NEGATION-AWARE RERANKING FRAMEWORK FOR ZERO-SHOT REFERRING IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot referring image segmentation reduces the reliance on task-specific expression–mask annotations by using pretrained vision–language models and class-agnostic mask proposal networks. However, existing methods show performance degradation on expressions containing negation, as CLIP-based holistic matching may select a candidate with high target compatibility despite violating the negation constraint imposed by the referring expression. To address this challenge, we introduce NARF, a training-free negation-aware framework that reranks candidate masks using a base score that preserves the base method's original assessment, a target-compatibility score that measures compatibility with the intended entity, and a negation-violation score that penalizes detected violations. We further construct a Negation-Focused Evaluation Set from existing RIS datasets for systematic evaluation. Across five zero-shot RIS methods, NARF improves aggregate oIoU and mIoU and provides further gains when combined with NegationCLIP, demonstrating the complementarity of negation-aware representation enhancement and candidate-level constraint verification.

1 INTRODUCTION

Referring image segmentation (RIS) aims to produce a pixel-level mask for the object described by a natural language expression Hu et al. (2016). Unlike supervised RIS methods Liu et al. (2017); Ye et al. (2019); Ding et al. (2021); Kim et al. (2022); Wang et al. (2022a); Yang et al. (2022); Li et al. (2024) that rely heavily on paired referring expressions and dense mask annotations, zero-shot RIS methods often combine pretrained vision–language models (VLMs) Radford et al. (2021) with class-agnostic mask proposal networks Kirillov et al. (2023) to avoid task-specific training. Global-Local CLIP Yu et al. (2023) introduces a candidate-based pipeline for zero-shot RIS: it first generates class-agnostic mask proposals, and then ranks the resulting candidate masks using global–local region–text matching. Subsequent methods further improve mask selection through richer descriptions Li et al. (2026), spatial guidance Liu & Li (2025), or contextual reasoning Tong et al. (2026).

Despite this progress, accurately grounding *negated referring expressions* remains challenging for existing zero-shot RIS methods. As shown in Figure 1(a), the five representative zero-shot RIS methods evaluated consistently exhibit lower performance on the selected subsets restricted to referring expressions containing negation than on their corresponding full splits, highlighting a common challenge in handling negation across different frameworks. One possible explanation lies in the CLIP-based vision–language representations widely used in zero-shot RIS. Recent studies Dobreva & Keller (2021); Singh et al. (2025); Park et al. (2025); Alhamoud et al. (2025) have shown that CLIP and related VLMs are often insensitive to negation, a limitation associated in part with the scarcity of explicitly negated image–text pairs in their pretraining data Park et al. (2025); Alhamoud et al. (2025). To address this limitation, negation-aware fine-tuning methods such as NegationCLIP Park et al. (2025) aim to learn negation-aware vision–language representations. **However, improving vision–language representations alone does not guarantee that the selected mask satisfies the negation constraint, which must also be verified during mask selection for candidate-based zero-shot RIS.**

Such candidate-level verification requires different types of visual evidence depending on the nature of the negation. Specifically, we distinguish two negation types: **Candidate-Local Negation**

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

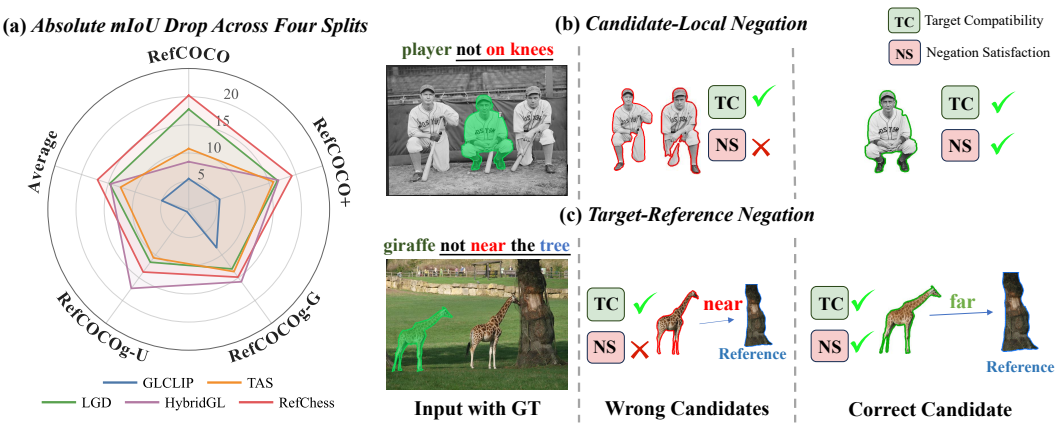


Figure 1: Motivation for explicit negation constraint verification in zero-shot RIS. (a) The absolute mIoU drop of five zero-shot RIS methods on negation-containing subsets relative to their corresponding full splits. (b) Candidate-Local Negation. (c) Target–Reference Negation.

and **Target–Reference Negation**. Candidate-Local Negation can be verified from the local visual content of a target candidate itself. As illustrated in Figure 1(b), for the expression "*player not on knees*," each displayed candidate is compatible with the target category *player*, whereas candidates depicting kneeling players violate the negation constraint *not on knees*. In contrast, Target–Reference Negation requires jointly reasoning about a target candidate, a reference object, and their relation. In Figure 1(c), both giraffe candidates match the target category in "*giraffe not near the tree*," but the incorrect candidate close to the referenced tree violates the negated spatial relation. These examples reveal that target compatibility alone does not guarantee negation satisfaction. For negated referring expressions, reliable mask selection in zero-shot RIS therefore requires evaluating both aspects, with negation satisfaction assessed using type-specific visual evidence.

To address this challenge, we introduce NARF, a **N**egation-**A**ware **R**eranking **F**ramework for zero-shot RIS. NARF operates on the candidate masks and scores produced by an existing zero-shot RIS method, refining their ranking without modifying the base method. For each candidate mask, it derives a reranking score from three complementary terms: the base score S_{base} preserves the candidate score provided by the base method, the target-compatibility score S_{target} measures compatibility with the target entity, and the negation-violation score S_{negative} penalizes violations of the negation constraint. Depending on the visual evidence required for verification, NARF computes S_{negative} using Direct Constraint Verification (DCV) for Candidate-Local Negation or Relational Constraint Verification (RCV) for Target–Reference Negation. This design integrates target compatibility and negation satisfaction into the candidate ranking while retaining the scoring information from the base method.

To enable systematic evaluation, we construct a Negation-Focused Evaluation Set from existing RIS datasets and further analyze its prevalence and negation type distribution. Experiments across five zero-shot RIS methods demonstrate improvements in aggregate performance and additional gains when NARF is combined with NegationCLIP Park et al. (2025). The main contributions can be summarized as follows:

- We construct a Negation-Focused Evaluation Set from existing RIS datasets, characterize the distribution of its negated expressions, and organize these expressions according to the visual evidence required for negation-aware zero-shot RIS.
- We introduce NARF, a **N**egation-**A**ware **R**eranking **F**ramework that combines scores from the base method with explicit target-compatibility and negation-violation scores computed through Direct and Relational Constraint Verification.
- Experiments on five zero-shot RIS methods show aggregate oIoU and mIoU improvements of varying magnitude, including additional gains when NARF is combined with NegationCLIP.

2 Related Work

2.1 Zero-Shot Referring Image Segmentation

Supervised RIS methods Liu et al. (2017); Ye et al. (2019); Ding et al. (2021); Kim et al. (2022); Wang et al. (2022a); Yang et al. (2022); Li et al. (2024) typically depend on pixel-level masks paired with referring expressions. To avoid the intensive label costs, zero-shot RIS methods exploit pretrained foundation models for referent localization and segmentation. Global-Local CLIP Yu et al. (2023) introduces a representative zero-shot RIS pipeline based on class-agnostic mask proposals and vision-language matching. Ref-Diff Ni et al. (2026) combines diffusion cross-attention with discriminative vision–language cues for referent localization, and Hybrid Global-Local Liu & Li (2025) enhances proposal representations through local–global feature fusion and spatial guidance. These methods primarily strengthen referent localization or positive proposal–expression correspondence. Other approaches additionally exploit auxiliary contrastive or distractor-aware evidence. TAS Suo et al. (2023) uses BLIP-2 Li et al. (2023) captions to generate a full-image caption and mines noun phrases as distracting negative expressions. BMS Li et al. (2025) contrasts each candidate with a negative mask formed by combining the remaining proposals. LGD Li et al. (2026) extracts semantically irrelevant noun phrases from MLLM-generated surrounding descriptions as negative samples, and RefChess Tong et al. (2026) evaluates proposal stability under sampled distractor contexts. Although these mechanisms improve general candidate discrimination, their negative signals represent caption-derived distractors, complementary visual regions, surrounding objects, or competing proposals rather than prohibited conditions specified by linguistic negation. In contrast, our framework parses the negated predicate and explicitly verifies negation constraint violations.

