HYPER-RIS: HYPERGRAPH CONSISTENCY FOR TRAINING-FREE REFERRING IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

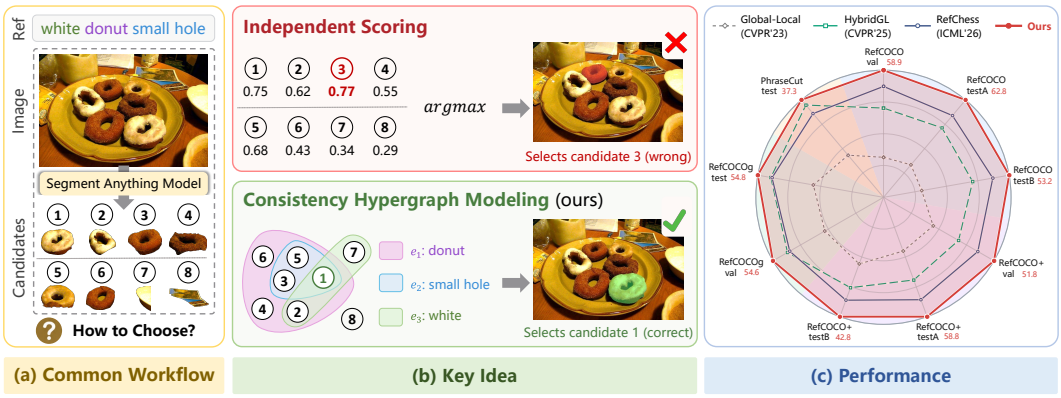

Figure 1: Overview and motivation of HYPER-RIS. (a) Training-free RIS commonly generates candidate masks with SAM and selects the referred object by ranking individual proposals. (b) Independent scoring can be confused by visually similar candidates and select a partially matched proposal. HYPER-RIS instead models visually related proposals as high-order groups and leverages hypergraph consistency to rerank them, recovering the correct referent. (c) This high-order reasoning consistently improves mIoU performance over existing methods across standard RIS benchmarks.

ABSTRACT

Training-free referring image segmentation (RIS) typically first generates class-agnostic mask candidates and then uses a pretrained vision–language model to select the candidate that best matches the referring expression. However, existing methods mainly score candidates independently or model only low-order interactions, making them prone to errors when multiple visually similar candidates compete for the same expression. We introduce HYPER-RIS, which improves candidate selection by explicitly modeling high-order consistency among candidate proposals. Our framework consists of two main components. First, Multi-Factor Scoring integrates complementary semantic cues from the subject, attributes, relations, and complete expression with dense spatial evidence to produce a unified score for each candidate. Second, Consistency Hypergraph Modeling uses the candidate scores and visual features to capture structured interactions among proposals. It first establishes pairwise visual context and then groups visually related candidates into hyperedges, where visually coherent groups provide consistency-weighted messages to their members. Candidate-wise reliability further suppresses conflicting group evidence, and the resulting high-order context is used to refine the proposal scores. Extensive experiments demonstrate that HYPER-RIS achieves state-of-the-art performance across multiple referring image segmentation benchmarks.

1 INTRODUCTION

Referring image segmentation (RIS) aims to identify, at the pixel level, the entity described by a free-form natural-language expression (Hu et al., 2016). Unlike category-level segmentation, RIS

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.

must distinguish the intended instance from other objects of the same category by jointly resolving its appearance, attributes, position, and relations to the surrounding scene. This requirement makes RIS an important interface between language understanding and dense visual perception, but it also makes supervision expensive: each training example requires both a precise mask and an unambiguous expression grounded to that mask (Yu et al., 2023). The annotation burden has consequently motivated zero-shot RIS, which seeks to perform referring segmentation without task-specific training on densely annotated RIS data.

The emergence of vision foundation models has made a training-free, proposal-based solution particularly attractive. CLIP provides transferable image–text representations learned from large-scale natural-language supervision (Radford et al., 2021), while the Segment Anything Model (SAM) can generate class-agnostic mask proposals without adapting to a particular referring dataset (Kirillov et al., 2023). As shown in Figure 1(a), a common training-free RIS pipeline uses SAM to generate candidate masks and then selects the referred candidate with a CLIP-based method, so improving the reliability of this selection becomes a core problem for training-free RIS.

Prior training-free RIS methods mainly improve the evidence assigned to each proposal and perform selection through *independent candidate scoring*. For example, Global-Local combines global and local visual–linguistic context (Yu et al., 2023), while TAS enriches SAM-based visual–text matching with positive and negative textual evidence and spatial rectification (Suo et al., 2023). Although these designs make individual proposal scores more informative, candidates are still largely evaluated in isolation, making it difficult to distinguish visually similar or overlapping alternatives that share strong partial matches with the referring expression. More recent methods have begun to recognize the importance of candidate interactions. HybridGL explicitly models pairwise spatial relations between target and reference-object candidates (Liu & Li, 2025), while RefChess evaluates proposals under sampled distractor contexts to favor candidates that remain stable during contextual search (Tong et al., 2026). These advances demonstrate the benefit of relational reasoning, yet their interactions remain primarily *low-order*: either pairwise between candidates or implicitly induced through sampled contexts. They therefore do not explicitly model a group of mutually related proposals as a unified reasoning unit, leaving higher-order candidate associations underexplored in ambiguous scenes where multiple competing proposals jointly determine the correct referent.

To address this limitation, we introduce HYPER-RIS, a training-free framework that explicitly models higher-order consistency among candidate proposals, as conceptually illustrated in Figure 1(b). We realize this idea through a training-free framework that formulates proposal selection as consistency-aware higher-order reasoning. Hypergraphs provide a natural representation for this purpose because a hyperedge can connect an arbitrary set of nodes, retaining relations that would be lost when complex interactions are compressed into pairwise edges (Zhou et al., 2006). Given SAM proposals, HYPER-RIS first performs *Multi-Factor Scoring*: it matches shared candidate visual features to complementary linguistic views of the expression, including the subject, attributes, relations, and the complete sentence, and combines these semantic cues with dense spatial heatmap evidence to obtain one fused score per candidate. It then performs *Consistency Hypergraph Modeling* using only the fused scores and candidate visual features. Pairwise visual relations establish a local context baseline, visually related candidates are organized into hyperedges, and each hyperedge aggregates a group-level score message weighted by its internal visual consistency. A candidate-wise reliability gate attenuates conflicting group messages, after which the resulting higher-order context is introduced as a residual correction to the original and pairwise-context scores. The final mask is selected from the reranked proposals, with no task-specific parameter learning.

Our contributions are threefold:

- We introduce HYPER-RIS, a training-free RIS framework that combines Multi-Factor Scoring and Consistency Hypergraph Modeling for reliable proposal selection.
- To the best of our knowledge, we are the first to explicitly model high-order associations among candidate proposals for training-free RIS, enabling group-wise consistency reasoning beyond independent and pairwise candidate modeling.
- As shown in Figure 1(c), HYPER-RIS achieves state-of-the-art performance across multiple RIS benchmarks, outperforming previous best training-free methods by 3.15–4.86 points in mIoU and 0.94–4.41 points in oIoU across the eight RefCOCO-family splits.

2 Related Work

2.1 Referring Image Segmentation

Referring image segmentation (RIS) aims to predict the pixel-level mask of an object described by a natural-language expression (Hu et al., 2016). Supervised methods improve cross-modal alignment through language-aware visual encoding, as in LAVT (Yang et al., 2022), or by transferring CLIP representations to dense prediction, as in CRIS (Wang et al., 2022). MRaCL further explores contrastive learning over fused image–text embeddings to better ground action-based descriptions (Kim et al., 2026). To reduce learning costs, DETRIS adopts parameter-efficient cross-modal adaptation (Huang et al., 2025), while WeakMCN jointly learns referring comprehension and segmentation with weak supervision (Cheng et al., 2025). Despite reducing adaptation or annotation requirements, these approaches still require task-specific training.

2.2 Training-Free Referring Image Segmentation

Training-free RIS combines pretrained visual and linguistic models to perform referring segmentation without task-specific training. A common paradigm separates mask proposal generation from region–text matching, with CLIP providing semantic representations and SAM generating class-agnostic proposals (Radford et al., 2021; Kirillov et al., 2023). Global-Local combines global and local context for proposal ranking (Yu et al., 2023). TAS augments region–text matching with textual and spatial cues (Suo et al., 2023). B2G exploits mid-layer vision–language representations to construct spatial maps for proposal reranking (An et al., 2026). RefAM derives grounding maps from diffusion-transformer cross-attention (Kukleva et al., 2026). These methods mainly strengthen the semantic or spatial evidence used for target localization and candidate selection.

More recent approaches further incorporate contextual information among candidates. HybridGL combines local and global visual context with spatial relations between target and reference-object candidates (Liu & Li, 2025). RefChess evaluates candidate stability under sampled distractor contexts using Monte Carlo Tree Search (Tong et al., 2026). These methods demonstrate the value of contextual reasoning for proposal selection. In contrast, Hyper-RIS explicitly groups visually related proposals into hyperedges and aggregates their scores through consistency- and reliability-aware high-order reasoning. Multi-Factor Scoring provides the initial semantic and spatial evidence, while Consistency Hypergraph Modeling refines proposal ranking through group-level messages.

