoints, and dataset annotations are open-sourced here.

## 1 Introduction

Referring segmentation [29, 47, 38, 30, 55, 8] aims to predict pixel-level masks for the targets described by a natural-language query, coupling language understanding with dense visual grounding. This fine-grained capability underpins applications such as interactive image and video editing [35, 58], industrial inspection [20], AR/VR interaction [71], and embodied perception [18, 84]. Building on large vision-language models (LVLMs), a growing line of work [32, 52, 75, 59] extends referring segmentation beyond single-expression grounding toward more *reasoning-intensive* settings. A representative paradigm [32, 52, 75, 59, 68, 54] appends a special `[SEG]` token to the LVLM, whose hidden state is decoded into a mask, providing an **implicit** interface between language reasoning and pixel prediction. This design yields strong results on standard referring and reasoning segmentation benchmarks [29, 47, 32], which are nevertheless dominated by *single-target* queries.

However, real-world queries are often *open-ended*, requiring reasoning over a heterogeneous set of plausible targets across multiple categories (e.g., picking ingredients for a salad), rather than locating

[*] Equal Contribution [‡] Project Lead [†] Corresponding Author

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

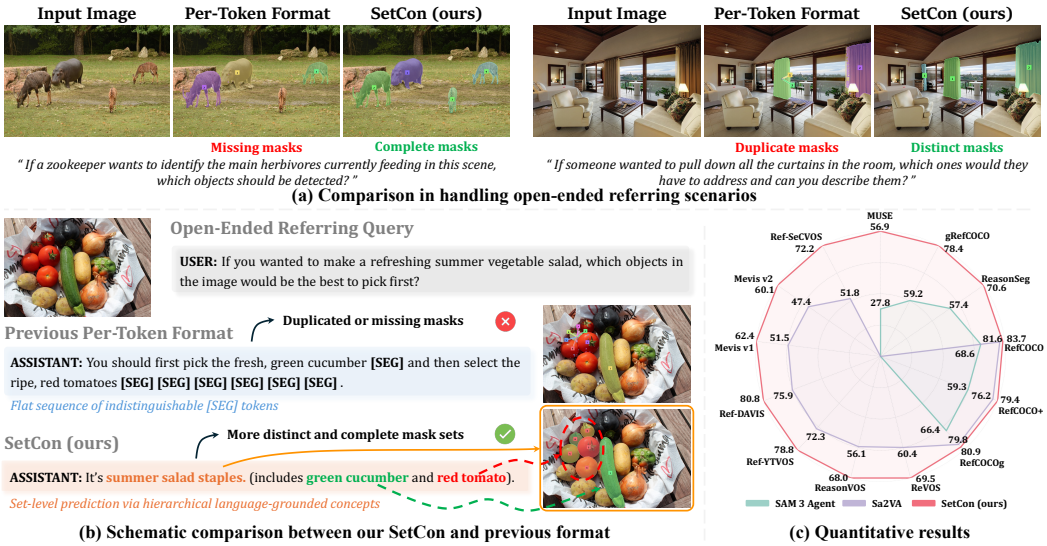


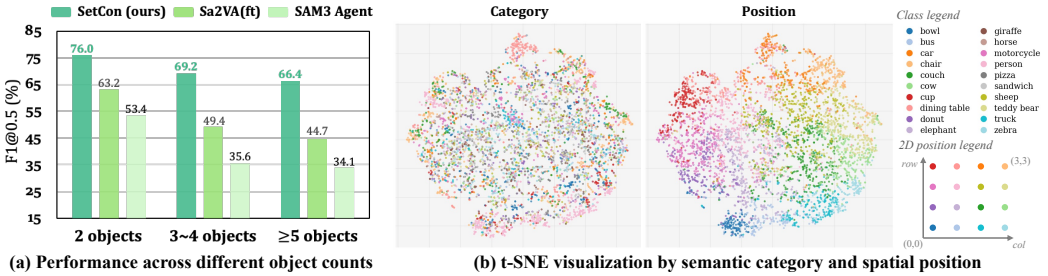

Figure 1: Overview of **Set-Con**cept Segmentation (**SETCON**). **(a)** Compared to previous per-token formulation, SETCON produces more distinct and complete mask sets in open-ended scenarios. **(b)** SETCON predicts interpretable, hierarchical set-level concepts and uses them as semantic conditions for mask-set decoding instead of indistinguishable special tokens. **(c)** Quantitative results show that SETCON achieves the best performance on image and video referring segmentation benchmarks.

a single object. Extending the prevailing paradigm to such scenarios remains challenging. Existing methods [68, 54, 32, 52] typically represent referred targets with one `[SEG]` token per target and decode the corresponding masks independently. This **per-token** formulation treats targets as separate outputs rather than a coherent semantic set, offering little incentive to capture set-level properties such as *completeness*, *cardinality*, and *mutual exclusivity*, and may yield duplicated or missing masks and unstable target-to-mask assignment (Fig. 1(a)). To probe this limitation, we conduct a pilot study by extending a representative LVLM-based segmentation model, Sa2VA [75], to the multi-target setting on the MUSE benchmark [54]. We observe that performance degrades sharply as the number of referred targets grows, and that the projected `[SEG]` embeddings cluster predominantly by *spatial position* rather than *semantic category*, suggesting that the implicit token interface encodes spatial layout more than the semantic structure of the referred set (see Sec. 3.1 for the detailed analysis).

Motivated by these findings, we reformulate open-ended referring segmentation as *explicit set-level concept prediction*. We present **Set-Con**cept **S**egmentation (**SETCON**), an end-to-end framework that couples LVLM-side language reasoning with set-level mask decoding through a **language-grounded concept interface** (Fig. 1 (b)). Given a language query and an image, SETCON produces a textual response in which the referred targets are surfaced as explicit natural-language concepts; the hidden states of these concept spans are then projected into the segmentation decoder as semantic conditions, and the full target mask set is decoded jointly. By replacing segmentation-specific tokens learned during finetuning with concepts drawn from the LVLM's own vocabulary, SETCON grounds every mask in an interpretable semantic anchor and reuses the model's *pretrained semantic geometry*, rather than relearning it from a newly initialized token. To handle cross-category scenarios, the concepts are organized via a **hierarchical semantic decomposition**: SETCON first emits a *set-level concept* that summarizes the overall target scope, and then decomposes it into *fine-grained sub-concepts*, one per target subset. This **coarse-to-fine** structure delineates the boundary of the target set while separating its sub-categories, improving both coverage and inter-category discrimination.

SETCON extends naturally to **referring video segmentation** under a detect-and-track formulation [53, 4]: the LVLM is invoked once per video to produce motion-aware concepts, which are shared across frames as fixed semantic conditions for per-frame decoding and memory-based mask propagation. This design keeps temporal inference efficient while preserving identity consistency across frames.

Training the concept interface requires linguistic supervision at both the set and the sub-category level, which existing reasoning segmentation datasets [32, 54] do not provide: their annotations are instance-level and rely on closed-vocabulary labels. To bridge this gap, we build *hierarchical*




(a) Performance across different object counts (b) t-SNE visualization by semantic category and spatial position

Figure 2: **Pilot study motivating explicit set-level concept prediction.** **(a)** special-token baselines degrade sharply as the target count grows, while SETCON remains comparatively stable. **(b)** t-SNE projection of `[SEG]` representations from the Sa2VA-based baseline, colored by semantic category and 2D spatial position; clusters align more clearly with position than with category.

*semantic supervision* via a two-stage annotation pipeline: every target is annotated with a free-form *sub-category phrase* grounded in its instance masks and the original query, and these phrases are organized under a global *set-level concept* that summarizes the target scope. Candidate annotations are further cleaned by rule-based filtering and multiple rounds of manual spot-checking. The resulting corpus has 236,396 samples and 784,809 concept phrases, with over 80% of the samples involving more than one sub-category, providing the set-level signal that prior corpora lack and supplying explicit concept-to-mask correspondence for open-ended multi-target and cross-category grounding.