2.2 Negation Understanding in Vision–Language Models

Vision–language models learn a shared cross-modal representation space through large-scale image-text pretraining. Representative models such as CLIP Radford et al. (2021) and ALIGN Jia et al. (2021) adopt contrastive learning objectives and exhibit strong zero-shot transferability across image classification, cross-modal retrieval, and open-vocabulary visual recognition. However, the holistic semantic alignment encouraged by conventional contrastive learning does not necessarily equip these models to distinguish an affirmative description from its negated counterpart. This limitation has been examined by VALSE Parcalabescu et al. (2022), which includes an existence diagnostic contrasting presence and absence; CREPE Ma et al. (2023), which incorporates negation foils among its compositional hard negatives; and NegBench Alhamoud et al. (2025), which evaluates negation through retrieval and multiple-choice tasks across image, video, and medical domains. To improve this capability, prior studies have introduced negation-aware data and training objectives for vision–language representation learning. CLIP-BNL Wang et al. (2022b) incorporates partially negated captions and a bidirectionally constrained learning objective. CC-Neg and CoN-CLIP Singh et al. (2025) provide image–caption–negated-caption triplets and a fine-grained contrastive objective for adapting CLIP's text encoder. Know No Better Park et al. (2025) constructs visually aligned negation-inclusive data and introduces NegationCLIP, together with NegRefCOCOg for assessing VLMs' negation comprehension in an image-text matching task. Although these approaches improve representation-level sensitivity to negation, they are not able to explicitly determine whether a candidate mask violates the negation constraint in zero-shot RIS.

3 Method

3.1 Problem Formulation

Given an image I and a referring expression Q, zero-shot referring image segmentation aims to identify a specific object described by Q without task-specific training. Although existing methods differ in candidate generation and region–text matching, their inference process can be abstracted as follows. A zero-shot RIS method $\mathcal{B}$ first produces a set of N candidate masks $\mathcal{M} = \{m_i\}_{i=1}^{N}$ and then evaluates these candidates using its inference and scoring mechanism. Let $S_{\mathcal{B}}(m_i)$ denote the resulting method-specific score assigned to candidate m_i. The prediction of $\mathcal{B}$ is obtained by selecting the candidate with the highest score:

$$m_{\mathcal{B}}^{*} = \arg\max_{m_i \in \mathcal{M}} S_{\mathcal{B}}(m_i). \tag{1}$$

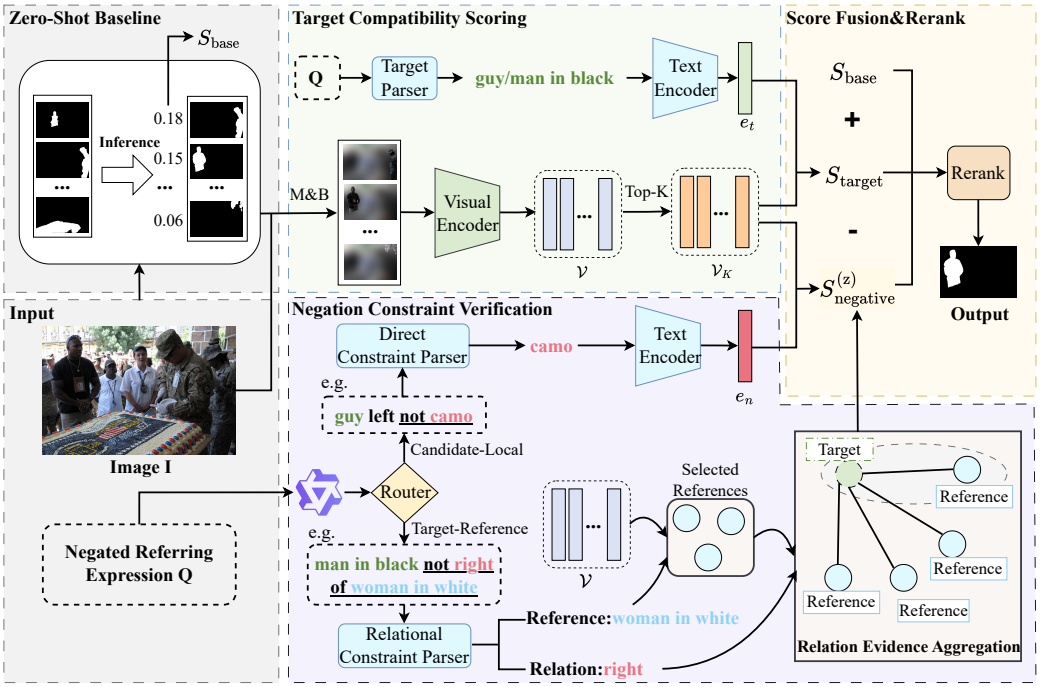


Figure 2: Overview of the proposed framework.

This selection procedure relies on the scoring logic of the zero-shot method to account for the semantic conditions expressed by Q.

3.2 Framework Overview

The overall structure of our framework is shown in Figure 2. First, the base zero-shot RIS method $\mathcal{B}$ produces the candidate set $\mathcal{M}$ and the method-specific scores, both of which are then passed to NARF. To control the reranking depth, we select the top-K masks according to their method-specific scores as the target candidates, denoted by $\mathcal{C}_K = \{m_i\}_{i=1}^{K}$.

Second, we extract the target description from Q to obtain a target prompt t. Each retained candidate in $\mathcal{C}_K$ is then compared with t to compute its target-compatibility score S_{target}.

Third, we determine how the negation constraint should be verified. Different negation types require different scopes of visual evidence. We therefore use a large language model Φ_{LLM} to predict a predefined negation subtype $\tau = \Phi_{\text{LLM}}(Q)$ for the referring expression Q (see Appendix A.3). For supported subtypes, we then map the predicted subtype τ to its corresponding negation type $z = g(\tau) \in \{\text{CL}, \text{TR}\}$, where CL and TR denote Candidate-Local and Target–Reference Negation, respectively. Based on z, the framework routes the candidates to the corresponding verification process to compute type-dependent negation-violation score $S^{(z)}_{\text{negative}}$.

Finally, we treat the original method-specific score $S_{\mathcal{B}}$ as the base score S_{base} and combine the three scores to obtain the final reranking score:

$$S_{\text{final}}(m_i) = S_{\text{base}}(m_i) + w_{\text{tar}} S_{\text{target}}(m_i) - w^{(z)}_{\text{neg}} S^{(z)}_{\text{negative}}(m_i), \tag{2}$$

where w_{tar} and $w^{(z)}_{\text{neg}}$ are non-negative weights. Because different zero-shot RIS methods may produce scores with different numerical scales and distributions, we standardize and normalize each score before combining them (see Appendix B.4). The base score preserves the candidate assessment made by $\mathcal{B}$, the target-compatibility score rewards candidates matching the intended entity, and the negation-violation score penalizes candidates that violate the prohibited condition. For a base method with a downstream candidate-selection rule, the reranked scores are passed to this rule to

produce the final prediction; otherwise, the final mask is selected according to:

$$m^* = \arg\max_{m_i \in \mathcal{C}_K} S_{\text{final}}(m_i). \tag{3}$$

For unsupported negation subtypes or expressions that cannot be parsed reliably, the framework retains the prediction of the base method.

3.3 Target Compatibility Scoring

The target-compatibility score measures how well each retained candidate matches the entity specified by the referring expression. Given an expression Q, the target parser extracts the phrase describing the intended entity and converts it into a compact target prompt t.

For each $m_j \in \mathcal{M}$, we apply mask and blurring strategy Suo et al. (2023) to obtain the candidate region I_j^{MB} (see Appendix B.2). We encode the candidate region and target prompt as

$$v_j = E_{\text{visual}}\big(I_j^{\text{MB}}\big), \qquad e_t = E_{\text{text}}(t), \qquad \mathcal{V} = \{v_j\}_{j=1}^{N}. \tag{4}$$

where E_{visual} and E_{text} denote the CLIP visual and text encoders, respectively, and the visual feature bank $\mathcal{V}$ contains the visual features of all candidate masks. From $\mathcal{V}$, we select the features corresponding to the target candidates in $\mathcal{C}_K$ to form the target visual feature set $\mathcal{V}_K = \{v_i \mid m_i \in \mathcal{C}_K\}$.

