DECO: EFFICIENT DECOUPLE-TO-COUPLE LEARNING FOR MULTI-TASK VISUAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-task visual grounding requires models to jointly understand linguistic semantics and perform accurate visual localization and segmentation. Despite the success of multimodal large language models, effectively adapting them to multiple grounding objectives remains challenging. Existing methods commonly enforce task cooperation through shared representations, while overlooking the intrinsic conflict between task-oriented feature interests. In this paper, we introduce **DeCo**, an efficient **De**couple-to-**Co**uple learning framework that resolves this dilemma through a two-stage paradigm: task-specific representation decoupling followed by complementary prior coupling. Specifically, we first propose Task-aware Semantic Decoupling (TSD) to route shared visual cues into individual features under salient word-level guidance, alleviating representation interference between localization and segmentation. Furthermore, we observe that segmentation naturally provides informative localization priors due to dense supervision. Based on this insight, we introduce Hybrid Prior Coupling (HPC), which integrates sentence-level semantic prior with mask-derived spatial prior for enhanced grounding. Built upon a frozen multimodal encoder, DeCo requires lightweight trainable parameters while achieving strong generalization across multiple grounding objectives. Extensive experiments on RefCOCO/+, G-Ref, ReferIt, Flickr, DIOR-RSVG, SARVG1.0, RRSIS-D, RIS-LAD, and RefDIOR demonstrate that DeCo achieves state-of-the-art performance on both natural and remote sensing benchmarks. The code and model will be publicly available.

1 INTRODUCTION

Multi-task visual grounding (MTVG) aims to simultaneously comprehend a natural language description and identify its corresponding visual target through multiple grounding objectives (Liu et al., 2023b; Xiao et al., 2024b; Yu et al., 2025; Dai et al., 2025a), primarily including referring expression comprehension (REC) (Deng et al., 2023; Xiao et al., 2024a; Su et al., 2024; Dai et al., 2024) and referring expression segmentation (RES) (Yang et al., 2022b; 2024; Mao et al., 2025; Chen et al., 2026a; Moon et al., 2026). By bridging linguistic concepts and visual content, MTVG serves as a fundamental capability for vision-language understanding (Xiao et al., 2025; Ding et al., 2025). Recent advances in multimodal large language models (MLLMs) (Liu et al., 2023a; Chen et al., 2024c; Wu et al., 2024b) have substantially improved grounding performance by leveraging informative, aligned cross-modal representations learned from extensive pre-training (Xia et al., 2024; Chen et al., 2024a; Wang et al., 2025). However, effectively adapting these foundation models to multiple grounding objectives remains challenging, especially when different tasks require distinct and sometimes conflicting visual representations (Lu et al., 2025; Ding et al., 2026).

Existing methods for MTVG can be broadly categorized into two paradigms. The first one leverages MLLM as a multimodal encoder, which typically regresses the coordinates of the referred object (Tang et al., 2026; Chen et al., 2026b; Kang et al., 2026) for REC, and employs various strategies, such as prompting a special token into the mask decoder (Lai et al., 2024; Liu et al., 2025b; Yang et al., 2026) of SAM (Kirillov et al., 2023; Ravi et al., 2025), regressing sampled contour points (Wang et al., 2023a; Pramanick et al., 2024; Wu et al., 2024a), or generating a text label corresponding to its image patch (Lan et al., 2025; 2026), for RES. These methods exploit the strong reasoning capability of MLLM but come with billions of parameters and additional post-processing procedures. The other one employs pretrained vision-language foundation models as

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

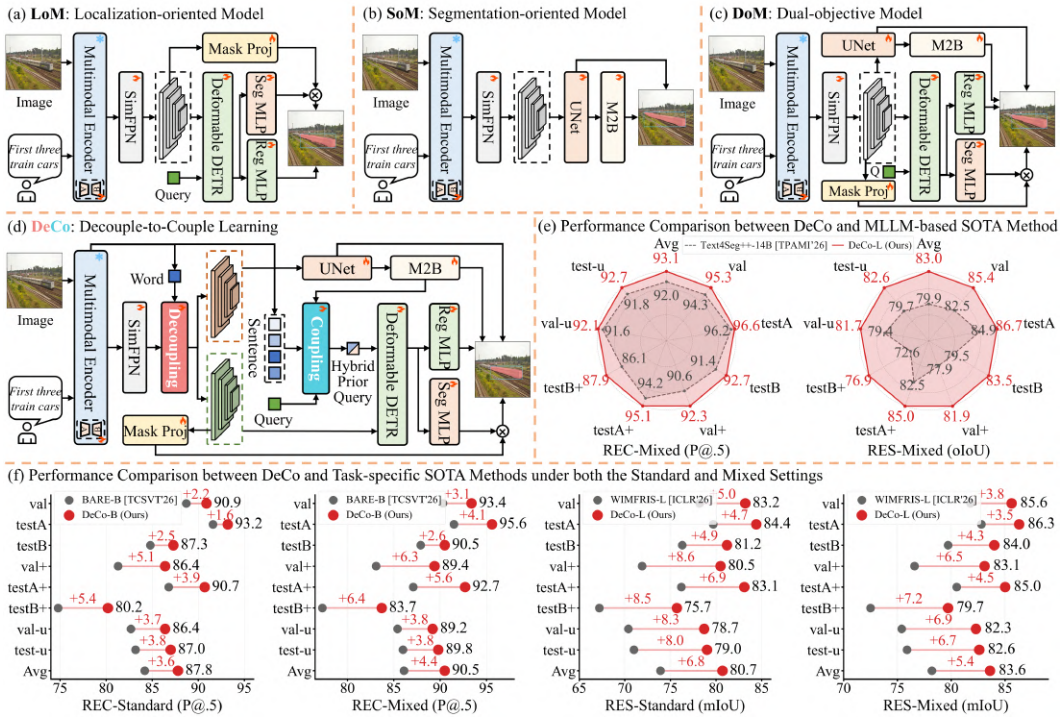

Figure 1: Overview. (a) LoM, which adopts a DETR-based detector with a query-based mask projection branch. (b) SoM, which employs a UNet decoder followed by a mask-to-box (M2B) operation. (c) DoM, which integrates LoM and SoM upon the shared multi-scale features. (d) DeCo, which follows a decouple-to-couple learning paradigm to disentangle discriminative representations and harness hybrid priors for MTVG. (e) Comparison between DeCo and MLLM-based method, Text4Seg++ (Lan et al., 2026), on both REC and RES. (f) Comparison between DeCo and REC-specific method, BARE (Li et al., 2026a), and RES-specific method, WIMFRIS (Moon et al., 2026).

multimodal encoders (Radford et al., 2021; Wang et al., 2023b) and designs task-specific architectures to perform REC and RES (Chen et al., 2024b; Dai et al., 2025b;c). This paradigm capitalizes on task-oriented preference modeling but typically requires fully training the entire model. However, both of them fundamentally rely on shared visual representations across distinct tasks. Despite this design facilitating knowledge transfer and reducing model complexity, it implicitly assumes that different objectives can benefit from a unified latent space. In practice, localization emphasizes high-level semantic correspondence and object-level spatial reasoning, whereas segmentation requires fine-grained spatial details and pixel-level discrimination (Lu et al., 2025; Ding et al., 2026).

To examine this assumption, we first construct three simple yet effective baselines based on a frozen multimodal encoder (Wang et al., 2023b) with low-rank adaptation (Hu et al., 2022), followed by a feature pyramid neck (Li et al., 2022) that expands a single-layer encoder output into multi-scale features for MTVG, as illustrated in Figure 1: (a) The Localization-oriented Model (LoM) interacts a randomly initialized object query with multi-scale features through a deformable decoder (Zhu et al., 2021) to perform object-level localization, while projecting the output query and multiplying it with the high-resolution feature to generate its corresponding pixel-level segmentation; (b) The Segmentation-oriented Model (SoM) feeds multi-scale features into a UNet decoder (Ronneberger et al., 2015) to predict a global mask, from which the minimal enclosing bounding box is subsequently derived; (c) By integrating LoM and SoM built upon shared multi-scale features, the Dual-objective Model (DoM) jointly optimizes box regression and mask prediction within a unified space. Although the results in Table 5 demonstrate that all three baselines already outperform recent methods BARE (Li et al., 2026a) and AMLRIS (Chen et al., 2026a) on REC and RES, respectively, we observe that LoM achieves stronger performance on REC but is relatively weaker on RES, whereas SoM exhibits the opposite trend, indicating that REC and RES have different feature preferences. More importantly, DoM largely inherits the respective strengths of both, yet fails to obtain substantial additional gains from multi-task learning. This finding suggests that simply combining multiple task heads is insufficient to fully exploit task complementarity. Instead, the shared representation itself may hinder each objective from developing its own task-oriented feature interests.

In this paper, we propose **DeCo**, an efficient **De**couple-to-**Co**uple learning framework to address the above limitation, as shown in Figure 1 (d). Unlike existing methods (Liu et al., 2023b; Chen et al., 2024b; Xiao et al., 2024b; Dai et al., 2025c; Yu et al., 2025; Dai et al., 2025a) which directly optimize multiple objectives over shared visual space, DeCo introduces a two-stage paradigm that separates task-specific representation learning from cross-objective prior exploitation, allowing each task to preserve its own feature preferences while benefiting from complementary knowledge. First, we propose **Task-aware Semantic Decoupling (TSD)** to transform shared multi-scale features into individual representations for localization and segmentation, respectively. TSD leverages a salient word-level feature to modulate visual features and employs lightweight experts to extract task-oriented feature interests of REC and RES. This design enables each task to learn characteristics better aligned with its preferences while retaining the efficiency of shared multi-scale encoding. Second, we introduce **Hybrid Prior Coupling (HPC)** to enrich the object query with complementary textual and spatial priors. Motivated by our observation that the minimal enclosing bounding box derived from RES rapidly achieves accurate localization during early training, HPC incorporates a global sentence-level feature to capture holistic semantic guidance and adaptively fuses the mask-based coordinate as an explicit reference into the query and its position embedding. This design enables REC to benefit from the spatial knowledge acquired through dense segmentation, while the localization objective in turn provides geometry constraints for refining the segmentation results. Overall, DeCo establishes a novel MTVG paradigm that first alleviates task interference through representation decoupling and then exploits task complementarity through prior coupling.

Extensive experiments on RefCOCO/+, G-Ref, ReferIt, Flickr, DIOR-RSVG, RRSIS-D, and other benchmarks demonstrate the effectiveness and generalization of our method. As shown in Figure 1 (e), DeCo consistently surpasses MLLM-based SOTA method (Lan et al., 2026) on both REC and RES, improving the average performance by 1.1% and 3.1%, respectively. Moreover, Figure 1 (f) illustrates that DeCo achieves significant improvements over task-specific SOTA methods (Li et al., 2026a; Moon et al., 2026) under both standard and mixed settings, with average gains of 3.6%/4.4% on REC and 6.8%/5.4% on RES, respectively. In summary, our main contributions are as follows:

- We identify and revisit the representation-sharing paradigm for MTVG, showing that REC and RES exhibit distinct visual preferences and that simply jointly optimizing them over shared representations is insufficient to fully exploit their complementarity.
- We propose DeCo, an efficient Decouple-to-Couple learning framework, which first separates task-oriented representation learning through Task-aware Semantic Decoupling (TSD) and subsequently exploits complementary semantic and spatial priors through Hybrid Prior Coupling (HPC), providing a simple and effective paradigm for MTVG.
- We establish strong empirical results with parameter-efficient adaptation, where DeCo introduces a small number of trainable parameters and achieves state-of-the-art performance across multiple REC and RES benchmarks, consistently outperforming both MLLM-based and task-specific competitors under different evaluation settings.

2 Related Work

Multi-Task Visual Grounding. Multi-task visual grounding (MTVG) (Luo et al., 2020) aims to jointly identify a language-referred object at both box- and pixel-level granularity. Early MTVG methods typically leverage separate visual and language encoders, such as Swin Transformer (Liu et al., 2021) and BERT (Devlin et al., 2019), to construct multimodal representations. These methods design multiple task-specific heads to perform REC and RES in parallel (Li & Sigal, 2021; Su et al., 2023; Chen et al., 2024b), or formulate both objectives in a unified point-sequence regression framework (Zhu et al., 2022; Liu et al., 2023b; Cheng et al., 2024). With the emergence of vision-language pre-trained models, subsequent works (Dai et al., 2025b; Yu et al., 2025; Dai et al., 2026) increasingly explore a unified multimodal encoder, such as CLIP (Radford et al., 2021) and BEiT3 (Wang et al., 2023b). OneRef (Xiao et al., 2024b) introduces a mask referring modeling paradigm to establish a unified visual-language feature space. C^3VG (Dai et al., 2025c) adopts a coarse-to-fine framework that uses bidirectional consistency constraints to improve the agreement between REC and RES. InstanceVG (Dai et al., 2025a) further employs instance queries with reference points to jointly predict instance-level boxes and masks. However, these methods generally require full tuning, resulting in considerable optimization costs. More recently, the rapid progress

of MLLMs has motivated researchers to exploit their strong reasoning and generalization capabilities (Liu et al., 2023a; Chen et al., 2024c) for MTVG. These methods typically perform REC through coordinate regression, while addressing RES via mask generation (Kirillov et al., 2023; Liu et al., 2025b; Yang et al., 2026), contour prediction (Wang et al., 2023a; Pramanick et al., 2024; Wu et al., 2024a), or semantic correspondence (Lan et al., 2025; 2026). Although these methods reduce the training cost through parameter-efficient adaptation (Hu et al., 2022; Houlsby et al., 2019) or instruction tuning (Ouyang et al., 2022; Che et al., 2026), inference remains computationally intensive due to billions of parameters. More importantly, all the aforementioned methods adopt a shared-representation paradigm, overlooking the distinct feature preferences of REC and RES. In contrast, DeCo explicitly separates representation learning for the two objectives and subsequently exploits their complementary priors. By combining low-rank adaptation (LoRA) with lightweight multi-task heads, DeCo achieves state-of-the-art performance with modest computational overhead.

Parameter-Efficient Fine-Tuning. Parameter-efficient fine-tuning (PEFT) aims to transfer frozen large foundation models to various downstream tasks while substantially reducing training costs (Han et al., 2024; Xin et al., 2026). Representative PEFT approaches, including adapter tuning (Houlsby et al., 2019; Chen et al., 2022), prompt tuning (Jia et al., 2022; Li et al., 2023), and low-rank reparameterization (Hu et al., 2022; Yin et al., 2023), have achieved remarkable success in visual perception (Yin et al., 2025; Liu et al., 2026b; He et al., 2026), language understanding (Hwang et al., 2026; Rocha Filho et al., 2026; Lin et al., 2026), and multimodal reasoning (Kang et al., 2026; Lee et al., 2026; Qian et al., 2026). Among them, LoRA (Hu et al., 2022) is one of the most widely adopted and effective methods, which introduces low-rank trainable matrices to parameterize task-specific updates. Recent studies have investigated PEFT for REC (Xiao et al., 2024a; Shi et al., 2025; Li et al., 2026a) and RES (Wang et al., 2024; Huang et al., 2025; Moon et al., 2026) individually, demonstrating competitive performance against fully trained methods. For REC, BARE (Li et al., 2026a) enhances referential understanding by mitigating multimodal bias and strengthening semantic reasoning, while WIMFRIS (Moon et al., 2026) improves RES through hierarchical multi-scale feature fusion and lightweight vision-language alignment. However, the effectiveness of PEFT for MTVG remains largely unexplored. DeCo bridges this gap by combining LoRA with our decouple-to-couple learning paradigm, where LoRA serves as a fundamental technique for enabling efficient MTVG adaptation rather than the methodological focus of this work.

3 METHODOLOGY

3.1 OVERVIEW

Given an image $I \in \mathbb{R}^{h\times w\times 3}$ and a linguistic expression T, MTVG aims to predict the corresponding bounding box $\mathbf{B}$ and segmentation mask $\mathbf{M}$. We employ BEiT3 (Wang et al., 2023b) as the frozen multimodal encoder with LoRA (Hu et al., 2022) adaptation to efficiently extract visual patch embeddings $\mathbf{V} = [\bar{\mathbf{v}}, \mathbf{v}_1, \cdots, \mathbf{v}_n] \in \mathbb{R}^{(n+1)\times d}$ and textual token embeddings $\mathbf{T} = [\bar{\mathbf{t}}, \mathbf{t}_1, \cdots, \mathbf{t}_m] \in \mathbb{R}^{(m+1)\times d}$, where $n = hw/p_{size}^2$ and m denote the numbers of visual patches and textual tokens, respectively, $\bar{\mathbf{v}}$ and $\bar{\mathbf{t}}$ represent the class tokens of the two modalities, and d is the embedding dimension. Subsequently, we leverage SimFPN (Li et al., 2022) to project $\mathbf{V}$ into low-dimensional multi-scale features $\mathcal{V} = \{\mathbf{V}_l\}_{l=1}^{L}$, where $\mathbf{V}_l \in \mathbb{R}^{h_l\times w_l\times c}$, $L = 4$, and $c < d$. Detailed preliminaries of LoRA and SimFPN are provided in Appendix A. Based on this multimodal encoding architecture, we initially construct three effective baseline models (*i.e.*, LoM, SoM, and DoM) to improve MTVG. As discussed in Section 1 and Figure 1 (a)-(c), LoM focuses on REC but is weaker on RES, whereas SoM exhibits the opposite behavior. By integrating both over the shared $\mathcal{V}$, DoM inherits the strengths of LoM and SoM, yet with only marginal additional gains. This observation motivates us to rethink the role of shared representations in MTVG.

