REFCON: LEARNING WHAT TO INCLUDE & EXCLUDE FOR MLLM SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have enabled language-guided image segmentation, including referring expression segmentation. However, MLLM-based segmentation methods remain referent-centric; when exclusion information is used, it is typically encoded as sparse point prompts rather than explicit non-referent image-patch representations maintained throughout decoding. Consequently, explicit non-referent image-patch evidence generally remains absent from the MLLM-generated representation used for dense prediction, leaving the referent represented primarily through what should be included. We introduce *RefCon*, a framework that represents each referent through complementary MLLM-generated referent and non-referent image-patch representations and maintains both throughout decoding for repeated contrast. Given an image and instruction, the MLLM autoregressively generates two image-grounded token sets: *Referent Tokens*, corresponding to referred-object patches, and *Contrast Tokens*, corresponding to non-referent patches. To exploit this complementary evidence, we introduce a *Referent–Contrast Decoder (RCD)* that repeatedly refines both branches with image features and explicitly contrasts their token and dense-feature representations throughout refinement. To structure these token sets for role separation and non-redundant coverage, we further propose a *Region-Aware Contrastive Transport (RCT) loss*: its contrastive component aligns Referent Tokens with foreground evidence and Contrast Tokens with non-referent evidence while separating their roles, whereas its optimal-transport component reduces within-set redundancy and encourages collective coverage without assigning predefined subregions to individual tokens. Together, RefCon introduces explicit referent–non-referent representation into MLLM segmentation, with complementary inclusion and exclusion evidence generated by the MLLM and repeatedly contrasted throughout dense decoding.

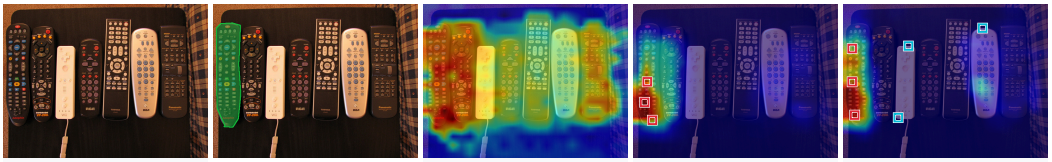
1 INTRODUCTION

Fine-grained image understanding has long been a central problem in computer vision. Early pixel-level prediction relied on specialized segmentation architectures, from convolutional models such as U-Net (Ronneberger et al., 2015) and DeepLabV3 (Chen et al., 2017) to more recent designs such as SegFormer (Xie et al., 2021) and SegNeXt (Guo et al., 2022). Referring expression segmentation (RES) extended this setting to language-guided prediction, requiring models to identify and segment objects specified by natural-language expressions. Methods such as RMI (Liu et al., 2017), LAVT (Yang et al., 2022), and CRIS (Wang et al., 2022) progressively strengthened vision–language interaction and text–pixel alignment, but remained largely dependent on task-specific architectures and supervision, limiting their ability to exploit the broader semantic knowledge and reasoning capabilities of large pretrained multimodal models.

The emergence of multimodal large language models (MLLMs) brought these capabilities into language-guided segmentation. LISA (Lai et al., 2024) introduced a dedicated `[SEG]` token for referring and reasoning segmentation, followed by extensions to multi-target reasoning (Ren et al., 2024), grounded conversational generation (Rasheed et al., 2024), generalized referring segmentation (Xia et al., 2024), and false-premise queries (Wu et al., 2024). In parallel, several works strengthened how referent information is transferred to dense prediction: READ (Qian et al., 2025) and UGround (Qian et al., 2026) derive spatial prompts from `[SEG]`–image similarity, while Claw-

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.

Machine (Ma et al., 2025) predicts multiple discrete visual tokens and PaDT (Su et al., 2026) dynamically incorporates current-image patch representations into the LLM vocabulary, enabling autoregressive prediction of referent-associated image patches alongside text. Together, these methods progressively enrich the explicit representation of the referent.



(a) Input image (b) Ground truth (c) UGround (d) PaDT (e) RefCon (ours)

Figure 1: **MLLM referent representation and token activation maps.** RefCon represents each referent using Referent Tokens (red boxes) & Contrast Tokens (blue boxes), yielding cleaner target-aligned activations.

Despite this progress in representing the referent, existing MLLM-based segmentation methods largely rely on referent-oriented representations, without pairing them with explicit non-referent image-patch representations and maintaining both for repeated contrast throughout decoding. This leaves the model relying primarily on referent-side evidence, without complementary non-referent cues that can sharpen referent discrimination and support more precise segmentation. READ (Qian et al., 2025) derives positive and negative points from a `[SEG]`–image similarity map, while SAM4MLLM (Chen et al., 2024) generates or classifies positive and negative points for SAM. These signals enter SAM as sparse prompt embeddings that encode spatial cues rather than local non-referent visual content, and do not form a separate non-referent representation repeatedly contrasted with the referent during decoding. PaDT (Su et al., 2026), in contrast, autoregressively generates image-patch representations, but only for the referent. GSVA (Xia et al., 2024) and SESAME (Wu et al., 2024) handle absent referents, but do not spatially represent non-referent regions when a valid referent exists. Thus, prior methods may provide sparse exclusion cues or explicit referent image-patch representations, but not a complementary non-referent representation that persists throughout decoding. Consequently, the MLLM-generated representation used for dense prediction remains primarily referent-oriented, limiting the model's ability to progressively distinguish the referent from non-referent content throughout dense prediction.

To address this gap, we introduce *RefCon*, which represents each referred object through complementary MLLM-generated *Referent Tokens* and *Contrast Tokens*, corresponding to referred-object and non-referent image patches, respectively. Unlike exclusion information supplied only as decoder prompts (Kirillov et al., 2023; Ravi et al., 2025; Qian et al., 2025; Chen et al., 2024), these representations are maintained throughout dense prediction, allowing the referent representation to be progressively refined relative to explicit non-referent evidence. We therefore introduce a *Referent–Contrast Decoder (RCD)* composed of successive *Referent–Contrast Refinement (RCR)* layers, each with two stages: *Independent Two-Way Refinement* and *Cross-Stream Contrast*. The first independently refines the referent and contrast branches with separate copies of the image features, allowing each to specialize to its corresponding evidence without exchanging information. The second explicitly contrasts the token representations and spatial image features of the two branches, allowing each to be updated relative to evidence from the other. Repeating this process across RCR layers progressively refines the referent stream relative to non-referent evidence rather than introducing exclusion information only once at the decoder input. Effective repeated contrast depends on the two token sets remaining faithful to their respective regions without becoming spatially redundant. We therefore introduce a *Region-Aware Contrastive Transport (RCT) loss*: its contrastive component aligns Referent Tokens with foreground evidence and Contrast Tokens with non-referent evidence while separating their roles, whereas its optimal-transport component reduces within-set redundancy and encourages collective coverage of the corresponding regions without assigning predefined sub-regions to individual tokens.

The main contributions of this work are summarized as follows:

- We propose **RefCon**, which represents each referred object through complementary MLLM-generated *Referent Tokens* and *Contrast Tokens*, explicitly capturing image-patch evidence from both referred-object and non-referent regions.
- We propose a **Referent–Contrast Decoder (RCD)** that maintains separate referent and contrast streams across successive RCR layers and repeatedly contrasts their token representations and spatial image features, using explicit non-referent evidence to strengthen referent–non-referent discrimination and support more precise segmentation.
- We introduce a **Region-Aware Contrastive Transport (RCT) loss** that encourages separation between Referent and Contrast representations through contrastive supervision, while optimal transport reduces within-set redundancy and encourages collective coverage of their corresponding regions.

2 Related Work

Pre-MLLM Visual Grounding and Segmentation: Before MLLMs, segmentation relied on specialized architectures such as U-Net (Ronneberger et al., 2015), DeepLabV3 (Chen et al., 2017), SegFormer (Xie et al., 2021), and SegNeXt (Guo et al., 2022). Self-supervised learning (Atito et al., 2025; Nandam et al., 2024; 2025b; Atito et al., 2023) and vision–language pretraining (Radford et al., 2021; Jia et al., 2021) further advanced few-shot, zero-shot, and open-vocabulary segmentation (Nandam et al., 2025a;c; 2026). For language-guided prediction, RMI (Liu et al., 2017), CRIS (Wang et al., 2022), and LAVT (Yang et al., 2022) improved vision–language interaction, while MDETR (Kamath et al., 2021), GLIP (Li et al., 2022), and Grounding DINO (Liu et al., 2024) extended grounding to free-form language. These approaches nevertheless relied largely on task-specific architectures rather than pretrained MLLMs.

MLLM-based Visual Grounding, Detection, and Segmentation: LISA (Lai et al., 2024) introduced MLLM-based segmentation through a dedicated `[SEG]` token. GSVA (Xia et al., 2024) extends this to multiple or absent referents, while SESAME (Wu et al., 2024) addresses false-premise queries. READ (Qian et al., 2025) and UGround (Qian et al., 2026) transfer `[SEG]` information to dense prediction through point and soft-mask prompts, respectively, with READ deriving points from `[SEG]`–image similarity and UGround using intermediate similarity features. SAM4MLLM (Chen et al., 2024) generates or classifies positive and negative point prompts. Seg-R1 (You and Wu, 2025) and CoPRS (Lu et al., 2026) explore reinforcement-learning-based spatial reasoning, while Shikra (Chen et al., 2023) and Kosmos-2 (Peng et al., 2023) represent locations through coordinates or location tokens. More recently, ClawMachine (Ma et al., 2025) predicts discrete visual tokens for referred regions, while PaDT (Su et al., 2026) incorporates current-image patch representations into the LLM vocabulary for autoregressive generation of referent-associated image patches.