For each $m_i \in \mathcal{C}_K$, the target-compatibility score is computed as

$$S_{\text{target}}(m_i) = \cos(v_i, e_t). \tag{5}$$

3.4 Negation Constraint Verification

Direct Constraint Verification (DCV). DCV is designed for negated expressions whose prohibited visual evidence can be assessed within each candidate region independently. The direct constraint parser identifies the linguistic scope of the negation and extracts a prohibited-content prompt n. The prohibited condition is converted into its positive visual form. For example, given the expression *left guy not camo*, the parser extracts $n =$ *camo*. Consequently, stronger visual evidence for *camo* indicates a greater violation of the negation constraint.

Using the visual feature v_i computed in Section 3.3, we encode the prohibited-content prompt and compute the direct constraint-violation score as

$$e_n = E_{\text{text}}(n), \qquad S^{\text{dir}}_{\text{negative}}(m_i) = \cos(v_i, e_n). \tag{6}$$

Relational Constraint Verification (RCV). RCV handles Target–Reference Negation, for which constraint verification requires jointly examining a target candidate and a distinct reference object. Given a referring expression Q, the relational parser extracts a reference prompt r and a prohibited relation ρ, and converts the negated relation into its positive canonical form. For example, *not right of* is represented as *right of*, since RCV measures visual evidence for the relation prohibited by the referring expression.

For each target candidate m_i, we construct the reference pool as $\mathcal{R}_i = \mathcal{M} \setminus \{m_i\}$. We encode the reference prompt as $e_r = E_{\text{text}}(r)$ and use the visual features in $\mathcal{V}$ to compute:

$$S_{\text{ref}}(m_j) = \cos(v_j, e_r), \qquad m_j \in \mathcal{R}_i, \tag{7}$$

We then retain the L candidates with the highest reference-compatibility scores:

$$\mathcal{J}_i = \text{TopL}_{m_j \in \mathcal{R}_i} S_{\text{ref}}(m_j). \tag{8}$$

RCV evaluates spatial and action relations using different visual evidence. Following Hybrid Global-Local Liu & Li (2025), spatial relations are assessed from the position and size of m_i and m_j, whereas action relations are evaluated from the union region Han et al. (2024): $I_{i,j}^{\text{union}} = \mathcal{U}(I, b_i, b_j)$. We define the pairwise relation evidence as

$$G_\rho(m_i, m_j) = \begin{cases} \Phi_\rho(b_i, b_j; W, H), & \text{if } \rho \text{ is spatial,} \\ \cos\big(E_{\text{visual}}\big(I_{i,j}^{\text{union}}\big), E_{\text{text}}(p_\rho(t, r))\big), & \text{if } \rho \text{ is an action,} \end{cases} \tag{9}$$

Table 1: Results on Negation-Focused Evaluation Set from RefCOCO, RefCOCO+, and RefCOCOg. Parenthesized rows report the absolute change relative to the corresponding base method. The better result is shown in bold.

Metric	Method	Backbone	RefCOCO	RefCOCO+	RefCOCOg-U	RefCOCOg-G	Overall
oIoU	GLCLIP	ViT-B/32	21.68	20.42	**27.59**	21.95	21.59
	GLCLIP+NARF	ViT-B/32	**22.70**	**21.04**	25.62	**22.00**	**21.91**
	(vs. GLCLIP)	–	+1.02	+0.62	-1.97	+0.05	+0.32
	TAS	ViT-B/32	23.44	18.49	25.29	23.65	20.92
	TAS+NARF	ViT-B/32	**25.64**	**21.46**	**26.51**	**28.78**	**24.05**
	(vs. TAS)	–	+2.20	+2.97	+1.22	+5.13	+3.13
	LGD	ViT-B/32	20.70	20.23	**26.64**	25.66	21.95
	LGD+NARF	ViT-B/32	**24.08**	**23.06**	24.61	**28.92**	**24.48**
	(vs. LGD)	–	+3.38	+2.83	-2.03	+3.26	+2.53
	HybridGL	ViT-B/16	**32.44**	16.54	29.09	**26.22**	21.72
	HybridGL+NARF	ViT-B/16	32.29	**18.99**	**30.40**	26.21	**23.31**
	(vs. HybridGL)	–	-0.15	+2.45	+1.31	-0.01	+1.59
	RefChess	ViT-B/16	28.78	19.35	**34.55**	32.48	24.48
	RefChess+NARF	ViT-B/16	**34.11**	**22.61**	32.46	**34.20**	**27.31**
	(vs. RefChess)	–	+5.33	+3.26	-2.09	+1.72	+2.83
mIoU	GLCLIP	ViT-B/32	20.79	21.83	**33.18**	**25.28**	23.58
	GLCLIP+NARF	ViT-B/32	**22.02**	**22.85**	30.65	24.74	**23.94**
	(vs. GLCLIP)	–	+1.23	+1.02	-2.53	-0.54	+0.36
	TAS	ViT-B/32	29.60	27.82	35.97	34.54	30.28
	TAS+NARF	ViT-B/32	**30.65**	**32.81**	**38.81**	**40.63**	**34.68**
	(vs. TAS)	–	+1.05	+4.99	+2.84	+6.09	+4.40
	LGD	ViT-B/32	27.63	30.66	35.71	36.24	31.86
	LGD+NARF	ViT-B/32	**30.36**	**35.87**	**37.44**	**41.34**	**36.32**
	(vs. LGD)	–	+2.73	+5.21	+1.73	+5.10	+4.46
	HybridGL	ViT-B/16	42.54	25.84	33.40	34.89	30.85
	HybridGL+NARF	ViT-B/16	**43.67**	**29.29**	**35.95**	**36.57**	**33.52**
	(vs. HybridGL)	–	+1.13	+3.45	+2.55	+1.68	+2.67
	RefChess	ViT-B/16	35.35	29.27	37.12	37.72	32.66
	RefChess+NARF	ViT-B/16	**42.86**	**34.11**	**38.93**	**41.22**	**37.30**
	(vs. RefChess)	–	+7.51	+4.84	+1.81	+3.50	+4.64

where b_i and b_j denote the bounding boxes of m_i and m_j, respectively, and W and H are the image width and height. Φ_ρ is a relation-specific geometric function (see Appendix B.3), and $p_\rho(t, r)$ is a positive prompt describing the prohibited action between the target and reference entities. For instance, the negated expression *man not holding beer* produces the positive relational prompt *man holding beer*.

Finally, we accumulate the relation evidence from the selected reference candidates:

$$\alpha_{i,j} = \frac{\exp(S_{\text{ref}}(m_j))}{\sum_{m_k \in \mathcal{J}_i} \exp(S_{\text{ref}}(m_k))}, \qquad S^{\text{rel}}_{\text{negative}}(m_i) = \sum_{m_j \in \mathcal{J}_i} \alpha_{i,j} G_\rho(m_i, m_j). \tag{10}$$

4 Experiments

4.1 Experimental Settings

Datasets and Metrics. We conduct all experiments on the Negation-Focused Evaluation Set from RefCOCO Nagaraja et al. (2016), RefCOCO+ Nagaraja et al. (2016), and RefCOCOg Mao et al. (2016) (see Appendix A). Following TAS Suo et al. (2023), we adopt two widely used metrics: mean Intersection over Union (mIoU) and overall Intersection over Union (oIoU). mIoU assigns equal weight to each expression, whereas oIoU places greater emphasis on samples with larger target regions.

Baselines. We evaluate NARF with five state-of-the-art zero-shot RIS methods: GLCLIP Yu et al. (2023), TAS Suo et al. (2023), LGD Li et al. (2026), HybridGL Liu & Li (2025), and RefChess Tong et al. (2026). These baselines cover different candidate-generation and scoring strategies, allowing us to assess whether NARF can complement diverse zero-shot RIS pipelines.

Implementation Details. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU. We use Qwen3-8B Qwen Team (2025) to route each referring expression. NARF reuses the vision–language model of each baseline with all model parameters frozen. We rerank the top $K = 10$ target candidates. For RCV, the reference pool comprises all masks generated by the base method, from which the top $L = 5$ candidates are retained for each target candidate according to their reference-compatibility scores. The base, target-compatibility, and negation-violation scores are z-standardized across candidates and normalized using softmax with a temperature of 6.0, while the reference-compatibility scores are converted into aggregation weights using a temperature of 1.0. We set $w_{\text{tar}} = 0.5$, $w_{\text{neg}}^{\text{dir}} = 1.0$, and $w_{\text{neg}}^{\text{rel}} = 0.5$. For TAS and LGD, whose spatial heuristics directly return a single candidate rather than candidate-level refined scores, NARF is inserted before their spatial selection stages.