To this end, we propose DeCo, which follows a decouple-to-couple learning paradigm. As illustrated in Figure 1 (d), DeCo first introduces a task-aware semantic decoupling module to route the shared $\mathcal{V}$ into localization- and segmentation-oriented multi-scale features, $\mathcal{V}^{loc}$ and $\mathcal{V}^{seg}$, under the guidance of the salient word-level feature $\mathbf{t}_s$. It guarantees that the two objectives learn distinct representations from their respective feature interests while preserving the efficiency of shared multi-scale encoding. Subsequently, DeCo further leverages a hybrid prior coupling module to exploit the complementary information across two tasks. Specifically, a Transformer layer is employed to inject text-based semantic prior into an object query, with the entire sentence-level feature $\mathbf{T}$ serving as the key and

value. Then, a mask-derived spatial prior is obtained from $\mathbf{M}$ through a mask-to-box strategy and adaptively incorporated into the query and its position embedding via a learnable gate. This design enables REC to benefit from the spatial knowledge acquired through dense segmentation, while the localization objective in turn provides geometric constraints for rectifying the result of RES.

3.2 Task-aware Semantic Decoupling

The widely adopted shared-representation paradigm overlooks the distinctiveness of REC and RES. To capture discriminative features for each objective, we propose Task-aware Semantic Decoupling (TSD), which decouples $\mathcal{V}$ via salient word-level modulation and lightweight task-specific experts.

Salient Word-level Modulation. A referred object is typically characterized by a subset of semantic attributes in the referring expression. Thus, we initially identify a salient word-level feature from the textual embeddings $\mathbf{T} = [\bar{\mathbf{t}}, \mathbf{t}_1, \ldots, \mathbf{t}_m]$. Specifically, we compute the cosine similarity between each token and the class token $\bar{\mathbf{t}}$, and select the token with the highest similarity as $\mathbf{t}_s$:

$$j^* = \arg\max_j \frac{\mathbf{t}_j^\top \bar{\mathbf{t}}}{\|\mathbf{t}_j\|_2 \|\bar{\mathbf{t}}\|_2}, \qquad j \in \{1, \cdots, m\}, \qquad \mathbf{t}_s = \mathbf{t}_{j^*} \in \mathbb{R}^{1\times d}. \tag{1}$$

This selection provides a compact semantic cue that adaptively captures the most discriminative target-related concept while avoiding the redundancy of the entire sentence representation. For each multi-scale feature $\mathbf{V}_l \in \mathbb{R}^{h_l \times w_l \times c}$, we flatten its spatial dimensions into visual tokens $\mathbf{X}_l \in \mathbb{R}^{h_l w_l \times c}$ and align $\mathbf{t}_s$ into the same feature space through a learnable projection $\mathbf{W}_t \in \mathbb{R}^{d\times c}$. We then adopt a standard cross-attention operator with $\mathbf{t}_s$ as the query and $\mathbf{X}_l$ as the key and value:

$$\mathbf{Q}_l = (\mathbf{t}_s \mathbf{W}_t)\mathbf{W}_q, \qquad \mathbf{K}_l = \mathbf{X}_l \mathbf{W}_k, \qquad \mathbf{U}_l = \mathbf{X}_l \mathbf{W}_v, \tag{2}$$

$$\mathbf{D}_l = \mathrm{softmax}\left(\frac{\mathbf{Q}_l \mathbf{K}_l^\top}{\sqrt{d_k}}\right), \qquad \mathbf{R}_l = \mathbf{D}_l \mathbf{U}_l, \tag{3}$$

where $\mathbf{W}_{q/k/v} \in \mathbb{R}^{c\times c}$ denotes a learnable query/key/value projection. Here, $\mathbf{D}_l \in \mathbb{R}^{1\times h_l w_l}$ provides a spatial relevance distribution over visual tokens with respect to the target semantic, while $\mathbf{R}_l \in \mathbb{R}^{1\times c}$ aggregates it into a word-conditioned global response.

Task-specific Experts. We exploit the two complementary signals from cross-attention to construct localization- and segmentation-oriented representations. Both experts adopt the same lightweight architecture, consisting of two 1×1 convolutional layers interleaved with GELU activation and LayerNorm. For localization, the word-conditioned global response $\mathbf{R}_l$ is broadcastly connected with the original visual tokens, enhancing semantic cues that facilitate object-level correspondence:

$$\mathbf{V}_l^{loc} = \mathcal{E}_l^{loc}\big(\mathrm{reshape}(\mathbf{X}_l + \mathbf{R}_l)\big), \tag{4}$$

where $\mathcal{E}_l^{loc}$ denotes the REC-specific expert network and $\mathrm{reshape}(\cdot)$ restores the spatial dimensions of the visual tokens to $h_l \times w_l$. For segmentation, we instead use the spatial relevance distribution $\mathbf{D}_l$ to reweight the shared visual tokens, highlighting target-relevant regions while preserving their fine-grained spatial structure, which can be mathematically formulated as follows:

$$\mathbf{V}_l^{seg} = \mathcal{E}_l^{seg}\big(\mathrm{reshape}(\mathbf{X}_l \odot \mathbf{D}_l^\top)\big), \tag{5}$$

where $\odot$ represents element-wise multiplication and $\mathcal{E}_l^{seg}$ is the RES-specific expert network. This asymmetric design allows localization to incorporate explicit word-aware semantic responses while segmentation focuses on spatially relevant visual regions. The resulting task-specific multi-scale representations $\mathcal{V}^{loc} \in \{\mathbf{V}_l^{loc}\}_{l=1}^L$ and $\mathcal{V}^{seg} \in \{\mathbf{V}_l^{seg}\}_{l=1}^L$ are subsequently delivered to their respective branches, thereby alleviating the mutual competition arising from shared representations and allowing each task to preserve its preferred visual cues.

3.3 Hybrid Prior Coupling

While TSD ensures task-oriented features, localization and segmentation remain inherently complementary. In particular, dense pixel supervision enables RES to produce a reliable object proposal earlier in training than sparse box supervision used in REC. To exploit this property, we introduce Hybrid Prior Coupling (HPC), which integrates sentence-level semantic prior and mask-derived spatial prior into the localization query, thereby establishing a collaborative coupling mechanism.

Sentence-level Semantic Prior. Given the sentence-level feature $\mathbf{T}$, we initially project it into the query embedding space via a linear layer $\mathbf{W}_{\mathrm{T}} \in \mathbb{R}^{d\times c}$. A Transformer layer is then employed to incorporate the textual semantic information into the randomly initialized object query $\mathbf{q} \in \mathbb{R}^{1\times c}$. Specifically, the object query first interacts with itself through self-attention and subsequently aggregates global linguistic contexts via cross-attention, followed by a feed-forward network (FFN):

$$\tilde{\mathbf{q}} = \mathrm{FFN}\Big(\mathrm{CrossAttn}\big(\mathrm{SelfAttn}(\mathbf{q}+\mathbf{p})+\mathbf{p}, \mathbf{T}\mathbf{W}_{\mathrm{T}}, \mathbf{T}\mathbf{W}_{\mathrm{T}}\big)\Big), \tag{6}$$

where $\mathbf{p} \in \mathbb{R}^{1\times c}$ denotes the position embedding of $\mathbf{q}$. Unlike TSD, which employs the salient word-level feature to provide a compact discriminative cue for feature decoupling, HPC utilizes the entire sentence-level feature to capture holistic linguistic semantics for object localization, explicitly grounding the initial query in the semantic space of the referring expression.

Mask-derived Spatial Prior. Pixel-level segmentation inherently provides informative object geometry that can complement the semantic query. To further transfer this prior knowledge, we convert the predicted binary mask $\mathbf{M}$ into a normalized bounding box $\mathbf{B}_{m2b}$ through a mask-to-box strategy. Given coordinate grids x_{ij} and y_{ij}, we apply each grid with an extremal constant according to the mask value and aggregate the foreground extremes via global min and max reductions as follows:

$$\begin{aligned} x_{\min} &= \min_{i,j}\big(\mathbf{M}_{ij}x_{ij} + (1-\mathbf{M}_{ij})w\big), \qquad y_{\min} = \min_{i,j}\big(\mathbf{M}_{ij}y_{ij} + (1-\mathbf{M}_{ij})h\big), \\ x_{\max} &= \max_{i,j}\big(\mathbf{M}_{ij}x_{ij} - (1-\mathbf{M}_{ij})\big), \qquad y_{\max} = \max_{i,j}\big(\mathbf{M}_{ij}y_{ij} - (1-\mathbf{M}_{ij})\big), \end{aligned} \tag{7}$$

$$\mathbf{B}_{m2b} = \big[(x_{\min}+x_{\max})/2w,\ (y_{\min}+y_{\max})/2h,\ (x_{\max}-x_{\min})/w,\ (y_{\max}-y_{\min})/h\big], \tag{8}$$

which is further mapped into the query and position embedding spaces, respectively:

$$\mathbf{q}_{m2b} = \mathbf{B}_{m2b}\mathbf{W}^{q}_{m2b}, \qquad \mathbf{p}_{m2b} = \mathbf{B}_{m2b}\mathbf{W}^{p}_{m2b}, \tag{9}$$

where $\mathbf{W}^{q}_{m2b}, \mathbf{W}^{p}_{m2b} \in \mathbb{R}^{4\times c}$ are two projection matrices. To adaptively balance the contributions of semantic and spatial information, we design a lightweight gating function conditioned on both the semantic-enriched query and the mask-derived bounding box:

$$\gamma = \sigma\Big(\mathrm{MLP}\big(\mathrm{Concat}(\tilde{\mathbf{q}}; \mathbf{B}_{m2b})\big)\Big), \tag{10}$$

where $\mathrm{Concat}(\cdot;\cdot)$, $\mathrm{MLP}(\cdot)$, and $\sigma(\cdot)$ represent concatenation, two linear layers separated by a GELU activation, and the sigmoid function, respectively. The gate value $\gamma \in (0,1)$ is then utilized to get the final hybrid prior query and its position embedding as follows:

$$\hat{\mathbf{q}} = \gamma\tilde{\mathbf{q}} + (1-\gamma)\,\mathbf{q}_{m2b}, \qquad \hat{\mathbf{p}} = \gamma\mathbf{p} + (1-\gamma)\mathbf{p}_{m2b}. \tag{11}$$

In this way, HPC evolves the randomly initialized object query into a strong context-geometric anchor that accelerates the convergence of the localization branch, while the resulting detection loss in turn provides bounding constraints that further refine the segmentation mask.

3.4 Training Objectives

Given the hybrid prior query $\hat{\mathbf{q}}$, its position embedding $\hat{\mathbf{p}}$, and REC-specific features $\mathcal{V}^{loc}$, we first employ Deformable DETR (Zhu et al., 2021) to produce the output query $\mathbf{q}_{\mathrm{o}}$. Subsequently, a box FFN regresses $\mathbf{q}_{\mathrm{o}}$ into the predicted box $\mathbf{B}$, while a mask FFN projects $\mathbf{q}_{\mathrm{o}}$ into a mask embedding, which is multiplied with the high-resolution segmentation feature $\mathbf{V}^{seg}_{1}$ to obtain the corresponding instance mask $\mathbf{M}_{ins}$. In parallel, RES-specific features $\mathcal{V}^{seg}$ are fed into UNet (Ronneberger et al., 2015) to get the predicted mask $\mathbf{M}$. The overall prediction process can be formulated as:

$$\mathbf{q}_{\mathrm{o}} = \mathrm{DETR}(\hat{\mathbf{q}},\ \hat{\mathbf{p}},\ \mathcal{V}^{loc}),\ \mathbf{B} = \mathrm{FFN}_b(\mathbf{q}_{\mathrm{o}}),\ \mathbf{M}_{ins} = \mathrm{FFN}_m(\mathbf{q}_{\mathrm{o}}) \otimes \mathbf{V}^{seg}_{1},\ \mathbf{M} = \mathrm{UNet}(\mathcal{V}^{seg}) \tag{12}$$

where $\otimes$ denotes Einstein multiplication. For box regression, we adopt L1 loss and generalized IoU loss, while a combination of binary cross-entropy and Dice losses is used for mask prediction:

$$\begin{aligned} \mathcal{L}_{det} &= \mathcal{L}_{l1}(\mathbf{B},\mathbf{B}^*) + \mathcal{L}_{giou}(\mathbf{B},\mathbf{B}^*), \qquad \mathcal{L}_{ins} = \mathcal{L}_{bce}(\mathbf{M}_{ins},\mathbf{M}^*) + \mathcal{L}_{dice}(\mathbf{M}_{ins},\mathbf{M}^*), \\ \mathcal{L}_{seg} &= \mathcal{L}_{bce}(\mathbf{M},\mathbf{M}^*) + \mathcal{L}_{dice}(\mathbf{M},\mathbf{M}^*), \quad \mathcal{L}_{m2b} = \mathcal{L}_{l1}(\mathbf{B}_{m2b},\mathbf{B}^*) + \mathcal{L}_{giou}(\mathbf{B}_{m2b},\mathbf{B}^*), \end{aligned} \tag{13}$$

where $\mathbf{B}^*$ and $\mathbf{M}^*$ are the ground-truth bounding box and segmentation mask. Finally, the total training objective is defined as a simple weighted sum of the four loss functions with same weights:

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \mathcal{L}_{ins} + \mathcal{L}_{seg} + \mathcal{L}_{m2b}. \tag{14}$$

During inference, only $\mathbf{B}$ and $\mathbf{M}$ are retained as the final REC and RES predictions. The auxiliary outputs $\mathbf{M}_{ins}$ and $\mathbf{B}_{m2b}$ are involved in the forward process but are discarded from the final outputs, requiring no additional post-processing.

Table 1: Main results of **REC** models using **P@.5** on RefCOCO, RefCOCO+, and G-Ref datasets. The best results are highlighted in **bold**, and the second-best are underlined. * indicates that the backbone is frozen with only the additional introduced adaptation parameters being optimized.