Despite these advances, existing MLLM-based segmentation methods largely remain referent-centric. LISA-style approaches represent the referred object through a `[SEG]` token, while PaDT (Su et al., 2026) autoregressively generates image-patch representations only for the referent. READ (Qian et al., 2025) and SAM4MLLM (Chen et al., 2024) incorporate exclusion information through positive and negative point prompts, which encode sparse spatial cues rather than explicit non-referent visual content and do not form an explicit non-referent image-patch representation that is maintained alongside the referent throughout decoding. Without explicit non-referent image-patch representations generated by the MLLM, exclusion evidence is not carried into dense prediction as a complementary representation alongside the referent, limiting the model's ability to progressively distinguish the referent from non-referent content. GSVA (Xia et al., 2024) and SESAME (Wu et al., 2024) address referent absence, but do not spatially represent non-referent regions when a valid referent exists. Background information has also been explored in conventional segmentation through additional supervision in GRSer (Wu et al., 2023) and background prototypes in NTRE (Liu et al., 2022), but not as complementary MLLM-generated referent and non-referent image-patch representations maintained throughout decoding. In contrast, RefCon represents each referent through complementary MLLM-generated Referent and Contrast Tokens, maintains both throughout prediction, and repeatedly contrasts their token representations and spatial image features during refinement. Its RCT objective further separates their roles while reducing within-set redundancy and encouraging collective coverage of their corresponding regions.

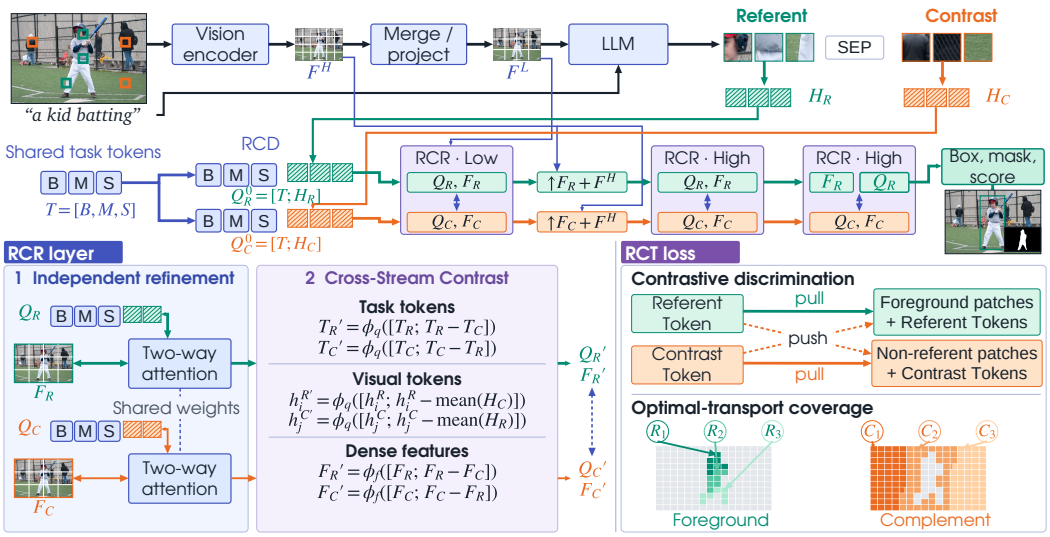


Figure 2: **Overview of RefCon.** Given an image and referring instruction, the MLLM autoregressively generates complementary Referent and Contrast Tokens, whose hidden states initialize the two-stream Referent–Contrast Decoder (RCD). Successive Referent–Contrast Refinement (RCR) layers perform *Independent Two-Way Refinement* and *Cross-Stream Contrast* over low- and high-resolution image features $\mathbf{F}^L$ and $\mathbf{F}^H$. The final referent task tokens $q_{\text{box},R}$, $q_{\text{mask},R}$, and $q_{\text{score},R}$ predict the bounding box, segmentation mask, and confidence score, respectively. Region-Aware Contrastive Transport (RCT) aligns Referent and Contrast Tokens with their corresponding regions while reducing within-set redundancy.

3 Methodology

We propose *RefCon*, a framework for fine-grained referring image tasks that represents each referred object through complementary image-patch representations of referent and non-referent regions. Given an image and referring instruction, the MLLM autoregressively generates two sets of image-patch tokens: *Referent Tokens*, corresponding to referred-object patches, and *Contrast Tokens*, corresponding to non-referent patches. Existing formulations either generate image-patch representations only for the referent (Su et al., 2026) or encode negative information as sparse point prompts (Qian et al., 2025; Chen et al., 2024); neither represents a referred object through complementary referent and non-referent image-patch representations nor maintains both throughout decoding for repeated contrast. RefCon carries both representations into dense prediction, allowing non-referent evidence to repeatedly inform referent–non-referent discrimination throughout refinement and thereby support more precise segmentation.

Generating complementary referent and non-referent representations alone is insufficient unless the decoder can first ground each representation in its corresponding image evidence and then exploit their interaction during dense prediction. We therefore introduce a *Referent–Contrast Decoder (RCD)* composed of successive *Referent–Contrast Refinement (RCR)* layers, each containing *Independent Two-Way Refinement* followed by *Cross-Stream Contrast*. The first stage independently refines the referent and contrast branches with separate copies of the image features, allowing each to specialize to its corresponding evidence without prematurely mixing their roles. However, because the branches remain independent after this refinement, neither can use evidence from the other to disambiguate referent from non-referent regions. The second stage therefore explicitly contrasts their token representations and spatial image features, allowing each branch to be updated relative to evidence from the other. Repeating these two stages across RCR layers progressively strengthens referent–non-referent discrimination as the spatial representation is refined. Effective repeated contrast further requires the two token sets to remain faithful to their corresponding regions without becoming spatially redundant. We therefore introduce a *Region-Aware Contrastive Transport (RCT) loss*: its contrastive component separates the two token roles and aligns them with their corresponding regions, while its optimal-transport component reduces within-set redundancy and encourages

each token set to collectively cover its corresponding region without assigning predefined subregions to individual tokens. An overview of our architecture is illustrated in Fig. 2.

3.1 Referent–Contrast Generation and Decoding

Given an image $I \in \mathbb{R}^{h\times w\times 3}$ and a referring instruction q, we build on a Qwen-based MLLM whose vision encoder produces high-resolution features $\mathbf{F}^H \in \mathbb{R}^{H_H\times W_H\times d_H}$. Spatial merging and projection further produce low-resolution features $\mathbf{F}^L \in \mathbb{R}^{H_L\times W_L\times d_L}$, where $(H_H, W_H) = (h/p, w/p)$ and $(H_L, W_L) = (h/(sp), w/(sp))$; p and s denote the vision-encoder patch size and spatial merging factor, respectively, and d_H, d_L are the corresponding channel dimensions. $\mathbf{F}^L$ is passed to the LLM for multimodal reasoning, while $\mathbf{F}^H$ is retained for fine-grained decoding. To enable image patches to be autoregressively predicted alongside text, each low-resolution patch feature $f_n^L \in \mathbb{R}^{d_L}$ is associated with a visual token π_n whose embedding is $\mathrm{VP}(f_n^L)$, where VP denotes projection into the LLM embedding space, following PaDT (Su et al., 2026).

Existing MLLM segmenters primarily generate referent-oriented representations, either through a dedicated `[SEG]` token (Lai et al., 2024) or multiple referent-associated visual tokens (Su et al., 2026). RefCon instead represents each referred object through complementary image-patch evidence. The MLLM autoregressively generates *Referent Tokens* corresponding to referred-object patches and *Contrast Tokens* corresponding to non-referent patches. Representing both sides explicitly provides the decoder with local visual evidence for what belongs to the referent and what does not, rather than requiring exclusion information to be inferred only from the referent representation.

For supervision, let $M^L \in \{0,1\}^{H_L\times W_L}$ denote the ground-truth mask of a referred object on the low-resolution patch grid, flattened over $N = H_L W_L$ locations. We form $\Omega_R = \{\pi_n \mid M_n^L = 1\}$ and $\Omega_C = \{\pi_n \mid M_n^L = 0\}$, and randomly sample K_R Referent Tokens $\{r_i\}_{i=1}^{K_R}$ and K_C Contrast Tokens $\{c_j\}_{j=1}^{K_C}$ from the corresponding sets. Their autoregressive target and resulting final-layer LLM hidden states are

$$\begin{aligned} \mathcal{Y}_{\text{vis}} &= [r_1, \ldots, r_{K_R}, \texttt{<|CT_SEP|>}, c_1, \ldots, c_{K_C}], \\ \mathbf{H}_R &= [h_1^R, \ldots, h_{K_R}^R], \qquad \mathbf{H}_C = [h_1^C, \ldots, h_{K_C}^C]. \end{aligned} \tag{1}$$

Here, `<|CT_SEP|>` separates the two token sets. Random sampling exposes the model to different foreground and non-referent patches across training iterations rather than a fixed subset of locations.

Generating both representations, however, does not by itself make their complementary information useful during dense prediction. Each representation must first be grounded in its corresponding image evidence so that the referent and non-referent roles remain distinct; subsequently, information must be exchanged between them so that non-referent evidence can help disambiguate the referent. Existing SAM-based approaches provide negative information through point prompts (Qian et al., 2025; Chen et al., 2024), whereas PaDT (Su et al., 2026) decodes multiple generated image-patch representations only for the referent; neither maintains complementary referent and non-referent image-patch streams for such repeated interaction. We therefore introduce a *Referent–Contrast Decoder (RCD)* composed of successive *Referent–Contrast Refinement (RCR)* layers. Each RCR layer first independently grounds the two streams in their corresponding image evidence and then explicitly contrasts them, allowing the updated non-referent evidence to repeatedly help distinguish the referent as dense spatial representations are progressively refined.