4.2 Results

Performance on different baselines. As shown in Table 1, we evaluate the effect of the proposed framework on five zero-shot RIS methods. On the aggregate evaluation set, NARF improves both metrics for every baseline, with gains ranging from 0.32 to 3.13 points in oIoU and from 0.36 to 4.64 points in mIoU. TAS obtains the largest oIoU improvement of 3.13 points, whereas RefChess achieves the largest mIoU improvement of 4.64 points. At the split level, TAS improves on every split and metric. The remaining methods exhibit several split-specific declines: oIoU decreases on RefCOCOg-U for GLCLIP, LGD, and RefChess, and on RefCOCO and RefCOCOg-G for HybridGL; GLCLIP also shows mIoU decreases on both RefCOCOg splits. The results therefore indicate that NARF is generally beneficial for negation-focused evaluation, but its gains are not uniform across base architectures and dataset protocols.

Complementarity with CLIP fine-tuned for negation understanding. We further evaluate the complementarity between NARF and CLIP fine-tuned for negation understanding. Specifically, we select NegationCLIP Park et al. (2025) as a representative negation-aware CLIP model. Following *Know "No" Better* Park et al. (2025), we replace the original CLIP text encoder used by each baseline with the NegationCLIP text encoder. For each baseline, we evaluate two configurations: using NegationCLIP alone and combining NegationCLIP with NARF.

Table 2: Compatibility with NegationCLIP on the complete Negation-Focused Evaluation Set. "Neg." denotes NegationCLIP. Parentheses indicate gains over NegationCLIP.

Method	Config.	oIoU	mIoU
GLCLIP	Base	21.59	23.58
	Neg.	22.08	24.35
	Neg.+NARF	**22.57** (+0.49)	**24.79** (+0.44)
TAS	Base	20.92	30.28
	Neg.	25.25	35.76
	Neg.+NARF	**26.48** (+1.23)	**37.54** (+1.78)
LGD	Base	21.95	31.86
	Neg.	25.60	37.59
	Neg.+NARF	**27.80** (+2.20)	**40.28** (+2.69)
HybridGL	Base	21.72	30.85
	Neg.	23.49	33.64
	Neg.+NARF	**24.56** (+1.07)	**35.34** (+1.70)
RefChess	Base	24.48	32.66
	Neg.	24.59	34.29
	Neg.+NARF	**29.26** (+4.67)	**39.30** (+5.01)

As shown in Table 2, NegationCLIP alone improves both metrics for all five baselines on the complete Negation-Focused Evaluation Set, with particularly large gains for RefChess and LGD. The improvements nevertheless vary across methods, indicating that a more negation-sensitive text representation interacts differently with different candidate-scoring mechanisms. NARF provides further improvements when combined with NegationCLIP, yielding additional gains of 0.49–4.67 points in oIoU and 0.44–5.01 points in mIoU. RefChess obtains the largest additional improvements, reaching 4.67 oIoU and 5.01 mIoU points, while the gains for GLCLIP remain modest at 0.49 oIoU and 0.44 mIoU points. These results indicate that negation-sensitive representations and candidate-level constraint verification are complementary: improving the text embedding does not

Table 3: Ablation of reranking-score components using LGD. S_{b}, S_{t}, and S_{n} denote S_{base}, S_{target}, and S_{negative}, respectively. Ref, Ref+, G-G, and G-U denote RefCOCO, RefCOCO+, RefCOCOg-G, and RefCOCOg-U, respectively. Column-best results are shown in bold.

Metric	Setting	Ref	Ref+	G-G	G-U
oIoU	S_{b}	20.70	20.23	25.66	26.64
	$S_{\mathrm{b}} + S_{\mathrm{t}}$	20.74	21.89	25.54	**28.93**
	$S_{\mathrm{b}} + S_{\mathrm{n}}$	22.79	21.90	25.19	24.32
	Full	**24.08**	**23.06**	**28.92**	24.61
mIoU	S_{b}	27.63	30.66	36.24	35.71
	$S_{\mathrm{b}} + S_{\mathrm{t}}$	27.24	32.86	36.63	35.83
	$S_{\mathrm{b}} + S_{\mathrm{n}}$	29.61	33.38	37.81	35.62
	Full	**30.36**	**35.87**	**41.34**	**37.44**

Table 4: Ablation of verification processes using LGD. DCV and RCV denote Direct and Relational Constraint Verification, respectively.

Metric	Setting	Ref	Ref+	G-G	G-U
oIoU	Base	20.70	20.23	25.66	**26.64**
	DCV-only	23.19	**23.98**	28.62	23.23
	RCV-only	22.30	20.12	27.86	23.01
	Full	**24.08**	23.06	**28.92**	24.61
mIoU	Base	27.63	30.66	36.24	35.71
	DCV-only	29.98	**36.21**	40.11	35.93
	RCV-only	28.79	31.30	39.01	34.36
	Full	**30.36**	35.87	**41.34**	**37.44**

eliminate the need to verify whether individual candidates violate the negation constraint during selection.

4.3 Ablation Studies

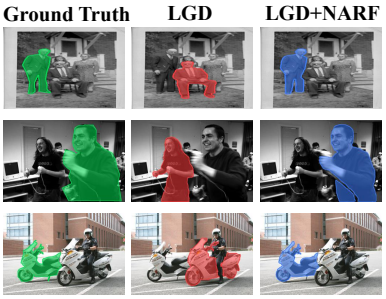
Effects of Reranking Score Components. Table 3 evaluates the contribution of each component in the reranking score. We evaluate four configurations:S_{b}, S_{b}+S_{t}, S_{b}+S_{n}, and S_{b}+S_{t}+S_{n} (Full). The S_{b} configuration directly uses the original candidate scores produced by LGD and therefore corresponds to the baseline prediction. Adding S_{t} overally improves the results because the target-compatibility score helps suppress candidates that are semantically inconsistent with the referred entity, particularly those belonging to a different object category. The S_{b}+S_{n} configuration also yields improvements overall, as S_{n} explicitly penalizes candidates that exhibit content or relations prohibited by the expression. Combining all three terms achieves the strongest overall performance. In this configuration, S_{t} distinguishes candidates that do not match the intended target, while S_{n} further differentiates among semantically similar candidates by penalizing those that violate the negation constraint.

Table 5: Effect of reranking depth K using LGD as the base method. Ref, Ref+, G-G, and G-U denote the four evaluation splits, while Overall is computed on the complete Negation-Focused Evaluation Set. Column-best results are shown in bold.

Metric	K	Evaluation Set				
		Ref	Ref+	G-G	G-U	Overall
oIoU	1	20.77	20.23	25.96	**27.36**	22.07
	3	**24.44**	21.67	28.47	25.76	23.79
	5	22.82	22.81	28.15	23.59	23.89
	10	24.08	**23.06**	**28.92**	24.61	**24.48**
	15	23.49	23.05	28.46	24.21	24.26
	20	23.55	23.05	28.46	24.21	24.27
mIoU	1	27.58	30.66	36.23	36.94	31.99
	3	**30.40**	34.06	39.51	**37.60**	34.99
	5	28.59	**36.06**	40.72	36.08	35.90
	10	30.36	35.87	**41.34**	37.44	**36.32**
	15	29.52	35.65	40.88	37.45	35.99
	20	29.52	35.65	40.88	37.45	35.99

Effects of Verification Processes. We compare the complete type-dependent framework with variants that retain only DCV or RCV. In each single process variant, the retained process is applied to every expression. As shown in Table 4, DCV-only improves mIoU on all four splits and improves oIoU on RefCOCO, RefCOCO+, and RefCOCOg-G. RCV-only improves mIoU on RefCOCO, RefCOCO+, and RefCOCOg-G and improves oIoU on RefCOCO and RefCOCOg-G. The full framework achieves the best results on RefCOCO and RefCOCOg-G for both metrics and the best mIoU on RefCOCOg-U. DCV remains slightly better on RefCOCO+, whereas the base method retains the highest RefCOCOg-U oIoU. These results show that type-dependent verification is generally more effective than relying on either process alone.

Effect of the Reranking Depth. Table 5 examines the effect of the reranking depth K. For LGD, NARF is inserted before the spatial heuristics, which remain part of the prediction pipeline. Thus,



Referring Expression: old man standing not sitting.
Baseline Mask Score: S_{base}=0.1423, S_{target}=0.1405, $S_{negative}$=0.1460
Ours Mask Score: S_{base}=0.1200, S_{target}=0.1087, $S_{negative}$=0.0969

Referring Expression: man no glasses.
Baseline Mask Score: S_{base}=0.1420, S_{target}=0.0939, $S_{negative}$=0.1168
Ours Mask Score: S_{base}=0.1029, S_{target}=0.0883, $S_{negative}$=0.0725

Referring Expression: a motorcycle without a person on it.
Baseline Mask Score: S_{base}=0.1800, S_{target}=0.1699, $S_{negative}$=0.1517
Ours Mask Score: S_{base}=0.1784, S_{target}=0.1718, $S_{negative}$=0.1448

Figure 3: Qualitative comparisons between LGD and LGD with NARF. The scores of the masks selected by LGD and NARF are reported on the right. Ground-truth, baseline, and reranked masks are shown in green, red, and blue, respectively.