Models	Vision / Language Backbone	RefCOCO			RefCOCO+			G-Ref		Avg.
		val	testA	testB	val+	testA+	testB+	val-u	test-u	
MLLM-based Methods										
GSVA-13B (Xia et al., 2024)	SAM-H / Llama2*	89.2	92.1	87.2	79.7	84.5	73.4	85.5	86.2	84.7
SegLLM-7B (Wang et al., 2025)	CLIP-L* / Vicuna-v1.5	90.0	92.1	86.2	82.2	85.5	76.1	83.9	85.9	85.2
Text4Seg-13B (Lan et al., 2025)	SAM-H / LLaVa-v1.5*	91.2	94.3	88.0	85.7	90.8	80.1	85.6	85.5	87.7
UniPixel-7B (Liu et al., 2025b)	SAM2-B+ / Qwen2.5-VL*	92.0	94.4	88.1	87.2	91.9	82.1	88.6	88.7	89.1
UFO-8B (Tang et al., 2025)	InternViT* / InternLM2.5	93.1	94.8	89.2	87.7	92.1	82.3	88.2	89.2	89.6
SIFThinker-7B (Chen et al., 2026b)	Qwen2.5-VL / Qwen2.5-VL	93.8	95.1	90.4	86.0	90.7	80.5	90.4	90.6	89.7
VGent-16B (Kang et al., 2026)	Qwen2.5-VL / Qwen2.5-VL*	92.4	94.7	89.8	88.1	92.2	83.3	90.4	90.1	90.1
Text4Seg++-14B (Lan et al., 2026)	SAM-H / Qwen2.5-VL*	**94.3**	**96.2**	**91.4**	**90.6**	**94.2**	**86.1**	**91.6**	**91.8**	**92.0**
Task-specific Methods (Standard Setting)										
HiVG (Xiao et al., 2024a)	CLIP-B* / CLIP-B*	87.3	89.9	83.3	78.1	83.8	68.1	78.3	78.8	80.9
SimVG (Dai et al., 2024)	BEiT3-B / BEiT3-B	87.6	90.2	84.0	78.7	83.4	71.8	80.4	80.5	82.1
OneRef (Xiao et al., 2024b)	BEiT3-B / BEiT3-B	88.8	91.0	85.3	80.4	86.5	74.3	83.7	83.5	84.2
PropVG (Dai et al., 2025b)	BEiT3-B / BEiT3-B	89.0	91.6	85.7	83.7	88.0	76.6	83.5	84.4	85.3
Latent-VG (Yu et al., 2025)	BEiT3-B / BEiT3-B	89.8	92.4	85.9	83.2	87.7	76.5	82.6	83.1	85.1
SwimVG (Shi et al., 2025)	DINOv2-B* / CLIP-B*	88.3	90.4	84.9	77.9	83.2	70.0	79.1	80.1	81.7
OneSVG (Shi et al., 2026)	BEiT3-B* / BEiT3-B*	88.8	91.0	85.6	81.7	86.6	75.6	82.4	82.1	84.2
BARE (Li et al., 2026a)	BEiT3-B* / BEiT3-B*	88.7	91.6	84.8	81.3	86.8	74.8	82.7	83.2	84.2
DeCo (Ours)	BEiT3-B* / BEiT3-B*	**90.9**	**93.2**	**87.3**	**86.4**	**90.7**	**80.2**	**86.4**	**87.0**	**87.8**
HiVG (Xiao et al., 2024a)	CLIP-L* / CLIP-L*	88.1	91.1	83.7	80.1	86.8	70.5	80.8	80.3	82.7
SimVG (Dai et al., 2024)	BEiT3-L / BEiT3-L	90.6	92.5	87.7	85.4	89.6	79.7	86.0	86.8	87.3
OneRef (Xiao et al., 2024b)	BEiT3-L / BEiT3-L	92.9	94.0	90.2	88.0	91.6	83.7	88.1	89.3	89.7
OneSVG (Shi et al., 2026)	BEiT3-L* / BEiT3-L*	91.1	93.3	88.3	87.4	90.9	81.2	87.5	87.6	88.4
BARE (Li et al., 2026a)	BEiT3-L* / BEiT3-L*	92.4	93.8	89.2	87.9	91.8	83.2	88.8	88.9	89.5
DeCo (Ours)	BEiT3-L* / BEiT3-L*	**93.8**	**95.1**	**91.1**	**90.6**	**94.1**	**85.1**	**90.5**	**91.0**	**91.4**
Task-specific Methods (Mixed Setting)										
HiVG (Xiao et al., 2024a)	CLIP-B* / CLIP-B*	90.6	92.6	87.2	83.1	89.2	76.7	84.5	85.6	86.2
SimVG (Dai et al., 2024)	BEiT3-B / BEiT3-B	91.5	93.7	87.9	84.8	88.9	79.1	86.3	87.3	87.4
OneRef (Xiao et al., 2024b)	BEiT3-B / BEiT3-B	91.9	94.3	88.6	86.4	90.4	79.5	86.8	87.3	88.2
C^3VG (Dai et al., 2025c)	BEiT3-B / BEiT3-B	92.5	94.6	88.7	87.4	90.7	81.4	87.7	88.3	88.9
PropVG (Dai et al., 2025b)	BEiT3-B / BEiT3-B	92.7	95.1	89.6	87.3	90.9	81.3	88.2	88.3	89.2
Latent-VG (Yu et al., 2025)	BEiT3-B / BEiT3-B	91.8	94.6	88.6	86.4	90.6	80.6	87.0	87.1	88.3
InstanceVG (Dai et al., 2025a)	BEiT3-B / BEiT3-B	92.4	95.1	88.8	87.1	90.8	80.7	87.2	88.1	88.8
BARE (Li et al., 2026a)	BEiT3-B* / BEiT3-B*	90.3	91.5	87.9	83.1	87.1	77.3	85.4	86.0	86.1
DeCo (Ours)	BEiT3-B* / BEiT3-B*	**93.4**	**95.6**	**90.5**	**89.4**	**92.7**	**83.7**	**89.2**	**89.8**	**90.5**
HiVG (Xiao et al., 2024a)	CLIP-L* / CLIP-L*	90.8	92.9	88.0	86.8	89.9	78.0	86.6	86.6	87.5
SimVG (Dai et al., 2024)	BEiT3-L / BEiT3-L	92.9	94.7	90.3	87.3	91.6	82.4	88.0	89.2	89.6
OneRef (Xiao et al., 2024b)	BEiT3-L / BEiT3-L	93.2	95.4	90.1	88.4	92.1	82.7	87.8	88.8	89.8
InstanceVG (Dai et al., 2025a)	BEiT3-L / BEiT3-L	94.4	96.0	92.4	90.1	92.9	85.9	89.6	90.6	91.5
BARE (Li et al., 2026a)	BEiT3-L* / BEiT3-L*	92.8	94.3	90.8	88.4	91.0	84.2	90.6	90.7	90.3
DeCo (Ours)	BEiT3-L* / BEiT3-L*	**95.3**	**96.6**	**92.7**	**92.3**	**95.1**	**87.9**	**92.1**	**92.7**	**93.1**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets & Metrics. We evaluate DeCo on ten widely adopted benchmarks, including RefCOCO/+, G-Ref, ReferIt, Flickr, DIOR-RSVG, RRSIS-D, and others. Following previous works (Xiao et al., 2024b; Dai et al., 2025a; Chen et al., 2026a), we report P@.5 for REC, while oIoU and mIoU are for RES. More descriptions about datasets and metrics are provided in Appendix B.1 and B.2.

Implementation Details. For most experiments, we train DeCo using the Adam optimizer for 35 epochs, with the learning rate reduced by a factor of 10 at the 28-th epoch. The input image is resized to 448×448, and the maximum expression length is limited to 20 tokens. The initial learning rate is set to 1×10^{-4}, with the batch size of 32. More implementation details are provided in Appendix B.3.

4.2 MAIN RESULTS

Referring Expression Comprehension. As evaluated in Table 1 and Table 3, DeCo consistently achieves state-of-the-art performance compared to all MLLM-based and task-specific methods across both standard and mixed settings. With frozen BEiT3-B/L (Wang et al., 2023b) backbones, DeCo sets top results on all eight splits under the mixed setting (90.5%/93.1% average P@.5) and delivers gains of up to 3.9%, 3.3%, and 0.8% on G-Ref (val-g), ReferIt, and Flickr, respectively. These results validate that decoupling task-preferred representations and coupling complementary priors effectively eliminates task interference while generalizing across diverse domains.

Referring Expression Segmentation. As shown in Table 2 and Table 4, DeCo sets new RES records across all benchmarks under the standard setting. With BEiT3-B/L, DeCo outperforms OneRef by

Table 2: Main results of **RES** models using **oIoU** on RefCOCO, RefCOCO+, and G-Ref datasets.

Models	Vision / Language Backbone	RefCOCO			RefCOCO+			G-Ref		Avg.
		val	testA	testB	val+	testA+	testB+	val-u	test-u	
MLLM-based Methods										
LISA-13B (Lai et al., 2024)	SAM-H / Llama2*	76.3	78.7	72.4	66.2	71.0	59.3	70.1	71.1	70.6
SegLLM-7B (Wang et al., 2025)	CLIP-L* / Vicuna-v1.5	80.2	81.5	75.4	70.3	73.0	62.5	72.6	73.6	73.6
Text4Seg-13B (Lan et al., 2025)	SAM-H / LLaVa-v1.5*	80.2	82.7	77.3	73.7	78.6	67.6	74.0	75.1	76.2
UniPixel-7B (Liu et al., 2025b)	SAM2-B+ / Qwen2.5-VL*	80.8	83.0	77.4	75.3	80.1	70.0	76.4	77.1	77.5
UFO-8B (Tang et al., 2025)	InternViT* / InternLM2.5	81.0	82.6	78.6	77.1	80.4	72.6	76.7	77.3	78.3
FCLM-7B (Yang et al., 2026)	SAM-H* / Qwen2.5-VL	82.6	83.9	79.9	77.4	80.5	71.3	79.4	**80.5**	79.4
STAMP-7B (Liu et al., 2026a)	SAM-H* / Qwen2-VL*	**83.1**	84.5	**80.8**	**79.4**	**82.8**	**74.6**	**79.9**	80.4	**80.7**
Text4Seg++-14B (Lan et al., 2026)	SAM-H / Qwen2.5-VL*	82.5	**84.9**	79.5	77.9	82.5	72.6	79.4	79.7	79.9
Task-specific Methods (Standard Setting)										
BarLeRIa Wang et al. (2024)	CLIP-B* / CLIP-B*	75.0	77.1	71.2	68.6	73.2	61.2	65.9	66.4	69.8
OneRef (Xiao et al., 2024b)	BEiT3-B / BEiT3-B	77.6	81.0	73.5	70.8	74.5	64.1	70.7	70.6	72.8
MaskRIS (Lee et al., 2025)	Swin-B / BERT-B	76.5	79.0	74.0	67.5	74.5	59.4	65.6	66.5	70.4
CMIRNet (Xu et al., 2025)	Swin-B / BERT-B	74.0	77.1	71.3	64.2	70.4	56.4	63.5	63.9	67.6
DETRIS (Huang et al., 2025)	DINOv2-B* / CLIP-B*	74.3	77.6	71.4	65.6	72.0	56.5	65.2	66.4	68.6
PropVG (Dai et al., 2025b)	BEiT3-B / BEiT3-B	76.8	79.6	73.7	70.2	74.3	63.4	69.3	70.5	72.2
SaFiRe (Mao et al., 2025)	Swin-B / BERT-B	75.3	78.5	71.4	66.4	72.0	58.5	67.1	66.9	69.5
AMLRIS (Chen et al., 2026a)	Swin-B / BERT-B	75.5	78.3	72.1	67.4	73.2	59.5	65.7	66.8	69.8
DeCo (Ours)	BEiT3-B* / BEiT3-B*	**79.0**	**81.3**	**75.7**	**73.1**	**77.5**	**66.3**	**73.4**	**73.0**	**74.9**
BarLeRIa Wang et al. (2024)	CLIP-L* / CLIP-L*	76.8	79.0	74.0	71.5	76.2	65.4	68.7	69.7	72.7
OneRef (Xiao et al., 2024b)	BEiT3-L / BEiT3-L	80.5	82.8	78.3	74.3	78.4	69.9	74.9	77.4	77.0
DETRIS (Huang et al., 2025)	DINOv2-L* / CLIP-B*	76.1	78.3	73.2	67.9	73.3	60.2	66.8	68.0	70.5
DeCo (Ours)	BEiT3-L* / BEiT3-L*	**82.7**	**84.4**	**80.2**	**79.2**	**82.9**	**73.0**	**77.9**	**78.7**	**79.9**
Task-specific Methods (Mixed Setting)										
BarLeRIa Wang et al. (2024)	CLIP-B* / CLIP-B*	77.6	79.4	75.3	71.7	75.7	66.0	70.9	71.4	73.5
OneRef (Xiao et al., 2024b)	BEiT3-B / BEiT3-B	81.1	83.1	77.8	72.2	77.3	67.1	75.1	77.2	76.4
MaskRIS (Lee et al., 2025)	Swin-B / BERT-B	78.7	80.6	75.1	70.3	75.2	62.8	69.1	71.1	72.9
LAVT-RS (Yang et al., 2025)	Swin-B / BERT-B	79.2	80.7	75.4	71.7	75.6	64.3	72.1	74.6	74.2
CMIRNet (Xu et al., 2025)	Swin-B / BERT-B	77.7	79.8	74.4	68.3	74.1	61.1	69.2	71.9	72.1
C^3VG (Dai et al., 2025c)	BEiT3-B / BEiT3-B	80.9	83.2	77.9	74.7	78.0	69.0	74.4	76.4	76.8
PropVG (Dai et al., 2025b)	BEiT3-B / BEiT3-B	81.8	83.7	79.3	74.8	78.7	69.2	75.5	77.4	77.6
SaFiRe (Mao et al., 2025)	Swin-B / BERT-B	81.4	83.6	78.2	74.8	78.9	68.3	75.8	76.8	77.2
InstanceVG (Dai et al., 2025a)	BEiT3-B / BEiT3-B	81.2	83.4	78.5	74.8	78.3	68.4	74.9	76.2	77.0
AMLRIS (Chen et al., 2026a)	Swin-B / BERT-B	77.2	79.5	73.6	68.6	73.8	61.6	68.8	70.0	71.9
DeCo (Ours)	BEiT3-B* / BEiT3-B*	**82.4**	**84.5**	**80.0**	**77.1**	**81.0**	**71.5**	**77.7**	**78.5**	**79.1**
BarLeRIa Wang et al. (2024)	CLIP-L* / CLIP-L*	79.0	80.8	77.0	74.2	77.8	68.3	72.7	73.3	75.4
OneRef (Xiao et al., 2024b)	BEiT3-L / BEiT3-L	82.0	83.6	79.6	74.6	79.5	70.7	76.5	79.2	78.2
CMIRNet (Xu et al., 2025)	Swin-L / BERT-B	78.2	80.4	75.2	69.8	75.3	62.1	70.4	72.1	72.9
DETRIS (Huang et al., 2025)	DINOv2-L* / CLIP-B*	80.7	82.2	77.9	73.1	77.5	66.0	73.2	74.7	75.7
InstanceVG (Dai et al., 2025a)	BEiT3-L / BEiT3-L	**85.8**	86.7	**84.1**	80.4	82.9	75.0	79.9	81.4	82.0
DeCo (Ours)	BEiT3-L* / BEiT3-L*	85.4	**86.7**	83.5	**81.9**	**85.0**	**76.9**	**81.7**	**82.6**	**83.0**

2.1%/2.9% average oIoU while keeping backbones frozen. Compared with MLLM-based methods, DeCo surpasses Text4Seg++-14B by 3.1% average oIoU and the strongest competitor STAMP-7B on all eight splits. Furthermore, DeCo exceeds DETRIS and WIMFRIS by significant margins of 13.4% oIoU and 9.8% mIoU on G-Ref (val-g), highlighting robust generalization to complex expressions. More RES results with mIoU and other remote sensing experiments are listed in Appendix C.

4.3 Ablation Studies

Effects of Core Components. As reported in Table 5, LoM and SoM validate dedicated task-oriented representations, while DoM shows limited gains from naive joint optimization. Adding TSD or HPC individually improves both objectives, confirming that decoupling and prior coupling address distinct interference sources. Combining both, DeCo achieves the best performance, exceeding DoM by 2.2%. Qualitative analysis in Figure 2 intuitively confirms the role of each component.
Effects of PEFT and LoRA r Settings. We next examine the efficiency of PEFT in Table 6. Full tuning yields only 61.4% despite 221.3M trainable parameters, whereas freezing drops to 54.1%, highlighting the necessity of lightweight adaptation. LoRA at $r = 4$ already reaches 75.0% with only 0.6M parameters and rises steadily to 76.6% at $r = 32$, which is set to the trade-off setting.
Effects of LoRA Insertion Positions. As listed in Table 7, query-only adaptation yields 72.9% average performance, while value and query-value adaptation rise to 75.0% and 75.8%, suggesting important roles of information retrieval and feature transformation in PEFT. Adapting the complete attention block (Q-K-V-O) further improves to 76.6%, providing a favorable balance across settings.
Effects of Segmentation Heads. As shown in Table 8, UNet achieves the best average performance with only a 0.4M/2.6G burden, while other classic or multimodal segmentors offer no gains at much higher costs, indicating the efficiency primarily comes from our decouple-to-couple learning rather than a complex decoder. More ablation studies and qualitative results are provided in Appendix D.

Table 3: Main results of **REC** models using **P@.5** on G-Ref, ReferIt, and Flickr datasets.

Models	Vision / Language Backbone	G-Ref val-g	ReferIt test	Flickr test
SimVG (Dai et al., 2024)	BEiT3-B / BEiT3-B	78.8	74.8	82.0
OneRef (Xiao et al., 2024b)	BEiT3-B / BEiT3-B	-	77.2	83.6
MAP (Yao et al., 2025)	DINOv2-B* / BERT-B*	76.7	75.0	81.3
SwimVG (Shi et al., 2025)	DINOv2-B* / CLIP-B*	79.1	-	83.1
OneSVG (Shi et al., 2026)	BEiT3-B* / BEiT3-B*	81.5	-	-
BARE (Li et al., 2026a)	BEiT3-B* / BEiT3-B*	-	75.3	82.3
DeCo (Ours)	BEiT3-B* / BEiT3-B*	**85.4**	**80.5**	**84.4**
HiVG (Xiao et al., 2024a)	CLIP-L* / CLIP-L*	-	76.2	82.2
SimVG (Dai et al., 2024)	BEiT3-L / BEiT3-L	80.4	78.8	83.2
OneRef (Xiao et al., 2024b)	BEiT3-L / BEiT3-L	-	81.1	84.8
OneSVG (Shi et al., 2026)	BEiT3-L* / BEiT3-L*	87.2	-	-
BARE (Li et al., 2026a)	BEiT3-L* / BEiT3-L*	-	80.0	83.3
DeCo (Ours)	BEiT3-L* / BEiT3-L*	**90.3**	**83.7**	**86.7**

Table 4: Main results of **RES** models using **oIoU** and **mIoU** on the G-Ref dataset.