2.3 Hypergraph Computation in Computer Vision

Hypergraphs capture high-order associations by connecting groups of vertices through hyperedges (Zhou et al., 2006). In computer vision, contextual hypergraph modeling captures relationships among pixels or regions for salient object detection (Li et al., 2013), while HGNN uses hyperedge convolution to learn representations for visual object recognition (Feng et al., 2019). HNN-HM models higher-order feature correspondences through an association hypergraph for visual matching (Liao et al., 2021). Recent works further apply hypergraph computation to model high-order visual relationships in object detection with YOLOv13 (Lei et al., 2025), image classification and retrieval with HgVT (Fixelle, 2025), and point cloud completion with Hyper-PCN (Li et al., 2026). These works motivate us to construct a consistency hypergraph over candidate masks, capturing high-order consistency among visually related candidates to guide referring mask selection.

3 Method

3.1 Overview

Given an image I and a referring expression q, Hyper-RIS selects the referred region from SAM proposals $\mathcal{M} = \{M_i\}_{i=1}^{N}$. For each proposal, we retain its pixels and fill the background with the CLIP pixel mean, preserving the full-image coordinate frame. A frozen vision–language encoder extracts a normalized feature $\mathbf{f}_i$ from each masked image, shared by semantic matching and consistency modeling. As shown in Figure 2, **Multi-Factor Scoring** combines semantic matching and spatial evidence into a proposal score s_i (Section 3.2). **Consistency Hypergraph Modeling**

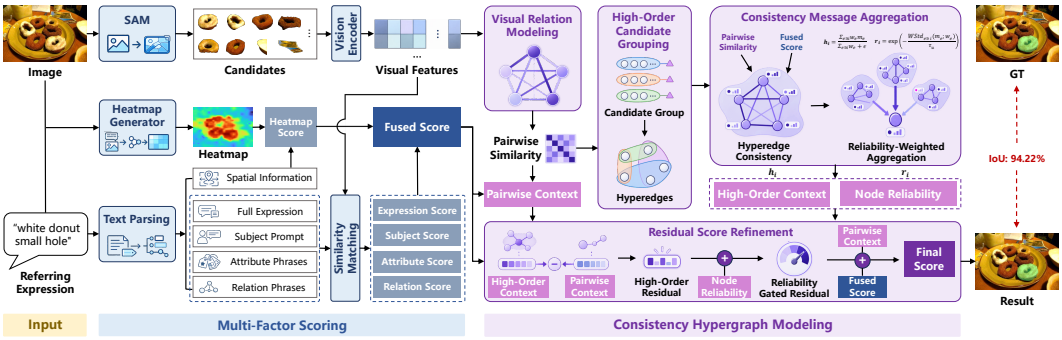

Figure 2: Overview of HYPER-RIS. Multi-Factor Scoring combines four semantic factors with a spatial heatmap score. Consistency Hypergraph Modeling constructs visually related candidate groups, aggregates their score messages, and applies a reliability-gated residual to refine the ranking.

then uses these scores and visual features to aggregate pairwise and group-level context and rerank the proposals (Section 3.3). The highest-scoring mask is returned without task-specific parameter learning.

3.2 MULTI-FACTOR SCORING

Semantic matching. Complementary linguistic views help distinguish candidates that only partially match the expression. We use a dependency parser to extract the subject head t_{sub}, attribute phrases $\mathcal{T}_{\text{attr}}$, and relation phrases $\mathcal{T}_{\text{rel}}$. The subject is encoded as the extracted head word, while attribute and relation phrases retain the corresponding modifiers and predicate–argument content. The full expression is retained as a separate factor to preserve its compositional context. For $x \in \mathcal{X} = \{\text{sub}, \text{attr}, \text{rel}, \text{exp}\}$, let $\mathcal{T}_{\text{sub}} = \{t_{\text{sub}}\}$ and $\mathcal{T}_{\text{exp}} = \{q\}$. Using normalized text features $\mathbf{e}(t)$ and the encoder temperature τ_{vl}, we compute

$$S_i^x = \frac{1}{\max(1, |\mathcal{T}_x|)} \sum_{t \in \mathcal{T}_x} \frac{\mathbf{f}_i^\top \mathbf{e}(t)}{\tau_{\text{vl}}}. \tag{1}$$

An empty attribute or relation set contributes zero. All four factors reuse the same visual feature; relation phrases are matched semantically rather than evaluated as explicit geometric predicates.

Dense heatmap generation. We complement region–text matching with a dense heatmap that localizes expression-relevant image content. Let $\mathcal{H}_{\text{GEM}}$ denote the spatial response produced by GEM. The TIPSv2 branch encodes the full image into normalized patch features $\{\mathbf{p}_u\}$ and computes patch–text responses $d_u = \mathbf{p}_u^\top \mathbf{e}(q)$. We reshape these responses into a square grid $\mathcal{D}$ and min–max normalize them to obtain $\mathcal{H}_{\text{TIPS}}$. The TIPSv2 map is bilinearly aligned to the GEM response and combined with it by equal averaging:

$$\mathcal{H} = \tfrac{1}{2}\big(\mathcal{H}_{\text{GEM}} + \text{Align}(\mathcal{H}_{\text{TIPS}})\big). \tag{2}$$

Position-aware spatial evidence. The parser maps location words to coarse tags, which modulate the heatmap through a multiplicative prior $\mathcal{B}_q$. For normalized image coordinates $(\xi, \eta) \in [0,1]^2$, the left and right tags contribute $1-\xi$ and ξ, while the top and bottom tags contribute $1-\eta$ and η. A center tag contributes $(1-2|\xi-\frac{1}{2}|)_+(1-2|\eta-\frac{1}{2}|)_+$, where $(\cdot)_+ = \max(\cdot, 0)$. Front/back tags are mapped to bottom/top as a coarse image-plane heuristic. Multiple tag responses are multiplied; without location tags, $\mathcal{B}_q = 1$. We then pool the modulated response within each candidate:

$$\widehat{\mathcal{H}} = \mathcal{H} \odot \mathcal{B}_q, \qquad \bar{S}_i^{\text{heat}} = \frac{\sum_{x,y} \widehat{\mathcal{H}}(x,y) M_i(x,y)}{\max(a_i, \epsilon)}, \quad a_i = \sum_{x,y} M_i(x,y). \tag{3}$$

Masks are resized to the heatmap resolution by nearest-neighbor interpolation, so a_i denotes area on the same grid. Here, $\epsilon > 0$ is a small constant for numerical stability. Mask averaging avoids increasing a score solely because the proposal covers more pixels.

Conditional size adjustment. When the expression specifies a large or small target, we adjust the heatmap score using relative mask area. Let $\widetilde{a}_i = (a_i - a_{\min})/\max(a_{\max} - a_{\min}, \epsilon)$, with extrema taken over the current candidates. Let $b_{\max} > 1$ control the maximum strength of the size adjustment. For proposals whose pooled response is at least the proposal median, we use

$$b_i = \begin{cases} 1 + (b_{\max} - 1)\widetilde{a}_i, & \text{large-target preference,} \\ b_{\max} - (b_{\max} - 1)\widetilde{a}_i, & \text{small-target preference,} \end{cases} \qquad S_i^{\text{heat}} = b_i \bar{S}_i^{\text{heat}}. \tag{4}$$

Other proposals use $b_i = 1$, keeping the size preference conditional on spatial relevance. For an even proposal count, we use the lower median. The adjustment is skipped if the area or response range is numerically negligible, or no unambiguous size preference is parsed. The size lexicon includes direct size terms and proxy cues such as closer/further.

Score fusion. The semantic factors and spatial evidence are combined as

$$s_i = \sum_{x \in \mathcal{X}} \lambda_x S_i^x + \lambda_{\text{heat}} S_i^{\text{heat}}. \tag{5}$$

Here, λ_x and λ_{heat} denote the fusion weights. We collect the fused proposal scores as $\mathbf{s} = [s_1, \ldots, s_N]^\top$, which is passed to Consistency Hypergraph Modeling.

3.3 Consistency Hypergraph Modeling

Pairwise context. We first establish a visual-context baseline for proposal scores. Let $\mathcal{Z}$ denote standardization using the population mean and standard deviation, returning zero when the standard deviation is below ϵ. This puts proposal scores on a common scale within each expression before contextual propagation. We set $\mathbf{z} = \mathcal{Z}(\mathbf{s})$ and compute cosine similarities $V_{ij} = \mathbf{f}_i^\top \mathbf{f}_j$. For $N > 1$, off-diagonal row normalization gives

$$A_{ij} = \frac{\exp(V_{ij})}{\sum_{k \neq i} \exp(V_{ik})} \quad (j \neq i), \qquad A_{ii} = 0. \tag{6}$$

Let $\mathbf{A} = [A_{ij}]$ denote the resulting pairwise adjacency matrix, and $\mathbf{g} = \mathbf{A}\mathbf{z}$ the corresponding pairwise context. The message g_i summarizes the scores of other proposals, with larger visual similarities receiving greater weight. Excluding the diagonal separates this contextual message from the candidate's own evidence. The pairwise baseline is $\mathbf{Q}^G = \mathbf{z} + \alpha\mathbf{g}$, where α controls the contribution of pairwise context.