RCD maintains separate referent and contrast streams. We define shared initial task tokens $\mathbf{Q}_{\text{task}}^0 = [q_{\text{box}}^0, q_{\text{score}}^0, q_{\text{mask}}^0]$ for bounding-box, confidence-score, and mask prediction, and initialize

$$\mathbf{Q}_R^0 = [\mathbf{Q}_{\text{task}}^0, \mathbf{H}_R], \qquad \mathbf{Q}_C^0 = [\mathbf{Q}_{\text{task}}^0, \mathbf{H}_C], \qquad \mathbf{F}_R^0 = \mathbf{F}_C^0 = \mathbf{F}^L. \tag{2}$$

For RCR layer $l = 1, \ldots, L$, the first stage, *Independent Two-Way Refinement*, independently grounds each stream in its corresponding image evidence using the shared two-way cross-attention block $\mathcal{T}^l$:

$$(\bar{\mathbf{Q}}_R^l, \bar{\mathbf{F}}_R^l) = \mathcal{T}^l(\mathbf{Q}_R^{l-1}, \mathbf{F}_R^{l-1}), \qquad (\bar{\mathbf{Q}}_C^l, \bar{\mathbf{F}}_C^l) = \mathcal{T}^l(\mathbf{Q}_C^{l-1}, \mathbf{F}_C^{l-1}). \tag{3}$$

Although initialized from the same image features, different generated tokens condition the two streams to specialize to referent and non-referent evidence without prematurely mixing their roles. Independent refinement alone, however, cannot use evidence from the opposite stream to resolve referent–non-referent ambiguity. The second stage, *Cross-Stream Contrast*, therefore updates each



(a) Input / Ground Truth (b) Without RCT (c) + OT (d) + OT + Contrastive

Figure 3: **Effect of RCT supervision.** Optimal-transport supervision improves spatial coverage and reduces within-set redundancy, while adding contrastive supervision further separates Referent (R) and Contrast (C) Tokens over referent and non-referent regions.

stream relative to the other. We define $\mathcal{C}_\phi(A, B) = \phi([A; A - B])$, where $[\cdot;\cdot]$ denotes feature concatenation and ϕ is a learnable projection. Here, A preserves the current representation while $A - B$ exposes its directional difference from the other stream. We apply it to task tokens, generated visual tokens, and dense image features, updating the prediction state, token representations, and spatial representation relative to evidence from the opposite stream.

Let $\bar{\mathbf{Q}}^l_{R,\text{task}}$ and $\bar{\mathbf{Q}}^l_{C,\text{task}}$ denote the task-token portions of the intermediate streams. Since their prediction roles correspond directly, they are contrasted element-wise:

$$\mathbf{Q}^l_{R,\text{task}} = \mathcal{C}_{\phi^l_q}(\bar{\mathbf{Q}}^l_{R,\text{task}}, \bar{\mathbf{Q}}^l_{C,\text{task}}), \qquad \mathbf{Q}^l_{C,\text{task}} = \mathcal{C}_{\phi^l_q}(\bar{\mathbf{Q}}^l_{C,\text{task}}, \bar{\mathbf{Q}}^l_{R,\text{task}}). \tag{4}$$

Let $\bar{\mathbf{H}}^l_R$ and $\bar{\mathbf{H}}^l_C$ denote the Referent- and Contrast-Token portions of the intermediate streams. Because their numbers may differ and no one-to-one correspondence is assumed, each token is contrasted against the mean of the opposite set:

$$h^{R,l}_i = \mathcal{C}_{\phi^l_q}\left(\bar{h}^{R,l}_i, \text{mean}(\bar{\mathbf{H}}^l_C)\right), \qquad h^{C,l}_j = \mathcal{C}_{\phi^l_q}\left(\bar{h}^{C,l}_j, \text{mean}(\bar{\mathbf{H}}^l_R)\right). \tag{5}$$

With $\mathbf{H}^l_R = [h^{R,l}_1, \ldots, h^{R,l}_{K_R}]$ and $\mathbf{H}^l_C = [h^{C,l}_1, \ldots, h^{C,l}_{K_C}]$, the updated streams are $\mathbf{Q}^l_R = [\mathbf{Q}^l_{R,\text{task}}, \mathbf{H}^l_R]$ and $\mathbf{Q}^l_C = [\mathbf{Q}^l_{C,\text{task}}, \mathbf{H}^l_C]$. Because the dense image representations share the same spatial layout, they are contrasted directly at corresponding locations:

$$\mathbf{F}^l_R = \mathcal{C}_{\phi^l_f}(\bar{\mathbf{F}}^l_R, \bar{\mathbf{F}}^l_C), \qquad \mathbf{F}^l_C = \mathcal{C}_{\phi^l_f}(\bar{\mathbf{F}}^l_C, \bar{\mathbf{F}}^l_R), \tag{6}$$

where ϕ^l_q and ϕ^l_f are the learnable projections for token and dense-feature contrasting, respectively. Thus, each RCR layer first specializes both streams to their corresponding evidence and then uses cross-stream differences to update the prediction state and spatial representation. Repeating this interaction lets non-referent evidence continue to disambiguate the referent as both representations evolve through decoding. Between the first and second RCR layers, we inject the retained high-resolution features:

$$\hat{\mathbf{F}}^1_R = \mathbf{F}^H + \text{Up}(\mathbf{F}^1_R), \qquad \hat{\mathbf{F}}^1_C = \mathbf{F}^H + \text{Up}(\mathbf{F}^1_C), \tag{7}$$

where $\text{Up}(\cdot)$ denotes bilinear interpolation from $H_L \times W_L$ to $H_H \times W_H$. These features replace $\mathbf{F}^1_R$ and $\mathbf{F}^1_C$ as the dense inputs to the second RCR layer, allowing subsequent referent–contrast refinement to operate on finer spatial detail. After the final RCR layer, the referent-stream task tokens $q^L_{\text{box},R}$, $q^L_{\text{score},R}$, and $q^L_{\text{mask},R}$ produce the bounding-box, confidence-score, and mask predictions, respectively; the contrast stream contributes only through refinement and produces no independent output.

3.2 Region-Aware Contrastive Transport Loss

Autoregressive supervision determines which image-patch tokens are generated as Referent and Contrast Tokens, but does not explicitly structure their MLLM hidden states according to these roles. We therefore impose a contrastive objective that aligns Referent Tokens with foreground evidence and Contrast Tokens with non-referent evidence while separating the two sets. Let $\mathbf{H}_R = \{h^R_i\}_{i=1}^{K_R}$ and $\mathbf{H}_C = \{h^C_j\}_{j=1}^{K_C}$. We define

$$\begin{aligned} \mathcal{P}_R(i) &= \{h^R_k \mid k \neq i\} \cup \Omega_R, \qquad & \mathcal{N}_R(i) &= \mathbf{H}_C \cup \Omega_C, \\ \mathcal{P}_C(j) &= \{h^C_k \mid k \neq j\} \cup \Omega_C, \qquad & \mathcal{N}_C(j) &= \mathbf{H}_R \cup \Omega_R. \end{aligned} \tag{8}$$

Table 1: Referring Expression Segmentation (RES) results in cIoU (%). RefCon uses Qwen2.5-VL backbones.

Model	Params	RefCOCO			RefCOCO+			RefCOCOg		Overall
		val	testA	testB	val	testA	testB	val	test	
VLT (Ding et al., 2021)	–	67.5	70.5	65.2	56.3	61.0	50.1	55.0	57.7	60.4
CRIS (Wang et al., 2022)	–	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4	64.3
LAVT (Yang et al., 2022)	7B	72.7	75.8	68.8	62.1	68.4	55.1	65.0	66.0	66.7
Seg-R1 (You and Wu, 2025)	3B	69.9	76.0	64.9	59.1	66.8	50.9	67.3	67.9	65.4
PaDT (Su et al., 2026)	3B	81.3	81.5	82.2	77.6	79.4	76.3	78.1	78.5	79.4
SAMTok (Zhou et al., 2026)	3B	82.4	83.9	79.9	78.4	81.8	74.8	79.0	79.1	79.9
RefCon (Ours)	**3B**	**85.1**	**85.4**	**85.6**	**81.7**	**83.3**	**80.3**	**82.4**	**82.8**	**83.3**
LISA (Lai et al., 2024)	7B	74.1	76.5	71.1	62.4	67.5	56.5	66.4	68.5	67.9
PixelLM (Ren et al., 2024)	7B	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5	69.2
GSVA (Xia et al., 2024)	7B	77.2	78.9	73.5	65.9	69.6	59.8	72.7	73.3	71.4
READ (Qian et al., 2025)	7B	78.1	80.2	73.2	68.4	73.7	60.4	70.1	71.4	71.9
SegLLM (Wang et al., 2025)	7B	80.2	81.5	75.4	70.3	73.0	62.5	72.6	73.6	73.6
GLaMM (Rasheed et al., 2024)	7B	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	75.6
Text4Seg + SAM (Lan et al., 2025)	7B	78.0	80.9	74.6	71.6	77.3	66.0	74.8	74.7	74.7
UGround (Qian et al., 2026)	7B	80.6	83.5	77.7	72.8	77.5	65.6	74.7	76.1	76.1
PaDT (Su et al., 2026)	7B	86.0	86.1	86.4	82.5	84.1	80.7	83.5	83.3	84.1
RefCon (Ours)	**7B**	**88.2**	**88.3**	**88.9**	**84.5**	**86.6**	**83.1**	**86.6**	**86.4**	**86.6**