We evaluate SETCON on six image and seven video referring segmentation benchmarks (see Fig. 1 (c)). On multi-object image segmentation, SETCON achieves SOTA performance on gRefCOCO and MUSE, surpassing the previous best by **+3.3 gIoU** on gRefCOCO *val* and **+12.1 gIoU** on the challenging MUSE *val*, with the margin widening as the number of referred targets grows. On standard single-target benchmarks (RefCOCO/+/g and ReasonSeg), SETCON remains competitive and attains the best average, indicating that our method does not regress on conventional settings. On video, SETCON achieves new SOTA results across all *seven* referring video segmentation benchmarks, with substantial gains on the challenging MeViS (**+10.9** $\mathcal{J}\&\mathcal{F}$) and Ref-SeCVOS (**+12.4** $\mathcal{J}\&\mathcal{F}$) settings, where stable semantic anchors are most beneficial for long-horizon temporal association.

Our contributions are: **1)** Through a pilot study on a representative LVLM-based segmentation model, we identify two limitations of the prevailing `[SEG]`-token paradigm in open-ended referring scenarios. **2)** We reformulate open-ended referring segmentation as *explicit set-level concept prediction* and propose SETCON, an end-to-end framework that couples LVLM-side reasoning with set-level mask decoding through a language-grounded concept interface, organized hierarchically into a set-level concept and fine-grained sub-concepts. **3)** SETCON achieves SOTA performance on six image and seven video referring segmentation benchmarks, with the largest gains on multi-target and cross-category settings, while remaining competitive on conventional single-target benchmarks.

## 2 Related Work

**Open-vocabulary Grounding.** Open-vocabulary grounding aims to localize visual entities specified by natural-language concepts, and has been advanced by vision-language pretraining, text-conditioned detection, and region-text alignment [28, 81, 34, 49, 40]. Subsequent segmentation models extend this capability from boxes or regions to open-vocabulary mask prediction with text or visual prompts [45, 36, 86, 85, 67]. Recent promptable foundation models further broaden this paradigm across image and video domains, enabling promptable and concept-level segmentation [31, 53, 14, 4]. However, these methods typically treat each prompt or concept separately, leaving open-ended cross-category scenarios underexplored. In contrast, SETCON predicts hierarchical concepts as explicit semantic conditions to decompose open-ended target sets and guide joint mask-set decoding.

**Referring Segmentation.** Referring segmentation aims to localize pixel-level masks for targets described by natural-language expressions [29, 47]. Early methods ground expressions through cross-modal fusion between visual features and linguistic queries [24, 39, 74], while later works improve language-aware dense prediction with stronger vision-language representations [73, 10, 86]. Subsequently, some works extend the task to more generalized referring scenarios [38, 25] and part-level formulations [61]. More recently, LVLM-based methods [32, 6, 54] further broaden the

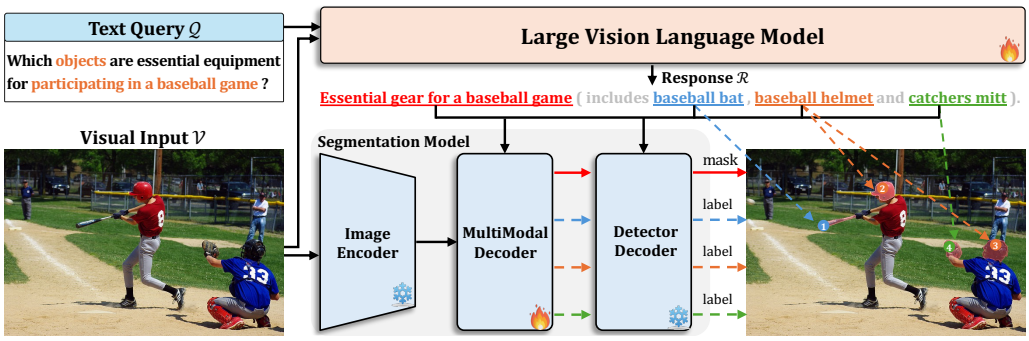



Figure 3: **Architecture of SETCON.** Given a text query $\mathcal{Q}$ and visual input $\mathcal{V}$, the LVLM produces a response $\mathcal{R}$ containing a global set-level concept and its decomposed sub-category concepts. The multimodal decoder is trainable while the image encoder and detector decoder remain frozen, to jointly predict the mask set with per-target labels.

task toward reasoning segmentation with open-ended instructions. In the video domain, referring video object segmentation additionally requires maintaining temporal consistency across evolving frames [30, 55, 8, 13, 51, 15, 16]. In this work, we provide a unified concept-based framework for open-ended image and video referring segmentation.

**LVLMs for Fine-grained Perception.** Large vision-language models (LVLMs) [56, 57, 2, 1, 11, 12, 21, 70] have recently been applied to fine-grained perception to bridge semantic understanding and dense prediction. Pioneered by LISA [32], many methods introduce a dedicated `[SEG]` token as an implicit interface for mask decoding, and further extend this design to grounded conversation [76, 52], visual understanding [75, 59], multi-target segmentation [54, 68, 27], and temporal scenarios [3, 72, 79]. Recent works also explore more explicit reasoning forms, such as reformulating segmentation as text generation [33] or producing intermediate reasoning steps and geometric prompts [44, 42, 43]. Instead of relying on special tokens or external geometric prompting, we adopt hierarchically organized natural-language concepts as semantic conditions for mask-set prediction, leveraging the pretrained semantic structure of LVLMs to better handle open-ended referring segmentation.

# 3 Methodology

## 3.1 Preliminary Study

Given a visual input $\mathcal{V}$ (an image or a video clip) and a natural-language query $\mathcal{Q}$, referring segmentation aims to predict the binary mask set $\mathcal{M}$ of all targets that satisfy $\mathcal{Q}$. A common practice in recent LVLM-based methods [32, 54, 68, 75, 59] is to introduce a special `[SEG]` token into the LVLM vocabulary and use it as an implicit interface for mask prediction. In this paradigm, the model generates a flat sequence of `[SEG]` tokens for each referred target, and the hidden state $\mathbf{h}_{\texttt{seg}}$ of each token is separately fed into a mask decoder $\mathcal{D}$ to produce a mask $m = \mathcal{D}(\mathcal{V}, \mathbf{h}_{\texttt{seg}})$.

However, it is highly counter-intuitive to encode the distinct semantic and spatial properties of multiple diverse targets into the hidden states of identical, repetitively generated `[SEG]` tokens. Such a trivial design inevitably limits the representation capacity and risks feature ambiguity. To further investigate the limitation of the implicit `[SEG]`-token paradigm under open-ended scenarios, we conduct a preliminary study on a representative model Sa2VA [75]. Following [54], we extend it to the open-ended multi-target setting by sequentially predicting multiple special `[SEG]` tokens within a single response. We finetune the model on the MUSE [54] training set and evaluate it on the test set.

We categorize samples by the number of referred targets and report the F1 score at an IoU threshold of 0.5. As shown in Fig. 2(a), performance degrades sharply as the target count grows, indicating the per-token formulation is ill-posed to more challenging multi-target segmentation settings, where the model has to reason about the target set holistically to ensure completeness and avoid duplicates. Besides, the absence of explicit semantic distinctions among repetitively generated tokens makes their alignment with ground-truth instances inherently ambiguous, which hinders model optimization.

To further investigate what these implicit tokens encode, we apply t-SNE to the projected `[SEG]` representations on a deduplicated, class-balanced subset of RefCOCO. As illustrated in Fig. 2(b),

the embeddings reveal clearer cluster structure under spatial-position coloring than under semantic-category coloring, and spatially adjacent regions remain close in the embedding space. This indicates that the implicit [SEG] representations are organized predominantly by spatial layout rather than by semantics. Consequently, although sufficient for mask decoding in few-target cases, they form a suboptimal interface for more complex, open-ended scenarios.

### 3.2 SETCON

To overcome these inherent limitations, we propose a paradigm shift from implicit token generation to explicit semantic grounding, directly utilizing natural language as an interpretable and discriminative interface for complex multi-target segmentation. We propose SETCON, an end-to-end framework that jointly grounds an open-ended target set with explicit, interpretable semantics. An overview of the architecture is shown in Fig. 3.