Models	Vision / Language Backbone	G-Ref (val-g) oIoU	mIoU
BarLeRIa (Wang et al., 2024)	CLIP-B* / CLIP-B*	61.6	-
LQMFormer (Shah et al., 2024)	Swin-B / BERT-B	63.0	-
Prompt-RIS (Shang et al., 2024)	SAM-B* / CLIP-B*	64.8	69.2
ASDA (Yue et al., 2024)	CLIP-B / CLIP-B	63.6	-
CMANet (Pan et al., 2025)	Swin-B / BERT-B	64.1	-
DETRIS (Huang et al., 2025)	DINOv2-B* / CLIP-B*	63.2	65.9
MiCA (Li et al., 2026c)	DINOv2-B* / CLIP-B*	-	66.7
WIMFRIS (Moon et al., 2026)	DINOv2-B* / CLIP-B*	-	67.1
DeCo (Ours)	BEiT3-B* / BEiT3-B*	**72.3**	**73.7**
DETRIS (Huang et al., 2025)	DINOv2-L* / CLIP-B*	65.0	67.9
MiCA (Li et al., 2026c)	DINOv2-L* / CLIP-B*	-	68.6
WIMFRIS (Moon et al., 2026)	DINOv2-L* / CLIP-B*	-	69.3
DeCo (Ours)	BEiT3-L* / BEiT3-L*	**78.4**	**79.1**

Table 5: Ablation on core components of DeCo.

test-u / Models	Trainable Head Params.	FLOPs	REC P@.5	P@.5:.95	RES oIoU	mIoU	Avg.
Previous SOTA	-	-	83.2	-	66.8	69.7	-
LoM	7.9M	15.9G	85.5	67.2	71.2	72.0	74.0
SoM	**3.4M**	**13.3G**	84.8	66.7	71.6	72.5	73.9
DoM	8.3M	17.6G	85.6	67.4	71.8	72.6	74.4
DoM *w/* TSD	10.6M	24.2G	86.3	69.1	72.5	73.6	75.4
DoM *w/* HPC	9.3M	17.6G	86.5	70.5	72.2	73.3	75.6
DeCo	11.6M	24.2G	**87.0**	**71.7**	**73.0**	**74.5**	**76.6**

Table 6: Ablation on PEFT and LoRA r settings.

test-u / Methods	Trainable Backbone Params.	FLOPs	REC P@.5	P@.5:.95	RES oIoU	mIoU	Avg.
Full Tuning	221.3M	81.0G	72.6	53.8	57.8	61.2	61.4
Frozen	0.0M	0.0G	68.1	43.5	51.8	53.0	54.1
LoRA ($r = 4$)	**0.6M**	**0.3G**	86.3	69.0	72.0	72.8	75.0
LoRA ($r = 8$)	1.2M	0.5G	86.5	69.7	72.2	73.2	75.4
LoRA ($r = 16$)	2.4M	1.0G	86.6	70.3	72.7	73.9	75.9
LoRA ($r = 32$)	4.7M	1.9G	87.0	71.7	**73.0**	**74.5**	**76.6**
LoRA ($r = 64$)	9.4M	3.8G	**87.1**	**71.8**	73.0	74.4	76.6

Table 7: Ablation on LoRA inserted positions.

test-u / Positions	Trainable Backbone Params.	FLOPs	REC P@.5	P@.5:.95	RES oIoU	mIoU	Avg.
Q	1.2M	0.5G	84.3	65.9	70.4	71.0	72.9
V	**1.2M**	**0.5G**	86.1	69.0	71.9	72.8	75.0
Q-V	2.4M	1.0G	86.7	70.1	72.6	73.9	75.8
Q-K-V-O	4.7M	1.9G	87.0	71.7	73.0	74.5	76.6
FFN	5.9M	2.4G	86.5	70.6	72.7	74.2	76.0
Q-K-V-O-FFN	10.6M	4.3G	**87.1**	**71.8**	**73.2**	**74.8**	**76.7**

Table 8: Ablation on various segmentation heads.

test-u / Models	Trainable Head Params.	FLOPs	REC P@.5	P@.5:.95	RES oIoU	mIoU	Avg.
UNet	**0.4M**	**2.6G**	**87.0**	**71.7**	73.0	74.5	**76.6**
SegFormer	0.5M	5.0G	86.5	70.9	72.9	74.1	76.1
UPerNet	7.5M	41.6G	86.7	71.2	73.3	74.6	76.5
ReMamber	15.0M	62.1G	86.7	71.0	72.7	74.4	76.2
LAVT-RS	5.3M	29.7G	86.6	71.3	**73.3**	**74.7**	76.5
DETRIS	23.9M	37.2G	86.3	71.1	72.7	74.2	76.1



Figure 2: Qualitative results on core components of DeCo on G-Ref (test-u).

5 Conclusion

We presented DeCo, an efficient decouple-to-couple learning framework for unified referring expression comprehension and segmentation. By first decoupling shared visual representations into task-specific features and then coupling complementary semantic and spatial priors, DeCo enables each task to exploit its preferred visual cues while benefiting from cross-task knowledge. Extensive experiments demonstrate consistent improvements over existing methods across multiple referring expression benchmarks and settings, with only lightweight adaptation of the pretrained backbone. Our results suggest that explicitly managing the tension between task-specificity and task complementarity provides an effective and efficient paradigm for multi-task visual grounding.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, grammar error checking, and supporting qualitative and quantitative data analysis. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models, and other required disclosure tasks are not applicable to this work.

ETHICS STATEMENT

This submission adheres to the ICLR Code of Ethics and Code of Conduct. Our work does not involve human subjects or animal experiments; thus, no IRB approval is required. All datasets used are publicly available for research under their respective licenses, which we follow without redistributing restricted content.

REPRODUCIBILITY STATEMENT

To ensure full reproducibility of our experimental results, we provide detailed algorithmic and architectural specifications in both the main paper and the supplementary material. After the review process, we will release the complete source code, training and testing scripts, and all trained model checkpoints. We believe that this combination of paper and code supports end-to-end reproducibility and facilitates fair comparison with future work.

REFERENCES

Jiamin Cao, Lingling Li, Xiaoqiang Lu, Licheng Jiao, Fang Liu, Wenping Ma, Xu Liu, and Long Sun. Coarse-to-fine collaborative visual grounding framework for remote sensing imagery. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Chang Che, Ziqi Wang, Pengwan Yang, Cheems Wang, Hui Ma, and Zenglin Shi. Lora in lora: Towards parameter-efficient architecture expansion for continual visual instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 19978–19986, 2026.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26540–26550, 2024a.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

Tongfei Chen, Shuo Yang, Yuguang Yang, Linlin Yang, Runtang Guo, Changbai Li, He Long, Chunyu Xie, Dawei Leng, and Baochang Zhang. Amlris: Alignment-aware masked learning for referring image segmentation. In *International Conference on Learning Representations*, 2026a.

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

Zhangquan Chen, Ruihui Zhao, Chuwei Luo, Mingze Sun, Xinlei Yu, Yangyang Kang, and Ruqi Huang. Sifthinker: Spatially-aware image focus for visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20436–20444, 2026b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024c.

Zesen Cheng, Kehan Li, Peng Jin, Siheng Li, Xiangyang Ji, Li Yuan, Chang Liu, and Jie Chen. Parallel vertex diffusion for unified visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1326–1334, 2024.

Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26563–26573, 2024.

Yubin Cho, Hyunwoo Yu, and Suk-Ju Kang. Cross-aware early fusion with stage-divided vision and language transformer encoders for referring image segmentation. *IEEE Transactions on Multimedia*, 26:5823–5833, 2023.

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

Ming Dai, Wenxuan Cheng, Jiang-Jiang Liu, Lingfeng Yang, Zhenhua Feng, Wankou Yang, and Jingdong Wang. Improving generalized visual grounding with instance-aware joint learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48:448–465, 2025a.

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

Ming Dai, Jian Li, Jiedong Zhuang, Xian Zhang, and Wankou Yang. Multi-task visual grounding with coarse-to-fine consistency constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2618–2626, 2025c.

Ming Dai, Kai Chen, Wenxuan Cheng, Jiedong Zhuang, Zhenhua Feng, Pengfei Zhu, and Wankou Yang. Gc3vg: Generalized multi-task visual grounding with coarse-to-fine consistency constraints. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(4):4827–4841, 2026.

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

Jiajun Deng, Zhengyuan Yang, Daqing Liu, Tianlang Chen, Wengang Zhou, Yanyong Zhang, Houqiang Li, and Wanli Ouyang. Transvg++: End-to-end visual grounding with language conditioned vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (11):13636–13652, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

Henghui Ding, Song Tang, Shuting He, Chang Liu, Zuxuan Wu, and Yu-Gang Jiang. Multimodal referring segmentation: A survey. *arXiv preprint arXiv:2508.00265*, 2025.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Yu-Gang Jiang. Grex: Generalized referring expression segmentation, comprehension, and generation. *International Journal of Computer Vision*, 134(2):79, 2026.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Transactions on Machine Learning Research*, 2024.

Jing He, Licheng Jiao, Lingling Li, Xiaoqiang Lu, Xu Liu, Wenping Ma, Fang Liu, and Long Sun. Dsgcr: Decomposed spectral geometry-aware cross-modal semantic representation for 3d visual grounding. In *International Conference on Machine Learning*, 2026.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

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

Zibo Hu, Kun Gao, Xiaodian Zhang, Zhijia Yang, Mingfeng Cai, Zhenyu Zhu, and Wei Li. Efficient grounding dino: Efficient cross-modality fusion and efficient label assignment for visual grounding in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025.

Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. Look before you leap: Learning landmark features for one-stage visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16883–16892, 2021.

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.

Junseo Hwang, Wonguk Cho, and Taesup Kim. Pica: Parameter-efficient fine-tuning with column space projection. In *International Conference on Learning Representations*, pp. 114267–114287, 2026.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727, 2022.

Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. Pseudo-q: Generating pseudo language queries for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15492–15502, 2022.

Weitai Kang, Gaowen Liu, Mubarak Shah, and Yan Yan. Segvg: Transferring object bounding box to segmentation for visual grounding. In *European Conference on Computer Vision*, pp. 57–75, 2024.

Weitai Kang, Jason Kuen, Mengwei Ren, Zijun Wei, Yan Yan, and Kangning Liu. Vgent: Visual grounding via modular design for disentangling reasoning and prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41160–41170, 2026.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 787–798, 2014.

Seoyeon Kim, Minguk Kang, Dongwon Kim, Jaesik Park, and Suha Kwak. Extending clip's image-text alignment to referring image segmentation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4611–4628, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3992–4003, 2023.

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

Meng Lan, Fu Rong, Hongzan Jiao, Zhi Gao, and Lefei Zhang. Language query-based transformer with multiscale cross-modal alignment for visual grounding on remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

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

Mengcheng Lan, Chaofeng Chen, Jiaxing Xu, Zongrui Li, Yiping Ke, Xudong Jiang, Yingchen Yu, Yunqing Zhao, and Song Bai. Text4seg++: Advancing image segmentation via generative language modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Donghyun Lee, Yuhang Li, Ruokai Yin, Shiting Xiao, and Priyadarshini Panda. Memba: Membrane-driven parameter-efficient fine-tuning for mamba. In *International Conference on Learning Representations*, pp. 97118–97141, 2026.

Minhyun Lee, Seungho Lee, Song Park, Dongyoon Han, Byeongho Heo, and Hyunjung Shim. Maskris: Semantic distortion-aware data augmentation for referring image segmentation. *Transactions on Machine Learning Research*, 2025.

Sen Lei, Xinyu Xiao, Tianlin Zhang, Heng-Chao Li, Zhenwei Shi, and Qing Zhu. Exploring fine-grained image-text alignment for referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–11, 2025.

Sen Lei, Shuai Li, Xinyu Xiao, Heng-Chao Li, Zhenwei Shi, and Qing Zhu. S2clnet: Structure-constrained semantic contrastive learning for referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 64:5636114–5636114, 2026.

Hongbing Li, Linhui Xiao, Zihan Zhao, Qi Shen, Yixiang Huang, Bo Xiao, and Zhanyu Ma. Bare: Towards bias-aware and reasoning-enhanced one-tower visual grounding. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026a.

Jiayuan Li, Zhen Wang, Xiao Sun, Yiming Yao, Nan Xu, Zhuhong You, and Deshuang Huang. A hierarchical vision-language model-guided feature fusion framework for referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–15, 2026b.

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

Ke Li, Di Wang, Haojie Xu, Haodi Zhong, and Cong Wang. Language-guided progressive attention for visual grounding in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in Neural Information Processing Systems*, 34:19652–19664, 2021.

Tianyang Li, Chao Wang, Sirui Tian, Bo Zhang, Fan Wu, Yixian Tang, and Hong Zhang. Tacmt: Text-aware cross-modal transformer for visual grounding on high-resolution sar images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 222:152–166, 2025.

Yang Li, Zitong Feng, Tingrui Wang, Chenyu Wang, and Xin Zhou. Mica: Intra-modal integration and cross-modal alignment adapters for parameter-efficient referring image segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026c.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pp. 280–296, 2022.

Yongfeng Li, Ying Wei, Jiapeng Zhang, and Jiaming Liu. Hierarchical feature-modulated transformer with introspective interactions for visual grounding in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–15, 2026d.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Zheng Lin, Yuxin Zhang, Zhe Chen, Zihan Fang, Xianhao Chen, Praneeth Vepakomma, Wei Ni, Jun Luo, and Yue Gao. Hsplitlora: A heterogeneous split parameter-efficient fine-tuning framework for large language models. *IEEE Transactions on Mobile Computing*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18653–18663, 2023b.

Jiazhen Liu, Mingkuan Feng, and Long Chen. Better, stronger, faster: Tackling the trilemma in mllm-based segmentation with simultaneous textual mask prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33121–33130, 2026a.

Maofu Liu, Xin Jiang, and Xiaokang Zhang. Cadformer: Fine-grained cross-modal alignment and decoding transformer for referring remote sensing image segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:14557–14569, 2025a.

Shih-Wen Liu, Yen-Chang Chen, Wei-Ta Chu, Fu-En Yang, and Yu-Chiang Frank Wang. Frequency switching mechanism for parameter-efficient multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b.

Sihan Liu, Yiwei Ma, Xiaoqing Zhang, Haowei Wang, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. Rotated multi-scale interaction network for referring remote sensing image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26648–26658, 2024.

Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. Caris: Context-aware referring image segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 779–788, 2023c.

Ye Liu, Zongyang Ma, Junfu Pu, Zhongang Qi, Yang Wu, Ying Shan, and Chang Chen. Unipixel: Unified object referring and segmentation for pixel-level visual reasoning. *Advances in Neural Information Processing Systems*, 38:126078–126108, 2025b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9992–10002, 2021.

Xiaoqiang Lu, Long Sun, Lingling Li, Licheng Jiao, Yuting Yang, Zhongjian Huang, Jinming Chai, Xu Liu, Fang Liu, Wenping Ma, et al. Rrsecs: Referring remote sensing expression comprehension and segmentation. *IEEE Geoscience and Remote Sensing Magazine*, 13(3):440–467, 2025.

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

Qin Ma, Lingling Li, Xiaoqiang Lu, Licheng Jiao, Fang Liu, Wenping Ma, Xu Liu, and Long Sun. Lscf: Long-term semantic-guidance convformer for referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–13, 2025.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11–20, 2016.

Zhenjie Mao, Yang Yuhuan, Chaofan Ma, Dongsheng Jiang, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Safire: Saccade-fixation reiteration with mamba for referring image segmentation. *Advances in Neural Information Processing Systems*, 38:7122–7148, 2025.

Seunghun Moon, Hyunwoo Yu, Haeuk Lee, and Suk-Ju Kang. Wimfris: Window mamba fusion and parameter efficient tuning for referring image segmentation. In *International Conference on Learning Representations*, 2026.

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, 2016.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Xiong Pan, Xuemei Xie, Jianxiu Yang, Xiaodan Song, and Guangming Shi. Cmanet: Context-aware mutual attention network for referring image segmentation. *IEEE Transactions on Multimedia*, 2025.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2641–2649, 2015.

Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks master of many: Designing general-purpose coarse-to-fine vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14076–14088, 2024.

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

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, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, pp. 28085–28128, 2025.