After ℓ_2 normalization, for token type $s \in \{R, C\}$,

$$\mathcal{L}_i^s = -\frac{1}{|\mathcal{P}_s(i)|} \sum_{p \in \mathcal{P}_s(i)} \log \frac{\exp(\mathrm{sim}(h_i^s, p)/\tau)}{\sum_{a \in \mathcal{P}_s(i) \cup \mathcal{N}_s(i)} \exp(\mathrm{sim}(h_i^s, a)/\tau)}, \qquad \tau = 0.07, \tag{9}$$

where $\mathrm{sim}(\cdot, \cdot)$ denotes cosine similarity. The overall contrastive objective is

$$\mathcal{L}_{\mathrm{ctr}} = \frac{1}{K_R} \sum_{i=1}^{K_R} \mathcal{L}_i^R + \frac{1}{K_C} \sum_{j=1}^{K_C} \mathcal{L}_j^C. \tag{10}$$

While $\mathcal{L}_{\mathrm{ctr}}$ establishes role-consistent Referent and Contrast representations, it does not structure how tokens within each role distribute over their corresponding region; multiple tokens may therefore concentrate on the same discriminative evidence and become spatially redundant. We therefore add an optimal-transport objective that matches each token set to its corresponding region. Let $\mathbf{C}_s = 1 - \mathrm{sim}(\mathbf{H}_s, \Omega_s)$ denote the pairwise cosine-distance cost matrix. We define

$$\mathcal{L}_{\mathrm{OT}} = \frac{1}{2} \left[\mathcal{L}_{\mathrm{IPOT}}(\mathbf{C}_R) + \mathcal{L}_{\mathrm{IPOT}}(\mathbf{C}_C) \right], \tag{11}$$

where $\mathcal{L}_{\mathrm{IPOT}}$ denotes the IPOT objective (Xie et al., 2020). This optimal-transport loss reduces within-set redundancy and encourages collective coverage of broader spatial evidence without assigning predefined subregions to individual tokens. The resulting Region-Aware Contrastive Transport objective is

$$\mathcal{L}_{\mathrm{RCT}} = w_{\mathrm{ctr}} \mathcal{L}_{\mathrm{ctr}} + w_{\mathrm{OT}} \mathcal{L}_{\mathrm{OT}}. \tag{12}$$

After the final RCR layer, the referent-stream task tokens $q^L_{\mathrm{box},R}$, $q^L_{\mathrm{score},R}$, and $q^L_{\mathrm{mask},R}$ produce the bounding-box, confidence-score, and mask predictions. The box and score tokens predict the box coordinates and confidence score, while the mask token interacts with the final referent-conditioned features $\mathbf{F}_R^{\mathrm{out}}$ to produce the segmentation mask. These outputs are supervised by $\mathcal{L}_{\mathrm{bbox}}$, $\mathcal{L}_{\mathrm{score}}$, and $\mathcal{L}_{\mathrm{mask}}$ and jointly optimized with $\mathcal{L}_{\mathrm{RCT}}$; detailed prediction-head and loss formulations are provided in the supplementary material.

4 Experimental Results

We evaluate RefCon on referring expression segmentation (RES) using RefCOCO, RefCOCO+, and RefCOCOg; generalized referring expression segmentation (GRES) on gRefCOCO; and false-premise referring segmentation on FP-RefCOCO(+/g). We report cIoU for RES, gIoU, cIoU, and no-target accuracy (N-acc.) for GRES, and *See* accuracy and *Segment* cIoU for FP-RefCOCO(+/g).

Table 2: Generalized Referring Expression Segmentation (GRES) results on gRefCOCO (%).

Model	Params	Validation			TestA			TestB		
		gIoU	cIoU	N-acc.	gIoU	cIoU	N-acc.	gIoU	cIoU	N-acc.
CRIS (Wang et al., 2022)	–	56.27	55.34	–	63.42	63.82	–	51.79	51.04	–
LAVT (Yang et al., 2022)	–	58.40	57.64	49.32	65.90	65.32	49.25	55.83	55.04	48.46
ReLA (Liu et al., 2023)	–	63.60	62.42	56.37	70.03	69.26	59.02	61.02	59.88	58.40
LISA (Vicuna) (Lai et al., 2024)	7B	61.63	61.76	54.67	66.27	68.50	50.01	58.84	60.63	51.91
GSVA (Vicuna) (Xia et al., 2024)	7B	66.47	63.29	62.43	71.08	69.93	65.31	62.23	60.47	60.56
SAM4MLLM (Chen et al., 2024)	8B	71.86	67.83	66.08	74.15	72.22	63.92	65.29	63.42	59.99
UGround (LLaVA-1.5) (Qian et al., 2026)	7B	72.46	65.56	74.53	74.29	70.87	73.93	66.85	61.84	71.22
SAMTok (Zhou et al., 2026)	7B	78.00	72.10	79.50	78.50	77.30	75.90	69.60	67.70	68.60
RefCon (Ours)	**3B**	79.69	82.09	**81.90**	77.03	78.53	**76.75**	75.39	79.03	**72.37**
RefCon (Ours)	**7B**	**80.22**	**82.56**	80.60	**78.70**	**79.61**	76.15	**77.03**	**79.85**	71.17



(a) Input + GT (b) UGround (c) PaDT (d) RefCon (Ours)

Figure 4: **Qualitative comparison on RefCOCO val.** Ground truth is shown in green and predictions in red. RefCon better captures the referred region than UGround and PaDT.

For training, we jointly use COCO2017, RefCOCO, RefCOCO+, RefCOCOg, Referring Image Captioning (RIC), gRefCOCO, and FP-RefCOCO(+/g). RefCon is instantiated with Qwen2.5-VL 3B and 7B backbones. We use $K_R = K_C = 5$ Referent and Contrast Tokens. The RCD comprises three RCR layers: the first operates on low-resolution features, followed by high-resolution feature injection and two high-resolution RCR layers. We set $\tau = 0.07$, $w_{\text{ctr}} = 0.1$, and $w_{\text{OT}} = 1.0$. Training uses AdamW with a learning rate of 2×10^{-5} and a global batch size of 128, together with gradient checkpointing and bfloat16 mixed precision. Additional implementation details, including training and inference prompt formats, REC results, and complete baseline comparisons are provided in the supplementary material.

Referring Expression Segmentation. Table 1 reports results on RefCOCO, RefCOCO+, and RefCOCOg. RefCon consistently improves over prior methods across all evaluation splits and both model scales. With the 3B backbone, RefCon achieves an overall cIoU of 83.3, outperforming PaDT by 3.9 points; scaling to 7B further improves the overall cIoU to 86.6, exceeding PaDT by 2.5 points. The consistent gains across RefCOCO, RefCOCO+, and RefCOCOg indicate that explicitly modeling both referent and non-referent evidence benefits fine-grained referring segmentation.

Generalized Referring Expression Segmentation. Table 2 evaluates RefCon on gRefCOCO, where models must handle multiple or absent referents. RefCon-7B achieves the best gIoU and cIoU on all three splits, reaching 80.22/82.56 on Validation, 78.70/79.61 on TestA, and 77.03/79.85 on TestB, with cIoU gains over SAMTok of 10.46, 2.31, and 12.15 points, respectively. RefCon-3B also obtains the highest N-acc. on all three splits. Since no Referent/Contrast tokens are decoded for no-target cases, N-acc. reflects the model's no-target recognition ability rather than the Referent–Contrast decoding process.

False-Premise Referring Segmentation. We further evaluate robustness to invalid referring expressions on the augmented FP-RefCOCO(+/g) benchmarks in Table 3. RefCon improves both premise recognition and segmentation quality. RefCon-7B achieves Segment cIoU scores of 74.13, 70.52, and 74.03 on FP-RefCOCO, FP-RefCOCO+, and FP-RefCOCOg, improving over UGround by 11.33, 14.49, and 15.48 points, respectively. Since invalid-premise cases do not decode Referent/Contrast tokens, See reflects premise recognition; the segmentation gains on valid referents are consistent with explicitly maintaining non-referent evidence during decoding.

Component Ablation. Table 3(b) isolates the effects of referent–contrast representation, Cross-Stream Contrast, and RCT. The 5- and 10-token referent-only settings use a single referent stream in the decoder for prediction. Within this setting, increasing the number of referent tokens from 5 to

Table 3: **False-premise results and analysis.** (a) Results on FP-RefCOCO (FP-RC), FP-RefCOCO+ (FP-RC+), and FP-RefCOCOg (FP-RCg), reporting *See* accuracy and *Segment* cIoU (%). (b) Component ablations on RefCOCO+ (RC+) testB and RefCOCOg (RCg) val. (c) Post-hoc SAM2 refinement; *Pred.* denotes the final RefCon prediction and PM denotes a pseudo mask.