Candidate grouping. To aggregate evidence from candidate groups, we form one hyperedge around each proposal and its most visually similar neighbors:

$$e_i = \{i\} \cup \text{TopK}_{j \neq i}(V_{ij}), k = \min\big(N-1, \max(1, \lceil \rho(N-1) \rceil)\big). \tag{7}$$

Duplicate hyperedges are removed, yielding $\mathcal{E}$ with incidence $B_{ie} = \mathbf{1}[i \in e]$. The relative-neighbor ratio ρ adapts group size to the number of available proposals. Each candidate belongs to its centered group, ensuring nonzero incidence even after duplicate groups are removed. Grouping uses visual similarity alone; expression relevance enters through the member scores. Figure 3 illustrates two complementary grouping patterns captured by the constructed hyperedges. Hyperedge 1 primarily connects candidates from the same semantic category, while Hyperedge 2 connects candidates sharing common visual attributes, including apparent scale and spatial location, relative distance, and appearance cues such as jersey color. Such grouping allows the hypergraph to represent both category-level and attribute-level consistency among multiple proposals, providing group-level context for subsequent high-order reasoning.

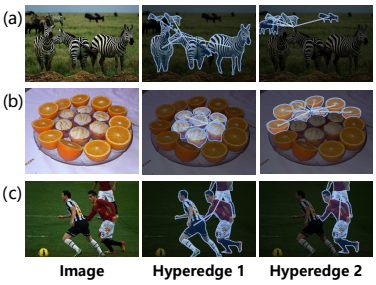

Figure 3: Visualization of two representative high-consistency hyperedges in three scenes. Hyperedge 1 connects same-category objects, while Hyperedge 2 connects objects sharing similar attributes: (a) small background animals, (b) food items farther from the camera, and (c) players with matching jersey colors.

Consistency-weighted messages. We weight each group by its internal visual coherence and summarize its member scores with a robust mean:

$$c_e = \frac{\sum_{i<j,\, i,j\in e} V_{ij}}{\binom{|e|}{2}}, \qquad w_e = \mathrm{sigmoid}\left(\frac{\mathcal{Z}(\mathbf{c})_e}{\tau_{\mathrm{edge}}}\right), \qquad m_e = \mathrm{TrimMean}(\{z_i : i \in e\}). \tag{8}$$

Here, $\mathbf{c} = (c_e)_{e\in\mathcal{E}}$ collects group consistencies, and $\mathcal{Z}(\mathbf{c})$ standardizes them across hyperedges. The temperature $\tau_{\mathrm{edge}} > 0$ controls the sensitivity of hyperedge weights to relative visual coherence. Groups with higher relative visual coherence receive larger weights. The operator $\mathrm{TrimMean}$ removes one minimum and one maximum score when $|e| \geq 4$, using the ordinary mean otherwise; this limits the influence of extreme member scores. For a singleton hyperedge, we define $c_e = 1$. Each candidate receives the weighted average of its incident messages:

$$D_i = \sum_e B_{ie} w_e, \qquad h_i = \frac{\sum_e B_{ie} w_e m_e}{D_i + \epsilon}. \tag{9}$$

We collect these messages as $\mathbf{h} = [h_1, \ldots, h_N]^\top$, referred to as the high-order context. To attenuate conflicting group evidence, we estimate the dispersion of incoming messages and convert it into a node reliability gate:

$$u_i = \sqrt{\left[\frac{\sum_e B_{ie} w_e m_e^2}{D_i + \epsilon} - h_i^2\right]_+}, \qquad r_i = \exp(-u_i/\tau_{\mathrm{node}}). \tag{10}$$

Here, $\tau_{\mathrm{node}} > 0$ controls the sensitivity of the reliability gate to message disagreement. We collect the node reliabilities as $\mathbf{r} = [r_1, \ldots, r_N]^\top$. The quantity u_i measures the weighted dispersion of incoming group messages, and $r_i \in (0, 1]$ reduces their influence when they disagree. We set $u_i = 0$ when only one hyperedge is incident to a candidate.

Residual refinement. We refine the pairwise baseline $\mathbf{Q}^G$ using the difference between the high-order context $\mathbf{h}$ and the pairwise context $\mathbf{g}$. The node reliability $\mathbf{r}$ modulates this residual elementwise, yielding the final proposal scores:

$$\mathbf{Q} = \mathbf{Q}^G + \beta\, \mathbf{r} \odot (\mathbf{h} - \mathbf{g}) = \mathbf{z} + \alpha\mathbf{g} + \beta\, \mathbf{r} \odot (\mathbf{h} - \mathbf{g}). \tag{11}$$

Here, β controls the residual strength. For each proposal, $h_i - g_i$ determines the contextual correction, while r_i scales its contribution according to the agreement among incident hyperedge messages. The baseline score is preserved when $h_i = g_i$, and a smaller r_i attenuates the correction. We return $\widehat{M} = M_{i^\star}$, where $i^\star = \arg\max_i Q_i$, resolving score ties by choosing the smallest proposal index. For $N = 1$, the pairwise adjacency is zero and the single proposal is selected.

4 Experiments

4.1 Settings

Datasets. We evaluate Hyper-RIS on RefCOCO, RefCOCO+ (Yu et al., 2016), RefCOCOg (Mao et al., 2016), and PhraseCut (Wu et al., 2020), using their standard evaluation splits.

Metrics. We report mean intersection over union (mIoU) and overall intersection over union (oIoU), following prior training-free RIS work. mIoU averages the per-expression mask IoU and therefore gives each referring expression equal weight, whereas oIoU divides the total intersection area by the total union area over the complete evaluation split. All results are percentages.

Implementation Details. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU. We generate class-agnostic proposals with SAM ViT-H (Kirillov et al., 2023) and encode masked candidates with TIPSv2-B/14 (Cao et al., 2026) at 448×448 resolution. Dense spatial evidence combines GEM ViT-B/16 (Bousselham et al., 2024) and TIPSv2-B/14 heatmaps. For consistency hypergraph inference, we set the relative-neighbor ratio to $\rho = 0.7$, the pairwise propagation coefficient to $\alpha = 1.5$, and the higher-order residual coefficient to $\beta = 1.0$. The full implementation details are provided in Appendix A.

Table 1: Comparisons with SOTA referring image segmentation methods on standard benchmarks. The best and second best results among zero-shot methods without additional training are highlighted with red shading and blue shading, respectively. * indicates the extra dataset used to train the model.

Metric	Method	Vision Backbone	Pre-trained Model	RefCOCO			RefCOCO+			RefCOCOg	
				val	testA	testB	val	testA	testB	val	test
oIoU	*zero-shot methods w/ additional training*										
	Pseudo-RIS (Yu et al., 2024)	ViT-B	SAM, CoCa, CLIP	37.33	43.43	31.90	40.19	46.43	33.63	41.63	43.52
	VLM-VG (Wang et al., 2025)	R101	COCO*, VLM-VG*	45.40	48.00	41.40	37.00	40.70	30.50	42.80	44.10
	zero-shot methods w/o additional training										
	Grad-CAM (Selvaraju et al., 2017)	R50	SAM, CLIP	23.44	23.91	21.60	26.67	27.20	24.84	23.00	23.91
	MaskCLIP (Zhou et al., 2022)	R50	SAM, CLIP	20.18	20.52	21.30	22.06	22.43	24.61	23.05	23.41
	Global-Local (Yu et al., 2023)	R50	FreeSOLO, CLIP	24.58	23.38	24.35	25.87	24.61	25.61	30.07	29.83
	Global-Local (Yu et al., 2023)	R50	SAM, CLIP	24.55	26.00	21.03	26.62	29.99	22.23	28.92	30.48
	Global-Local (Yu et al., 2023)	ViT-B	SAM, CLIP	21.71	24.48	20.51	23.70	28.12	21.86	26.57	28.21
	Ref-Diff (Ni et al., 2023)	ViT-B	SAM, SD, CLIP	35.16	37.44	34.50	35.56	38.66	31.40	38.62	37.50
	TAS (Suo et al., 2023)	ViT-B	SAM, BLIP2, CLIP	29.53	30.26	28.24	33.21	38.77	28.01	35.84	36.16
	HybridGL (Liu & Li, 2025)	ViT-B	SAM, CLIP	41.81	44.52	38.50	35.74	41.43	30.90	42.47	42.97
	B2G (An et al., 2026)	ViT-B	SAM, CLIP	38.47	40.12	36.95	34.58	39.24	30.03	44.03	44.02
	RefAM (Kukleva et al., 2026)	DiT	SAM, FLUX	42.67	46.21	40.89	34.43	38.70	30.53	39.58	41.13
	RefChess (Tong et al., 2026)	ViT-B	SAM, FG-CLIP	48.47	52.50	43.58	41.20	48.19	32.73	42.85	44.21
	HYPER-RIS	ViT-B	SAM, TIPSv2	**51.12**	**56.91**	**44.52**	**42.84**	**51.05**	**34.34**	**45.34**	**46.24**
	zero-shot methods w/ additional training										
	Pseudo-RIS (Yu et al., 2024)	ViT-B	SAM, CoCa, CLIP	41.05	48.19	33.48	44.33	51.42	35.08	45.99	46.67
	VLM-VG (Wang et al., 2025)	R101	COCO*, VLM-VG*	49.90	53.10	46.70	42.70	47.30	36.20	48.00	48.50
	zero-shot methods w/o additional training										
	Grad-CAM (Selvaraju et al., 2017)	R50	SAM, CLIP	30.22	31.90	27.17	33.96	25.66	32.29	33.05	32.50
	MaskCLIP (Zhou et al., 2022)	R50	SAM, CLIP	25.62	26.66	25.17	27.49	28.49	30.47	30.13	30.15
	Global-Local (Yu et al., 2023)	R50	FreeSOLO, CLIP	26.70	24.99	26.48	28.22	26.54	27.86	33.02	33.12
	Global-Local (Yu et al., 2023)	R50	SAM, CLIP	31.83	34.93	28.64	34.97	37.11	30.61	40.66	40.94
	Global-Local (Yu et al., 2023)	ViT-B	SAM, CLIP	33.12	36.52	29.58	35.29	39.58	31.89	40.08	40.74
	CaR (Sun et al., 2024)	ViT-B/ViT-L	CLIP	33.57	35.56	30.51	34.22	36.03	31.02	36.67	36.57
	Ref-Diff (Ni et al., 2023)	ViT-B	SAM, SD, CLIP	37.21	38.40	37.19	37.29	40.51	33.01	44.02	44.51
	TAS (Suo et al., 2023)	ViT-B	SAM, BLIP2, CLIP	39.84	41.08	36.24	43.63	49.13	36.54	46.62	46.80
	HybridGL (Liu & Li, 2025)	ViT-B	SAM, CLIP	49.48	53.37	45.19	43.40	49.13	37.17	51.25	51.59
	B2G (An et al., 2026)	ViT-B	SAM, CLIP	45.17	46.82	43.65	41.28	45.94	36.72	50.75	50.74
	RefAM (Kukleva et al., 2026)	DiT	SAM, FLUX	52.17	55.60	49.64	43.06	48.58	37.55	48.11	48.57
	RefChess (Tong et al., 2026)	ViT-B	SAM, FG-CLIP	55.19	57.94	50.04	48.63	54.64	39.63	50.63	51.32
	HYPER-RIS	ViT-B	SAM, TIPSv2	**58.87**	**62.80**	**53.21**	**51.81**	**58.80**	**42.78**	**54.60**	**54.81**