$K = 1$ does not necessarily reproduce the LGD baseline prediction. Increasing K from 1 to 10 substantially improves aggregate oIoU and mIoU, indicating that a sufficiently broad candidate set is important for correcting the original ranking. Performance decreases at $K = 15$ and $K = 20$, suggesting that lower-ranked masks provide limited additional target evidence and may introduce uncertainty into the reranking process. Although the optimal depth varies across individual splits, $K = 10$ achieves the best aggregate results and performs best on both RefCOCOg splits. We therefore adopt $K = 10$ as the default setting.

4.4 Qualitative Results

Figure 3 presents three examples in which LGD selects a candidate that matches the target entity but violates the negation constraint. The first two rows illustrate Candidate-Local Negation handled by DCV. For the expression *"old man standing not sitting,"* LGD selects a seated man, whereas NARF selects the intended standing man. Although the mask selected by NARF has lower base and target-compatibility scores, its substantially lower negation-violation score allows it to be favored after score fusion. For the expression *"man no glasses,"* LGD selects a man wearing glasses and therefore violates the specified constraint. NARF instead selects a man without glasses, again supported by a lower negation-violation score. The last row illustrates Target–Reference Negation through the expression *"a motorcycle without a person on it,"* which is handled by RCV. Verifying this expression requires considering each motorcycle candidate together with potential person candidates and determining whether the prohibited *on* relation holds. LGD selects a motorcycle associated with a person on it, whereas NARF selects the intended motorcycle with a lower negation-violation score.

5 Conclusion

This paper investigates negation understanding in zero-shot referring image segmentation, where compatibility-based mask selection may identify the correct object category while overlooking constraints intended to exclude particular instances. To facilitate focused evaluation, we construct a Negation-Focused Evaluation Set from RefCOCO, RefCOCO+, and RefCOCOg. We further introduce a negation-aware candidate reranking framework that operates on the outputs of existing zero-shot RIS methods. The framework combines the original candidate scores with target-compatibility and negation-violation scores, and employs direct and relational verification to evaluate constraint violations. Experiments with five zero-shot RIS baselines show that the framework improves overall oIoU and mIoU for every evaluated method. Its additional improvements when combined with NegationCLIP further indicate that representation-level negation sensitivity and decision-level constraint verification address complementary aspects of the problem. The current framework remains sensitive to candidate-mask quality, and the localization of fine-grained reference content. Future work will explore more reliable scope resolution and relation modeling to extend the coverage of negation-aware RIS.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, literature retrieval and discovery, and research brainstorming, including discussion of methodological and experimental alternatives. AI tools were also used to assist with code editing and debugging. All AI-assisted content and identified references were reviewed and verified by the authors against the original sources. All experiments and reported results were conducted and validated by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

Our Negation-Focused Evaluation Set is derived from the publicly available RefCOCO, RefCOCO+, and RefCOCOg benchmarks. It is constructed by selecting and manually categorizing existing validation examples, without collecting new data from human participants or introducing demographic or sensitive-attribute annotations. Because the evaluation data and NARF rely on existing datasets and pretrained vision–language and language models, they may inherit representational, linguistic, and annotation biases. Accordingly, results on the Negation-Focused Evaluation Set should not be interpreted as evidence of demographic fairness or reliability in high-stakes deployment settings.

REPRODUCIBILITY STATEMENT

Section 3 specifies the NARF inference procedure, including candidate selection, target-compatibility scoring, type-dependent Direct and Relational Constraint Verification, and score fusion. Section 4 documents the evaluation datasets and metrics, evaluated baselines and backbones, hardware and model configurations, candidate and reference selection depths, normalization temperatures, and fusion weights. Appendix A describes the construction and manual subtype annotation of the Negation-Focused Evaluation Set, while Appendix B provides the exact negation classification prompt and implementation details for mask and blurring strategy, relation-specific geometric functions, and score standardization and normalization.

REFERENCES

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip HS Torr, Yoon Kim, and Marzyeh Ghassemi. Vision-language models do not understand negation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29612–29622. IEEE, 2025.

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 (ICCV)*, pp. 16321–16330, October 2021.

Radina Dobreva and Frank Keller. Investigating negation in pre-trained vision-and-language models. In *Proceedings of the fourth blackboxnlp workshop on analyzing and interpreting neural networks for nlp*, pp. 350–362, 2021.

Zeyu Han, Fangrui Zhu, Qianru Lao, and Huaizu Jiang. Zero-shot referring expression comprehension via structural similarity between images and captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14364–14374. IEEE, 2024.

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

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*, pp. 4904–4916, 2021.

Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. ReSTR: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18145–18154, 2022.

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)*, pp. 4015–4026, 2023.

Jiachen Li, Qing Xie, Xiaojun Chang, Jinyu Xu, and Yongjian Liu. Mutually-guided hierarchical multi-modal feature learning for referring image segmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(12):393:1–393:18, 2024. doi: 10.1145/3698771.

Jiachen Li, Qing Xie, Renshu Gu, Jinyu Xu, Yongjian Liu, and Xiaohan Yu. LGD: Leveraging generative descriptions for zero-shot referring image segmentation. *Pattern Recognition*, 172: 112549, 2026. doi: 10.1016/j.patcog.2025.112549.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, 2023.

Wenhui Li, Chao Pang, Weizhi Nie, Hongshuo Tian, and An-An Liu. Bidirectional mask selection for zero-shot referring image segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1):911–921, 2025. doi: 10.1109/TCSVT.2024.3460874.

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 (ICCV)*, pp. 1271–1280, Oct 2017.

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 (CVPR)*, pp. 29634–29643, 2025.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10910–10921. IEEE, 2023.

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 (CVPR)*, pp. 11–20, 2016.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European conference on computer vision*, pp. 792–807. Springer, 2016.

Minheng Ni, Yabo Zhang, Kailai Feng, Xiaoming Li, Yiwen Guo, and Wangmeng Zuo. Ref-diff: Zero-shot referring image segmentation with generative models. *Science China Information Sciences*, 69(8):182104, 2026.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8253–8280, 2022.

Junsung Park, Jungbeom Lee, Jongyoon Song, Sangwon Yu, Dahuin Jung, and Sungroh Yoon. Know "no" better: A data-driven approach for enhancing negation awareness in clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2825–2835. IEEE, 2025.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learning the power of "no": Foundation models with negations. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 7991–8001, February 2025.

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, 2023.

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.

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 (CVPR)*, pp. 11686–11695, 2022a.

Ziyue Wang, Aozhu Chen, Fan Hu, and Xirong Li. Learn to understand negation in video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 434–443, 2022b.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. LAVT: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18155–18165, 2022.

Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10502–10511, June 2019.

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 (CVPR)*, pp. 19456–19465, 2023.

A Appendix: Negation-Focused Evaluation Set

A.1 Data Source

We construct the Negation-Focused Evaluation Set from the validation splits of three widely-used benchmarks for the referring image segmentation task, including RefCOCO Nagaraja et al. (2016), RefCOCO+ Nagaraja et al. (2016), and RefCOCOg Mao et al. (2016). For RefCOCOg, we consider both the Google and UMD partitions. Specifically, RefCOCO typically contains short expressions and often includes explicit spatial words, which are banned in RefCOCO+, and RefCOCOg contains longer and more detailed expressions, with richer contextual and relational descriptions. The expressions in Negation-Focused Evaluation Set are selected from the existing dataset annotations rather than synthetically generated.

To identify negation-containing expressions from original datasets, we first employ a large language model to retrieve potentially relevant expressions from the validation data according to whether a referring expression contains explicit negation cues. We then manually inspect the retrieved candidates and remove clear noise and retrieval errors. After this verification process, the final evaluation data comprises 705 negation-containing expressions with their original image and target annotations preserved.

A.2 Data Distribution

We next examine how these expressions are distributed across the three source datasets. As summarized in Table 6, RefCOCO contains 10,834 validation expressions, of which 102 contain negation, accounting for 0.94%. RefCOCO+ contains 388 negation-containing expressions among 10,758 expressions, corresponding to the highest proportion of 3.61%. Under the two RefCOCOg partition protocols, RefCOCOg-G contains 139 negation-containing expressions out of 9,536 expressions (1.46%), whereas RefCOCOg-U contains 76 out of 4,896 expressions (1.55%). Together, the four evaluated partitions provide 705 negation-containing expressions for the subsequent evaluation.