**Explicit Set-Level Concept Prediction.** Instead of relying on implicit [SEG] tokens as a latent signal for mask decoding, SETCON reads the target concepts directly from its own textual output. Specifically, the model first generates a response $\mathcal{R}$ in which each referred concept $\mathcal{C}$ is expressed as a free-form language phrase and delimited by special markers <ref>...</ref>. Each concept is bound to an entire *set* of semantically coherent instances rather than to a single target, so that $\mathcal{R}$ explicitly encodes the granularity of set-level mask prediction. The hidden states $\mathbf{H}$ of the concept tokens are then projected into the decoder feature space as semantic conditions $\tilde{\mathbf{H}}$, and the segmentation module $\mathcal{D}$ produces the associated mask set $\mathcal{M} = \mathcal{D}(\mathcal{V}, \tilde{\mathbf{H}})$. Compared with implicit-token interfaces, this formulation not only attaches an interpretable semantic anchor to each predicted target set, but also reuses the LVLM's pretrained semantic representations rather than learning them from scratch with newly initialized tokens. Moreover, by expressing concepts in free-form language instead of a fixed taxonomy, it naturally supports open-ended, cross-category scenarios in which the target cardinality, identity, and semantic scope are dynamically determined by the query and visual context.

**Hierarchical Semantic Decomposition.** Many real-world queries refer not to a single category but to a group spanning several sub-categories, e.g. "all animals in the picture". Enumerating every sub-category as a flat sequence of concepts tends to yield incomplete target sets. SETCON therefore organizes the predicted concepts hierarchically: it first emits a set-level concept that summarizes the shared semantics of the entire target group, and then decomposes it into a list of fine-grained sub-category concepts. Concretely, the first predicted concept serves as the global representation $\tilde{\mathbf{H}}_0$ for joint mask generation, while the remaining concepts $\{\tilde{\mathbf{H}}_i\}_{i=1}^{N}$ correspond to the $N$ semantic subsets used for label assignment, with $N$ freely determined by the LVLM rather than fixed in advance. During mask decoding, each sub-category representation is fused with $\tilde{\mathbf{H}}_0$ so that fine-grained predictions are conditioned on the shared semantic scope. This coarse-to-fine design is intended to improve both coverage and discrimination: the global concept delineates what belongs to the target set, while the sub-category concepts separate the subsets within it.

**Training and Inference.** SETCON is trained end-to-end with an auto-regressive language-modeling loss on the response $\mathcal{R}$ and a DETR-style set-prediction loss [5] on the mask predictions of each concept group. For video clips, we randomly sample sparse frames per clip and share a single concept response across them, as the referred semantics remain consistent over time. For no-target samples, we replace the ground-truth concepts with "no target" in $50\%$ of the cases and keep the original concepts in the rest, encouraging the LVLM and segmentation model $\mathcal{D}$ to learn abstention from complementary signals. At inference time, $\mathcal{D}$ separately decodes the global representation $\tilde{\mathbf{H}}_0$ and the fused sub-category representations $\{\tilde{\mathbf{H}}_i\}_{i=1}^{N}$, producing a primary mask set $\mathcal{M}_0$ together with $N$ sub-category mask sets $\{\mathcal{M}_i\}_{i=1}^{N}$. We then take the union $\bigcup_i \mathcal{M}_i$ and align it with $\mathcal{M}_0$ via mask-level Hungarian matching, attaching a fine-grained sub-category label to each primary mask. For video inference, we adopt the detect-and-track pipeline of SAM 3 [4], where the LVLM is invoked only once per clip and its generated concepts are broadcast to all frames as semantic anchors for per-frame detection, while memory-based mask propagation is delegated to the off-the-shelf tracker without any video-specific modification.

### 3.3 Hierarchical Semantic Annotation

Existing referring segmentation datasets mainly focus mainly on instance-level annotation and provide little set-level annotation. To fill this gap, we augment existing reasoning segmentation

Table 1: **Comparison with prior work on multi-object referring segmentation benchmarks**, gRefCOCO [38] and MUSE [54], reporting gIoU, cIoU, and F1@0.5 where available. SETCON obtains the **best** score on every column. † denotes results reproduced by us.

| Model | gRefCOCO | | | | | | MUSE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | val | | testA | | testB | | val | | | test | | |
| | gIoU | cIoU | gIoU | cIoU | gIoU | cIoU | gIoU | cIoU | F1@0.5 | gIoU | cIoU | F1@0.5 |
| LISA-7B [32] | 61.6 | 61.8 | 66.3 | 68.5 | 58.8 | 60.6 | 42.0 | 46.1 | - | 38.9 | 44.4 | - |
| LISA-Llama2-13B [32] | 63.5 | 63.0 | 68.2 | 69.7 | 61.8 | 62.2 | 43.6 | 50.2 | - | 41.9 | 50.5 | - |
| GSVA [68] | 66.5 | 63.3 | 71.1 | 69.9 | 62.2 | 60.5 | - | - | - | - | - | - |
| PixelLM-7B [54] | - | - | - | - | - | - | 42.6 | 50.7 | - | 39.2 | 46.3 | - |
| PixelLM-Llama2-13B [54] | - | - | - | - | - | - | 44.8 | 54.1 | - | 42.3 | 51.0 | - |
| SAM4MLLM [6] | 71.9 | 67.8 | 74.2 | 72.2 | 65.3 | 63.4 | - | - | - | - | - | - |
| Text4Seg [33] | 74.4 | 69.1 | 75.1 | 73.8 | 67.3 | 66.6 | - | - | - | - | - | - |
| MLLMSeg [60] | 75.1 | 71.6 | 77.0 | 76.9 | 69.7 | 68.5 | - | - | - | - | - | - |
| SAM3 Agent [4] | 59.2† | 49.1† | 63.4† | 61.8† | 58.5† | 55.3† | 27.8† | 24.2† | 41.0† | 23.8† | 26.3† | 38.1† |
| Visionreasoner [43] | 41.5 | 48.3 | 58.2† | 64.2† | 48.5† | 51.1† | 42.4† | 41.6† | 54.5† | 36.1† | 36.8† | 49.2† |
| **SETCON (ours)** | **78.4** | **72.0** | **78.5** | **78.1** | **73.1** | **72.4** | **56.9** | **59.5** | **71.2** | **53.2** | **60.0** | **69.3** |

datasets [32, 54] with hierarchical semantic annotations through a two-stage pipeline. We adopt Qwen3-VL-235B-A22B [2] as the annotator throughout the pipeline for its strong visual grounding and instruction-following ability. In the first stage, starting from the source annotations in which each target is tagged with a category from a closed list, we overlay the instance masks of that category on the image and prompt an LVLM, conditioned on the original label, to produce a free-form sub-category phrase that captures the targets' appearance, attributes, and role in context; the prompt is anchored to the original label to prevent semantic drift. In the second stage, conditioned on the image, the original query, and the collected sub-category phrases, the LVLM is prompted to produce a global set-level concept that is faithful to the query semantics and inclusive of all referred sub-targets. The resulting annotations are further cleaned by rule-based filtering and multiple rounds of manual spot-checking to remove malformed or low-quality cases.

The final corpus contains 236,396 samples and 784,809 concept phrases, averaging 2.32 sub-categories per sample and reaching up to 16 in the long tail cases. Over $80\%$ of the samples involve more than one sub-category, supplying the set-level supervision that prior corpora lack. The concept phrases average 3.5 words, indicating natural-language expressions rather than closed-set labels. The full annotation procedure, filtering criteria, and detailed statistics are provided in Appendix B.

# 4 Experiments

## 4.1 Experimental Setup

**Implementation Details.** We adopt Qwen3-VL-8B-Instruct [2] as the LVLM backbone and SAM 3 [4] as the segmentation model. During training, we jointly optimize the multimodal decoder, the projection module, the token embeddings and language-modeling head of the LVLM, and LoRA adapters with rank 128, $\alpha = 256$, and dropout 0.05 attached to all linear layers of the LVLM language model. We train the model for one epoch on 8 NVIDIA H200 GPUs using AdamW with a learning rate of $4 \times 10^{-5}$, a cosine annealing schedule, and a batch size of 64.

**Datasets and Benchmarks.** Our training set comprises both image and video segmentation datasets. For images, we use RefCOCO/+/g [29, 47], gRefCOCO [38], ReasonSeg [32], and MUSE [54]; for videos, we use MeViS [8, 9], Ref-DAVIS [30], Ref-YouTube-VOS [55], and ReVOS [72]. Referring datasets are used with their original natural-language queries, while reasoning segmentation datasets (ReasonSeg and MUSE) are augmented with hierarchical semantic annotations via the two-stage pipeline in Sec. 3.3. We evaluate on the test splits of these training datasets, and additionally report results on ReasonVOS [3] and SeCVOS [79] to assess performance in complex video scenarios. Detailed implementation and evaluation metrics are provided in Appendix C.