4.2 RESULTS

Results on the RefCOCO series. Table 1 compares HYPER-RIS with representative training-free zero-shot RIS methods using the result organization of RefChess (Tong et al., 2026). Among zero-shot methods without additional training, HYPER-RIS improves the previous best result on every split, with gains of 0.94–4.41 percentage points in oIoU and 3.15–4.86 points in mIoU. In particular, it obtains the highest oIoU and mIoU across all eight evaluation splits within this setting, demonstrating consistent improvements under both aggregate pixel-level and expression-balanced evaluation. As shown in Figure 4, HYPER-RIS better distinguishes the referred targets from competing objects in scenes with complex contextual relations, producing masks that align more closely with the ground truth than the compared methods. These results support the benefit of high-order candidate consistency in resolving ambiguity in multi-object scenes.

Results on PhraseCut. Table 2 follows the compact PhraseCut presentation of HybridGL (Liu & Li, 2025). HYPER-RIS reaches an average score of 38.08, exceeding HybridGL by 0.39 percentage points and achieving the highest oIoU and Avg. Ref-Diff retains the highest mIoU, while HYPER-RIS ranks second on this metric and improves the RefChess result by 2.07 mIoU points. The complementary outcome across mIoU and oIoU shows that HYPER-RIS improves aggregate mask overlap while remaining competitive under expression-balanced evaluation.

4.3 ABLATION STUDIES

Component analysis. Table 3 evaluates the two main components on the validation splits of RefCOCO, RefCOCO+, and RefCOCOg, and the PhraseCut test split. Removing both components reduces mIoU by 1.82–15.26 points and oIoU by 4.37–18.41 points, whereas removing Consistency Hypergraph Modeling alone reduces mIoU by 0.50–4.88 points and oIoU by 0.60–3.32 points. The full model outperforms both module-level ablations on every dataset and metric, confirming that the two components provide complementary improvements.

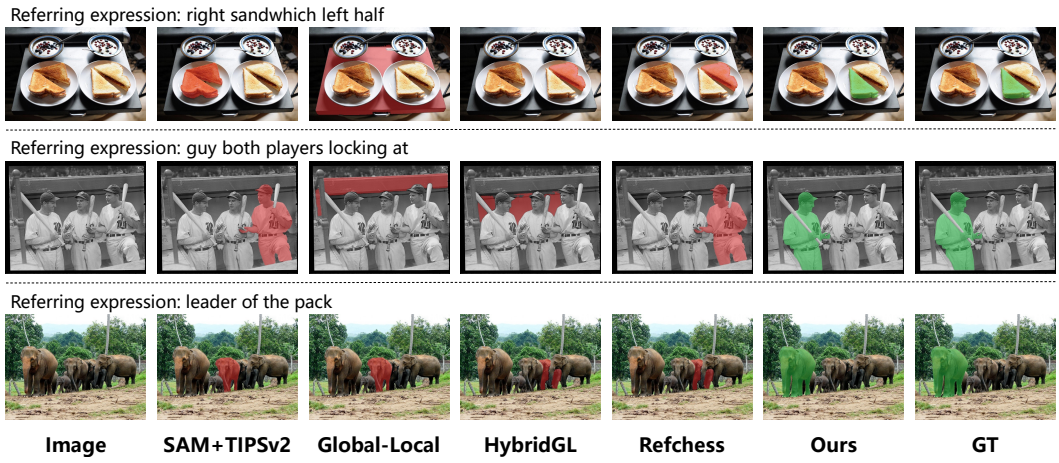

Figure 4: Qualitative comparison in multi-object scenes with complex contextual relations. HYPER-RIS identifies the referred targets more accurately than the compared methods.

Table 2: Comparison on the PhraseCut test split. The best and second-best values are highlighted with red and blue shading, respectively.

Method	oIoU	mIoU	Avg
Global-Local (Yu et al., 2023)	23.64	–	23.64
TAS (Suo et al., 2023)	25.64	24.66	25.15
Ref-Diff (Ni et al., 2023)	29.42	**41.75**	35.59
HybridGL (Liu & Li, 2025)	38.39	36.98	37.69
RefChess (Tong et al., 2026)	38.28	35.25	36.77
HYPER-RIS	**38.84**	37.32	**38.08**

Table 3: Module ablation results. M: Multi-Factor Scoring; C: Consistency Hypergraph Modeling; ✓: enabled. Best results are bold.

M	C	RefCOCO		RefCOCO+		RefCOCOg		PhraseCut	
		mIoU	oIoU	mIoU	oIoU	mIoU	oIoU	mIoU	oIoU
–	–	43.61	32.71	49.99	38.47	49.73	36.37	34.08	33.84
✓	–	54.97	48.38	50.43	40.35	49.72	42.02	36.82	38.24
✓	✓	**58.87**	**51.12**	**51.81**	**42.84**	**54.60**	**45.34**	**37.32**	**38.84**

Table 4: Scoring-factor ablation results. $\lambda_x = 0$ removes factor x.

Setting	RefCOCO		RefCOCO+		RefCOCOg		PhraseCut	
	mIoU	oIoU	mIoU	oIoU	mIoU	oIoU	mIoU	oIoU
Full	**58.87**	**51.12**	**51.81**	**42.84**	**54.60**	**45.34**	37.32	38.84
$\lambda_{\text{sub}} = 0$	57.01	50.07	51.43	42.71	53.36	43.68	36.56	38.34
$\lambda_{\text{attr}} = 0$	58.64	51.05	51.34	42.32	54.13	44.87	37.25	38.85
$\lambda_{\text{rel}} = 0$	58.75	51.04	51.68	42.68	54.21	45.18	**37.34**	**38.91**
$\lambda_{\text{exp}} = 0$	55.09	48.71	47.33	39.28	50.21	42.00	35.88	38.35
$\lambda_{\text{heat}} = 0$	44.83	34.30	49.96	38.72	51.03	38.80	34.90	34.95

Table 5: Consistency Hypergraph Modeling ablation results. HC: high-order context; PC: pairwise context; NR: node reliability; HG→G: replacing hypergraph with graph.

Setting	RefCOCO		RefCOCO+		RefCOCOg		PhraseCut	
	mIoU	oIoU	mIoU	oIoU	mIoU	oIoU	mIoU	oIoU
Full	**58.87**	**51.12**	51.81	**42.84**	**54.60**	**45.34**	**37.32**	**38.84**
w/o HC	55.58	49.32	50.76	40.69	50.27	42.49	36.39	36.79
w/o PC	58.01	50.66	51.60	42.74	54.33	45.20	37.26	38.83
w/o NR	58.81	51.07	**51.84**	42.82	54.46	45.12	36.84	37.49
HG→G	57.93	50.25	51.51	42.61	54.38	44.98	36.84	38.15

Multi-Factor Scoring components. Table 4 examines the contribution of each scoring signal. The heatmap and complete-expression matching score provide the largest improvements within Multi-Factor Scoring. Removing the heatmap produces the largest oIoU degradation on all three datasets. The complete-expression score is particularly important for expression-balanced evaluation, and its removal causes the largest mIoU drop on RefCOCO+ and RefCOCOg. Removing the subject, attribute, or relation score also consistently degrades performance on the RefCOCO series; on PhraseCut, removing the attribute or relation score yields small gains of up to 0.07 points.