(a) False-Premise Segmentation

Method	FP-RC		FP-RC+		FP-RCg	
	See	Seg.	See	Seg.	See	Seg.
ReLA (Liu et al., 2023)	65.11	47.47	64.00	37.61	58.15	35.15
SEEM (Zou et al., 2023)	51.36	37.92	51.32	40.01	51.25	35.87
LISA-7B (Lai et al., 2024)	51.36	44.00	51.32	39.62	51.25	39.64
Cascading (Wu et al., 2024)	75.59	55.18	75.03	48.64	76.07	49.98
SESAME (Wu et al., 2024)	79.84	57.93	80.00	50.81	81.78	53.79
READ (Qian et al., 2025)	82.87	61.50	83.51	54.54	84.67	56.12
RobustSeg (Li et al., 2026)	83.37	59.57	83.00	52.91	84.21	54.76
UGround (Qian et al., 2026)	85.86	62.80	85.10	56.03	86.86	58.55
RefCon-3B	**89.16**	72.28	**88.90**	68.87	**89.42**	71.42
RefCon-7B	88.65	**74.13**	87.73	**70.52**	89.39	**74.03**

(b) Component Ablation

Method	RC+ testB	RCg val
Ref-only (5 tok.)	80.7	83.5
Ref-only (10 tok.)	78.3	79.8
R-only + SAM2 pts.	77.2	79.7
Indep. R/C + M sub.	79.8	81.7
RCD w/o RCT	81.5	84.6
RCD + $\mathcal{L}_{\mathrm{OT}}$	82.2	85.6
RCD + RCT	**83.1**	**86.6**

(c) Post-hoc SAM2 Refinement

Split	Source	Pre	+SAM2
RC+ testB	Ours	83.1	**83.6**
	Ours PM	49.8	**57.6**
	PaDT PM	39.8	43.9
RCg val	Ours	86.6	**87.1**
	Ours PM	50.8	**59.6**
	PaDT PM	42.5	48.6

10 reduces cIoU from 80.7/83.5 to 78.3/79.8, showing that the gains do not arise simply from more tokens. Further, adding positive and negative SAM2 point prompts fails to improve this setting, indicating that sparse negative prompting is insufficient. Using five Referent and five Contrast Tokens with only Independent Two-Way Refinement and output-level mask subtraction obtains 79.8/81.7, showing that explicit non-referent representation alone is insufficient when the streams remain independent. RCD without RCT improves this to 81.5/84.6, demonstrating the benefit of Cross-Stream Contrast. Adding $\mathcal{L}_{\mathrm{OT}}$ reaches 82.2/85.6, while full RCT achieves 83.1/86.6. Fig. 3 further shows that OT improves spatial coverage and contrastive supervision separates Referent and Contrast representations. Further ablations on the RCD design, Referent/Contrast token count, and the loss weights w_{OT} and w_{ctr} are provided in the supplementary material.

Post-hoc SAM2 Refinement. Table 3(c) evaluates whether RefCon outputs can further benefit from SAM2 on RefCOCO+ testB and RefCOCOg val. Applying SAM2 to the final RefCon prediction gives modest gains from 83.1/86.6 to 83.6/87.1. We also evaluate decoder-free pseudo masks constructed directly from generated-token activations. For PaDT, we average the heatmaps of its five predicted referent tokens and threshold the result. For RefCon, we separately average the five Referent-Token and five Contrast-Token heatmaps and compute $\max(\mathcal{H}_R - \mathcal{H}_C, 0)$, followed by normalization and thresholding. RefCon's pseudo masks achieve 49.8/50.8 cIoU, compared with 39.8/42.5 for PaDT, and improve to 57.6/59.6 after SAM2 refinement. Fig. 1 qualitatively shows the same effect through cleaner target-aligned token activations.

5 Conclusion

We introduced RefCon, which represents each referred object using explicit image-patch evidence from both referent and non-referent regions. This representation is carried throughout prediction: the MLLM generates Referent and Contrast Tokens, the Referent–Contrast Decoder repeatedly contrasts the two streams during dense refinement, and Region-Aware Contrastive Transport supervision encourages both role separation and spatial coverage. RefCon delivers consistent gains across RES, GRES, and false-premise segmentation, while ablations show that these improvements cannot be reduced to additional visual tokens or conventional positive/negative prompting. More broadly, our results show that fine-grained MLLM segmentation benefits from explicitly representing non-referent image patches and using them throughout decoding to strengthen referent–non-referent discrimination, rather than relying only on referent evidence.

Reproducibility Statement

We provide the architectural and training details required to reproduce RefCon throughout Sec. 3 and the supplementary material. The main paper specifies the Referent–Contrast generation procedure, Referent–Contrast Decoder, Region-Aware Contrastive Transport objective, training datasets, model configurations, and principal optimization settings. The supplementary material provides further implementation details, including the prediction heads and complete training objective, training and inference prompt formats, and additional experimental and ablation settings. We report results using standard publicly available benchmarks and their established evaluation protocols.

AI Use Statement

Generative AI tools were used to assist with manuscript drafting and editing, software implementation and debugging, and the execution of experiments. AI-assisted code and experimental procedures were reviewed and verified by the authors, and all reported results were obtained, checked, and validated by the authors. Generative AI was not used to develop the research methodology, design the experiments, interpret the results, or generate or modify the datasets used in this work. The authors take full responsibility for the final content of the paper and all reported results.

References

Sara Atito, Muhammad Awais, Srinivasa Nandam, and Josef Kittler. Gmml is all you need. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2125–2129. IEEE, 2023. doi: 10.1109/ICIP49359.2023.10222150.

Sara Atito, Josef Kittler, Imran Razzak, and Muhammad Awais. Cg-ssl: Concept-guided self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4823.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic, 2023. URL https://arxiv.org/abs/2306.15195.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. URL https://arxiv.org/abs/1706.05587.

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.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021.

Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0084.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021.

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.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, October 2023.

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.

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 *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vkakKdznFS.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

Xinzhuo Li, Adheesh Juvekar, Jiaxun Zhang, Xingyou Liu, Muntasir Wahed, Kiet A Nguyen, Yifan Shen, Tianjiao Yu, and Ismini Lourentzou. Counterfactual segmentation reasoning: Diagnosing and mitigating pixel-grounding hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 7450–7460, 2026.

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. doi: 10.1109/CVPR52729.2023.02259.

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.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55, 2024. doi: 10.1007/978-3-031-72970-6_3.

Yuanwei Liu, Nian Liu, Qinglong Cao, Xiwen Yao, Junwei Han, and Ling Shao. Learning non-target knowledge for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11573–11582, 2022.

Zhenyu Lu, Liupeng Li, Jinpeng Wang, Yan Feng, Bin Chen, Ke Chen, and Yaowei Wang. Coprs: Learning positional prior from chain-of-thought for reasoning segmentation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4048c740bec4c682b17c640f584d2762-Abstract-Conference.html.

Tianren Ma, Lingxi Xie, Yunjie Tian, Boyu Yang, and Qixiang Ye. Clawmachine: Learning to fetch visual tokens for referential comprehension. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TOtk9dTYGG.

Srinivasa Rao Nandam, Sara Atito, Zhenhua Feng, Josef Kittler, and Muhammad Awais. Pseudo labelling for enhanced masked auto encoders. In *35th British Machine Vision Conference 2024, BMVC 2024*. BMVA, 2024. URL https://papers.bmvc2024.org/0737.pdf.

Srinivasa Rao Nandam, Sara Atito, Zhenhua Feng, Josef Kittler, and Muhammad Awais. Enhanced weakly supervised few-shot classification & segmentation. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025a. doi: 10.1109/ICASSP49660.2025.10890308.

Srinivasa Rao Nandam, Sara Atito, Zhenhua Feng, Josef Kittler, and Muhammad Awais. Investigating self-supervised methods for label-efficient learning. *International Journal of Computer Vision*, 133:4522–4537, 2025b. doi: 10.1007/s11263-025-02397-4.

Srinivasa Rao Nandam, Sara Atito, Zhenhua Feng, Josef Kittler, and Muhammad Awais. Text augmented correlation transformer for few-shot classification & segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25357–25366, 2025c. doi: 10.1109/CVPR52734.2025.02361.

Srinivasa Rao Nandam, Sara Atito, Zhenhua Feng, Josef Kittler, and Muhammad Awais. Object-centric refinement for enhanced zero-shot segmentation. In *International Conference on Learning Representations*, volume 2026, pages 77790–77809, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/7db2ffcbfd0bd361d47b7fa612bd2ba2-Paper-Conference.pdf.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023. URL https://arxiv.org/abs/2306.14824.

Rui Qian, Xin Yin, and Dejing Dou. Reasoning to attend: Try to understand how ¡seg¿ token works. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24722–24731, 2025.

Rui Qian, Xin Yin, Chuanhang Deng, Zhiyuan Peng, Jian Xiong, Wei Zhai, and Dejing Dou. Uground: Towards unified visual grounding with unrolled transformers. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2510.03853.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

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.

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, pages 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, 2015. doi: 10.1007/978-3-319-24574-4_28.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model, 2025. URL https://arxiv.org/abs/2504.07615.

Yongyi Su, Haojie Zhang, Shijie Li, Nanqing Liu, Jingyi Liao, Junyi Pan, Yuan Liu, Xiaofen Xing, Chong Sun, Chen Li, Nancy F. Chen, Shuicheng Yan, Xulei Yang, and Xun Xu. Patch-as-decodable-token: Towards unified multi-modal vision tasks in mllms. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.01954.

Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities, 2023. URL https://arxiv.org/abs/2305.11172.

XuDong Wang, Shaolun Zhang, Shufan Li, Konstantinos Kallidromitis, Kehan Li, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Segllm: Multi-round reasoning segmentation with large language models. In *International Conference on Learning Representations*, 2025.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11686–11695, 2022.

Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E. Gonzalez, and Trevor Darrell. See, say, and segment: Teaching lmms to overcome false premises. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13459–13469, 2024.

Yixuan Wu, Zhao Zhang, Chi Xie, Feng Zhu, and Rui Zhao. Advancing referring expression segmentation beyond single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2628–2638, 2023.

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.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact wasserstein distance. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 433–453. PMLR, 22–25 Jul 2020.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336, 2023.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. 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.

Zuyao You and Zuxuan Wu. Seg-r1: Segmentation can be surprisingly simple with reinforcement learning, 2025. URL https://arxiv.org/abs/2506.22624.