## 4.2 Main Results

We conduct a comprehensive evaluation of SETCON on 6 image and 7 video segmentation benchmarks. The comparison includes both classical segmentation models and recent LVLM-based approaches.

Table 2: **Comparison with prior work on referring video object segmentation benchmarks** [72, 3, 55, 30, 8, 9, 79], reporting $\mathcal{J}\&\mathcal{F}$. SETCON obtains the **best** score on all seven benchmarks, with substantial gains on the more challenging MeViS and Ref-SeCVOS settings.

| Model | $\mathcal{J}\&\mathcal{F}$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | **ReVOS** | **ReasonVOS** | **Ref-YTVOS** | **Ref-DAVIS** | **MeViS v1** | **MeViS v2** | **Ref-SeCVOS** |
| LBDT [17] | - | - | 49.4 | 54.1 | 29.3 | 25.1 | - |
| ReferFormer [66] | 28.1 | 32.9 | 62.9 | 61.1 | 31.0 | 26.7 | - |
| VLT+TC [10] | - | - | 62.7 | 60.3 | 35.6 | 30.1 | - |
| HTML [22] | - | - | 63.4 | 62.1 | - | - | - |
| OnlineRefer [65] | - | 38.7 | 63.5 | 64.8 | 32.3 | - | - |
| LMPM [8] | 26.4 | - | - | - | 37.2 | 38.3 | - |
| SOC [46] | - | 35.9 | 66.0 | 64.2 | - | - | - |
| SgMg [48] | - | 36.2 | 65.7 | 63.3 | - | - | - |
| TrackGPT [82] | 45.0 | - | 59.5 | 66.5 | 41.2 | - | - |
| LISA [32] | 40.9 | 31.1 | 53.9 | 64.8 | 37.2 | - | - |
| VideoLISA [3] | - | 47.5 | 63.7 | 68.8 | 44.4 | - | 42.8 |
| VISA [72] | 50.9 | - | 63.0 | 70.4 | 44.5 | - | 59.5 |
| DsHmp [23] | - | - | 67.1 | 64.9 | 46.4 | 40.8 | - |
| DMVS [62] | - | - | 64.3 | 65.2 | 48.6 | - | - |
| VideoGLaMM [50] | - | - | - | 69.5 | 45.2 | - | - |
| ViLLa [80] | 57.0 | - | 67.5 | 74.3 | 49.4 | - | - |
| SAMWISE [7] | - | - | 69.2 | 70.6 | 49.5 | - | 54.0 |
| GLUS [37] | 54.9 | 49.9 | 67.3 | 73.9 | 51.3 | 46.5 | 59.8 |
| Sa2VA [75] | 60.4 | 56.1 | 72.3 | 75.9 | 51.5 | 47.4 | 51.8 |
| InstructSeg [64] | 54.5 | - | 67.5 | 71.1 | - | - | - |
| VRS-HQ [19] | 59.1 | - | 70.4 | 76.0 | 50.6 | - | - |
| SDAM [83] | 58.0 | 55.1 | 65.3 | 76.0 | 48.6 | - | - |
| **SETCON (ours)** | **69.5** | **68.0** | **78.8** | **80.8** | **62.4** | **60.1** | **72.2** |

Table 3: **Ablation on the proposed modules.** Adding set-level prediction and explicit conditioning each contributes a further improvement.

| Set-level Prediction | Concept Condition | gIoU | F1@0.5 |
|---|---|---|---|
| ✗ | ✗ | 43.6 | 59.2 |
| ✓ | ✗ | 51.4 | 67.8 |
| ✓ | ✓ | **53.2** | **69.3** |

Table 4: **Ablation on the proposed annotations.** Combining the "Diverse Labeling" and "Hierarchical Semantic" yields the best result.

| Diverse Labeling | Hierarchical Semantic | gIoU | LLM Score |
|---|---|---|---|
| ✗ | ✗ | 50.3 | 5.99 |
| ✓ | ✗ | 50.0 | 6.92 |
| ✓ | ✓ | **53.2** | **6.94** |

As shown in Tab. 1, SETCON achieves superior performance on the multi-object segmentation benchmarks gRefCOCO and MUSE, surpassing the previous state of the art by +12.1 gIoU on MUSE *val* and +3.3 gIoU on gRefCOCO *val*. Notably, as illustrated in Fig. 2(a), the performance margin becomes more pronounced as the number of target objects increases. This trend suggests that our explicit semantic set prediction design provides strong stability and scalability in challenging multi-target scenarios. Moreover, emphasizing set-level concepts does not compromise single-target segmentation accuracy. As reported in Tab. 5, SETCON also achieves the competitive performance across RefCOCO/+/g and ReasonSeg, demonstrating that the proposed set-level formulation remains compatible with conventional single-object referring segmentation.

Tab. 2 further demonstrates that SETCON establishes new state-of-the-art results across all seven referring video object segmentation benchmarks. In particular, the improvements are especially pronounced on the more challenging MeViS and Ref-SeCVOS benchmarks, where SETCON surpasses previous best methods by $+10.9$ and $+12.4$ $\mathcal{J}\&\mathcal{F}$, respectively. These substantial gains suggest that explicit semantic concepts serve as stable target anchors for long-horizon temporal association and semantically complex video understanding. As a result, SETCON can maintain more robust object tracking under challenging conditions such as occlusion, distractors, and appearance changes.

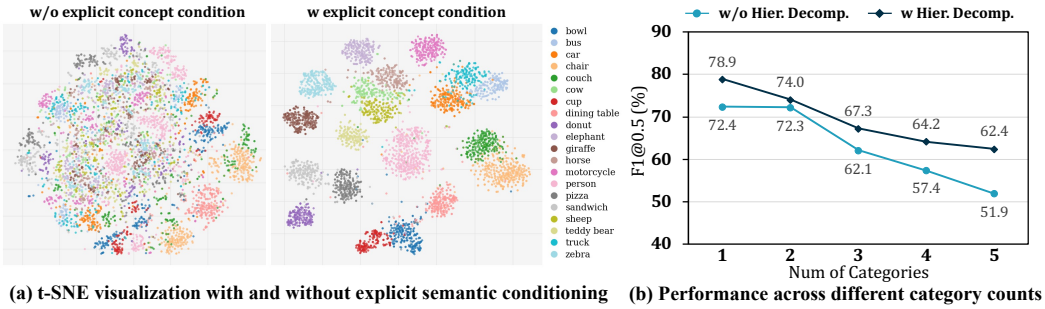
### 4.3 Ablation Study and Analysis

We conduct a series of ablation studies on the MUSE test set to isolate the contributions of our architectural design and annotation quality, with results reported in Tab. 3 and Tab. 4.

**Effectiveness of proposed modules.** Tab. 3 presents an ablation study evaluating the contributions of set-level prediction and explicit concept conditioning. Starting from a baseline that follows the prevailing implicit-token paradigm, introducing set-level prediction yields a substantial improvement,

Table 5: **Comparison with prior work on single-object referring and reasoning segmentation benchmarks** RefCOCO/+/g [29, 47] and ReasonSeg [32]. SETCON remains competitive on RefCOCO/+/g, achieves the best results on ReasonSeg, and obtains the best overall average, indicating that the set-level formulation **does not regress** on conventional single-target settings.