Consistency Hypergraph Modeling components. Table 5 isolates high-order context (HC), pairwise context (PC), and node reliability (NR). To distinguish hypergraph reasoning from repeated pairwise propagation, HG→G replaces hyperedge aggregation with an additional propagation over the same pairwise visual graph, with node reliability computed from weighted neighbor-message dispersion. Although HG→G consistently improves over removing HC, the full model further improves all eight metrics, showing that group-wise hyperedge aggregation provides complementary benefits beyond pairwise propagation. Removing PC decreases performance on all four benchmarks, while NR has a smaller dataset-dependent effect and is most beneficial on PhraseCut. Together, these

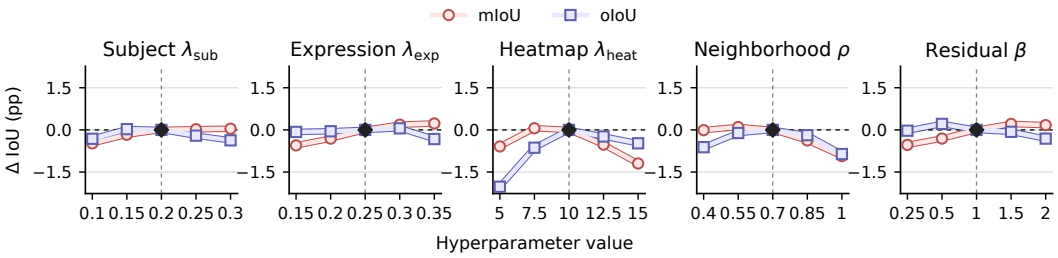

Figure 5: Sensitivity to selected hyperparameters on RefCOCO val. Curves show mIoU/oIoU changes in percentage points relative to the default configuration, marked by black diamonds.

Table 6: Encoder comparison on RefCOCO val (mIoU / oIoU, %).

Method	CLIP	FG-CLIP	TIPSv2
HybridGL (Liu & Li, 2025)	49.48 / 41.81	44.43 / 36.28	46.00 / 37.87
RefChess (Tong et al., 2026)	45.39 / 38.91	55.19 / 48.47	56.53 / 49.63
HYPER-RIS	**50.68 / 41.97**	**56.34 / 48.89**	**58.87 / 51.12**

Table 7: Inference efficiency on RefCOCO val.

Method	Inference time (ms)	mIoU	oIoU
Global-Local (Yu et al., 2023)	**511.54**	33.12	21.71
HybridGL (Liu & Li, 2025)	757.56	49.48	41.81
RefChess (Tong et al., 2026)	982.53	55.19	48.47
HYPER-RIS	942.02	**58.87**	**51.12**

results show that high-order context provides the dominant improvement, with pairwise context and node reliability offering complementary refinements.

Hyperparameter sensitivity. Figure 5 evaluates selected hyperparameters on RefCOCO val. The subject and expression weights and the residual coefficient yield relatively small performance changes across the tested ranges. The heatmap weight and neighborhood ratio have larger effects, particularly when the heatmap weight is too small or the neighborhood includes all candidates. Overall, performance remains stable near the default configuration. A more comprehensive analysis is provided in Appendix B.

4.4 Generality and Inference Efficiency

Generality across visual encoders. Table 6 evaluates whether the advantage of HYPER-RIS depends on a particular visual encoder. We replace the encoder in HybridGL, RefChess, and HYPER-RIS with CLIP ViT-B/16, FG-CLIP Base, or TIPSv2-B/14 while retaining the remaining method-specific design. HYPER-RIS achieves the highest mIoU and oIoU with every encoder. These consistent results indicate that the gain of consistency hypergraph modeling is not tied to a single encoder; within each method, the preferred encoder may differ, and TIPSv2 provides the best overall configuration for HYPER-RIS.

Inference efficiency. Table 7 compares average end-to-end inference time on RefCOCO validation expressions. HYPER-RIS requires 942.02 ms per expression on average, which is 4.12% lower than RefChess, while improving its full-split mIoU and oIoU by 3.68 and 2.65 percentage points, respectively. Although HybridGL and Global-Local are faster, HYPER-RIS provides substantially higher segmentation accuracy, demonstrating a favorable accuracy–efficiency trade-off among the strongest-performing training-free methods.

5 Conclusion

We presented HYPER-RIS, a training-free framework that improves referring image segmentation by explicitly modeling high-order consistency among candidate proposals. By combining complementary semantic and spatial evidence with consistency-aware hypergraph reasoning, HYPER-RIS moves beyond independent or pairwise candidate modeling and exploits visually coherent candidate groups to refine proposal ranking. Extensive experiments across standard RIS benchmarks demonstrate consistent improvements over existing training-free methods, while ablations and encoder comparisons verify the effectiveness and generality of high-order candidate reasoning. As the framework ultimately reranks a fixed proposal set, future work could extend high-order reasoning to jointly refine or generate candidate masks.

AI USE STATEMENT

Generative AI tools were used to aid or polish writing and for retrieval and discovery. They were used to improve the language and presentation of the manuscript, to enhance figures and tables by generating some visual elements for pipeline illustrations and adjusting table layouts, and to assist in the retrieval and review of related work and pretrained models. All AI-assisted content was reviewed and verified by the authors. All experimental results were obtained from our actual implementations and experiments, and the authors take full responsibility for the correctness of the methods, results, and claims presented in this work.

REPRODUCIBILITY STATEMENT

An anonymized implementation of HYPER-RIS is available at `https://anonymous.4open.science/r/Hyper-RIS-53F6`. Implementation details and hyperparameter settings are provided in Appendix A.

REFERENCES

Na Min An, Inha Kang, Minhyun Lee, and Hyunjung Shim. Blind to position, biased in language: Probing mid-layer representational bias in vision-language encoders for zero-shot language-grounded spatial understanding. In *European Conference on Computer Vision*, 2026. URL https://eccv.ecva.net/virtual/2026/poster/3141.

Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3828–3837, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Bousselham_Grounding_Everything_Emerging_Localization_Properties_in_Vision-Language_Transformers_CVPR_2024_paper.html.

Bingyi Cao, Koert Chen, Kevis-Kokitsi Maninis, Kaifeng Chen, Arjun Karpur, Ye Xia, Sahil Dua, Tanmaya Dabral, Guangxing Han, Bohyung Han, Joshua Ainslie, Alex Bewley, Mithun Jacob, René Wagner, Washington Ramos, Krzysztof Choromanski, Mojtaba Seyedhosseini, Howard Zhou, and André Araujo. TIPSv2: Advancing vision-language pretraining with enhanced patch-text alignment. *arXiv preprint arXiv:2604.12012*, 2026. URL https://arxiv.org/abs/2604.12012.

Silin Cheng, Yang Liu, Xinwei He, Sebastien Ourselin, Lei Tan, and Gen Luo. WeakMCN: Multi-task collaborative network for weakly supervised referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9175–9185, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Cheng_WeakMCN_Multi-task_Collaborative_Network_for_Weakly_Supervised_Referring_Expression_Comprehension_CVPR_2025_paper.html.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3558–3565, 2019. doi: 10.1609/aaai.v33i01.33013558. URL https://ojs.aaai.org/index.php/AAAI/article/view/4235.

Joshua Fixelle. Hypergraph vision transformers: Images are more than nodes, more than edges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9751–9761, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fixelle_Hypergraph_Vision_Transformers_Images_are_More_than_Nodes_More_than_CVPR_2025_paper.html.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision*, pp. 108–124, 2016. doi: 10.1007/978-3-319-46448-0_7. URL https://doi.org/10.1007/978-3-319-46448-0_7.

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*, volume 39, pp. 3653–3661, 2025. doi: 10.1609/aaai.v39i4.32380. URL https://ojs.aaai.org/index.php/AAAI/article/view/32380.

Chaeyun Kim, Seunghoon Yi, Yejin Kim, Yohan Jo, and Joonseok Lee. Towards motion-aware referring image segmentation. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 559–567. PMLR, 2026. URL https://proceedings.mlr.press/v300/kim26a.html.

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*, pp. 4015–4026, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html.

Anna Kukleva, Enis Simsar, Alessio Tonioni, Muhammad Ferjad Naeem, Federico Tombari, Jan Eric Lenssen, and Bernt Schiele. Refam: Attention magnets for zero-shot referral segmentation. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=5DTaxhAP4h.

Mengqi Lei, Siqi Li, Yihong Wu, Han Hu, You Zhou, Xinhu Zheng, Guiguang Ding, Shaoyi Du, Zongze Wu, and Yue Gao. YOLOv13: Real-time object detection with hypergraph-enhanced adaptive visual perception. *arXiv preprint arXiv:2506.17733*, 2025. URL https://arxiv.org/abs/2506.17733.