Itamar Rocha Filho, Rosie Zhao, Sham Kakade, Eran Malach, and Samy Jelassi. Parameter-efficient reinforcement learning using prefix optimization. In *International Conference on Learning Representations*, pp. 154801–154822, 2026.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, 2015.

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/CVF International Conference on Computer Vision*, pp. 618–626, 2017.

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

Chao Shang, Zichen Song, Heqian Qiu, Lanxiao Wang, Fanman Meng, and Hongliang Li. Prompt-driven referring image segmentation with instance contrasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4124–4134, 2024.

Fengyuan Shi, Ruopeng Gao, Weilin Huang, and Limin Wang. Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1181–1198, 2024.

Liangtao Shi, Ting Liu, Xiantao Hu, Yue Hu, Quanjun Yin, and Richang Hong. Swimvg: Step-wise multimodal fusion and adaption for visual grounding. *IEEE Transactions on Multimedia*, 2025.

Liangtao Shi, Ting Liu, Jinxia Xie, Ning Li, Bineng Zhong, and Richang Hong. Text-guided vision token reduction with low-rank adaptation for efficient visual grounding. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Wei Su, Peihan Miao, Huanzhang Dou, Gaoang Wang, Liang Qiao, Zheyang Li, and Xi Li. Language adaptive weight generation for multi-task visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10857–10866, 2023.

Wei Su, Peihan Miao, Huanzhang Dou, and Xi Li. Scanformer: Referring expression comprehension by iteratively scanning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13449–13458, 2024.

Yuxi Sun, Shanshan Feng, Xutao Li, Yunming Ye, Jian Kang, and Xu Huang. Visual grounding in remote sensing images. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 404–412, 2022.

Yuzhe Sun, Zhe Dong, Haochen Jiang, Yanfeng Gu, and Tianzhu Liu. Crobim-u: Uncertainty-driven referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–16, 2026.

Hao Tang, Chen-Wei Xie, Haiyang Wang, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Ufo: A unified approach to fine-grained visual perception via open-ended language interface. *Advances in Neural Information Processing Systems*, 38:83761–83791, 2025.

Jiajin Tang, Ge Zheng, Cheng Shi, and Sibei Yang. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23570–23580, 2023.

Wei Tang, Yanpeng Sun, Qinying Gu, and Zechao Li. Visual position prompt for mllm based visual grounding. *IEEE Transactions on Multimedia*, 2026.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36:61501–61513, 2023a.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19175–19186, 2023b.

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

Yaoming Wang, Li Jin, Xiaopeng Zhang, Bowen Shi, Chenglin Li, Wenrui Dai, Hongkai Xiong, and Qi Tian. Barleria: An efficient tuning framework for referring image segmentation. In *International Conference on Learning Representations*, pp. 33795–33810, 2024.

Yuhan Wang, Lingling Li, Xu Liu, Licheng Jiao, Xiaoqiang Lu, Fang Liu, and Wenping Ma. Causal prior–driven visual grounding for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37:69925–69975, 2024a.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024b.

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

Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. Hivg: Hierarchical multimodal fine-grained modulation for visual grounding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5460–5469, 2024a.

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

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Towards visual grounding: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3): 2749–2771, 2025.

Jun Xie, Bing Zhang, Zhengchao Chen, Xuan Yang, Yongqing Bai, Zhaoming Wu, and Yue Xu. Global vision-language feature interaction enhanced by object-context association for remote sensing visual grounding. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025.

Yi Xin, Jianjiang Yang, Siqi Luo, Yuntao Du, Qi Qin, Haoxing Chen, Kangrui Cen, Yangfan He, Bin Fu, Yuewen Cao, et al. Parameter-efficient fine-tuning for pre-trained vision models: A survey and benchmark. *International Journal of Computer Vision*, 134(6):304, 2026.

Mingzhu Xu, Tianxiang Xiao, Yutong Liu, Haoyu Tang, Yupeng Hu, and Liqiang Nie. Cmirnet: Cross-modal interactive reasoning network for referring image segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(4):3234–3249, 2025.

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

Jing Yang, Sen Yang, Boqiang Duan, Ming Dai, Wei Zhang, Xiao Tan, Kunbin Chen, Wei He, Jingdong Wang, and Hanli Wang. Hugging visual prompt and segmentation tokens: Consistency learning for fine-grained visual understanding in mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5175–5186, 2026.

Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9489–9498, 2022a.

Yuhuan Yang, Chaofan Ma, Jiangchao Yao, Zhun Zhong, Ya Zhang, and Yanfeng Wang. Remamber: Referring image segmentation with mamba twister. In *European Conference on Computer Vision*, pp. 108–126, 2024.

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*, pp. 18134–18144. IEEE, 2022b.

Zhao Yang, Jiaqi Wang, Xubing Ye, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. Language-aware vision transformer for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5238–5255, 2025.

Ruilin Yao, Yi Rong, Tianyu Zou, Bo Zhang, Jian Li, Shengwu Xiong, and Shili Xiong. Map: Parameter-efficient tuning for referring expression comprehension via multi-modal adaptive positional encoding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2264–2273, 2025.

Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15481–15491, 2022.

Kai Ye, YingShi Luan, Zhudi Chen, Guangyue Meng, Pingyang Dai, and Liujuan Cao. Ris-lad: A benchmark and model for referring image segmentation in low-altitude drone imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11937–11945, 2026.

Dongshuo Yin, Yiran Yang, Zhechao Wang, Hongfeng Yu, Kaiwen Wei, and Xian Sun. 1% vs 100%: Parameter-efficient low rank adapter for dense predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20116–20126, 2023.

Dongshuo Yin, Leiyi Hu, Bin Li, Youqun Zhang, and Xue Yang. 5%>100%: Breaking performance shackles of full fine-tuning on visual recognition tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20071–20081, 2025.

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.

Seonghoon Yu, Joonbeom Hong, Joonseok Lee, and Jeany Son. Latent expression generation for referring image segmentation and grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21374–21383, 2025.

Zhenghang Yuan, Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. Rrsis: Referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.

Pengfei Yue, Jianghang Lin, Shengchuan Zhang, Jie Hu, Yilin Lu, Hongwei Niu, Haixin Ding, Yan Zhang, Guannan Jiang, Liujuan Cao, et al. Adaptive selection based referring image segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1101–1110, 2024.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. Rsvg: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023.

Jingwen Zhang, Lingling Li, Licheng Jiao, Xu Liu, Fang Liu, Wenping Ma, and Shuyuan Yang. A multiscale vision–text collaborative dual encoder for referring rs image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–15, 2026a.

Xiaowen Zhang, Licheng Jiao, Lingling Li, Long Sun, Xu Liu, Fang Liu, Mengjia Wang, Wenping Ma, and Yuan Wu. Visprovg: Confidence-aware visual prompting for remote sensing visual grounding. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–14, 2026b.

Enyuan Zhao, Ziyi Wan, Ze Zhang, Jie Nie, Xinyue Liang, and Lei Huang. A spatial-frequency fusion strategy based on linguistic query refinement for rsvg. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024.

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

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

APPENDIX

A DETAILS OF EFFICIENT MULTIMODAL ENCODING

As introduced in Section 3.1, DeCo employs a frozen BEiT3 (Wang et al., 2023b) encoder with the classic Low-Rank Adaptation (LoRA) (Hu et al., 2022) followed by the Simple Feature Pyramid (SimFPN) (Li et al., 2022) to construct efficient multimodal and multi-scale representations. Here, we provide more architectural details of these two components.

Low-Rank Adaptation. To efficiently exploit the pretrained multimodal encoder while preserving its original knowledge, we keep all BEiT3 parameters frozen and inject LoRA into the linear projections of its Transformer layers. Formally, at the i-th layer, the visual patch embeddings $\mathbf{V}_i$ and linguistic token embeddings $\mathbf{T}_i$ are first concatenated into a unified multimodal representation $\mathbf{Z}_i$, which is then processed by the multi-head self-attention operator. Given a fixed projection matrix $\mathbf{W}_0 \in \mathbb{R}^{d\times d}$, LoRA models its task-specific update with a pair of low-rank projections:

$$\mathbf{W} = \mathbf{W}_0 + \Delta\mathbf{W}, \qquad \Delta\mathbf{W} = \frac{\alpha}{r}\mathbf{C}\mathbf{A}, \tag{15}$$

where $\mathbf{A} \in \mathbb{R}^{r\times d}$ and $\mathbf{C} \in \mathbb{R}^{d\times r}$ are two lightweight trainable matrices, $r \ll d$ denotes the low-rank dimension, and the scaling factor α controls the update ratio. Accordingly, the adapted projection of the multimodal tokens $\mathbf{Z}_i$ is formulated as follows:

$$\mathbf{Z}_i = \mathbf{Z}_i\mathbf{W}_0^\top + \frac{\alpha}{r}\mathbf{Z}_i\,\mathbf{A}^\top\mathbf{C}^\top. \tag{16}$$

In DeCo, LoRA is inserted into the query, key, value, and output projections of each Transformer attention layer. For BEiT3-B and BEiT3-L, the embedding dimension d is 768 and $1,024$, with 12 and 24 layers, respectively. Unless otherwise specified, we set both r and α to 32 for all experiments. Since these projections operate on the unified vision-language sequence, the low-rank updates efficiently enable multimodal interactions while avoiding retraining the original encoder.

Simple Feature Pyramid. The final visual representation produced by BEiT3 is a single-scale sequence of patch embeddings. To provide the multi-scale representations required by localization and segmentation, we adopt SimFPN to construct a four-level feature pyramid. Specifically, excluding the class token $\bar{\mathbf{v}}$, we first reshape the visual patch embeddings into a spatial feature map:

$$\mathbf{F} = \text{reshape}\left([\mathbf{v}_1, \ldots, \mathbf{v}_n]\right) \in \mathbb{R}^{\frac{h}{p_{size}}\times\frac{w}{p_{size}}\times d}, \tag{17}$$

where $p_{size} = 16$ represents the patch size. Then, $\mathbf{F}$ is processed by four parallel deconvolutions and convolutions to obtain features at different spatial resolutions, as follows:

$$\mathbf{V}_l = \phi_l\big(\mathcal{S}_l(\mathbf{F})\big) \in \mathbb{R}^{h_l\times w_l\times c}, \qquad l \in \{1,2,3,4\}, \tag{18}$$

where $\mathcal{S}_l(\cdot)$ denotes the corresponding rescaling operation, $\phi_l(\cdot)$ projects each feature map into a small channel dimension $c = 256$, and the four scale factors $s_l \in \{4, 2, 1, \frac{1}{2}\}$ yield different spatial resolutions $(h_l, w_l) = s_l\,(\frac{h}{p_{size}}, \frac{w}{p_{size}})$, resulting in multi-scale features $\mathcal{V} = \{\mathbf{V}_l\}_{l=1}^4$. This design bridges the single-scale representation of the frozen multimodal encoder and the multi-resolution requirements of localization and segmentation, while its lightweight modules avoid a costly hierarchical backbone and balance spatial representation capability with computational efficiency.

B MORE EXPERIMENTAL SETUP DETAILS

B.1 DATASETS

To investigate the effectiveness of DeCo, we conduct extensive experiments across diverse vision-language benchmarks covering natural and remote sensing images. The datasets differ in visual appearance, linguistic complexity, image modality, and task formulation, providing a comprehensive testbed for evaluating multi-task visual grounding. Detailed statistics are summarized in Table 9.

Table 9: Detailed statistics of the datasets used in our experiments. **Words** denote the average number of whitespace-separated words in each referring expression. **REC**, **RES**, **RSVG**, and **RRSIS** represent referring expression comprehension, referring expression segmentation, remote sensing visual grounding, and referring remote sensing image segmentation, respectively.

Datasets	Images	Expressions				Words	Tasks
Natural Image Benchmarks							
RefCOCO (Yu et al., 2016)	19,994	train: 120,642	val: 10,834	testA: 5,657	testB: 5,095	3.5	REC, RES
RefCOCO+ (Yu et al., 2016)	19,992	train: 120,191	val: 10,758	testA: 5,726	testB: 4,889	3.5	REC, RES
G-Ref UMD (Nagaraja et al., 2016)	25,799	train: 80,512	val-u: 4,896	test-u: 9,602	/	8.4	REC, RES
G-Ref Google (Mao et al., 2016)		train: 85,474	val-g: 9,536	/	/	8.4	
ReferIt (Kazemzadeh et al., 2014)	20,000	train: 54,127	val: 5,842	test: 60,103	/	3.4	REC
Flickr (Plummer et al., 2015)	31,783	train: 427,193	val: 14,433	test: 14,481	/	1.6	REC
Remote Sensing Image Benchmarks							
DIOR-RSVG (Zhan et al., 2023)	17,402	train: 26,991	val: 3,829	test: 7,500	/	7.5	RSVG
SARVG1.0 (Li et al., 2025)	2,465	train: 6,093	val: 761	test: 763	/	6.2	RSVG
RRSIS-D (Liu et al., 2024)	17,402	train: 12,181	val: 1,740	test: 3,481	/	6.8	RRSIS
RIS-LAD (Ye et al., 2026)	2,103	train: 9,712	val: 1,396	test: 2,763	/	19.6	RRSIS
RefDIOR (Lu et al., 2025)	17,380	train: 26,824	val: 3,832	test: 7,664	/	7.5	RSVG, RRSIS

B.1.1 Natural Image Benchmarks

We first evaluate DeCo on five widely used natural image benchmarks: **RefCOCO** (Yu et al., 2016) is derived from MSCOCO (Lin et al., 2014) and contains relatively concise referring expressions for identifying objects in diverse everyday scenes. **RefCOCO+** (Yu et al., 2016) follows the same image source but removes location-based expressions, placing greater emphasis on appearance-related attributes and therefore requiring finer-grained correspondence. **G-Ref** (Mao et al., 2016; Nagaraja et al., 2016) provides substantially longer and more descriptive expressions, with an average of 8.4 words per expression, making it a challenging benchmark for multi-task visual grounding. It provides two different partitions, UMD (Nagaraja et al., 2016) and Google (Mao et al., 2016), with the former adopted as the main training and evaluation protocol in our experiments. **ReferIt** (Kazemzadeh et al., 2014) contains referring expressions collected for objects and regions in natural images and provides a large test set for evaluating generalization beyond MSCOCO-based benchmarks. **Flickr** (Plummer et al., 2015) constructs large-scale image-phrase pairs with relatively short descriptions, providing an additional benchmark for evaluating REC in diverse natural scenes.

Following previous works (Xiao et al., 2024b; Dai et al., 2025b; Chen et al., 2026a), we conduct experiments on both the *Standard* and *Mixed* settings. In the standard setting, each dataset is trained and evaluated independently using its corresponding splits. In the mixed setting, we combine the training sets of RefCOCO, RefCOCO+, and the UMD partition of G-Ref to construct a unified training set, while keeping their original validation and test splits separate for evaluation.

B.1.2 Remote Sensing Image Benchmarks

To further examine the generalization of DeCo beyond natural images, we consider five remote sensing benchmarks, including optical, SAR, and UAV imagery. **DIOR-RSVG** (Zhan et al., 2023) is an optical visual grounding benchmark, where referring expressions are used to identify specific objects in aerial scenes. Compared with natural images, the large image scale, dense object distributions, and substantial appearance variations introduce additional challenges for localization. **SARVG1.0** (Li et al., 2025) focuses on visual grounding in SAR imagery. In contrast to optical imagery, SAR observations exhibit distinct scattering characteristics and substantially different visual patterns, providing a challenging testbed for evaluating the robustness of DeCo under various imaging modalities. **RRSIS-D** (Liu et al., 2024) is an optical referring image segmentation benchmark that associates textual description with pixel-level object masks. It therefore evaluates whether language-guided understanding can be translated from object-level localization to fine-grained spatial segmentation in aerial scenes. **RIS-LAD** (Ye et al., 2026) further considers UAV imagery and contains substantially longer referring expressions, with an average length of 19.6 words. Its fine-grained descriptions and aerial perspectives provide a more challenging setting for evaluating tiny object segmentation. **RefDIOR** (Lu et al., 2025) provides a unified benchmark covering both object

Table 10: Detailed implementation settings across benchmarks. BS: Batch Size. LR: Learning Rate.

Settings		RefCOCO	RefCOCO+	G-Ref	ReferIt	Flickr	DIOR-RSVG	SARVG1.0	RRSIS-D	RIS-LAD	RefDIOR
Input Size		448×448	448×448	448×448	448×448	448×448	640×640	512×512	512×512	640×640	512×512
Text Length		20	20	20	20	20	20	20	20	40	20
BS	DeCo-B	64	64	64	64	64	32	32	32	32	32
	DeCo-L	32	32	32	32	32	16	16	16	16	16
LR	DeCo-B	2×10^{-4}	2×10^{-4}	2×10^{-4}	2×10^{-4}	2×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
	DeCo-L	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}
Epochs		35	35	35	35	10	50	90	40	50	50
Milestones		28	28	28	28	8	36, 45	60	30	36, 45	36, 45

localization and pixel-level segmentation in optical remote sensing imagery. We use it to evaluate the unified referring remote sensing comprehension and segmentation capability of DeCo, thereby extending the joint REC and RES formulation to remote sensing scenarios. All the above remote sensing image benchmarks follow the standard setting for training, validation, and testing.