| Model | RefCOCO | | | RefCOCO+ | | | RefCOCOg | | ReasonSeg | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | val | testA | testB | val | testA | testB | val | test | val | test | |
| LISA [32] | 74.9 | 79.1 | 72.3 | 65.1 | 70.8 | 58.1 | 67.9 | 70.6 | 46.0 | 34.1 | 63.9 |
| PixelLM [54] | 73.0 | 76.5 | 68.2 | 66.3 | 71.7 | 58.3 | 69.3 | 70.5 | - | - | - |
| GSVA [68] | 79.2 | 81.7 | 77.1 | 70.3 | 73.8 | 63.6 | 75.7 | 77.0 | - | - | - |
| GLaMM [52] | 79.5 | 83.2 | 76.9 | 72.6 | 78.7 | 64.6 | 74.2 | 74.9 | 47.2 | - | - |
| SAM4MLLM [6] | 79.8 | 82.7 | 74.7 | 74.6 | 80.0 | 67.2 | 75.5 | 76.4 | 60.4 | - | - |
| GLEE [67] | 80.0 | - | - | 69.6 | - | - | 72.9 | - | - | - | - |
| UniLSeg [41] | 81.7 | 83.2 | 79.9 | 73.2 | 78.3 | 68.2 | 79.3 | 80.5 | - | - | - |
| EVF-SAM [77] | 82.4 | 84.2 | 80.2 | 76.5 | 80.0 | 71.9 | 78.2 | 78.3 | - | - | - |
| PSALM [78] | 83.6 | 84.7 | 81.6 | 72.9 | 75.5 | 70.1 | 73.8 | 74.4 | - | - | - |
| HyperSeg [63] | 84.8 | 85.7 | 83.4 | 79.0 | 83.5 | 75.2 | 79.4 | 78.9 | 56.7 | - | - |
| Text4Seg [33] | 79.2 | 81.7 | 75.6 | 72.8 | 77.9 | 66.5 | 74.0 | 75.3 | - | - | - |
| DETRIS [26] | 81.0 | 81.9 | 79.0 | 75.2 | 78.6 | 70.2 | 74.6 | 75.3 | - | - | - |
| MLLMSeg [60] | 81.0 | 82.4 | 78.7 | 76.4 | 79.1 | 72.5 | 79.9 | 80.8 | - | - | - |
| Sa2VA [75] | 81.6 | - | - | 76.2 | - | - | 78.7 | - | - | - | - |
| RICE [69] | 83.5 | 85.3 | 81.7 | 79.4 | 82.8 | 75.4 | 79.8 | 80.4 | - | - | - |
| X-SAM [59] | 85.1 | 87.1 | 83.4 | 78.0 | 81.0 | 74.4 | 83.8 | 83.9 | 32.9 | 41.0 | 73.1 |
| RSVP [44] | - | - | - | - | - | - | 65.5 | 66.4 | 56.7 | 50.7 | - |
| Seg-Zero [42] | - | 80.3 | - | - | 76.2 | - | - | 72.6 | 62.0 | 52.0 | - |
| Visionreasoner [43] | 76.8 | 78.9 | 72.5 | 70.9 | 74.9 | 64.6 | 72.9 | 71.3 | 66.3 | 63.6 | 71.3 |
| SAM3 Agent [4] | 68.6 | 72.3 | 63.9 | 59.3 | 64.5 | 55.6 | 66.4 | 66.9 | 57.4 | 67.3 | 64.2 |
| **SETCON (ours)** | 83.7 | 84.4 | 81.6 | 79.4 | 82.3 | 75.6 | 80.9 | 80.0 | 70.6 | 70.5 | **78.9** |




**(a) t-SNE visualization with and without explicit semantic conditioning** **(b) Performance across different category counts**

Figure 4: **Ablation analysis of SETCON on MUSE. (a)** t-SNE projection of segmentation conditions, colored by category: explicit concept conditioning yields tighter, more category-aligned clusters than the token-only variant. **(b)** F1@0.5 versus the number of referred categories (1–5): hierarchical decomposition consistently outperforms the flat variant, with the gap widening as cardinality grows.

suggesting that modeling referred targets as a set better aligns with the requirements of open-ended scenarios. Incorporating explicit concept conditioning further improves performance, indicating that grounding targets with interpretable semantic concepts provides additional discriminative cues beyond latent mask tokens. We further compare the learned feature spaces of token-based and concept-based representations in Fig. 4(a). Compared with special token representations, concept-based representations form clearer concept-level clusters, corroborating that explicit semantic anchoring reshapes the feature space toward semantic structure and makes it better suited to open-ended queries.

**Effectiveness of proposed datasets.** To evaluate the impact of our annotation pipeline, we ablate its two stages and use an LVLM-based judge to assess the quality of the resulting semantic labels. As shown in Tab. 4, replacing the original rigid labels with diverse natural-language descriptions improves the semantic expressiveness while preserving mask prediction quality. This indicates that richer descriptions provide more informative semantic cues without compromising localization accuracy. Further adding hierarchical semantic annotation leads to more accurate predictions, suggesting that organizing the enriched labels into a coarse-to-fine structure helps the model better exploit semantic supervision. Fig. 4(b) further shows that this advantage becomes more pronounced as the number of referred categories increases, highlighting the benefit of hierarchy-aware supervision for high-cardinality and compositionally complex queries.




Figure 5: **Qualitative comparison on in-the-wild images.** We show the user prompt, input image, the LISA-format baseline, and SETCON, with the predicted sub-category concepts listed below. Across reasoning-style queries spanning *multi-instance*, *cross-category*, and *open-ended scenarios*, SETCON tends to produce more complete and semantically coherent mask sets.

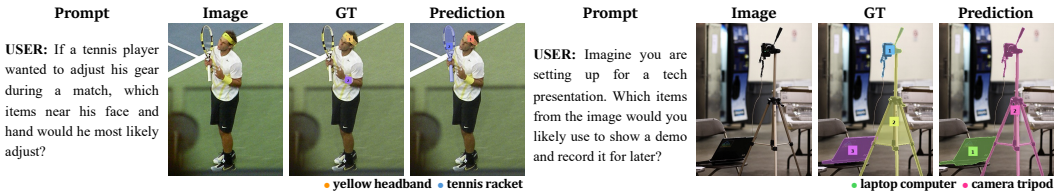



Figure 6: **Failure cases on the MUSE benchmark.** Open-ended queries expose limitations in target disambiguation and concept granularity, which may lead to incorrect or incomplete predictions.

**Qualitative Results.** To more intuitively demonstrate the segmentation performance of our framework in real-world scenarios, Fig. 5 presents a visual comparison between SETCON and a standard LISA-format baseline on in-the-wild images. As shown, our method consistently produces more accurate and semantically coherent masks across diverse visual scenes. Beyond simply localizing the referred targets, SETCON can distinguish multiple object instances based on fine-grained category semantics, as well as appearance and contextual attributes. Benefiting from our hierarchical semantic annotations, the model is better able to activate and ground the world knowledge encoded in the LVLM, enabling robust generalization to long-tail, cross-category, and open-ended scenarios. Additional qualitative results on video referring segmentation are provided in Appendix A.

**Failure Cases.** However, since SETCON captures the overall scene semantics, open-ended queries may introduce ambiguity in the intended targets, and the semantic boundaries between related objects can be blurred. As illustrated in Fig. 6, open-ended queries may cause ambiguity in target selection and concept granularity. Relational descriptions may refer to multiple plausible objects satisfying the same spatial relation, while functional descriptions may compose related objects into a common semantic concept, leading to mismatches with the annotated target set.

## 5 Conclusion

In this paper, we presented **Set**-**Con**cept Segmentation (**SETCON**), a framework that reformulates open-ended referring segmentation as explicit set-level concept prediction. Through an interpretable language-based concept interface and hierarchical semantic decomposition, SETCON jointly grounds target sets and decodes coherent mask set, enabling effective segmentation in multi-instance, cross-category, and open-ended scenarios. We further construct large-scale hierarchical supervision to enhance concept-to-mask learning and open-world generalization. Extensive experiments on image and video benchmarks show that SETCON achieves leading performance, particularly in challenging multi-target and cross-category settings. We hope this work provides a simple yet effective step toward more semantic, scalable, and general-purpose segmentation systems.

## Acknowledgments

This paper is funded in part by the National Science and Technology Major Project of China (Grant No. 2026ZD0128702) and by the Shanghai Innovation Institute.

## References

[1] Arash Ebrahim Abtahizadeh, Muna Aghamelu, Adil Ahmed, Juan C Acuna, Alexis Audran-Reiss, Asaf Avrahamy, Yoram Bachrach, Can Balioglu, Amar Budhiraja, Onur Çelebi, et al. Muse glimmer 30b: An on-device agentic model. *arXiv preprint arXiv:2610.24628*, 2026.

[2] 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.

[3] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Z Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, pages 6833–6859, 2024.

[4] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229, 2020.

[6] 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*, pages 323–340. Springer, 2024.

[7] Claudia Cuttano, Gabriele Trivigno, Gabriele Rosi, Carlo Masone, and Giuseppe Averta. Samwise: Infusing wisdom in sam2 for text-driven video segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3395–3405, 2025.

[8] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2694–2703, 2023.

[9] Henghui Ding, Chang Liu, Shuting He, Kaining Ying, Xudong Jiang, Chen Change Loy, and Yu-Gang Jiang. Mevis: A multi-modal dataset for referring motion expression video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[10] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 7900–7916, 2022.