Linfei Li, Pei Tan, Siqi Li, Changqing Zou, and Yue Gao. Hyper-PCN: Hypergraph-based point cloud completion via high-order correlation modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39121–39130, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Hyper-PCN_Hypergraph-Based_Point_Cloud_Completion_via_High-Order_Correlation_Modeling_CVPR_2026_paper.html.

Xi Li, Yao Li, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Contextual hypergraph modeling for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013. doi: 10.1109/ICCV.2013.413. URL https://openaccess.thecvf.com/content_iccv_2013/papers/Li_Contextual_Hypergraph_Modeling_2013_ICCV_paper.pdf.

Xiaowei Liao, Yong Xu, and Haibin Ling. Hypergraph neural networks for hypergraph matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1266–1275, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Liao_Hypergraph_Neural_Networks_for_Hypergraph_Matching_ICCV_2021_paper.html.

Ting Liu and Siyuan Li. Hybrid global-local representation with augmented spatial guidance for zero-shot referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29634–29643, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_Hybrid_Global-Local_Representation_with_Augmented_Spatial_Guidance_for_Zero-Shot_Referring_CVPR_2025_paper.html.

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*, pp. 11–20, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Mao_Generation_and_Comprehension_CVPR_2016_paper.html.

Minheng Ni, Yabo Zhang, Kailai Feng, Xiaoming Li, Yiwen Guo, and Wangmeng Zuo. Ref-Diff: Zero-shot referring image segmentation with generative models. *arXiv preprint arXiv:2308.16777*, 2023. URL https://arxiv.org/abs/2308.16777.

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*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74. URL https://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf.

Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. CLIP as RNN: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52733.2024.01251. URL https://openaccess.thecvf.com/content/CVPR2024/html/Sun_CLIP_as_RNN_Segment_Countless_Visual_Concepts_without_Training_Endeavor_CVPR_2024_paper.html.

Yucheng Suo, Linchao Zhu, and Yi Yang. Text augmented spatial aware zero-shot referring image segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1032–1043. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.73. URL https://aclanthology.org/2023.findings-emnlp.73/.

Shiyan Tong, Jinxia Zhang, Zhiyuan Wang, Hao Tian, Yingying Wang, Kanjian Zhang, and Haikun Wei. Refchess: Training-free contextual search for zero-shot referring image segmentation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=RESRoG7nvd.

Shijie Wang, Dahun Kim, Ali Taalimi, Chen Sun, and Weicheng Kuo. Learning visual grounding from generative vision and language model. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8057–8067. IEEE, 2025. URL https://www.openaccess.thecvf.com/content/WACV2025/papers/Wang_Learning_Visual_Grounding_from_Generative_Vision_and_Language_Model_WACV_2025_paper.pdf.

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*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/papers/Wang_CRIS_CLIP-Driven_Referring_Image_Segmentation_CVPR_2022_paper.pdf.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. PhraseCut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10216–10225, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wu_PhraseCut_Language-Based_Image_Segmentation_in_the_Wild_CVPR_2020_paper.html.

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*, pp. 18155–18165, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Yang_LAVT_Language-Aware_Vision_Transformer_for_Referring_Image_Segmentation_CVPR_2022_paper.html.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pp. 69–85, 2016. doi: 10.1007/978-3-319-46475-6_5. URL https://doi.org/10.1007/978-3-319-46475-6_5.

Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Zero-shot referring image segmentation with global-local context features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19456–19465, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yu_Zero-Shot_Referring_Image_Segmentation_With_Global-Local_Context_Features_CVPR_2023_paper.html.

Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Pseudo-RIS: Distinctive pseudo-supervision generation for referring image segmentation. In *European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-73113-6_2. URL https://doi.org/10.1007/978-3-031-73113-6_2.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *European Conference on Computer Vision*, pp. 696–712, 2022. doi: 10.1007/978-3-031-19815-1_40. URL https://doi.org/10.1007/978-3-031-19815-1_40.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper/2006/hash/dff8e9c2ac33381546d96deea9922999-Abstract.html.

A Implementation Details

Proposal generation and visual encoding. We generate class-agnostic proposals with SAM ViT-H (Kirillov et al., 2023), using 8 points per side, predicted-IoU and stability thresholds of 0.7, and a minimum mask area of 1,600 pixels. A smaller proposal is removed when at least 90% of its area is covered by a larger retained proposal. Pixels outside each proposal are filled with the CLIP pixel mean. TIPSv2-B/14 (Cao et al., 2026) encodes the masked candidates at a resolution of 448×448 and supplies both the candidate scores and the visual features used for consistency modeling.

Multi-factor scoring. The vision–language temperature τ_{vl} is inherited from the pretrained TIPSv2 model and is not tuned. The heatmap score averages a GEM ViT-B/16 heatmap (Bousselham et al., 2024) and a TIPSv2-B/14 patch–text heatmap with equal weights. The subject, attribute, relation, complete-expression, and heatmap weights are $(0.20, 0.20, 0.15, 0.25, 10.0)$, respectively. For conditional size adjustment, we set $b_{\max} = 3$.

Consistency hypergraph inference. We set the relative-neighbor ratio to $\rho = 0.7$, the pairwise propagation coefficient to $\alpha = 1.5$, and the higher-order residual coefficient to $\beta = 1.0$. The hyperedge and candidate reliability temperatures are both 1.0, and the numerical stability constant is 10^{-6}. Hyperedge reliability uses visual consistency without an additional score-disagreement penalty.

B Complete Hyperparameter Sensitivity Analysis

We provide a more comprehensive analysis of the hyperparameters used in Multi-Factor Scoring and Consistency Hypergraph Modeling. Specifically, we evaluate the five scoring weights λ_{sub}, λ_{attr}, λ_{rel}, λ_{exp}, and λ_{heat}, together with the neighborhood ratio ρ, high-order residual coefficient β, pairwise coefficient α, hyperedge temperature τ_{edge}, and node-reliability temperature τ_{node}. For each experiment, we vary one parameter while keeping all remaining parameters fixed at the default configuration reported in Appendix A. All plots report changes in mIoU and oIoU, in percentage points, relative to the default configuration, which is indicated by the black diamonds.

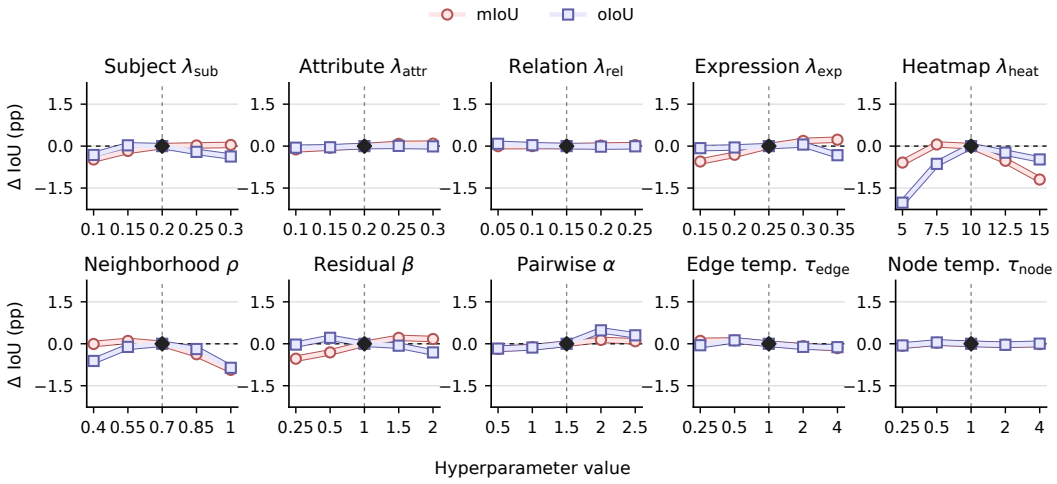

Figure 6: Complete hyperparameter sensitivity analysis on RefCOCO val. Each hyperparameter is varied independently while all others are fixed to their default values. Curves report changes in mIoU and oIoU relative to the default configuration, marked by black diamonds.

Multi-Factor Scoring. Figures 6–8 show that the semantic scoring weights are generally stable under moderate perturbations around the default configuration. The attribute and relation weights cause relatively small variations on RefCOCO and RefCOCO+, whereas the preferred directions of some semantic weights differ on RefCOCOg. The expression and heatmap weights exhibit larger

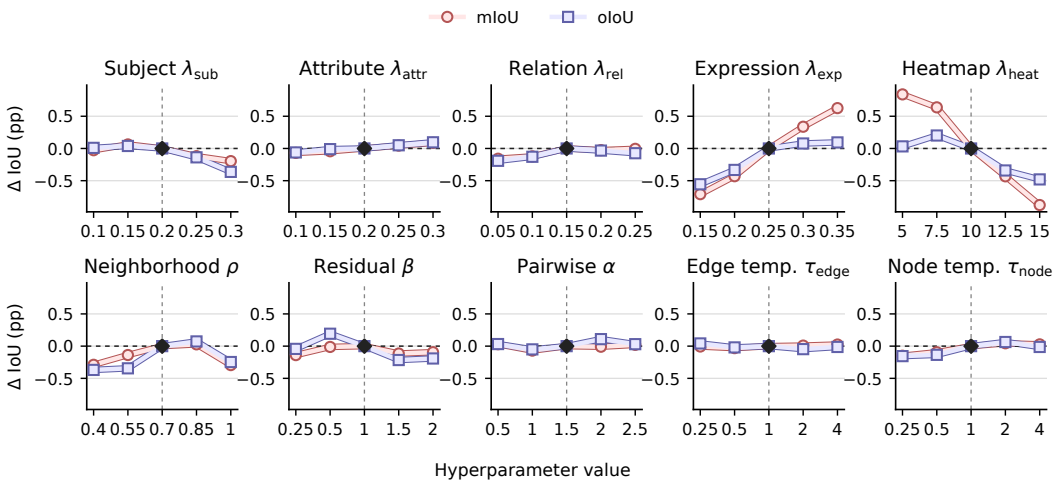

Figure 7: Complete hyperparameter sensitivity analysis on RefCOCO+ val, using the same parameter ranges and default configuration as on RefCOCO.