B.2 Evaluation Metrics

In the main experiments, we evaluate DeCo using commonly used metrics for REC, RES, and their corresponding remote sensing variants, RSVG and RRSIS. For REC, we adopt Precision@IoU threshold of 0.5 (**P@.5**), which measures the percentage of predictions whose IoU with the ground-truth box exceeds 0.5. Specifically, given the i-th predicted bounding box $\mathbf{B}_i$ and its corresponding ground-truth box $\mathbf{B}_i^*$, P@.5 can be calculated as follows:

$$\mathrm{IoU}(\mathbf{B}_i, \mathbf{B}_i^*) = \frac{|\mathbf{B}_i \cap \mathbf{B}_i^*|}{|\mathbf{B}_i \cup \mathbf{B}_i^*|}, \qquad \mathrm{P@.5} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\big[\,\mathrm{IoU}(\mathbf{B}_i, \mathbf{B}_i^*) \geq 0.5\big], \tag{19}$$

where N denotes the number of referring expressions and $\mathbb{I}[\cdot]$ is the indicator function. For RES, we report overall IoU (**oIoU**) and mean IoU (**mIoU**). Given the predicted mask $\mathbf{M}_i$ and the ground-truth mask $\mathbf{M}_i^*$, oIoU aggregates the intersection and union across all samples, while mIoU averages the per-sample IoU, which can be formulated as follows:

$$\mathrm{oIoU} = \frac{\sum_{i=1}^{N} |\mathbf{M}_i \cap \mathbf{M}_i^*|}{\sum_{i=1}^{N} |\mathbf{M}_i \cup \mathbf{M}_i^*|}, \qquad \mathrm{mIoU} = \frac{1}{N}\sum_{i=1}^{N} \frac{|\mathbf{M}_i \cap \mathbf{M}_i^*|}{|\mathbf{M}_i \cup \mathbf{M}_i^*|}. \tag{20}$$

For a more precise evaluation in the ablation studies, we additionally introduce **P@.5:.95** for REC, which averages the precision over IoU thresholds from 0.5 to 0.95 with an interval of 0.05. The metrics used for RSVG and RRSIS follow similar computation protocols described above. All metrics are reported in percentage form, with higher values indicating better performance.

B.3 Implementation Details

We instantiate DeCo with two model scales, termed **DeCo-B** and **DeCo-L**, using frozen BEiT3-B/L (Wang et al., 2023b) with LoRA (Hu et al., 2022) adaptation as the multimodal encoders, respectively. For localization, we adopt a decoder-only Deformable DETR (Zhu et al., 2021) consisting of four Transformer decoder layers, with a feature dimension of 256 and an FFN hidden dimension of $1,024$. For segmentation, we employ a lightweight UNet (Ronneberger et al., 2015) with an input feature dimension of 256 and a hidden embedding dimension of 32. For datasets without mask annotations, including ReferIt, Flickr, DIOR-RSVG, and SARVG1.0, we use box-filled binary masks derived from the ground-truth bounding boxes as pseudo-mask supervision. We use Adam (Kingma & Ba, 2015) as the optimizer and MultiStepLR as the learning rate scheduler for all benchmarks. The dataset-specific training configurations are summarized in Table 10. The maximum text length is set to 20 tokens for all benchmarks except RIS-LAD, for which we use 40 tokens due to its substantially longer referring expressions. All experiments are conducted on eight NVIDIA RTX 4090 GPUs. Different batch sizes and learning rates are adopted for DeCo-B and DeCo-L according to their model scales and the input resolutions of various benchmarks.

Table 11: Main results of **RES** models using **mIoU** on RefCOCO, RefCOCO+, and G-Ref datasets.

Models	Vision / Language Backbone	RefCOCO val	RefCOCO testA	RefCOCO testB	RefCOCO+ val+	RefCOCO+ testA+	RefCOCO+ testB+	G-Ref val-u	G-Ref test-u	Avg.
Task-specific Methods (Standard Setting)										
EEVG (Chen et al., 2024b)	ViTDet-B / BERT-B	78.2	79.3	76.6	69.0	72.7	62.3	69.2	70.0	72.2
RISCLIP (Kim et al., 2024)	CLIP-B* / CLIP-B*	75.7	78.0	72.5	69.2	73.5	60.7	67.6	68.0	70.6
OneRef (Xiao et al., 2024b)	BEiT3-B / BEiT3-B	77.6	79.1	75.1	71.3	75.4	65.5	69.4	69.7	72.9
MaskRIS (Lee et al., 2025)	Swin-B / BERT-B	78.4	80.2	76.1	71.7	76.7	64.5	69.3	69.4	73.3
CMIRNet (Xu et al., 2025)	Swin-B / BERT-B	76.0	78.1	73.1	68.2	73.0	61.3	66.7	66.5	70.4
DETRIS (Huang et al., 2025)	DINOv2-B* / CLIP-B*	76.0	78.2	73.5	68.9	74.0	61.5	67.9	68.1	71.0
PropVG (Dai et al., 2025b)	BEiT3-B / BEiT3-B	78.0	79.8	75.3	72.9	76.5	67.2	71.3	72.1	74.1
Latent-VG (Yu et al., 2025)	BEiT3-B / BEiT3-B	77.7	79.5	75.5	73.2	76.2	67.4	71.1	71.5	74.0
MiCA (Li et al., 2026c)	DINOv2-B* / CLIP-B*	76.6	78.9	73.8	69.8	74.7	63.5	68.7	68.8	71.9
AMLRIS (Chen et al., 2026a)	Swin-B / BERT-B	77.9	79.5	75.0	71.3	75.6	64.6	69.2	69.7	72.9
WIMFRIS (Moon et al., 2026)	DINOv2-B* / CLIP-B*	77.3	79.2	74.8	70.2	75.1	63.5	69.2	69.4	72.3
DeCo (Ours)	BEiT3-B* / BEiT3-B*	**80.3**	**81.8**	**77.5**	**75.8**	**79.2**	**70.3**	**74.0**	**74.5**	**76.7**
RISCLIP (Kim et al., 2024)	CLIP-L* / CLIP-L*	78.9	81.5	75.4	74.4	78.8	66.8	71.8	71.7	74.9
OneRef (Xiao et al., 2024b)	BEiT3-L / BEiT3-L	80.1	82.2	77.5	75.2	79.4	70.2	73.2	73.8	76.4
DETRIS (Huang et al., 2025)	DINOv2-L* / CLIP-B*	77.3	79.0	75.2	70.8	75.3	64.7	69.3	70.2	72.7
MiCA (Li et al., 2026c)	DINOv2-L* / CLIP-B*	78.0	79.6	75.8	72.2	76.7	65.9	70.5	70.7	73.7
WIMFRIS (Moon et al., 2026)	DINOv2-L* / CLIP-B*	78.2	79.7	76.3	71.9	76.2	67.2	70.4	71.0	73.9
DeCo (Ours)	BEiT3-L* / BEiT3-L*	**83.2**	**84.4**	**81.2**	**80.5**	**83.1**	**75.7**	**78.7**	**79.0**	**80.7**
Task-specific Methods (Mixed Setting)										
PolyFormer (Liu et al., 2023b)	Swin-B / BERT-B	76.0	77.1	73.2	70.7	74.5	64.6	69.4	69.9	71.9
RISCLIP (Kim et al., 2024)	CLIP-B* / CLIP-B*	76.0	78.6	71.9	69.7	74.3	61.4	69.6	69.6	71.4
EEVG (Chen et al., 2024b)	ViTDet-B / BERT-B	79.5	81.0	77.4	71.9	76.7	66.3	73.6	73.5	75.0
OneRef (Xiao et al., 2024b)	BEiT3-B / BEiT3-B	79.8	81.9	77.0	74.7	77.9	69.6	74.1	74.9	76.2
CMIRNet (Xu et al., 2025)	Swin-B / BERT-B	78.5	80.2	76.0	71.2	75.3	65.0	71.5	72.8	73.8
C^3VG (Dai et al., 2025c)	BEiT3-B / BEiT3-B	81.4	82.9	79.1	77.1	79.6	72.4	76.3	77.1	78.2
PropVG (Dai et al., 2025b)	BEiT3-B / BEiT3-B	82.0	83.6	80.0	77.1	79.8	72.2	77.0	77.7	78.7
Latent-VG (Yu et al., 2025)	BEiT3-B / BEiT3-B	81.0	82.3	79.8	76.9	79.5	73.0	76.1	76.5	78.1
SaFiRe (Mao et al., 2025)	Swin-B / BERT-B	82.1	83.7	80.1	77.4	80.5	72.9	77.3	78.1	79.0
InstanceVG (Dai et al., 2025a)	BEiT3-B / BEiT3-B	81.4	83.1	79.3	76.6	79.5	71.6	75.9	76.6	78.0
DeCo (Ours)	BEiT3-B* / BEiT3-B*	**82.8**	**84.3**	**80.9**	**79.3**	**81.8**	**74.7**	**78.2**	**78.9**	**80.1**
PolyFormer (Liu et al., 2023b)	Swin-L / BERT-B	76.9	78.5	74.8	72.2	75.7	66.7	71.2	71.2	73.4
RISCLIP (Kim et al., 2024)	CLIP-L* / CLIP-L*	79.5	82.1	75.8	74.9	78.9	68.1	73.5	74.5	75.9
UNINEXT (Yan et al., 2023)	ConvNeXt-L / BERT-B	81.2	82.4	79.1	72.6	76.3	66.8	74.1	75.4	76.0
OneRef (Xiao et al., 2024b)	BEiT3-L / BEiT3-L	81.3	83.1	79.5	76.6	80.2	73.0	75.7	76.8	78.2
CMIRNet (Xu et al., 2025)	Swin-L / BERT-B	79.0	80.7	76.6	72.5	77.0	66.9	72.6	73.2	74.8
DETRIS (Huang et al., 2025)	DINOv2-L* / CLIP-B*	81.0	81.9	79.0	75.2	78.6	70.2	74.6	75.3	77.0
MiCA (Li et al., 2026c)	DINOv2-L* / CLIP-B*	80.8	82.7	80.1	75.6	79.3	71.5	75.2	75.7	77.6
InstanceVG (Dai et al., 2025a)	BEiT3-L / BEiT3-L	**86.3**	**87.1**	**85.3**	82.5	84.3	79.2	81.4	82.3	83.5
WIMFRIS (Moon et al., 2026)	DINOv2-L* / CLIP-B*	81.8	82.8	79.7	76.6	80.5	72.5	75.4	75.9	78.2
DeCo (Ours)	BEiT3-L* / BEiT3-L*	85.6	86.3	84.0	**83.1**	**85.0**	**79.7**	**82.3**	**82.6**	**83.6**

C More Experimental Results

C.1 Main Results for Referring Expression Segmentation using mIoU

Table 11 reports the RES results using mIoU on RefCOCO, RefCOCO+, and G-Ref. Overall, DeCo remains highly competitive under this sample-wise segmentation metric, showing a consistent trend with the oIoU results reported in the main paper. Under the standard setting, DeCo-B achieves an average mIoU of 76.7%, outperforming the strongest competing method, PropVG (Dai et al., 2025b), by 2.6%. Scaling to DeCo-L further increases the average mIoU to 80.7%, yielding a substantial 4.3% gain over OneRef (Xiao et al., 2024b). Notably, DeCo-L consistently ranks first across all eight evaluation splits, demonstrating that the proposed decouple-to-couple learning effectively improves segmentation quality across samples rather than benefiting only from large foreground regions.

Under the mixed setting, DeCo-B obtains an average mIoU of 80.1%, surpassing the previous best result of SaFiRe (Mao et al., 2025) by 1.1% and achieving the best performance on all evaluation splits. DeCo-L further reaches an average mIoU of 83.6%. Although InstanceVG (Dai et al., 2025a) performs better on the RefCOCO splits, DeCo-L consistently delivers superior results on the more challenging RefCOCO+ and G-Ref datasets, leading to the best overall average performance. The advantage also extends to the G-Ref Google split, where DeCo-B and DeCo-L achieve 73.7% and 79.1% mIoU, exceeding WIMFRIS (Moon et al., 2026) by 6.6% and 9.8%, as listed in Table 4.

Moreover, Figure 3 provides qualitative comparisons across various model scales and training settings, showing that DeCo consistently produces well-aligned localization boxes and accurate masks

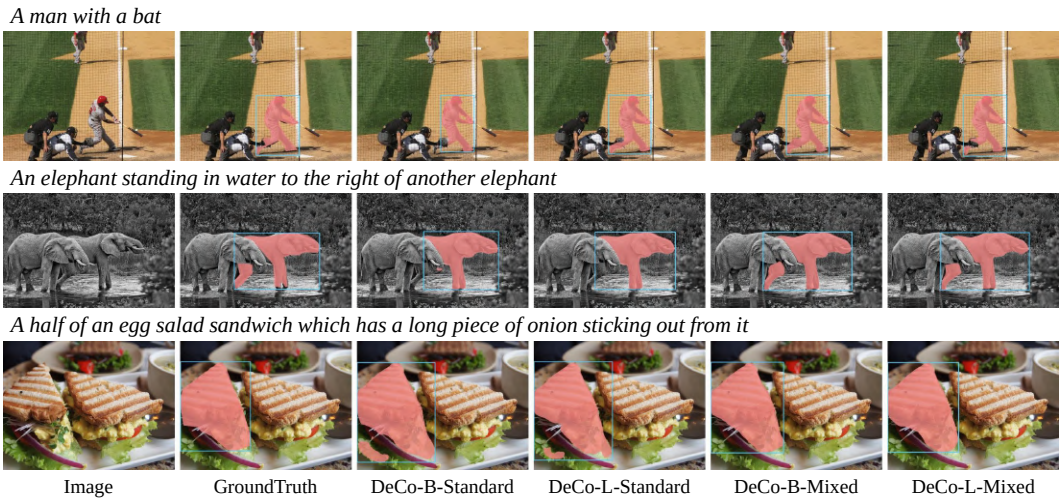

Figure 3: Qualitative results of the proposed DeCo across different settings on G-Ref (test-u).

for both simple and challenging referring expressions. These results confirm the effectiveness and robustness of DeCo under both dataset-specific and mixed-data training settings.

C.2 Generalization to Remote Sensing Scenarios

Beyond the standard natural image benchmarks, we further investigate the generalization of DeCo to diverse remote sensing scenarios. Compared with natural images, remote sensing imagery exhibits substantial domain shifts caused by aerial viewpoints, large-scale variations, dense object distributions, and modality-specific visual characteristics, posing additional challenges to multi-task visual grounding. In this domain, REC and RES are extended to remote sensing visual grounding (RSVG) (Sun et al., 2022) and referring remote sensing image segmentation (RRSIS) (Yuan et al., 2024), respectively. To provide a comprehensive evaluation, we consider multiple benchmarks spanning optical, SAR, and UAV imagery, enabling us to assess whether the proposed method remains effective under substantially different visual distributions and imaging modalities.

Main results for RSVG on DIOR-RSVG. Table 12 reports the results on the DIOR-RSVG test set. DeCo-B already achieves competitive performance, obtaining 85.2% P@.5 and 76.0% mIoU, which are on par with the strongest existing methods. Scaling to DeCo-L further yields consistent improvements across all evaluation metrics. Specifically, DeCo-L achieves 86.9%, 84.3%, 79.0%, 69.2%, and 47.6% under P@.5–P@.9, respectively, while reaching 86.2% oIoU and 78.6% mIoU. Compared with the recent advanced method, VisproVG (Zhang et al., 2026b), DeCo-L improves P@.5, P@.7, P@.9, oIoU, and mIoU by 1.7%, 0.8%, 0.7%, 1.5%, and 2.6%, respectively. The consistent gains across increasingly strict IoU thresholds indicate that DeCo not only identifies the referred remote sensing objects reliably but also produces more precise localization.

Main results for RSVG on SARVG1.0. Table 13 lists the results on the SARVG1.0 test set, which further examines the generalization of DeCo from optical imagery to SAR modality. Despite the distinct scattering characteristics and visual appearance of SAR, both DeCo-B and DeCo-L achieve strong performance across all evaluation metrics. DeCo-B reaches 89.6% P@.5 and 83.8% mIoU, already surpassing the previous best method, TACMT (Li et al., 2025), on all reported precision thresholds. DeCo-L further improves the performance to 91.2% P@.5 and 86.3% mIoU. The advantage becomes more evident under stricter localization criteria. At P@.9, DeCo-B and DeCo-L achieve 73.5% and 75.1%, outperforming TACMT by 5.2% and 6.8%, respectively. These consistent gains across various IoU thresholds demonstrate that DeCo can effectively transfer its language-guided localization capability to SAR imagery while maintaining high localization precision.