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*, pages 37852–37863, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A Training and Implementation Details

We instantiate RefCon with Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct. We follow the full-parameter supervised fine-tuning recipe of (Su et al., 2026) on COCO2017, RefCOCO, RefCOCO+, RefCOCOg, and Referring Image Captioning (RIC), using 4 epochs for the 3B model and 6 epochs for the 7B model. We then perform one additional epoch of LoRA adaptation on the

LLM with rank $r = 16$ and $\alpha = 32$, augmenting the training mixture with gRefCOCO and FP-RefCOCO(+/g), to support GRES and false-premise tasks, while retaining the same optimization settings. We use a global batch size of 128, gradient checkpointing, and bfloat16 mixed precision. Training is performed on AMD Instinct MI250X GPUs with 64 GB of HBM. A single RefCon checkpoint is used for evaluation across all tasks and benchmarks.

Unless otherwise specified, we use $K_R = K_C = 5$ Referent and Contrast Tokens, sampled randomly from their corresponding patch sets during training. The RCD contains three RCR layers: the first operates on low-resolution image features, followed by high-resolution feature injection and two high-resolution RCR layers. For RCT, we use $\tau = 0.07$, $w_{\text{ctr}} = 0.1$, and $w_{\text{OT}} = 1.0$. For $\mathcal{L}_{\text{OT}}$, we use the default solver settings of the authors' public IPOT implementation (Xie et al., 2020), with $\beta = 2$, $L = 1$ inner iteration, and 1000 outer iterations. At inference, we use greedy autoregressive decoding for visual-token generation, without explicitly enforcing distinct patch-token predictions. Table 4 summarizes the task-specific question and answer formats used for training.

For conventional referring expression segmentation, we evaluate on RefCOCO, RefCOCO+, and RefCOCOg using the standard evaluation protocol adopted by prior RES methods, including VLT (Ding et al., 2021), CRIS (Wang et al., 2022), LAVT (Yang et al., 2022), LISA (Lai et al., 2024), GSVA (Xia et al., 2024), GLaMM (Rasheed et al., 2024), READ (Qian et al., 2025), SegLLM (Wang et al., 2025), UGround (Qian et al., 2026), PaDT (Su et al., 2026), and SAMTok (Zhou et al., 2026). These benchmarks and splits are kept consistent with the comparisons reported in Table 1. We additionally evaluate generalized referring segmentation on gRefCOCO (Liu et al., 2023) and false-premise referring segmentation on FP-RefCOCO(+/g) (Wu et al., 2024).

Table 4: **Task-specific prompt formats.** *[Referent Tokens]* and *[Contrast Tokens]* denote the generated image-patch tokens corresponding to referent and non-referent regions, respectively.

Task	Question format	Answer format
REC / RES	`Please carefully check the image and detect the object this sentence describes: "{expression}".`	`The "{expression}" refers to` *[Referent Tokens]* `<|CT_SEP|>` *[Contrast Tokens]*
gRefCOCO: target present	`Locate the region described by the expression "{expression}" and return the complete matching region as a single segmentation result. If the expression does not match anything in the image, state that it cannot be found and stop.`	`The complete region matching "{expression}" is` *[Referent Tokens]* `<|CT_SEP|>` *[Contrast Tokens]*
gRefCOCO: no target	Same as above.	`The "{expression}" cannot be found in this image.`
FP-RefCOCO: valid premise	`Check whether the referring expression "{expression}" correctly describes an object in the image. If it does, locate and segment that object. If it does not, state that it cannot be found and provide the corrected referring expression.`	`The expression "{expression}" is valid and refers to` *[Referent Tokens]* `<|CT_SEP|>` *[Contrast Tokens]*
FP-RefCOCO: false premise	Same as above.	`The expression "{expression}" does not match an object in this image. The corrected referring expression is "{corrected_expression}".`

Table 5: Referring Expression Comprehension (REC) results in AP@0.5 (%).

Model	Params	RefCOCO			RefCOCO+			RefCOCOg		Overall
		val	testA	testB	val	testA	testB	val	test	
Grounding-DINO-L (Liu et al., 2024)	–	90.6	93.2	88.2	82.8	89.0	75.9	86.1	87.0	86.6
UNINEXT-H (Yan et al., 2023)	–	92.6	94.3	91.5	85.2	89.6	79.8	88.7	89.4	88.9
ONE-PEACE (Wang et al., 2023)	–	92.6	94.2	89.3	88.8	92.2	83.2	89.2	89.3	89.9
Qwen2.5-VL (Bai et al., 2025)	3B	89.1	91.7	84.0	82.4	88.0	74.1	85.2	85.7	85.0
VLM-R1 (Shen et al., 2025)	3B	90.1	92.3	85.2	84.2	89.4	76.8	85.6	86.8	86.3
PaDT (Su et al., 2026)	3B	96.0	95.5	95.0	91.8	94.8	88.4	93.6	94.0	93.6
RefCon (Ours)	**3B**	**96.2**	**97.4**	**95.1**	**92.4**	**95.1**	**89.2**	**93.7**	**94.1**	**94.2**
ClawMachineX (Ma et al., 2025)	7B	89.7	92.5	86.9	84.4	88.9	78.0	86.7	87.1	86.8
Qwen2.5-VL (Bai et al., 2025)	7B	90.0	92.5	85.4	84.2	89.1	76.9	87.2	87.2	86.6
InternVL3 (Zhu et al., 2025)	8B	92.5	94.6	88.0	88.2	92.5	81.8	89.6	90.0	89.6
PaDT (Su et al., 2026)	7B	96.6	**97.4**	95.6	**92.8**	95.2	89.4	94.6	94.2	94.5
RefCon (Ours)	**7B**	**96.7**	**97.4**	**96.0**	92.7	**95.6**	**89.8**	**94.8**	**94.5**	**94.7**
InternVL3 (Zhu et al., 2025)	78B	93.4	95.4	90.3	90.1	93.8	85.3	91.5	91.5	91.4

A.1 Prediction Heads and Training Objective

Following standard practice in referring segmentation and prompt-based mask prediction (Kirillov et al., 2023; Su et al., 2026), we use task-specific heads for box, confidence, and mask prediction together with the corresponding structured supervision objectives. After the final RCR layer, the referent-stream task tokens $q^L_{\mathrm{box},R}$, $q^L_{\mathrm{score},R}$, and $q^L_{\mathrm{mask},R}$ are used for box, confidence, and mask prediction, respectively. The box token is passed through a three-layer MLP with GELU activations after the first two layers and a sigmoid after the final four-dimensional layer, yielding the normalized box $\hat{\mathbf{b}} = [c_x, c_y, w, h]$. The score token is mapped through a linear layer to the scalar confidence prediction $\hat{s}$. For segmentation, the mask token is mapped through a three-layer MLP to a mask embedding. The final referent-conditioned dense features $\mathbf{F}^{\mathrm{out}}_R$ are upsampled through two learned $2\times$ upsampling stages, and their inner product with the mask embedding produces the segmentation logits $\hat{\mathbf{M}}$.

We likewise use standard structured objectives for mask, confidence, and bounding-box supervision. Given the ground-truth mask $\mathbf{M}$ and box $\mathbf{b}$, the segmentation loss combines Dice and focal losses:

$$\mathcal{L}_{\mathrm{mask}} = \mathcal{L}_{\mathrm{Dice}}\left(\hat{\mathbf{M}}, \mathbf{M}\right) + \mathcal{L}_{\mathrm{Focal}}\left(\hat{\mathbf{M}}, \mathbf{M}\right). \tag{13}$$

The confidence prediction is supervised using the detached GIoU between the predicted and ground-truth boxes:

$$\mathcal{L}_{\mathrm{score}} = \left\| \hat{s} - \mathrm{stopgrad}\left(\mathrm{GIoU}(\hat{\mathbf{b}}, \mathbf{b})\right) \right\|_2^2. \tag{14}$$

For bounding-box regression, we combine GIoU and CIoU with an IoU-weighted ℓ_1 term:

$$\mathcal{L}_{\mathrm{FL1}} = \left(1 - \mathrm{IoU}(\hat{\mathbf{b}}, \mathbf{b})\right)^2 \left\| \hat{\mathbf{b}} - \mathbf{b} \right\|_1, \tag{15}$$

and define

$$\mathcal{L}_{\mathrm{bbox}} = \mathcal{L}_{\mathrm{GIoU}}(\hat{\mathbf{b}}, \mathbf{b}) + \mathcal{L}_{\mathrm{CIoU}}(\hat{\mathbf{b}}, \mathbf{b}) + \mathcal{L}_{\mathrm{FL1}}. \tag{16}$$

For autoregressive visual-token supervision, we use the robust per-token cross-entropy formulation of Su et al. (2026), applied separately to Referent and Contrast Tokens. At a Referent-Token position, the valid image-patch set is Ω_R, while at a Contrast-Token position it is Ω_C. Let y_t denote the sampled target token and $z_{t,v}$ the logit for vocabulary token v at autoregressive step t. For a visual-token position of type $s \in \{R, C\}$, we define

$$\tilde{z}_{t,v} = \begin{cases} -\infty, & v \in \Omega_s \setminus \{y_t\}, \\ z_{t,v}, & \text{otherwise.} \end{cases} \tag{17}$$

Thus, other valid patches from the same region are excluded from the softmax competition rather than being treated as negatives. Textual positions, including <|CT_SEP|>, use standard cross-entropy supervision. We denote the resulting autoregressive objective by $\mathcal{L}_{\mathrm{SFT}}$.

Table 6: **RCD design ablations.** RES cIoU (%) on RefCOCO+ (RC+) testB and RefCOCOg (RCg) val. All variants use the same RefCon objective, including RCT.