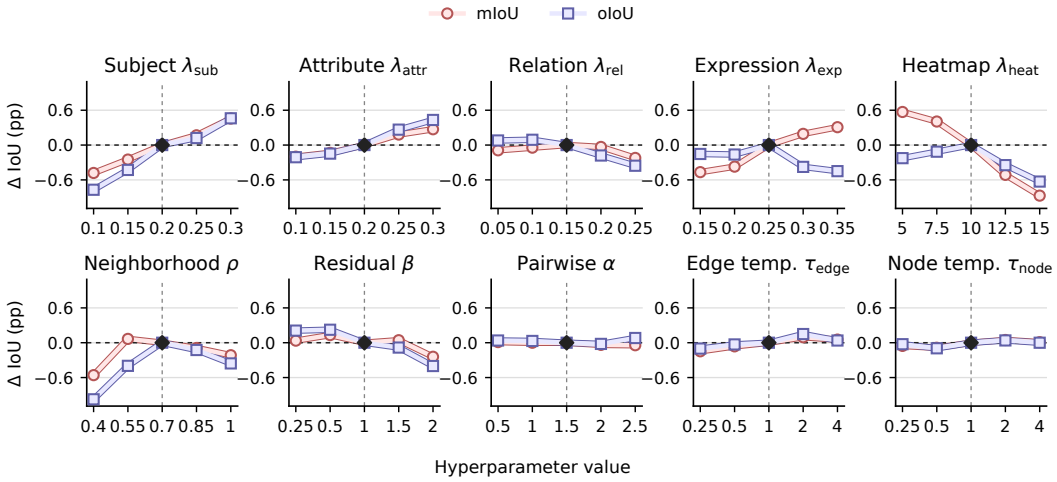

Figure 8: Complete hyperparameter sensitivity analysis on RefCOCOg val, using the same parameter ranges and default configuration as on RefCOCO.

dataset-dependent changes, consistent with their stronger contributions observed in the component ablation. In particular, overly weak or strong heatmap weighting can degrade performance, indicating that spatial evidence should complement rather than dominate semantic matching.

Consistency Hypergraph Modeling. The hypergraph-related parameters are generally less sensitive. The neighborhood ratio ρ has the clearest effect: very small neighborhoods provide insufficient group context, while including nearly all candidates can weaken the locality and consistency of the constructed groups. In contrast, moderate changes to β and α lead to relatively small variations around the default configuration. Performance is also stable over broad ranges of τ_{edge} and τ_{node}, indicating that the consistency weighting and node-reliability mechanism do not depend on narrowly tuned temperature values.

Cross-dataset robustness. Importantly, all three datasets use exactly the same default configuration rather than dataset-specific hyperparameters. Although the locally preferred value of an individual parameter may differ across datasets or between mIoU and oIoU, no alternative setting consistently dominates the default configuration across RefCOCO, RefCOCO+, and RefCOCOg.

Most curves remain stable in a neighborhood of the default values. These results indicate that the performance of HYPER-RIS is not dependent on a narrowly tuned parameter configuration and that the selected configuration generalizes across different referring-expression distributions.

C STAGE-WISE INFERENCE TIME ANALYSIS

To further characterize the inference efficiency of HYPER-RIS, we report a stage-wise runtime breakdown on RefCOCO val. The experiment is conducted in an isolated process with CUDA synchronization, and model loading time is excluded. Table 8 reports the mean runtime of each inference stage.

Table 8: Stage-wise inference-time breakdown of HYPER-RIS on RefCOCO val. The percentage denotes the fraction of the total end-to-end inference time. Model loading is excluded.

Stage	Runtime (ms)	E2E (%)
SAM proposal generation	475.21	50.45
Language parsing	13.68	1.45
Heatmap generation	117.72	12.50
Proposal feature encoding	279.60	29.68
Solver initialization	36.57	3.88
Selection	19.22	2.04
Method-specific computation	466.80	49.55
End-to-end	**942.02**	**100.00**

The breakdown shows that the main computational cost of HYPER-RIS comes from visual processing rather than the subsequent candidate-selection procedure. Proposal feature encoding requires 279.60 ms per expression and heatmap generation requires 117.72 ms, together accounting for 85.1% of the method-specific runtime. In comparison, language parsing, solver initialization, and final selection together require only 69.48 ms. SAM proposal generation accounts for 50.45% of the complete end-to-end runtime, while all method-specific computation accounts for the remaining 49.55%. These results indicate that the overall inference cost is dominated by proposal generation and pretrained visual feature extraction.

D ADDITIONAL QUALITATIVE RESULTS

Figures 9–12 provide additional qualitative comparisons on RefCOCO, RefCOCO+, RefCOCOg, and PhraseCut. Each row corresponds to one referring expression, and we compare the predictions of the initial SAM+TIPSv2 scoring, Global-Local, HybridGL, RefChess, and HYPER-RIS with the ground-truth mask. The examples cover diverse forms of ambiguity, including visually similar instances, relative positions, object attributes, inter-object relations, and compositional descriptions.

On the RefCOCO-family benchmarks, the selected examples illustrate that HYPER-RIS can better distinguish the referred instance when several semantically or visually related candidates compete for similar scores. This behavior is particularly visible for expressions requiring relative position, size, appearance, or relations to surrounding objects. The PhraseCut examples further include broader object and region categories, such as structural elements and background regions, showing that the improvement is not restricted to compact foreground instances. Together, these visualizations complement the quantitative results and illustrate the benefit of consistency-aware high-order candidate reasoning in challenging referring scenarios.

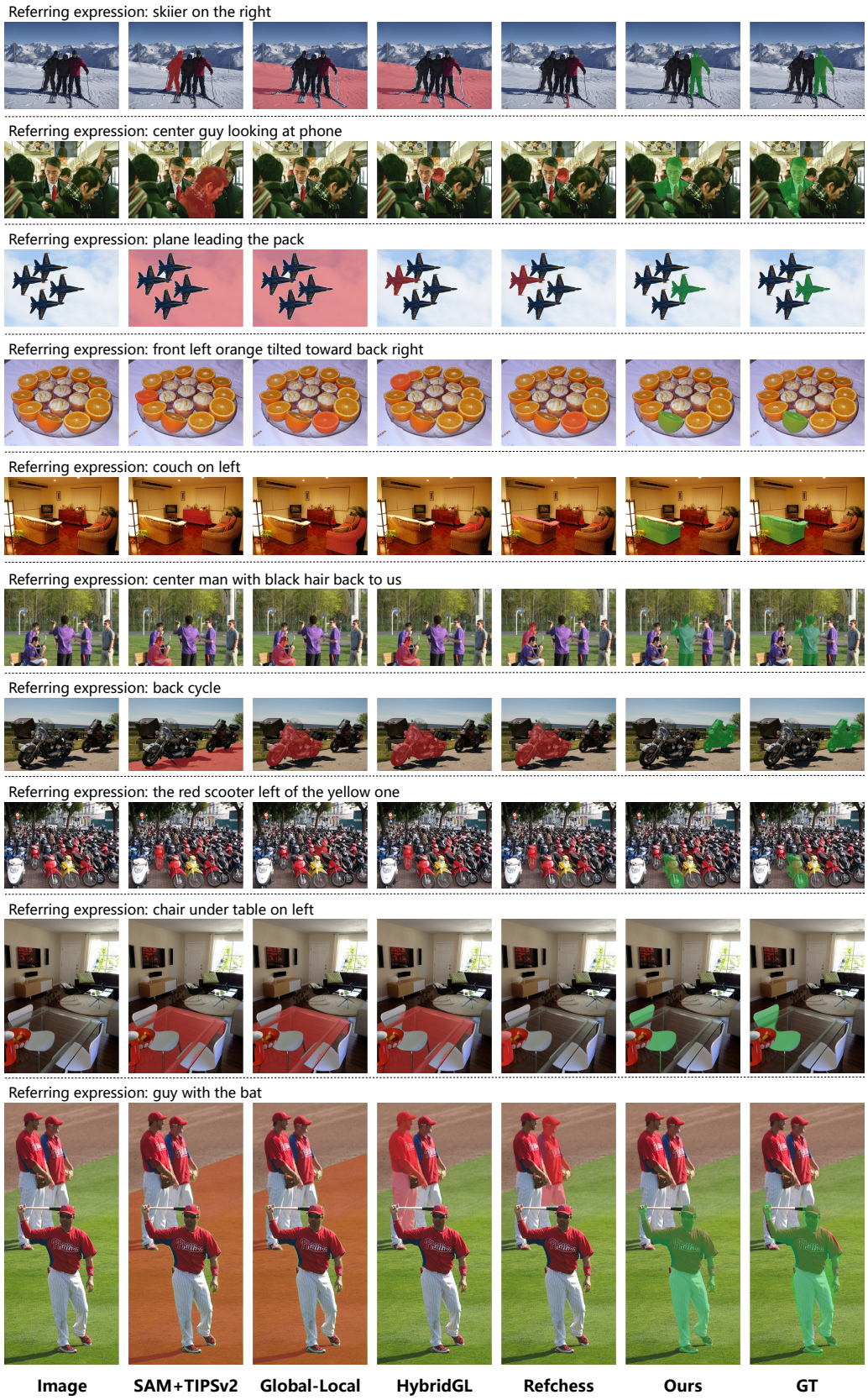

Figure 9: Additional qualitative comparisons on RefCOCO.