Main results for RSVG on RefDIOR. Table 14 summarizes the localization results on the RefDIOR test set. RefDIOR is designed for multi-task visual grounding in remote sensing scenes, making it a more comprehensive benchmark for evaluating our DeCo. DeCo-B achieves 83.1%

Table 12: Main results of **REC** and **RSVG** models on the DIOR-RSVG test set.

Models	Vision / Language	P@.5	P@.6	P@.7	P@.8	P@.9	oIoU	mIoU
TransVG (Deng et al., 2021)	ResNet-50 / BERT-B	72.4	67.4	60.1	49.1	27.8	76.3	63.6
VLTVG (Yang et al., 2022a)	ResNet-50 / BERT-B	69.4	65.2	58.4	46.6	24.4	72.0	60.0
SegVG (Kang et al., 2024)	ResNet-50 / BERT-B	79.6	73.7	65.0	51.1	25.7	80.9	68.5
SimVG (Dai et al., 2024)	BEiT3-B / BEiT3-B	79.1	73.3	66.0	51.9	26.6	74.0	69.5
MGVLF (Zhan et al., 2023)	ResNet-50 / BERT-B	76.8	72.7	66.7	56.4	35.1	78.4	68.0
LQVG (Lan et al., 2024)	ResNet-50 / BERT-B	83.4	81.0	75.9	65.5	43.5	82.2	74.0
FQRNet (Zhao et al., 2024)	ResNet-50 / BERT-B	77.2	73.2	67.2	56.9	35.4	78.8	68.4
EG-DINO (Hu et al., 2025)	ResNet-50 / BERT-B	83.1	80.1	75.0	63.6	42.3	81.1	73.4
GLAED (Xie et al., 2025)	ResNet-50 / BERT-B	84.4	81.5	73.5	62.9	43.4	82.9	73.7
CPVG (Wang et al., 2026)	BEiT3-B / BEiT3-B	82.6	78.4	71.9	59.1	33.6	76.9	73.2
CFCVG (Cao et al., 2026)	Swin-T / BERT-B	84.7	82.2	77.6	66.3	44.7	83.3	75.2
HMIT (Li et al., 2026d)	HGNetv2 / BERT-B	84.4	82.2	78.1	68.2	45.8	83.1	75.3
VisproVG (Zhang et al., 2026b)	Swin-T / BERT-B	85.2	83.0	78.2	68.2	46.9	84.7	76.0
DeCo (Ours)	BEiT3-B* / BEiT3-B*	85.2	82.2	76.6	66.7	44.6	84.1	76.0
DeCo (Ours)	BEiT3-L* / BEiT3-L*	**86.9**	**84.3**	**79.0**	**69.2**	**47.6**	**86.2**	**78.6**

Table 13: Main results of **REC** and **RSVG** models on the SARVG1.0 test set.

Models	Vision / Language	P@.5	P@.6	P@.7	P@.8	P@.9	oIoU	mIoU
TransVG (Deng et al., 2021)	ResNet-50 / BERT-B	82.2	76.6	69.5	50.5	15.9	-	68.8
RefTR (Li & Sigal, 2021)	ResNet-50 / BERT-B	77.2	74.4	69.8	59.9	27.0	-	68.4
LBYLNet (Huang et al., 2021)	ResNet-50 / BERT-B	77.1	76.4	75.9	74.8	71.6	-	72.7
VLTVG (Yang et al., 2022a)	ResNet-50 / BERT-B	82.4	80.7	73.1	62.4	29.8	-	72.4
SeqTR (Zhu et al., 2022)	ResNet-50 / BERT-B	85.7	84.3	80.4	69.8	34.2	-	74.8
TACMT (Li et al., 2025)	ResNet-50 / BERT-B	89.4	88.6	87.0	83.8	68.3	-	82.8
DeCo (Ours)	BEiT3-B* / BEiT3-B*	89.6	89.1	88.0	84.0	73.5	80.6	83.8
DeCo (Ours)	BEiT3-L* / BEiT3-L*	**91.2**	**90.2**	**88.5**	**85.0**	**75.1**	**83.5**	**86.3**

Table 14: Main results of **REC** and **RSVG** models on the RefDIOR test set.

Models	Vision / Language	P@.5	P@.6	P@.7	P@.8	P@.9	oIoU	mIoU
TransVG (Deng et al., 2021)	Swin-T / BERT-B	53.7	46.3	36.1	21.3	5.3	67.1	46.7
VLTVG (Yang et al., 2022a)	Swin-T / BERT-B	59.4	51.5	41.2	26.4	7.8	70.2	51.7
PseudoQ (Jiang et al., 2022)	Swin-T / BERT-B	57.1	48.7	37.6	23.1	5.4	69.2	49.7
QRNet (Ye et al., 2022)	Swin-T / BERT-B	57.5	49.4	39.1	23.7	5.8	69.5	50.1
D-MDETR (Shi et al., 2024)	Swin-T / BERT-B	60.9	53.3	43.6	28.6	8.1	69.9	52.5
LQVG (Lan et al., 2024)	Swin-T / BERT-B	78.7	75.6	70.5	60.8	42.9	81.4	70.7
LPVA (Li et al., 2024)	Swin-T / BERT-B	60.5	52.6	41.9	27.1	7.3	71.1	52.7
CCFormer (Lu et al., 2025)	Swin-T / BERT-B	82.4	79.1	73.5	64.1	45.8	82.4	74.1
DeCo (Ours)	BEiT3-B* / BEiT3-B*	83.1	79.7	74.2	65.0	47.0	83.1	75.0
DeCo (Ours)	BEiT3-L* / BEiT3-L*	**85.3**	**82.4**	**77.3**	**67.5**	**50.0**	**84.2**	**77.0**

P@.5 and 75.0% mIoU, surpassing the previous best method, CCFormer (Lu et al., 2025), across all reported metrics. DeCo-L further improves the results to 85.3% P@.5 and 77.0% mIoU. The superiority of DeCo remains consistent under stricter IoU thresholds. Compared with CCFormer, DeCo-L improves P@.5, P@.7, and P@.9 by 2.9%, 3.8%, and 4.2%, respectively, while gaining 1.8% in oIoU and 2.9% in mIoU. These results indicate that the benefits of modeling task-specificity and cross-task complementarity also extend to unified multi-task learning in remote sensing scenarios.

Main results for RRSIS on RRSIS-D. Different from natural referring segmentation, RRSIS requires language-guided pixel-level understanding under overhead viewpoints and complex object layouts, making it a challenging testbed for evaluating DeCo in remote sensing segmentation. As shown in Table 15, DeCo-B achieves 77.3% P@.5, 78.3% oIoU, and 66.4% mIoU, outperforming the previous best method, MCD-Net (Zhang et al., 2026a), on the major evaluation metrics. DeCo-L further improves the performance to 81.3% P@.5, 79.7% oIoU, and 68.9% mIoU. Compared with the strongest prior results (Lei et al., 2026; Zhang et al., 2026a), DeCo-L improves P@.5, P@.6, and P@.7 by 5.7%, 5.7%, and 3.8%, respectively, while increasing oIoU and mIoU by 1.6% and 3.8%. These results show that the decouple-to-couple design transfers effectively to remote sensing referring segmentation and provides more accurate region-level alignment.

Table 15: Main results of **RES** and **RRSIS** models on the RRSIS-D test set.

Models	Vision / Language	P@.5	P@.6	P@.7	P@.8	P@.9	oIoU	mIoU
LAVT (Yang et al., 2022b)	Swin-B / BERT-B	69.7	63.9	54.1	42.0	25.0	77.7	61.3
CrossVLT (Cho et al., 2023)	Swin-B / BERT-B	66.4	59.4	49.8	38.7	23.3	75.5	58.5
CARIS (Liu et al., 2023c)	Swin-B / BERT-B	71.5	63.5	52.9	40.9	23.9	77.2	62.2
DMMI (Hu et al., 2023)	Swin-B / BERT-B	68.7	61.0	50.3	38.4	21.6	76.2	60.1
LGCE (Yuan et al., 2024)	Swin-B / BERT-B	69.9	64.0	53.6	40.8	24.2	77.2	60.9
RMSIN (Liu et al., 2024)	Swin-B / BERT-B	74.2	67.1	55.6	41.7	23.1	77.5	64.0
FIANet (Lei et al., 2025)	Swin-B / BERT-B	74.5	67.0	56.3	42.8	24.1	76.9	64.0
LSCF (Ma et al., 2025)	Swin-B / BERT-B	74.3	67.7	56.3	43.1	25.7	77.4	64.3
CroBIM-U (Sun et al., 2026)	ConvNeXt-B / BERT-B	75.6	67.7	56.5	42.6	24.2	76.7	65.1
VSPNet (Li et al., 2026b)	MambaVision / CLIP	75.1	67.6	55.4	44.6	25.4	77.8	64.7
S2CLNet (Lei et al., 2026)	Swin-B / BERT-B	74.9	68.9	56.7	44.1	25.6	77.7	64.5
MCD-Net (Zhang et al., 2026a)	Swin-B / BERT-B	75.6	68.7	57.5	44.2	**26.0**	78.1	65.1
DeCo-B (Ours)	BEiT3-B* / BEiT3-B*	77.3	70.4	58.0	43.5	23.9	78.3	66.4
DeCo-L (Ours)	BEiT3-L* / BEiT3-L*	**81.3**	**74.6**	**61.3**	**45.8**	25.7	**79.7**	**68.9**

Table 16: Main results of **RES** and **RRSIS** models on the RIS-LAD test set.

Models	Vision / Language	P@.5	P@.6	P@.7	P@.8	P@.9	oIoU	mIoU
LAVT (Yang et al., 2022b)	Swin-B / BERT-B	31.7	27.0	23.0	17.6	8.6	42.0	30.1
ASDA (Yue et al., 2024)	Swin-B / BERT-B	33.2	26.9	20.2	12.8	3.9	37.5	33.3
LGCE (Yuan et al., 2024)	Swin-B / BERT-B	26.4	21.8	17.6	13.3	7.1	40.8	26.2
RMSIN (Liu et al., 2024)	Swin-B / BERT-B	43.4	38.1	32.4	25.1	13.8	48.8	39.6
FIANet (Lei et al., 2025)	Swin-B / BERT-B	40.2	35.2	29.9	23.7	13.2	43.4	37.4
CADFormer (Liu et al., 2025a)	Swin-B / BERT-B	42.2	37.0	31.9	23.6	12.6	46.5	39.3
SAARN (Ye et al., 2026)	Swin-B / BERT-B	45.0	39.3	33.4	26.2	15.3	49.6	41.7
DeCo (Ours)	BEiT3-B* / BEiT3-B*	51.1	46.3	40.8	31.9	15.5	50.3	47.1
DeCo (Ours)	BEiT3-L* / BEiT3-L*	**56.1**	**51.6**	**45.9**	**35.5**	**18.9**	**54.8**	**50.2**

Table 17: Main results of **RES** and **RRSIS** models on the RefDIOR test set.

Models	Vision / Language	P@.5	P@.6	P@.7	P@.8	P@.9	oIoU	mIoU
LAVT (Yang et al., 2022b)	Swin-T / BERT-B	67.8	62.1	54.2	44.7	29.9	79.0	60.8
CGFormer (Tang et al., 2023)	Swin-T / BERT-B	67.8	61.8	54.9	45.3	30.1	79.3	61.3
DMMI (Hu et al., 2023)	Swin-T / BERT-B	70.5	64.6	57.0	47.0	31.0	79.9	63.3
ReMamber (Yang et al., 2024)	Swin-T / BERT-B	71.4	66.1	58.9	48.7	32.9	80.2	63.9
MagNet (Chng et al., 2024)	Swin-T / BERT-B	72.5	66.7	58.8	49.0	33.1	80.8	64.8
LGCE (Yuan et al., 2024)	Swin-T / BERT-B	69.8	64.0	56.6	46.9	31.2	79.6	62.3
RMSIN (Liu et al., 2024)	Swin-T / BERT-B	68.6	62.9	55.4	45.8	30.5	79.0	61.6
CCFormer (Lu et al., 2025)	Swin-T / BERT-B	80.1	74.9	66.6	54.8	36.8	80.9	71.0
DeCo-B (Ours)	BEiT3-B* / BEiT3-B*	81.6	76.0	67.2	55.3	36.8	81.8	71.6
DeCo-L (Ours)	BEiT3-L* / BEiT3-L*	**83.2**	**77.4**	**69.5**	**57.8**	**38.6**	**83.8**	**73.4**

Main results for RRSIS on RIS-LAD. Table 16 provides the comparison on the RIS-LAD test set. RIS-LAD is a challenging UAV-based benchmark with substantially longer textual expressions and complex small objects, which impose greater demands on fine-grained language understanding and precise spatial localization. Despite these challenges, DeCo-B achieves 51.1% P@.5, 50.3% oIoU, and 47.1% mIoU, outperforming the previous best method, SAARN (Ye et al., 2026), by 6.1%, 0.7%, and 5.4%, respectively. DeCo-L further improves these results to 56.1% P@.5, 54.8% oIoU, and 50.2% mIoU, while retaining its advantage under stricter IoU thresholds. Such substantial gains demonstrate the ability of DeCo to preserve semantic discrimination and spatial precision in challenging UAV scenarios with long referring expressions and small targets.

Main results for RRSIS on RefDIOR. Table 17 presents the segmentation results on the RefDIOR test set. As a unified multi-task remote sensing benchmark, RefDIOR provides a challenging setting for jointly evaluating localization and segmentation. DeCo-B achieves 81.6% P@.5, 81.8% oIoU, and 71.6% mIoU, outperforming CCFormer (Lu et al., 2025) by 1.5%, 0.9%, and 0.6%, respectively. DeCo-L further improves these results to 83.2% P@.5, 83.8% oIoU, and 73.4% mIoU, bringing impressive gains of 3.1%, 2.9%, and 2.4% over CCFormer, respectively. Together with the RSVG results on the same benchmark, all the improvements demonstrate that our decouple-to-couple paradigm is well suited to unified multi-task learning in remote sensing scenarios.

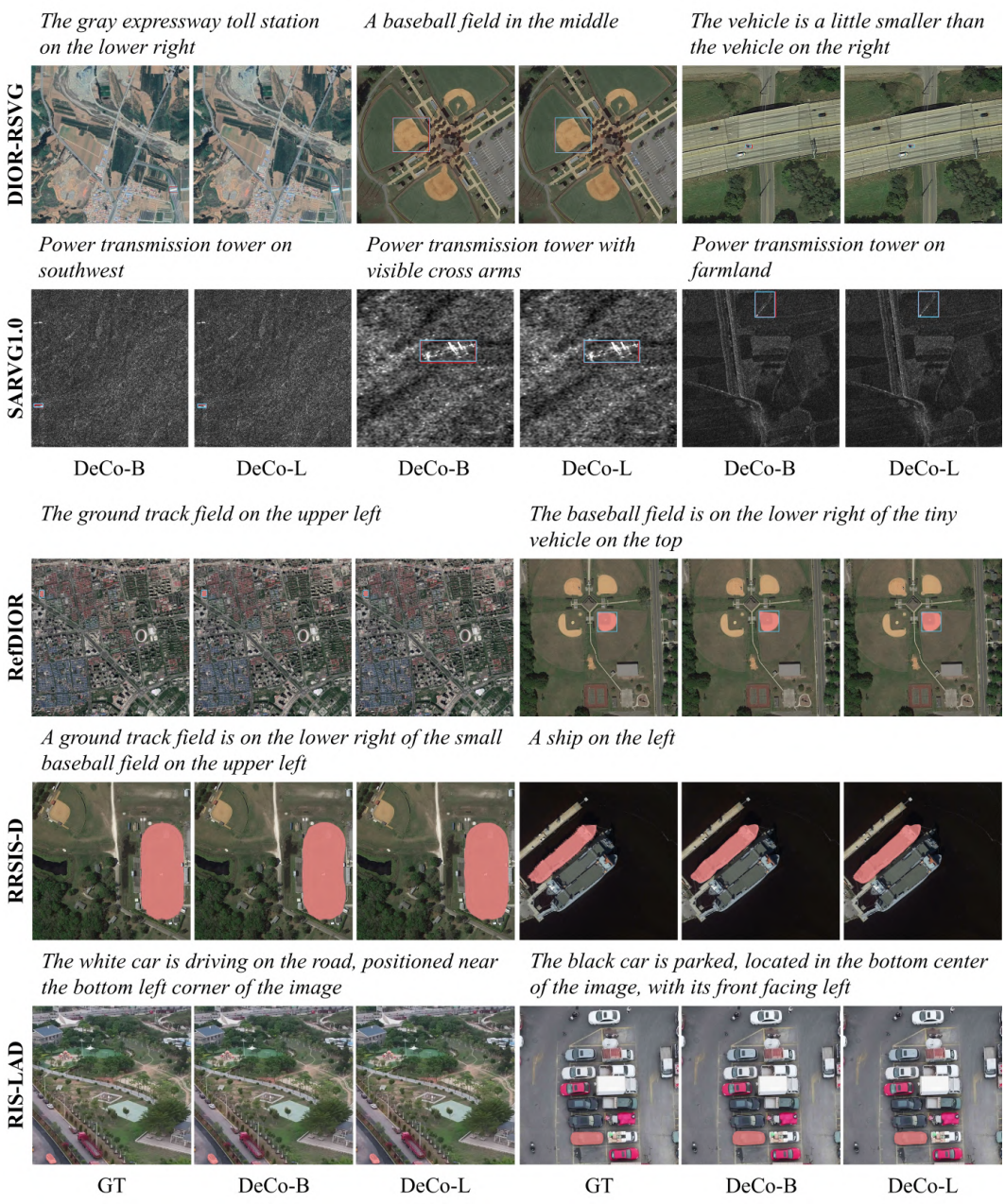

Figure 4: Qualitative results of our DeCo across different remote sensing image benchmarks.