[11] Shengyuan Ding, Xinyu Fang, Ziyu Liu, Yuhang Zang, Yuhang Cao, Xiangyu Zhao, Haodong Duan, Xiaoyi Dong, Jianze Liang, Bin Wang, et al. Arm-thinker: Reinforcing multimodal generative reward models with agentic tool use and visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22205, 2026.

[12] Shengyuan Ding, Xilin Wei, Xinyu Fang, Haodong Duan, Dahua Lin, Jiaqi Wang, and Yuhang Zang. Beyond the current observation: Evaluating multimodal large language models in controllable non-markov games. *arXiv preprint arXiv:2606.19338*, 2026.

[13] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9706–9716. IEEE, 2022.

[14] Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13614–13624, 2025.

[15] Shuangrui Ding, Rui Qian, Haohang Xu, Dahua Lin, and Hongkai Xiong. Betrayed by attention: A simple yet effective approach for self-supervised video object segmentation. In *European Conference on Computer Vision*, pages 215–233. Springer, 2024.

[16] Shuangrui Ding, Peisen Zhao, Xiaopeng Zhang, Rui Qian, Hongkai Xiong, and Qi Tian. Prune spatio-temporal tokens by semantic-aware temporal accumulation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16899–16910. IEEE, 2023.

[17] Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, and Si Liu. Language-bridged spatial-temporal interaction for referring video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4964–4973, 2022.

[18] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8469–8488, 2023.

[19] Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, and Huchuan Lu. The devil is in temporal token: High quality video reasoning segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29183–29192, 2025.

[20] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1932–1940, 2024.

[21] Feng Han, Zhixiong Zhang, Zheming Liang, Yibin Wang, and Jiaqi Wang. Lomo: Local modality substitution for deeper vision-language fusion. *arXiv preprint arXiv:2605.30265*, 2026.

[22] Mingfei Han, Yali Wang, Zhihui Li, Lina Yao, Xiaojun Chang, and Yu Qiao. Html: Hybrid temporal-scale multimodal learning framework for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13414–13423, 2023.

[23] Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13332–13341, 2024.

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

[25] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4067–4077, 2023.

[26] Jiaqi Huang, Zunnan Xu, Ting Liu, Yong Liu, Haonan Han, Kehong Yuan, and Xiu Li. Densely connected parameter-efficient tuning for referring image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3653–3661, 2025.

[27] Donggon Jang, Yucheol Cho, Suin Lee, Taehyeon Kim, and Daeshik Kim. Mmr: A large-scale benchmark dataset for multi-target and multi-granularity reasoning segmentation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[28] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021.

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

[30] Anna Khoreva, Anna Rohrbach, and Brent Schiele. Video object segmentation with referring expressions. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[31] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[32] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9579–9589, 2024.

[33] Mengcheng Lan, Chaofeng Chen, Yue Zhou, Jiaxing Xu, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Text4seg: Reimagining image segmentation as text generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[34] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022.

[35] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22511–22521, 2023.

[36] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7061–7070, 2023.

[37] Lang Lin, Xueyang Yu, Ziqi Pang, and Yu-Xiong Wang. Glus: Global-local reasoning unified into a single large language model for video segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8658–8667, 2025.

[38] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023.

[39] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1271–1280, 2017.

[40] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55, 2024.

[41] Yong Liu, Cairong Zhang, Yitong Wang, Jiahao Wang, Yujiu Yang, and Yansong Tang. Universal segmentation at arbitrary granularity with language instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3459–3469, 2024.

[42] 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.

[43] Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified reasoning-integrated visual perception via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

[44] Yi Lu, Jiawang Cao, Yongliang Wu, Bozheng Li, Licheng Tang, Yangguang Ji, Chong Wu, Jay Wu, and Wenbo Zhu. Rsvp: Reasoning segmentation via visual prompting and multi-modal chain-of-thought. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 14699–14716, 2025.

[45] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096, 2022.

[46] Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. Soc: Semantic-assisted object cluster for referring video object segmentation. *Advances in Neural Information Processing Systems*, pages 26425–26437, 2023.

[47] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[48] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 920–930, 2023.

[49] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pages 728–755, 2022.

[50] Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19036–19046, 2025.

[51] Rui Qian, Shuangrui Ding, Xian Liu, and Dahua Lin. Semantics meets temporal correspondence: Self-supervised object-centric learning in videos. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16629–16641. IEEE, 2023.

[52] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024.

[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. In *The Thirteenth International Conference on Learning Representations*, 2025.

[54] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024.

[55] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European conference on computer vision*, pages 208–223, 2020.

[56] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[57] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[58] Yuanpeng Tu, Hao Luo, Xi Chen, Sihui Ji, Xiang Bai, and Hengshuang Zhao. Videoanydoor: High-fidelity video object insertion with precise motion control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025.

[59] 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*, pages 26187–26196, 2026.

[60] 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*, 2025.

[61] Wenxuan Wang, Tongtian Yue, Yisi Zhang, Longteng Guo, Xingjian He, Xinlong Wang, and Jing Liu. Unveiling parts beyond objects: Towards finer-granularity referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12998–13008, 2024.

[62] Zanyi Wang, Dengyang Jiang, Liuzhuozheng Li, Sizhe Dang, Chengzu Li, Harry Yang, Guang Dai, Mengmeng Wang, and Jingdong Wang. Deforming videos to masks: Flow matching for referring video segmentation. *arXiv preprint arXiv:2510.06139*, 2025.

[63] Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Jie Hu, Dengjie Li, Zheng Zhao, and Yujiu Yang. Hyperseg: Hybrid segmentation assistant with fine-grained visual perceiver. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 8931–8941, June 2025.

[64] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20193–20203, 2025.

[65] Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2761–2770, 2023.

[66] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022.

[67] Junfeng Wu, Yi Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3783–3795, 2024.

[68] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869, 2024.

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

[70] Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. Caprl: Stimulating dense image caption capabilities via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pages 13066–13093, 2026.

[71] Yanming Xiu, Tim Scargill, and Maria Gorlatova. Viddar: Vision language model-based task-detrimental content detection for augmented reality. *IEEE transactions on visualization and computer graphics*, 2025.

[72] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pages 98–115, 2024.

[73] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18155–18165, 2022.

[74] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018.

[75] 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.

[76] 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, 2024.

[77] Yuxuan Zhang, Tianheng Cheng, Lianghui Zhu, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Evf-sam: Early vision-language fusion for text-prompted segment anything model. *arXiv preprint arXiv:2406.20076*, 2024.

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

[79] Zhixiong Zhang, Shuangrui Ding, Xiaoyi Dong, Songxin He, Jianfan Lin, Junsong Tang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Sec: Advancing complex video object segmentation via progressive concept construction. In *The Fourteenth International Conference on Learning Representations*, 2026.

[80] Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, and Hengshuang Zhao. Villa: Video reasoning segmentation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23667–23677, 2025.

[81] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022.

[82] Jiawen Zhu, Zhi-Qi Cheng, Jun-Yan He, Chenyang Li, Bin Luo, Huchuan Lu, Yifeng Geng, and Xuansong Xie. Tracking with human-intent reasoning. *arXiv preprint arXiv:2312.17448*, 2023.

[83] Zhengtong Zhu, Jiaqing Fan, Zhixuan Liu, and Fanzhang Li. Training-free spatio-temporal decoupled reasoning video segmentation with adaptive object memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14022–14030, 2026.

[84] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183, 2023.

[85] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15116–15127, 2023.

[86] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in neural information processing systems*, 36:19769–19782, 2023.