Although negation-containing expressions constitute only a small fraction of the original validation data, they provide a focused test of exclusion reasoning. A negated constraint can change which instance is the correct referent rather than merely add secondary descriptive information. Consequently, selecting a candidate that violates this constraint produces an incorrect target even when its object category is correct. Because non-negated expressions dominate the original validation

Table 6: Distribution of negation-containing expressions in the evaluated validation partitions. RefCOCOg-G and RefCOCOg-U follow the Google and UMD partition protocols, respectively.

Dataset	Total Expressions	Negative Expressions	Ratio (%)
RefCOCO	10,834	102	0.94
RefCOCO+	10,758	388	3.61
RefCOCOg-G	9,536	139	1.46
RefCOCOg-U	4,896	76	1.55

Table 7: Classification of the negation-containing expressions. Ratios are calculated over all 705 expressions.

Type	Subtype	Count	Ratio (%)	Example
Candidate-Local	Attribute negation	58	8.23	*man not in purple*
	Partial visibility	114	16.17	*giraffe not seen completely*
	Action negation	25	3.55	*guy in black not jumping*
	Target exclusion	117	16.60	*person walking near umbrella not the boy*
	Accessory absence	192	27.23	*man without hat black shirt*
Target–Reference	Spatial relation negation	87	12.34	*the dark chair with no light on it*
	Relational action negation	73	10.35	*man not holding beer*
Other/uncertain	–	39	5.53	–

sets, aggregate full-set metrics are driven primarily by conventional expressions and can therefore obscure failures specific to negation. We thus report performance on the negation-containing expressions separately from standard full-set evaluation.

A.3 Negation Classification

To characterize the reasoning required by negated expressions, we manually inspect the 705 expressions and assign each one to a fine-grained subtype according to the visual evidence required for constraint verification. The manual subtype annotations are used to characterize the evaluation data, and inference-time routing in NARF is performed automatically. As summarized in Table 7, this process yields seven negation subtypes.

Attribute negation specifies a prohibited property or state of the target, while partial visibility indicates that the target or one of its parts is not fully visible because of occlusion, image cropping, or viewpoint. Action negation specifies an action that the target does not perform without requiring an explicit reference object. Target exclusion directly rules out a competing entity or region, whereas accessory absence specifies that the target lacks an associated object. Because these five subtypes constrain the target itself without requiring a distinct reference entity, we group them as Candidate-Local Negation. Together, they account for 506 expressions. Accessory absence is the largest subtype with 192 expressions, followed by target exclusion with 117.

The remaining two subtypes require evidence involving both the target and another entity. Spatial relation negation specifies that the target does not satisfy a particular spatial configuration with respect to a reference object, whereas relational action negation denies an interaction between the target and a reference entity. We therefore group these subtypes as Target–Reference Negation. Spatial relation negation and relational action negation contain 87 and 73 expressions, respectively, for a total of 160 expressions. The remaining 39 expressions are labeled as other/uncertain because they cannot be assigned reliably to one of the seven subtypes.

B Appendix: Method Details

B.1 Negation Classification Prompt

We use a large language model to assign each referring expression to one of the predefined negation subtypes in NARF. The exact system prompt used in our experiments is provided below.

Negation Classification Prompt

```
You are a strict eight-class classifier for negative referring expressions. Return
exactly one label and nothing else.

Use the following decision procedure in order.

1. other_or_uncertain
Use this when the negative word is not a reliable visual constraint, the text is
conversational/noisy/incomplete, "no" means a number or appears in written
text, or the intended visual meaning cannot be classified confidently.
Examples:
"not sure which object" -> other_or_uncertain
"athlete with number no. 2" -> other_or_uncertain
"unclear target, cannot decide" -> other_or_uncertain
Do not force an unclear expression into another class.

2. attribute_negation
Use this for a negated color, material, appearance, shape, size, condition, or other
non-action and non-spatial property.
Examples:
"person not wearing a purple color" -> attribute_negation
"vehicle not red" -> attribute_negation
"object not clean" -> attribute_negation
Counterexample: "person no hat" is accessory_absence because a hat is a missing item.
Counterexample: "object not fully visible" is partial_visibility.
```

```
3. partial_visibility
Use this when the negative meaning says the target itself is only partially visible,
incomplete, occluded, outside the frame, not clearly seen, or when a body part such
as face, head, body of the target is not visible. This class is about the targets
visibility/completeness, not about a removable accessory.
Examples:
"object not fully visible" -> partial_visibility
"person with face unseen" -> partial_visibility
Important counterexamples:
"person with no hat" is accessory_absence, because a hat is a wearable accessory.
"container with no objects in it" is spatial_relation_negation, because the missing
objects are related contents of the container.

4. target_exclusion
Use this only when the speaker explicitly rejects a competing candidate and selects
another target: choose A, not B; A rather than B; A not B. There must be two
competing candidates, objects, or alternatives.
Examples:
"the cup not the plate" -> target_exclusion
Important counterexamples:
"person with no hat" is accessory_absence, not target_exclusion.
"object not blue" is attribute_negation, not target_exclusion.
A phrase containing "not" alone is never sufficient for target_exclusion.

5. accessory_absence
Use this only when the target lacks a wearable item, accessory, clothing item, or
carried/personal item. This includes without X, no X, does not have X, and missing
X, but only for accessories/clothing/portable possessions attached to or associated
with the target.
Do not use accessory_absence for body parts, target visibility, object contents or
other intrinsic properties. Those should be partial_visibility,
spatial_relation_negation, or attribute_negation depending on the meaning.
Examples:
"person no hat" -> accessory_absence
"person with no jacket" -> accessory_absence
Counterexamples:
"table with no cup on it" is spatial_relation_negation, not accessory_absence.
"person with face unseen" is partial_visibility, not accessory_absence.
An explicit negated interaction such as "not holding beer" belongs to the action
decision below.

6. spatial_relation_negation
Use this when a spatial or object-relation constraint between the target and another
object, group, region, or scene location is negated. Relations include in, on,
under, above, near, behind, beside, inside, within, with, among, close to, and
object/content relations such as no X in/on/with target, no one in/on target, or
nothing in/on target.
Examples:
"table with no cup on it" -> spatial_relation_negation
"food item not behind another food item" -> spatial_relation_negation
"container with no objects in it" -> spatial_relation_negation
Do not use this merely because the sentence also contains an unrelated location
phrase. If the missing item is a wearable/carried accessory of the target, use
accessory_absence instead.

7. Decide between the two action classes.
An action includes running, jumping, standing, sitting, grazing, eating, looking,
facing, smiling, participating, cutting, touching, feeding, playing, holding,
catching, or another event/pose.

referenced_action_negation:
The negated action directly acts on, contacts, or is defined relative to a named
concrete scene object/entity. The object must participate in the action. Use this
for not holding/carrying/touching/cutting/playing with/facing/eating a named object,
and for sitting/standing/lying on a named support object.
Examples:
"person not eating food" -> referenced_action_negation
"person not holding the drink" -> referenced_action_negation
"object not touching another object" -> referenced_action_negation

no_ref_action_negation:
```

```
The negated action has no named concrete scene object/entity as its direct
participant. Deictic words such as us/viewer do not count as a scene object. Use
this for bare actions or poses such as not running, not sitting, not jumping, not
standing, not grazing, or not eating when no concrete object is named as the action
participant.
Examples:
"person not running" -> no_ref_action_negation
"person not jumping" -> no_ref_action_negation
Do not infer an unmentioned object. For example, "not eating" is no-reference, while
"not eating food" has a referenced object.

Allowed labels:
attribute_negation
target_exclusion
accessory_absence
partial_visibility
referenced_action_negation
no_ref_action_negation
spatial_relation_negation
other_or_uncertain

Return only the exact label. Do not explain, paraphrase, extract entities, or
perform any other task.
```

B.2 Mask and Blurring Strategy

We adopt a mask and blurring strategy to construct the visual representation of each candidate mask. Given an image $I \in \mathbb{R}^{H\times W\times 3}$ and a binary candidate mask $m_i \in \{0,1\}^{H\times W}$, we first apply Gaussian blurring to the full image:

$$I^{\text{blur}} = G_r * I, \qquad r = \max\left(3, \frac{\min(H,W)}{30}\right), \tag{11}$$

where G_r denotes a Gaussian filter with blur radius r and $*$ denotes the filtering operation. The same radius is used for all candidate masks from an image.