Figure 10: Additional qualitative comparisons on RefCOCO+.

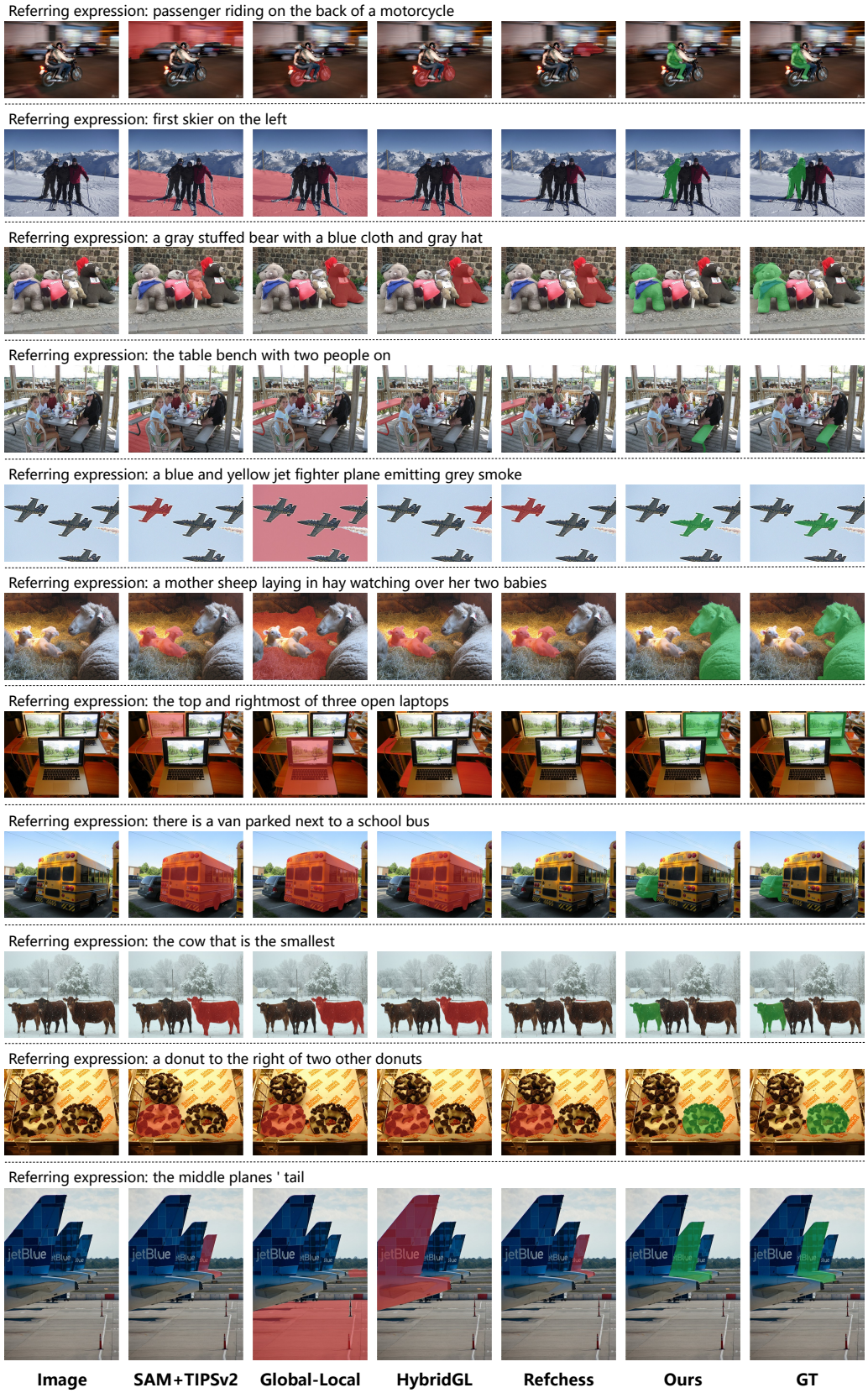

Figure 11: Additional qualitative comparisons on RefCOCOg.

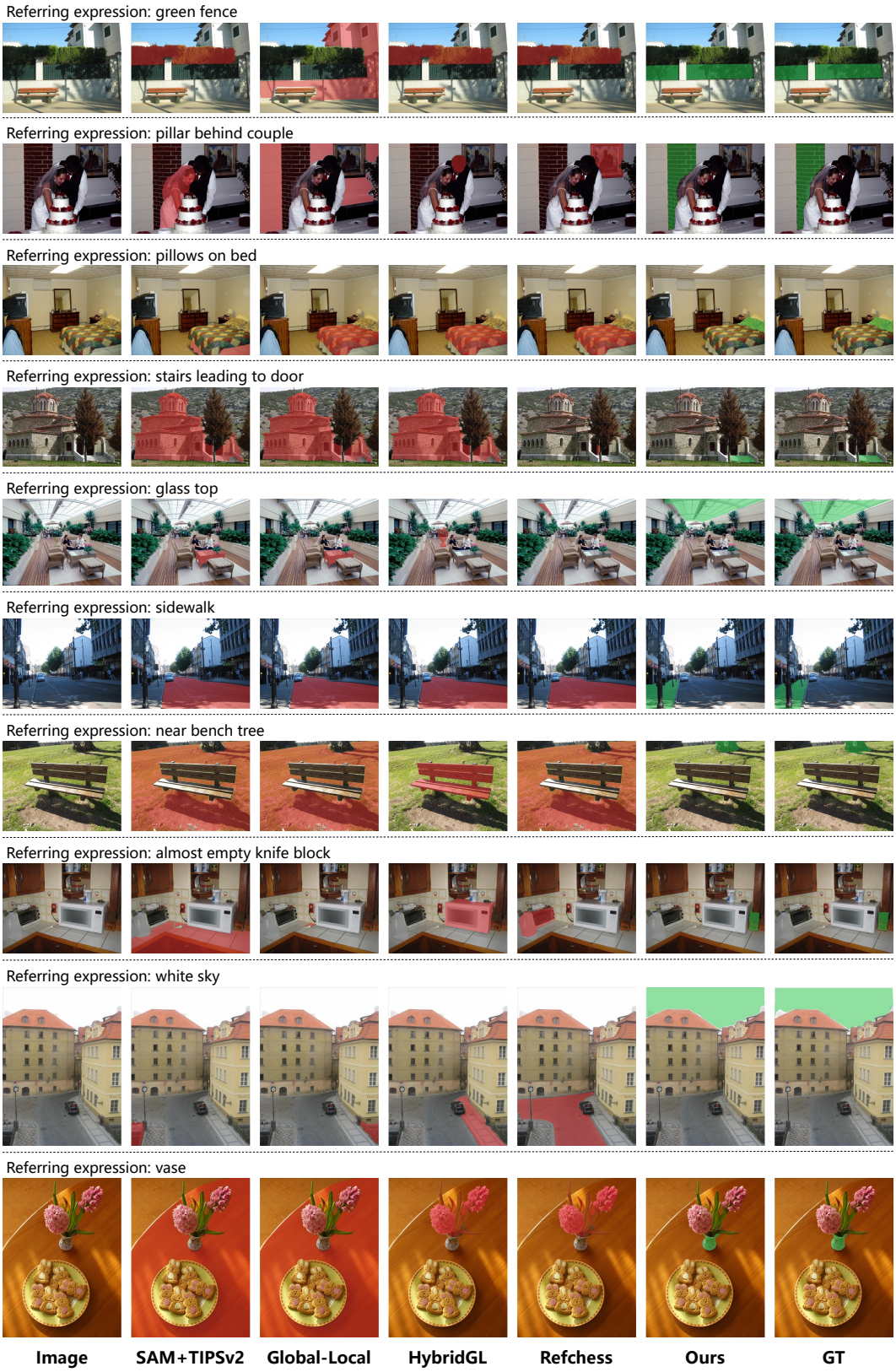

Figure 12: Additional qualitative comparisons on PhraseCut.