# A Additional Qualitative Results on Video

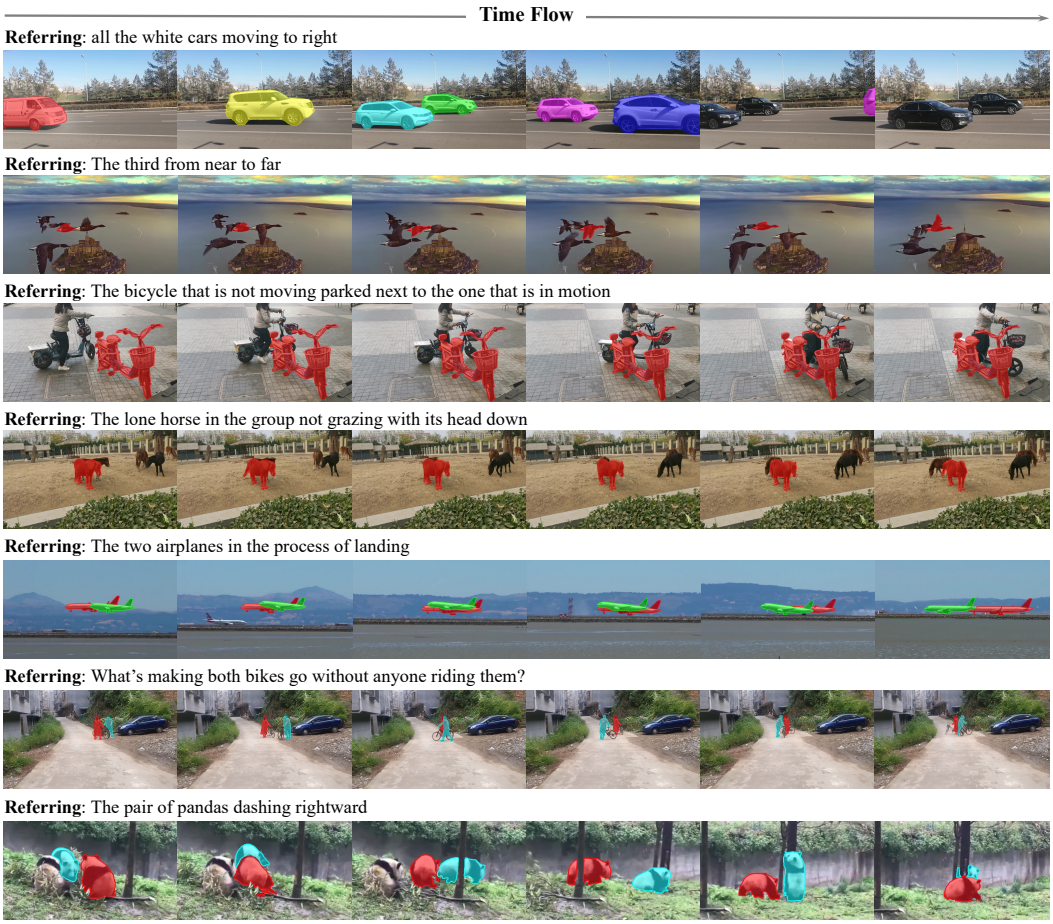



Figure 7: **Additional qualitative results on referring video object segmentation.** SETCON maintains stable target identities and temporally consistent masks across challenging video sequences with occlusion, distractors, and appearance changes, by using explicit semantic concepts as persistent anchors shared across frames.

We provide additional qualitative results on referring video object segmentation in Fig. 7, sampled from the challenging MeViS v2 [9]. The selected clips cover scenarios that are typically challenging for referring video segmentation, including long temporal horizons spanning hundreds of frames, fast or non-rigid object motion, and queries that refer to multiple targets simultaneously. SETCON generalizes naturally to these settings: a single set-level concept is generated once per clip and shared across all frames as a persistent semantic anchor, yielding mask sets that remain temporally coherent and identity-consistent even when individual frames are visually ambiguous in isolation.

# B Details of Hierarchical Semantic Annotation

To support the language-grounded concept interface of SETCON, we augment existing reasoning segmentation datasets [32, 54] with hierarchical semantic supervision. The pipeline consumes an image, the original natural-language query, the per-target instance masks, and the source closed-vocabulary labels, and produces **(i)** a free-form sub-category phrase for each target and **(ii)** a global set-level concept summarizing the target scope. We use `Qwen3-VL-235B-A22B` [2] as the annotator throughout and run it as a two-stage pipeline followed by multi-step quality control.

**Stage 1: Diverse Sub-category Labeling.** For each target, the annotator is shown two views, the original image and the same image overlaid with a colored mask highlighting the target, and is conditioned on the source closed-vocabulary label. It is instructed to produce a natural noun phrase of 1–6 words that may include visual attributes (e.g., color, material) when these are clearly visible,

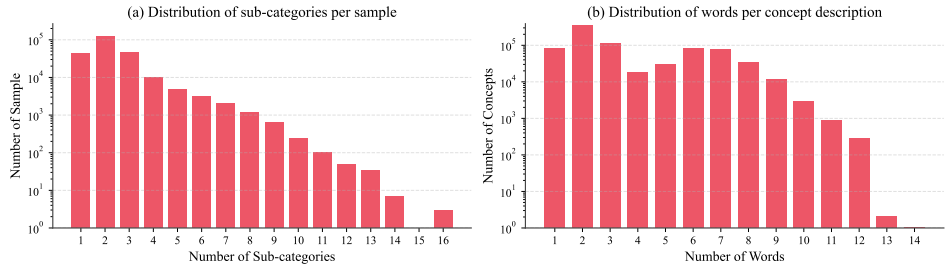



Figure 8: **Statistics of the training corpus** produced by the proposed two-stage hierarchical annotation pipeline, including the per-sample distribution of sub-categories, and the length distribution of concept phrases.

with the goal of replacing rigid category names (e.g., *light*) by more contextually accurate phrases (e.g., *traffic light*, *red bicycle*). The source label is supplied explicitly as an anchor to suppress semantic drift, so that the model refines, but never silently invents, a new category. We sample at a low temperature of $0.2$ to favor faithfulness over diversity.

**Stage 2: Set-level Concept Synthesis.** Given the image, the original query, and the sub-category phrases produced by Stage 1, the annotator is asked to summarize the overall scope of the referred targets into a single *set-level concept*, i.e., an 8–15 word free-form description of the scenario, behavior, or purpose that the targets collectively express (e.g., *the essentials for a carefree day by the water under strong sun*). The output is constrained to lie within $[5, 20]$ words and is truncated when over the upper bound; outputs falling below the lower bound are replaced by a query-conditioned fallback template. We sample at a moderate temperature of $0.7$ to encourage linguistic diversity. The set-level concept produced here, together with the per-target phrases from Stage 1, fully specifies the hierarchical supervision used by SETCON.

**Quality Control.** The two annotation stages already enforce light in-line cleaning, including stripping of Markdown markup, URLs, quotes, and boilerplate prefixes (e.g., *The answer is:*), as well as length capping and rejection of overly generic phrases (e.g., *object, thing, item*). On top of this, we run a three-step post-processing pipeline that explicitly targets the failure modes we observed in practice:

- **Label sanity check.** We identify broken sub-category phrases through two complementary scans. A rule-based pass first catches unambiguous degeneracies, including system-level error tokens (e.g., *OMPI*, *EPHIR*, *MPI_Init*), residual Markdown markup, URLs, leading exclamation marks, and outputs shorter than two characters. The remaining unique entries are then aggregated into a global vocabulary and reviewed by a text-only LLM judge, which flags non-English characters, gibberish, code snippets, full sentences, apologetic or meta-commentary phrases, and disallowed punctuation. Flagged entries are collected into a global blacklist for the next step.

- **Targeted re-generation.** For each sample whose sub-category phrases include a blacklisted entry, we re-invoke the visual annotator with the same image and mask overlay used in Stage 1 to regenerate, rather than wholesale reverting them to the source label. This step preserves the linguistic richness of Stage 1 wherever the original output was salvageable.

- **Within-sample de-duplication.** A common failure mode is that two distinct sub-category groups under the same query are independently relabeled to the same phrase (e.g., *two "coffee cup" groups on the same table*). For each detected duplicate, the annotator is asked to decide, given the image and the two highlighted mask groups, whether to *merge* them into a single sub-category or *split* them into two distinguishable ones. Merging applies when the source dataset has over-split a coherent semantic group, or when the two groups in fact share the same fine-grained semantics; splitting applies when the two groups are visually or functionally distinct but happen to share the same closed-vocabulary label. In the former case, the two mask groups are consolidated under one shared label; in the latter, the duplicate labels are replaced with distinctly worded alternatives. A second de-duplication pass is run afterwards to eliminate any residual collisions.

We additionally conduct multiple rounds of manual spot-checking on random subsets to verify annotation faithfulness; systematic failure modes observed during these rounds prompted iterative refinement of the prompts and filters above.