We then preserve the original pixels inside the candidate mask and replace the pixels outside it with their blurred counterparts:

$$I_i^{\text{mix}} = m_i \odot I + (1 - m_i) \odot I^{\text{blur}}, \tag{12}$$

where the binary mask is broadcast along the RGB channel dimension and $\odot$ denotes element-wise multiplication. Thus, the appearance of the candidate foreground remains unchanged, while fine-grained content outside the candidate is suppressed without completely removing its surrounding context.

Let b_i denote the tight bounding box enclosing the foreground pixels of m_i. We crop the composed image to this bounding box:

$$\widetilde{I}_i^{\text{MB}} = \text{Crop}\big(I_i^{\text{mix}}, b_i\big)\,. \tag{13}$$

Because the resulting crop may not be square, we center it on a square canvas whose side length is the larger dimension of the crop. The padding color is set to the mean RGB value of I_i^{mix}. We denote the resulting square candidate representation by

$$I_i^{\text{MB}} = \text{SquarePad}\Big(\widetilde{I}_i^{\text{MB}}, \mu\big(I_i^{\text{mix}}\big)\Big)\,, \tag{14}$$

where $\mu(\cdot)$ computes the channel-wise mean color.

B.3 Relation-Specific Geometric Functions

RCV uses relation-specific geometric functions to evaluate spatial relations between a target candidate m_i and a reference candidate m_j. Specifically, we use their corresponding bounding boxes b_i and b_j, together with the bounding-box centers $c_i = (x_i^c, y_i^c)$ and $c_j = (x_j^c, y_j^c)$, to compute the geometric relation evidence. Let A_i and A_j denote the areas of b_i and b_j, respectively, and let W

and H denote the image width and height. We define the normalized center distance as

$$d_{ij} = \sqrt{\left(\frac{x_i^c - x_j^c}{W}\right)^2 + \left(\frac{y_i^c - y_j^c}{H}\right)^2}. \tag{15}$$

The relation-specific geometric function is defined as

$$\Phi_\rho(b_i, b_j; W, H) = \begin{cases} \left[\frac{x_j^c - x_i^c}{W}\right]_+, & \rho \in \mathcal{R}_{\text{left}}, \\ \left[\frac{x_i^c - x_j^c}{W}\right]_+, & \rho \in \mathcal{R}_{\text{right}}, \\ \left[\frac{y_j^c - y_i^c}{H}\right]_+, & \rho \in \mathcal{R}_{\text{above}}, \\ \left[\frac{y_i^c - y_j^c}{H}\right]_+, & \rho \in \mathcal{R}_{\text{below}}, \\ \frac{|b_i \cap b_j|}{A_i}, & \rho \in \mathcal{R}_{\text{inside}}, \\ \exp\left(-\frac{d_{ij}}{\sigma}\right), & \rho \in \mathcal{R}_{\text{near}}, \\ \left[1 - \frac{A_i}{A_j}\right]_+, & \rho \in \mathcal{R}_{\text{smaller}}, \\ \left[1 - \frac{A_j}{A_i}\right]_+, & \rho \in \mathcal{R}_{\text{bigger}}, \end{cases} \tag{16}$$

where $[x]_+ = \max(0, x)$.σ controls the decay rate of the proximity score. We set $\sigma = 0.25$ in all experiments. ρ denotes the spatial relation expressed by the negated predicate, and each $\mathcal{R}_d$ denotes the semantic class associated with a canonical spatial relation d. The parser maps semantically equivalent relation expressions to their corresponding canonical classes before applying Φ_ρ. In equation 16, the relation is formulated in the target-to-reference direction; when the parsed relation is expressed in the reverse direction, we reorder the inputs as (b_j, b_i) before evaluating Φ_ρ.

B.4 Score Standardization and Normalization

NARF is designed to operate with different zero-shot RIS methods, whose method-specific candidate scores may vary substantially in numerical scale and distribution. Directly combining these scores with the target-compatibility and negation-constraint scores would cause the same fusion weights to have different effects across base methods. To facilitate consistent score fusion across different base methods, we independently apply z-score standardization followed by temperature-scaled softmax normalization to each score component over the target candidate set $\mathcal{C}_K$.

Let $S_q(m_i)$ denote the raw score of candidate m_i for component $q \in \{\text{base}, \text{target}, \text{negative}\}$. We first compute the mean and standard deviation of this component over the K target candidates:

$$\mu_q = \frac{1}{K} \sum_{m_i \in \mathcal{C}_K} S_q(m_i), \qquad \sigma_q = \sqrt{\frac{1}{K} \sum_{m_i \in \mathcal{C}_K} (S_q(m_i) - \mu_q)^2}. \tag{17}$$

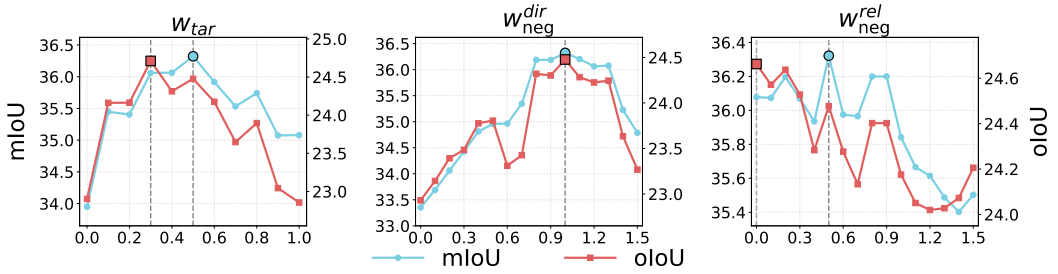


Figure 4: Parameter sensitivity analysis of w_{tar}, $w_{\text{neg}}^{\text{dir}}$, and $w_{\text{neg}}^{\text{rel}}$.

The standardized score is given by

$$Z_q(m_i) = \frac{S_q(m_i) - \mu_q}{\sigma_q}. \tag{18}$$

We then apply softmax normalization:

$$\widehat{S}_q(m_i) = \frac{\exp\left(Z_q(m_i)/T\right)}{\sum_{m_j \in \mathcal{C}_K} \exp\left(Z_q(m_j)/T\right)}, \tag{19}$$

where $T > 0$ is the softmax temperature. A smaller T produces a sharper distribution, whereas a larger T reduces the differences among candidates. We use $T = 6$ for all score components in our experiments.

The normalized scores $\widehat{S}_{\text{base}}$, $\widehat{S}_{\text{target}}$, and $\widehat{S}_{\text{negative}}$ are subsequently used in the score fusion.

C Appendix: Additional Experiments and Failure Cases

Parameter Sensitivity. We analyze the sensitivity of the framework to the three weights in the negation-aware reranking score. Specifically, we vary one parameter at a time while fixing the remaining parameters at their default values. As shown in Figure 4, for target-compatibility weights, oIoU reaches its maximum at $w_{\text{tar}} = 0.3$, whereas mIoU peaks at $w_{\text{tar}} = 0.5$. Performance generally decreases when w_{tar} approaches 1.0, suggesting that target compatibility provides useful complementary evidence but may dominate the base and negation-violation scores when assigned excessive importance. For $w_{\text{neg}}^{\text{dir}}$, both metrics improve over a broad intermediate range and attain their maximum at $w_{\text{neg}}^{\text{dir}} = 1.0$, before declining at larger values. This result indicates that direct verification requires a sufficiently strong penalty to suppress candidates exhibiting prohibited content, while an overly large weight may amplify uncertainty in the constraint-violation score. The relational verification weight w_{neg}^{rel} exhibits a slightly different optimum across the two metrics. The highest mIoU is obtained at $w_{\text{neg}}^{rel} = 0.5$, while slightly higher oIoU are observed at smaller values of w_{neg}^{rel}. Both metrics generally deteriorate when w_{neg}^{rel} exceeds 1.0, suggesting that excessive relational penalties may suppress valid target candidates. We therefore set $w_{\text{neg}}^{rel} = 0.5$ to achieve a favorable balance between the two metrics.

Additional qualitative results. Figures 5–8 present additional qualitative results using LGD and TAS as the base methods. Across all four figures, ground-truth masks, base method predictions, and predictions after NARF reranking are shown in green, red, and blue, respectively. The score components of the masks selected before and after reranking are reported below each example. Figures 5 and 6 show Candidate-Local examples processed by DCV, whereas Figures 7 and 8 show Target–Reference examples processed by RCV.

Failure Cases. Although the proposed framework consistently improves most zero-shot RIS methods on negative referring expressions, we observe several failure cases in which the reranking process produces an incorrect candidate. These failures mainly arise from unsupported constraint scopes and ambiguous visual evidence. The first row of Figure 9 illustrates a limitation of the constraint scope currently supported by NARF. In the expression *"girl just behind boy in yellow shirt and no jacket,"* the phrase *"no jacket"* specifies a negated property of the reference object, namely the boy, rather than the referred target. NARF currently verifies only negation constraints imposed on the referred target and does not support negated properties of reference objects. It therefore cannot correctly verify whether a candidate girl stands behind a boy satisfying the *"no jacket"* constraint, which may cause the reranking process to favor a candidate that does not correspond to the ground-truth target.