RCD variant	RC+ testB	RCg val
Indep. R/C + R pred.	79.8	81.8
Late Cross-Stream Contrast	78.0	78.9
Indep. R/C + mask sub.	79.8	81.7
2 RCR layers	80.4	82.3
4 RCR layers	82.6	85.7
RCD w/o projection	79.8	81.8
Raw pair $[A; B]$	83.0	86.4
Full RCD	**83.1**	**86.6**

Table 7: **Transfer across MLLM backbones** RES cIoU (%) on RefCOCO+ (RC+) testB and RefCOCOg (RCg) val.

Method	Backbone	RC+	RCg
PaDT	Qwen2.5-VL-3B	76.3	78.1
SAMTok	Qwen2.5-VL-3B	74.8	79.0
SAMTok	Qwen3-VL-4B	76.6	80.7
RefCon	Qwen2.5-VL-3B	80.3	82.4
RefCon	Qwen3.5-4B	**80.5**	**82.9**

The complete RefCon objective combines these standard prediction objectives with the proposed Region-Aware Contrastive Transport supervision:

$$\mathcal{L} = \mathcal{L}_{\mathrm{SFT}} + \mathcal{L}_{\mathrm{bbox}} + \mathcal{L}_{\mathrm{mask}} + \mathcal{L}_{\mathrm{score}} + w_{\mathrm{ctr}}\mathcal{L}_{\mathrm{ctr}} + w_{\mathrm{OT}}\mathcal{L}_{\mathrm{OT}}. \tag{18}$$

The autoregressive, box, mask, and score losses are used without additional scaling; only the two RCT components are explicitly weighted by w_{ctr} and w_{OT}.

B Referring Expression Comprehension

We additionally report Referring Expression Comprehension (REC) results on RefCOCO, RefCOCO+, and RefCOCOg. As shown in Table 5, RefCon maintains strong localization performance while introducing explicit Referent–Contrast representation for segmentation. RefCon-3B improves over PaDT-3B on all eight evaluation splits, increasing the overall average from 93.6 to 94.2. The largest gains occur on RefCOCO testA and RefCOCO+ testB, where RefCon improves by 1.9 and 0.8 points, respectively.

RefCon-7B achieves an overall average of 94.7, compared with 94.5 for PaDT-7B. It improves or matches PaDT on seven of the eight splits, including gains on RefCOCO testB, RefCOCO+ testA/testB, and both RefCOCOg splits. These results show that the additional Contrast representation used by RefCon does not trade off the MLLM's ability to identify the referred object, while providing substantially stronger dense segmentation performance.

C More ablations

We further ablate the design of the Referent–Contrast Decoder (RCD) using RefCon-7B on RefCOCO+ testB and RefCOCOg val. All variants use the same RefCon training objective, including RCT, and modify only the decoder design. Without Cross-Stream Contrast, independently refining the Referent and Contrast streams and predicting from the Referent stream alone obtains 79.8/81.8 cIoU. Combining the same independently refined streams only at the output through mask subtraction gives a similar 79.8/81.7, while applying Cross-Stream Contrast only once to the final decoder features further reduces performance to 78.0/78.9. These results show that simply maintaining separate Referent and Contrast representations is not sufficient; their interaction needs to occur progressively throughout decoding.

Decoder depth shows a similar trend. Using two RCR layers gives 80.4/82.3, whereas increasing the depth to four layers reaches 82.6/85.7 but remains below the three-layer design at 83.1/86.6. Removing the learned projection from the contrast operation also degrades performance to 79.8/81.8. Finally, replacing the explicit contrast input $[A; A - B]$ with the raw pair $[A; B]$ reaches 83.0/86.4, close to but still below the full RCD. This suggests that the main benefit comes from progressive learned interaction between the Referent and Contrast streams, while explicitly providing the feature difference offers a small additional advantage in our formulation.

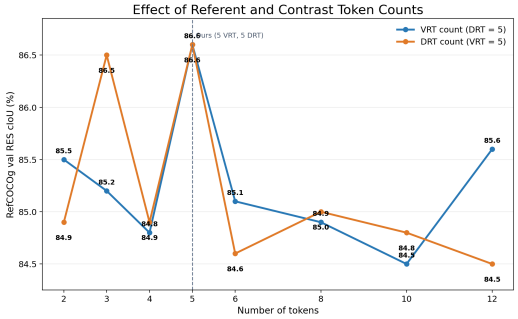


Figure 5: **Sensitivity to Referent and Contrast token counts.** RES cIoU (%) on RefCOCOg val when varying one token count while fixing the other to five.

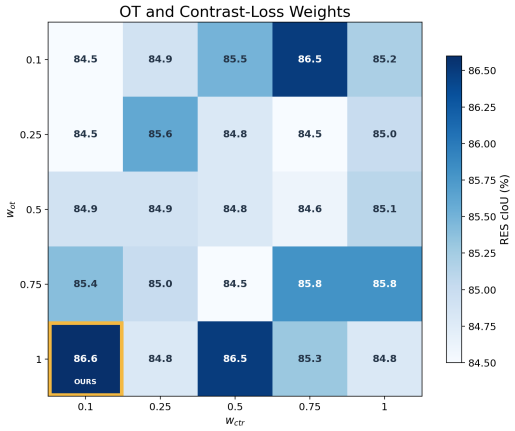


Figure 6: **Sensitivity to RCT loss weights.** RES cIoU (%) on RefCOCOg val for different w_{ctr} and w_{OT} settings. The main setting, $w_{\text{ctr}} = 0.1$ and $w_{\text{OT}} = 1.0$, is highlighted.

C.1 Transfer Across MLLM Backbones

We further evaluate RefCon with Qwen3.5-4B to examine whether its benefits transfer beyond the Qwen2.5-VL backbone. As shown in Table 7, we compare RefCon instantiated with Qwen2.5-VL-3B and Qwen3.5-4B against PaDT (Su et al., 2026) and SAMTok (Zhou et al., 2026) models of comparable scale.

RefCon transfers consistently to the Qwen3.5-4B backbone, achieving 80.5 and 82.9 cIoU on RefCOCO+ testB and RefCOCOg val, respectively. These results slightly improve over the Qwen2.5-VL-3B RefCon model at 80.3/82.4, showing that the gains are preserved when moving to a different MLLM family and scale. RefCon-4B also improves over SAMTok-4B by 3.9 and 2.2 points on the two splits, respectively, while RefCon-3B exceeds PaDT-3B by 4.0 and 4.3 points. These results support the applicability of the Referent–Contrast representation and decoding framework across different MLLM backbones.

Referent and Contrast Token Counts. Figure 5 studies the effect of varying the number of Referent or Contrast Tokens while fixing the other stream to five tokens. The 5/5 configuration achieves the best performance at 86.6 cIoU. Increasing the number of tokens does not consistently improve performance: for example, increasing the Referent count to 6, 8, or 10 gives 85.1, 84.9, and 84.5 cIoU, respectively, while increasing the Contrast count beyond five similarly provides no consistent gain. The results indicate that the improvement of RefCon does not arise simply from using more generated visual tokens, and we therefore use five Referent and five Contrast Tokens throughout the main experiments.

Sensitivity to RCT Loss Weights. Figure 6 evaluates the sensitivity of RefCon to the contrastive and optimal-transport loss weights w_{ctr} and w_{OT}. Performance remains strong across a broad range of settings, while the configuration used in our main experiments, $w_{\text{ctr}} = 0.1$ and $w_{\text{OT}} = 1.0$, achieves the best result of 86.6 cIoU. Increasing either weight does not yield a monotonic improvement, indicating that RefCon benefits from balancing contrastive separation with region-aware transport supervision. We therefore use $w_{\text{ctr}} = 0.1$ and $w_{\text{OT}} = 1.0$ throughout the main experiments.

C.2 Controlled Training-Mixture and Inference-Overhead Analysis

We further examine two potential confounding factors: whether the standard-RES improvements can be attributed to the additional training data used for generalized and false-premise referring segmentation, and whether the gains come at substantial parameter or inference cost. We define the *Base* training mixture as COCO2017, RefCOCO, RefCOCO+, RefCOCOg, and Referring Image Captioning (RIC), matching the training data used by Su et al. (2026). The *Extended* setting denotes

Table 8: Controlled training-mixture comparison on standard RES. Base denotes COCO2017, RefCOCO, RefCOCO+, RefCOCOg, and RIC. Results are cIoU (%).

Method	Training	RC+ testB	RCg val
PaDT	Base	80.7	83.5
RefCon	Base	83.0	86.5
RefCon	Extended	**83.1**	**86.6**

Table 9: Full-model parameter count and end-to-end evaluation time on RefCOCO+ testB (633 images) under the same 8-GPU setup.

Method	Params (B)	Full (s)	Per image (s)
PaDT	8.393	248.814	0.393
RefCon	8.412	258.358	0.408

the final RefCon setup, which additionally performs adaptation with gRefCOCO, FP-RefCOCO, FP-RefCOCO+, and FP-RefCOCOg.

Controlled training mixture. With the controlled Base mixture, RefCon achieves 83.0 and 86.5 cIoU on RefCOCO+ testB and RefCOCOg val, respectively, compared with 80.7 and 83.5 for PaDT. Thus, under the same training-data mixture, RefCon improves by 2.3 and 3.0 cIoU on the two splits. The subsequent adaptation with gRefCOCO and FP-RefCOCO(+/g) changes standard-RES performance only marginally, from 83.0/86.5 to 83.1/86.6. Hence, the standard-RES gains are already present before the additional generalized and false-premise adaptation and are not explained by the expanded training mixture.