Table 6: **Robustness to segmentation backends.** SETCON consistently outperforms previous methods across different segmentation backends.

| Model | MUSE val | | | MUSE test | | |
|---|---|---|---|---|---|---|
| | gIoU | cIoU | F1@0.5 | gIoU | cIoU | F1@0.5 |
| PixelLM | 42.6 | 50.7 | – | 39.2 | 46.3 | – |
| SAM 3 Agent | 27.8 | 24.2 | 41.0 | 23.8 | 26.3 | 38.1 |
| VisionReasoner | 42.4 | 41.6 | 54.5 | 36.1 | 36.8 | 49.2 |
| SETCON (GroundingDINO) | 50.6 | 51.6 | 63.6 | 50.1 | 56.6 | 64.4 |
| **SETCON (SAM 3)** | **56.9** | **59.5** | **71.2** | **53.2** | **60.0** | **69.3** |

Table 7: **Robustness to judge models.** Diverse labeling and hierarchical semantics yield consistent gains across all judges.

| Div. Label | Hier. Sem. | Qwen3-VL | GPT-5.5 | Gemini | Claude | Doubao |
|---|---|---|---|---|---|---|
| ✗ | ✗ | 5.99 | 5.99 | 5.75 | 5.97 | 5.80 |
| ✓ | ✗ | 6.89 | 6.71 | 6.57 | 6.80 | 6.40 |
| ✓ | ✓ | **6.93** | **6.76** | **6.85** | **6.86** | **6.69** |

**Data Statistics.** The final corpus contains $236{,}396$ samples and $784{,}809$ concept phrases, with an average of $2.32$ sub-categories per sample and a heavy tail reaching up to $16$ categories in the most complex cases. Over $80\%$ of the samples involve more than one sub-category, supplying the set-level supervision that prior referring corpora lack, and the average concept phrase has $3.5$ words, confirming that the resulting labels behave as natural-language descriptions rather than closed-vocabulary tokens. Detailed distributions over the per-sample number of sub-categories and the length of concept phrases are summarized in Fig. 8.

# C Evaluation Details

For RefCOCO/+/g and ReasonSeg, we use the highest-confidence prediction as the final mask and report cIoU. For the multi-object benchmarks gRefCOCO and MUSE, we discard predicted masks whose confidence is below $0.7$ before aggregation: for gRefCOCO, following [38], we merge the retained masks into a single foreground mask and report gIoU and cIoU; for MUSE, we align the retained mask set with the ground truth via Hungarian matching, report gIoU and cIoU for segmentation quality, and additionally report F1@0.5 for set-level detection performance. For video benchmarks, we track each predicted target over time, take the union of all tracked masks in each frame as the final foreground sequence, and report standard $\mathcal{J}\&\mathcal{F}$.

For the annotation-quality evaluation in Tab. 4, we adopt `Qwen3-VL-8B-Instruct` [2] as the judge model. The LLM Score evaluates how accurately and descriptively a predicted label represents its corresponding visual region. For each predicted mask, we match it to the labeled GT mask, highlight the matched region in the original image using a bounding box, and provide the judge with the image, predicted label, and the dataset's original closed-vocabulary label as a semantic reference. The scoring scale and evaluation criteria are defined as follows:

- **Scale**: The score ranges from 1 to 10, with 6 representing a correct but generic label.
- **Scoring rubric**: 9–10 — highly specific, rich, and visually grounded details; 7–8 — correct label with additional attributes, states, functions, relations, or scene context; 6 — correct subject with no additional visual detail; 4–5 — visually related to or confused with a similar category; 2–3 — incorrect object with some scene or attribute association; 1 — unrelated to the highlighted region.

# D Additional Experiments

## D.1 Robustness to Segmentation Backends

To investigate whether SETCON relies on a specific segmentation backend, we replace SAM 3 with GroundingDINO [40] while keeping the remaining pipeline unchanged. Since GroundingDINO produces bounding boxes rather than segmentation masks, we use SAM to convert the detected boxes into masks. As shown in Tab. 6, SETCON with GroundingDINO still consistently outperforms previous methods. This demonstrates that the improvements mainly originate from the proposed set-level concept formulation rather than a particular segmentation backend.

## D.2 Robustness to Judge Models

To examine whether the annotation-quality results in Tab. 4 are affected by same-family evaluation bias, we randomly sample 500 items from the MUSE test set and repeat the evaluation under the same protocol using Qwen3-VL-8B-Instruct, GPT-5.5, Gemini-3.1-Flash-Lite, Claude-Sonnet-4.6,

Table 8: **Image inference efficiency.** SETCON achieves substantially lower latency and computational cost than the compared methods.

| Method | Params | Memory (GiB) | Latency (ms) | FLOPs (TFLOPs) |
|---|---|---|---|---|
| PixelLM | 7.15B | 13.7 | 1,877 | 9.9 |
| SAM 3 Agent | 9.62B | 23.4 | 17,070 | – |
| VisionReasoner | 8.29B | 16.9 | 3,349 | 20.0 |
| **SETCON (ours)** | **9.62B** | **18.6** | **878** | **5.5** |

Table 9: **Video inference efficiency.** SETCON achieves competitive throughput close to native SAM 3 inference.

| Method | Params | Memory (GiB) | FPS |
|---|---|---|---|
| SAM 3 | 860M | 11.6 | 11.65 |
| GLUS | 7.29B | 18.8 | 9.22 |
| VISA | 7.70B | 28.3 | 4.63 |
| **SETCON (ours)** | **9.62B** | **22.1** | **10.16** |

and Doubao-Seed-2.0-lite as judges. As shown in Tab. 7, diverse labeling consistently improves semantic-label quality across all five model families. Hierarchical semantics provides a further gain for every judge, confirming that the trend is robust to the choice of evaluation model.

## E Efficiency Analysis

We conduct an efficiency analysis of SETCON against representative segmentation methods for both image and video tasks. All experiments are performed using one NVIDIA H200 GPU, 14 CPU cores, and 208 GiB of RAM. As reported in Tab. 8 and Tab. 9, SETCON achieves significantly lower latency and computational cost for image inference, while maintaining competitive throughput for video inference.

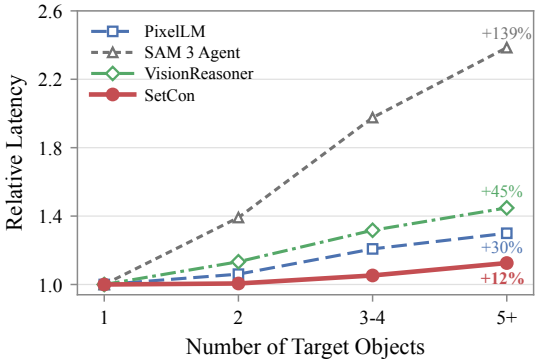



Figure 9: **Latency scaling with the number of target objects.** SETCON scales favorably as the number of target objects increases, indicating efficient inference in complex multi-object scenarios.

We further investigate the latency scaling behavior with respect to the number of target objects. As shown in Fig. 9, SETCON increases latency by only 12% when scaling from one object to five or more objects, compared with 30%, 45%, and 139% for PixelLM, VisionReasoner, and SAM 3 Agent, respectively. This favorable scaling is attributed to SETCON's unified prediction interface, which avoids object-by-object inference and enables efficient multi-object segmentation.

## F Limitations

Despite its promising results, our work still leaves room for improvement. First, open-ended queries can have ambiguous target boundaries; explicit concepts make the model's assumptions interpretable, and future work could incorporate interactive clarification to resolve such ambiguity. Second, our hierarchical annotations, built with model-assisted pipelines and quality control, may still contain noise and have limited coverage of rare long-tail cases. More diverse data sources and finer-grained human validation could further improve robustness and open-world generalization.

## G Broader Impact

SETCON improves fine-grained visual perception, with potential applications in accessibility, image and video editing, AR/VR interaction, and embodied perception. Like other vision-language systems, however, its deployment in surveillance, biometric identification, or other high-stakes settings raises concerns around privacy, consent, and fairness. The explicit concept interface introduced in this work makes grounding decisions more interpretable and easier to audit than implicit-token alternatives, but it does not eliminate biases inherited from large-scale pretraining or source segmentation datasets, which are largely based on web imagery and may underrepresent certain regions, demographics, and contexts. We encourage the responsible use of our dataset and method, and explicitly discourage any applications that may infringe upon personal privacy or be deployed for harmful purposes.

All training and evaluation datasets used in this work are publicly released for academic research. We use them strictly under their original licenses, and our hierarchical semantic annotations are derived only from publicly available samples. All annotations and experimental results were generated solely for research purposes and follow ethical guidelines for the use of public data in academic research.