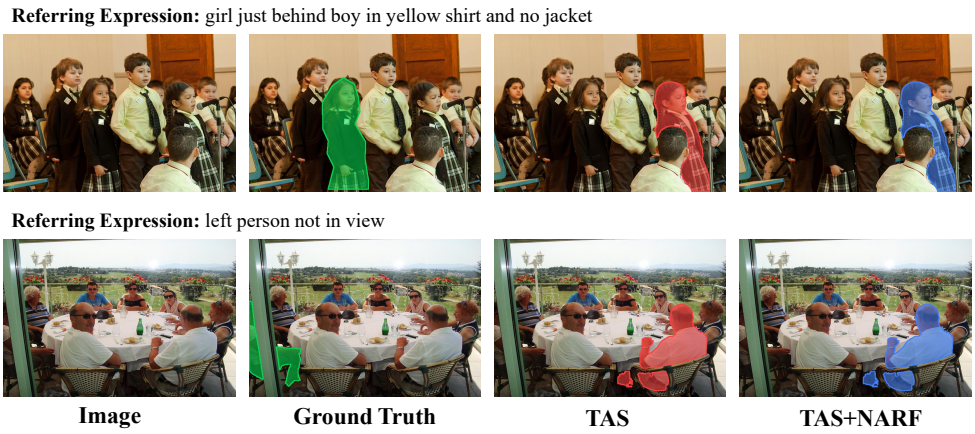


Figure 9: Failure cases of the proposed negation-aware reranking framework. Green, red, and blue masks denote the ground truth, TAS prediction, and TAS with NARF, respectively.

The second row illustrates a partial-visibility expression, *"left person not in view."* The ground-truth target is only partially visible at the image boundary, whereas TAS selects a fully visible person. Such expressions are difficult to verify because unobserved object content may result from occlusion or image cropping rather than an intrinsic visual property of the target. We therefore treat partial visibility as an unsupported subtype and retain the prediction of the base method in NARF.

Ground Truth	LGD	LGD+NARF



Referring Expression: the cute thing that doesnt have a hat.

Baseline Mask Score: S_{base}=0.1461, S_{target}=0.1258, $S_{negative}$=0.1069

NARF Mask Score: S_{base}=0.1253, S_{target}=0.1063, $S_{negative}$=0.0753



Referring Expression: tofu dog not the sign.

Baseline Mask Score: S_{base}=0.1273, S_{target}=0.1253, $S_{negative}$=0.1257

NARF Mask Score: S_{base}=0.1227, S_{target}=0.1150, $S_{negative}$=0.0854



Referring Expression: red car that is not the corvette.

Baseline Mask Score: S_{base}=0.1592, S_{target}=0.1457, $S_{negative}$=0.1688

NARF Mask Score: S_{base}=0.1373, S_{target}=0.1424, $S_{negative}$=0.0963



Referring Expression: pirota whatever not wearing cap guy.

Baseline Mask Score: S_{base}=0.2596, S_{target}=0.1644, $S_{negative}$=0.2268

NARF Mask Score: S_{base}=0.2009, S_{target}=0.1864, $S_{negative}$=0.1588

Figure 5: Additional qualitative results of DCV using LGD as the base method.

Ground Truth	TAS	TAS+NARF

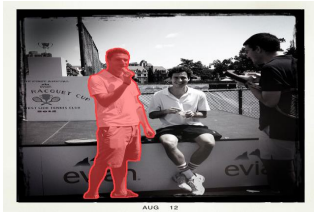




Referring Expression: guy without mask.

Baseline Mask Score: S_{base}=0.2970, S_{target}=0.2873, $S_{negative}$=0.3127

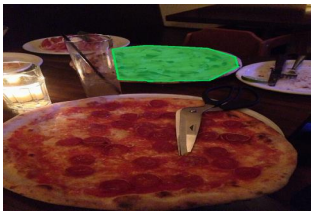
NARF Mask Score: S_{base}=0.1975, S_{target}=0.2941, $S_{negative}$=0.1958


Referring Expression: person not in white.

Baseline Mask Score: S_{base}=0.2354, S_{target}=0.2143, $S_{negative}$=0.2646

NARF Mask Score: S_{base}=0.2311, S_{target}=0.2485, $S_{negative}$=0.2118





Referring Expression: pizza not pepperoni.

Baseline Mask Score: S_{base}=0.1428, S_{target}=0.1302, $S_{negative}$=0.1447

NARF Mask Score: S_{base}=0.1213, S_{target}=0.1178, $S_{negative}$=0.1125





Referring Expression: white tshirt no hat.

Baseline Mask Score: S_{base}=0.1388, S_{target}=0.1133, $S_{negative}$=0.1346

NARF Mask Score: S_{base}=0.1188, S_{target}=0.1145, $S_{negative}$=0.0739

Figure 6: Additional qualitative results of DCV using TAS as the base method.



Figure 7: Additional qualitative results of RCV using LGD as the base method.

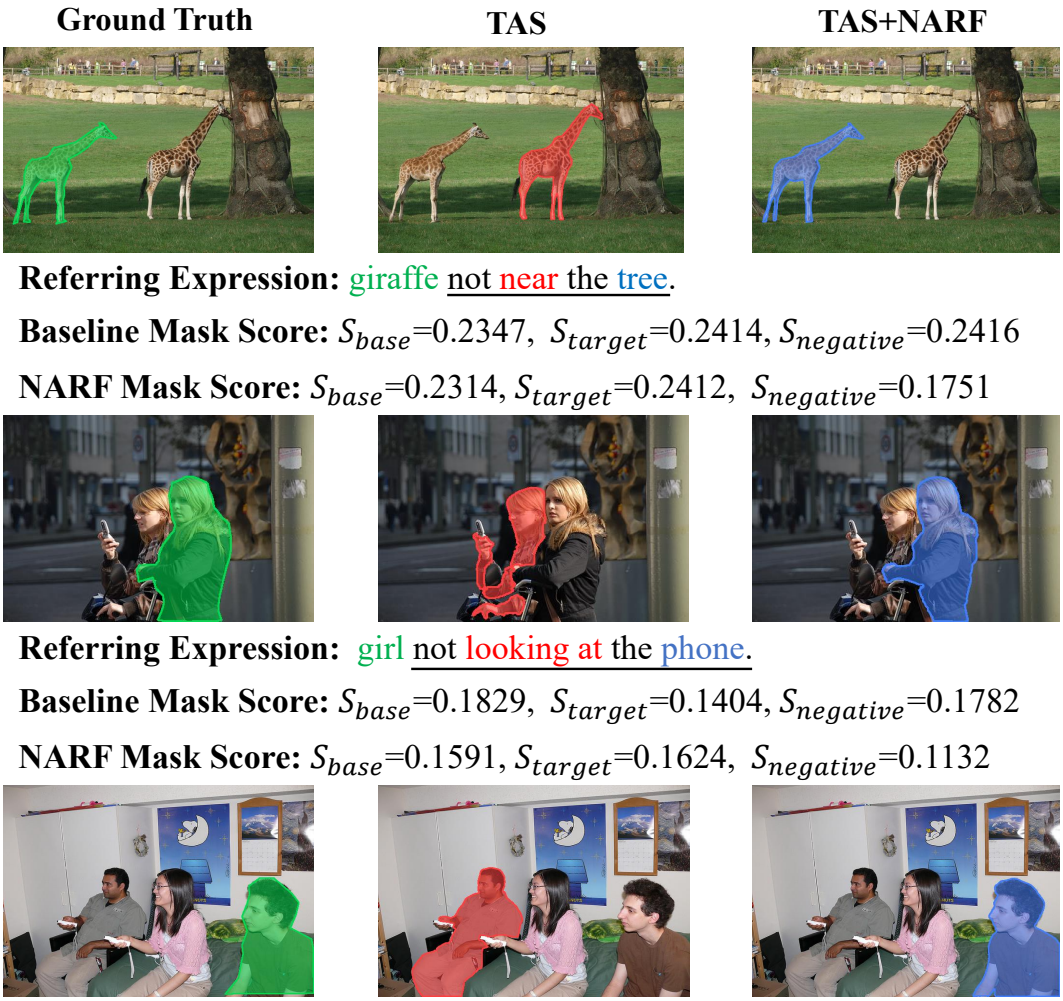


Figure 8: Additional qualitative results of RCV using TAS as the base method.