This controlled result complements the component ablation in Table 3(b). Increasing the referent-only configuration from five to ten tokens reduces performance from 80.7/83.5 to 78.3/79.8, while adding SAM2 point prompts gives 77.2/79.7. Independent Referent/Contrast streams with output-level mask subtraction obtain 79.8/81.7, whereas the full RCD with RCT reaches 83.1/86.6. Together, these results indicate that the improvement is not readily explained by additional training data, by simply generating more visual tokens, or by a trivial increase in referent-only capacity.

Benchmark scope. For conventional referring expression segmentation, we follow the established RefCOCO, RefCOCO+, and RefCOCOg evaluation protocol used for direct comparison with prior work in Table 1. Beyond standard RES, we separately evaluate generalized referring segmentation on gRefCOCO and false-premise referring segmentation on FP-RefCOCO(+/g). Sec. C.1 further evaluates whether the proposed formulation transfers across different MLLM backbones.

Parameter and inference overhead. RefCon contains 8.412B parameters in total, compared with 8.393B for PaDT, so the additional Referent–Contrast machinery changes the size of the complete MLLM only slightly. We also measure end-to-end evaluation time on all 633 RefCOCO+ testB images using the same compute node, eight GPUs, and identical evaluation setup. PaDT requires 248.814 s in total, or 0.393 s per image, while RefCon requires 258.358 s, or 0.408 s per image. Thus, the explicit Referent–Contrast representation and decoding introduce only a modest inference-time overhead while providing the segmentation gains reported above.

D Additional Qualitative Results

Qualitative Analysis of Mask Predictions. Figure 7 provides additional qualitative comparisons with UGround and PaDT. The examples include challenging spatial descriptions, multiple candidate instances, and references to objects that occupy relatively small or partially visible regions. Across these cases, RefCon more consistently identifies the intended referent and recovers its complete spatial extent. For example, it better preserves the structure of the referred chair and bowl, selects the correct person among multiple candidate instances, and follows relative spatial cues such as "closest to us" and "on left." In contrast, the competing methods more frequently select an incorrect region, produce fragmented regions, or include substantial non-referent areas. These examples further illustrate the benefit of explicitly maintaining non-referent evidence when resolving fine-grained referring expressions.

Qualitative Analysis of Token-Derived Pseudo Masks. Figure 1 further visualizes the spatial evidence encoded by the generated visual tokens. For PaDT, the pseudo mask is constructed by

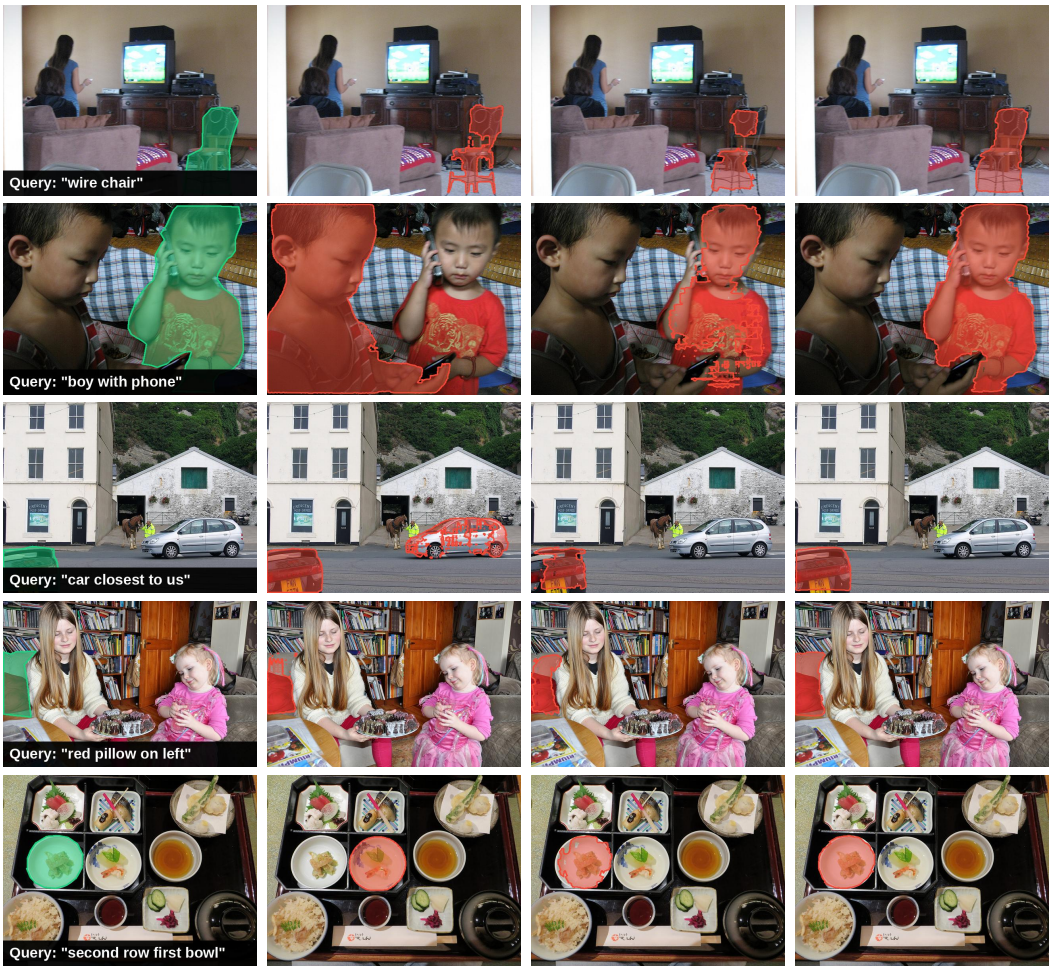


Figure 7: **Additional qualitative comparisons on RefCOCO val.** Each row shows the input image with the referring expression and ground-truth mask (green), followed by the predictions of UGround, PaDT, and RefCon (red). RefCon more consistently selects the intended referent while excluding non-referent regions.

averaging the activation maps of its five predicted referent tokens. For RefCon, we separately average the activation maps of the five Referent and five Contrast Tokens and form $\max(\mathcal{H}_R - \mathcal{H}_C, 0)$. Across the examples, PaDT often produces diffuse responses over non-referent objects or nearby regions, whereas RefCon yields more concentrated activation on the intended referent while suppressing non-referent regions. This is especially visible for the foreground chair and the right planter, where the Referent Tokens align with the target and the Contrast Tokens capture non-referent object and background evidence. These visualizations support the quantitative pseudo-mask results in Table 3(c), showing that explicit Referent–Contrast representation produces cleaner target-aligned spatial evidence even before the final segmentation decoder.

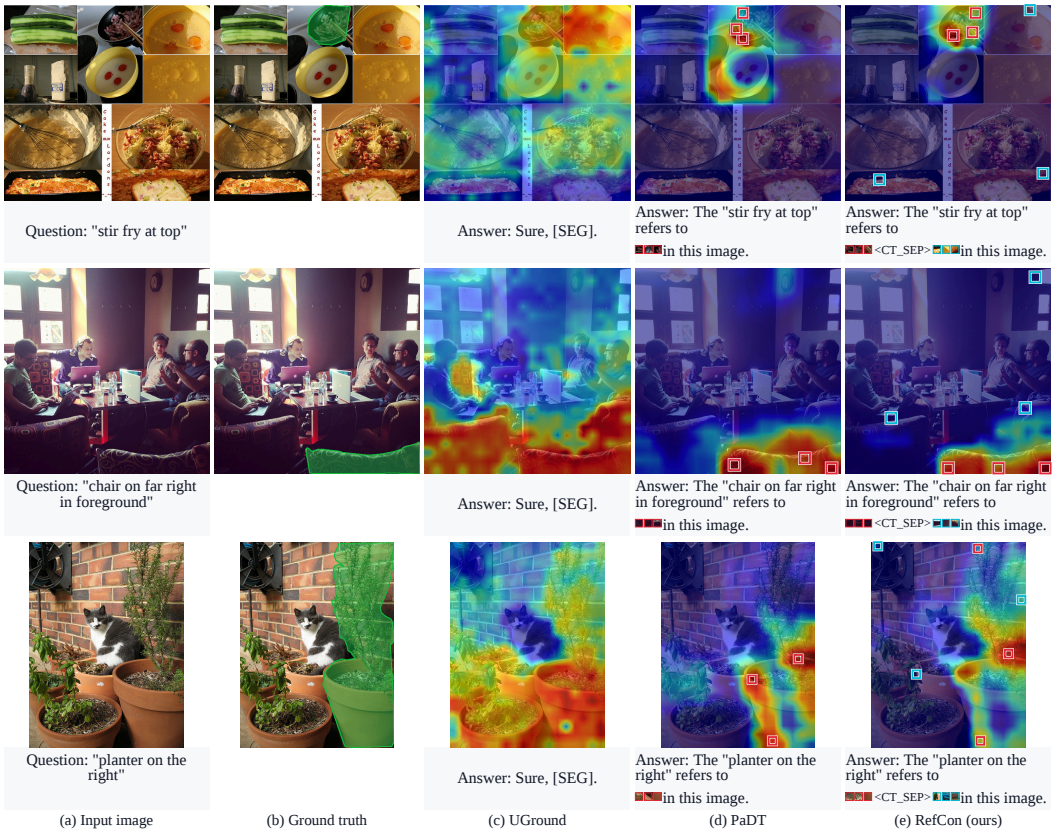


Figure 8: **Qualitative comparison of token-derived pseudo masks.** From left to right: input image, ground truth, UGround, PaDT, and RefCon. For RefCon, Referent-Token activations are contrasted with Contrast-Token activations to suppress non-referent regions, producing cleaner target-aligned spatial evidence.