Figure 4 shows qualitative results across diverse remote sensing benchmarks. DeCo reliably grounds the referred objects under optical, SAR, and UAV imagery, including small objects and relation-intensive expressions, while producing well-aligned localization boxes and segmentation masks. Overall, the above six quantitative results reveal that existing methods developed for natural images generally lag behind specialized remote sensing approaches, indicating a clear domain gap between the two scenarios. In contrast, DeCo consistently bridges this gap and achieves superior performance across various tasks and modalities. Moreover, existing remote sensing methods typically rely on full tuning, which introduces substantial optimization burden when processing high-resolution images. By adapting a frozen multimodal encoder with only lightweight trainable modules, DeCo demonstrates that parameter-efficient learning is also applicable to remote sensing while substantially reducing training costs. This property makes DeCo particularly attractive for high-resolution scenarios, where computational efficiency and visual understanding capability are both critical.

Table 18: Ablation on different feature decoupling and prior coupling. SWM: Salient Word-level Modulation.

Models \ test-u	Trainable Head		REC		RES		Avg.
	Params.	FLOPs	P@.5	P@.5:.95	oIoU	mIoU	
DeCo w/o TSD	9.3M	17.6G	86.5	70.5	72.2	73.3	75.6
w/ Dual SimFPN	12.1M	28.1G	86.8	71.3	72.8	74.2	76.3
w/ TSD (w/o SWM)	10.4M	22.0G	86.8	71.2	72.6	74.2	76.2
DeCo	11.6M	24.2G	**87.0**	**71.7**	**73.0**	**74.5**	**76.6**
DeCo w/o HPC	10.6M	24.2G	86.3	69.1	72.5	73.6	75.4
w/ HPC (Text)	11.6M	24.2G	86.5	70.2	72.7	74.2	75.9
w/ HPC (Mask)	10.6M	24.2G	86.8	71.1	72.8	74.4	76.3

Table 19: Ablation on the LoRA scaling factor α.

α \ test-u	REC		RES		Avg.
	P@.5	P@.5:.95	oIoU	mIoU	
$r = 32$, Trainable Backbone Param.: 4.7M, FLOPs: 1.9G					
8	86.7	70.8	72.6	73.9	76.0
16	86.5	71.4	**73.1**	74.2	76.3
32	**87.0**	**71.7**	73.0	**74.5**	**76.6**
64	86.3	71.4	72.6	74.0	76.1
128	85.5	70.2	71.8	73.3	75.2

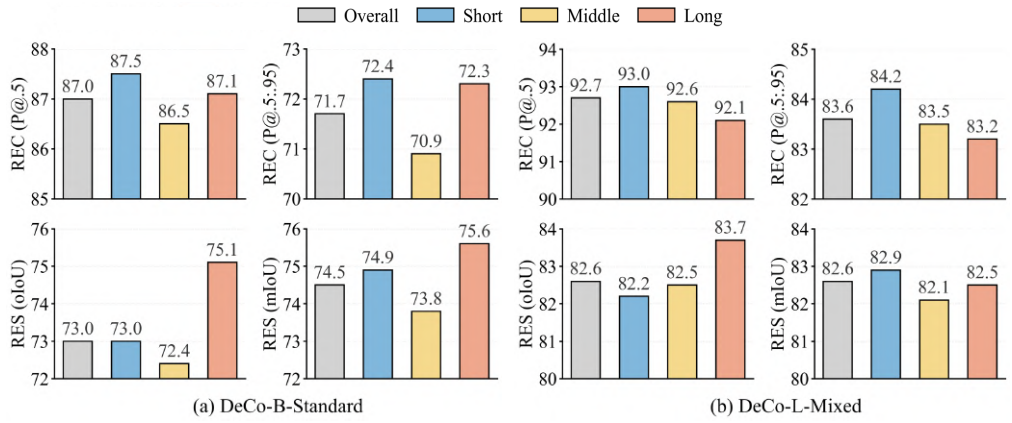

Figure 5: Performance across different expression lengths on G-Ref (test-u).

D More Ablation Studies and Qualitative Results

Effects of Task Decoupling and Prior Coupling. We further investigate the design choices of TSD and HPC in Table 18. Removing TSD reduces the average performance to 75.6%. A straightforward alternative is to duplicate SimFPN for the localization and segmentation branches, which improves the average score to 76.3%, confirming the benefit of task-specific representations, but at the cost of substantially higher parameters and FLOPs. In comparison, TSD achieves the best average score of 76.6% with lower computational burden. Removing SWM from TSD decreases the score to 76.2%, further validating the importance of word conditioned feature decoupling. For prior coupling, removing HPC decreases the average score from 76.6% to 75.4%. Introducing only the textual prior improves it to 75.9%, whereas using only the mask-derived spatial prior yields a larger gain to 76.3%. In particular, the mask prior improves P@.5:.95 from 69.1% to 71.1%, highlighting the value of dense segmentation cues for precise localization. Combining both priors achieves the best overall performance of 76.6%, demonstrating that sentence-level semantics and mask-derived geometry provide complementary information for cross-task coupling.

Effects of the LoRA Scaling Factor α. We then study the effect of the scaling factor α while fixing the rank to $r = 32$. As shown in Table 19, the overall performance gradually improves as α increases from 8 to 32, reaching the best average score of 76.6% at $\alpha = 32$. A smaller α may under-adapt the frozen multimodal encoder, whereas an excessively large value can amplify the low-rank residual and disturb the pretrained representations. This is also reflected by the performance drop to 75.2% when $\alpha = 128$. These results suggest that a moderate adaptation strength is crucial for preserving pretrained knowledge while sufficiently adjusting multimodal representations to the downstream grounding objectives. We therefore set $\alpha = 32$ by default.

Grounding across Expression Lengths. Since longer linguistic expressions typically contain richer attributes and relational cues, we examine the robustness of DeCo to expression length. According to the number of words, the 9,602 samples in G-Ref (test-u) are divided into *short* ($L \leq 6$), *middle* ($6 < L \leq 12$), and *long* ($L > 12$) groups, containing 3,415, 4,858, and 1,329 expressions, respectively. As shown in Figure 5, we first evaluate the default ablation configuration, DeCo-B under the standard setting (DeCo-B-Standard), which remains stable across different lengths, with no consistent degradation as expressions become longer. Notably, the long group even achieves the highest



Figure 6: Grounding results of DeCo with BEiT3-L across eight splits under the mixed setting.

RES scores of 75.1% oIoU and 75.6% mIoU. We further examine our strongest model, DeCo-L under the mixed setting (DeCo-L-Mixed), and observe the same trend. Its REC P@.5 changes only from 93.0% for short expressions to 92.1% for long ones, while P@.5:.95 varies from 84.2% to 83.2%. Meanwhile, RES oIoU increases from 82.2% to 83.7%. These results indicate that DeCo



Figure 7: Grad-CAM visualization of task-specific feature preferences and decoupling. LoM ($\mathbf{V}_l^{LoM}$) and SoM ($\mathbf{V}_l^{SoM}$) exhibit distinct localization- and segmentation-oriented responses, while the shared representations of DoM ($\mathbf{V}_l^{DoM}$) and DeCo ($\mathbf{V}_l$) mix these preferences. TSD effectively decouples them into $\mathbf{V}_l^{loc}$ and $\mathbf{V}_l^{seg}$. The salient word selected for modulation is highlighted in red.

is largely insensitive to expression length and preserves both localization and segmentation quality under increasingly complex descriptions. This robustness is further illustrated by the representative examples in Figure 6. For each length group, we select examples from all eight splits where both REC and RES achieve high IoU. Specifically, we use IoU thresholds of 0.98, 0.95, and 0.90 for short, middle, and long expressions, respectively, to account for the increasing difficulty of processing longer expressions. DeCo accurately identifies the referred targets across different expression lengths, highlighting its consistent performance regardless of expression length.

Visualization of Task-specific Feature Preferences and Decoupling. To further confirm the representation conflict between REC and RES, we visualize the third-level ($l = 3$) feature responses of different models using Grad-CAM (Selvaraju et al., 2017). As illustrated in Figure 7, we first compare the task-oriented models and observe different activation patterns: LoM tends to focus on sparse yet discriminative regions that facilitate object localization, whereas SoM distributes its attention more broadly over the target extent to support dense segmentation. This observation provides qualitative evidence that REC and RES favor distinct visual cues. When both objectives share the same representation, as in DoM, the resulting activations tend to either resemble those of one dominant task or exhibit a mixture of the two task preferences. A similar behavior is observed in the shared feature $\mathbf{V}_l$ of DeCo before TSD, demonstrating the limitation of directly coupling different objectives within a common representation. In contrast, after TSD, the decoupled localization-oriented feature $\mathbf{V}_l^{loc}$ recovers the sparse discriminative responses preferred by REC, while the segmentation-oriented feature $\mathbf{V}_l^{seg}$ highlights more complete target regions for RES. Notably, the salient words highlighted in red cover diverse discriminative semantics, including object categories, actions, and descriptive modifiers, rather than being restricted to class attributes. This indicates that the salient-

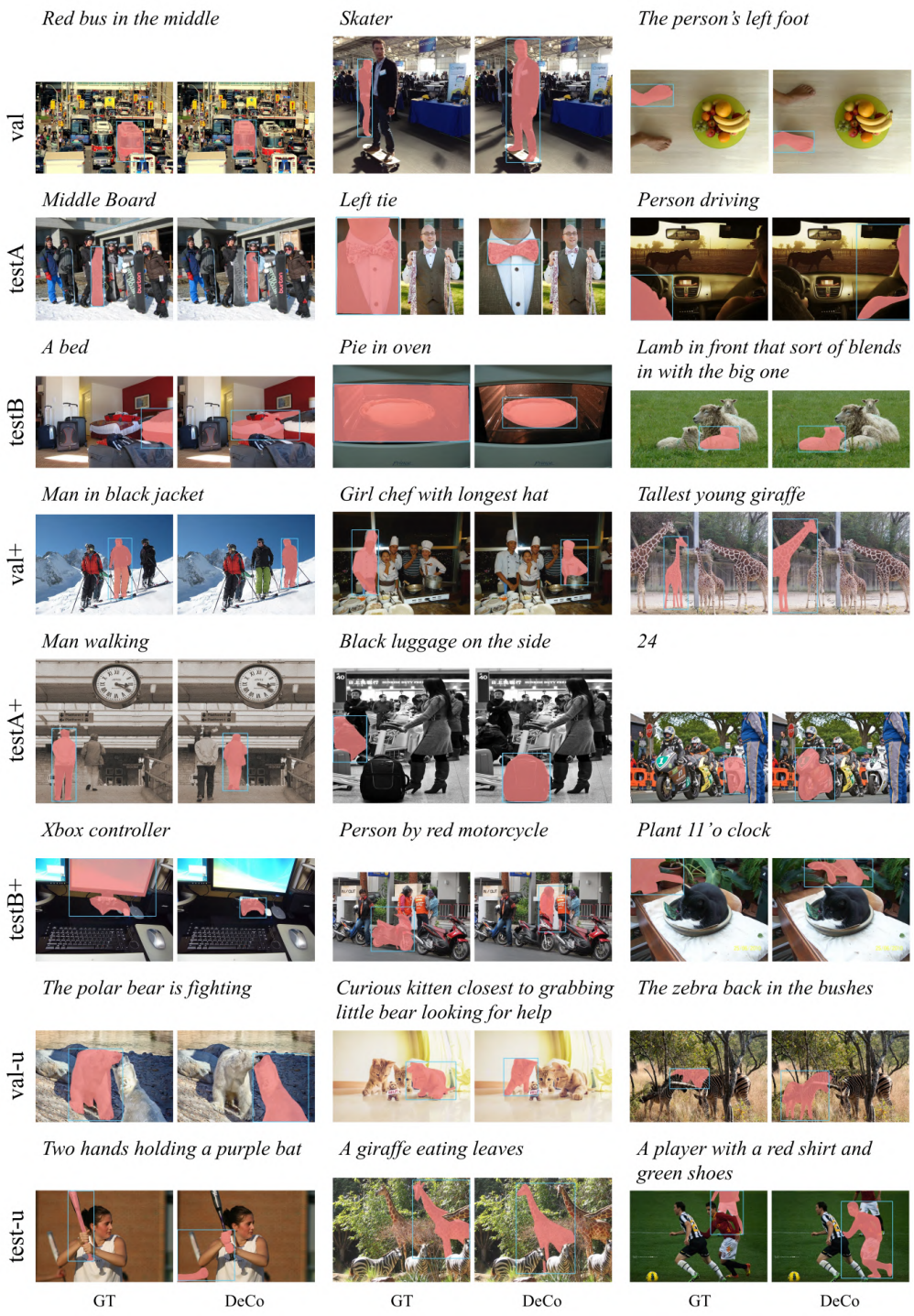

Figure 8: Failure cases of DeCo with BEiT3-L across eight splits under the mixed setting.

word selection can adaptively identify informative linguistic cues for task-specific feature modulation. These visualizations further support our motivation for decoupling shared representations according to the distinct visual interests of REC and RES before cross-task coupling.

Failure Cases. Figure 8 shows representative failure cases of DeCo with BEiT3-L across eight splits under the mixed setting. We identify a sample as a failure case when its IoU is below 0.3 for both REC and RES. Our analysis reveals three major sources of failure: (i) *Annotation ambiguity* occurs when multiple visual instances can match the same referring expression, making the single-

target annotation inherently non-unique; (ii) *Annotation errors* arise when the annotated target does not match the textual description, indicating noise in the ground truth rather than a genuine model failure; (iii) *Model limitations* account for cases where the prediction requires complex logical relations or relies on subtle and fine-grained visual cues. These observations suggest that some apparent failures originate from annotation ambiguity and noise, while the remaining examples highlight the need for stronger relational reasoning and fine-grained visual understanding.

Open-World Generalization. Despite the impressive in-domain performance of DeCo, real-world application requires robust generalization beyond the training distribution. We therefore examine its open-world cognition on unseen images from diverse scenarios, including *optical*, *SAR*, *indoor*, *anime*, *pixel art*, *comics*, *computer graphics*, *painting*, *sketch*, and *artwork*. Free-form referring expressions involving *spatial position*, *visual attributes*, *inter-object relations*, and *logical reasoning* are used to assess fine-grained vision-language alignment under substantial distribution shifts. As shown in Figure 9, DeCo generalizes well across these diverse visual domains. It can distinguish similar instances by fine-grained attributes, localize small targets, and resolve relational descriptions, while remaining effective under strong style shifts from anime and pixel art to sketch and classical artwork. These results indicate that the learned vision-language correspondence transfers beyond the benchmark distributions and retains semantic discrimination and relational reasoning in the wild.

E Limitations and Future Works

Despite its strong performance and generalization, DeCo still has several limitations that motivate further investigation. First, the current formulation assumes single-target referring expressions. The localization branch employs a single object query, while the mask-to-box operation derives only one enclosing box through global spatial extrema, making DeCo less suitable for expressions referring to multiple objects or spatially disconnected regions. A natural extension is to introduce multiple object queries together with instance-aware mask decomposition, where each instance mask is independently converted into a spatial prior. This would enable multiple referred objects to be separately represented, localized, and coupled across REC and RES. Second, although HPC adaptively integrates semantic and mask-derived spatial priors, the quality of the latter still depends on the intermediate segmentation prediction. In challenging cases with severe occlusion, tiny objects, or ambiguous expressions, inaccurate masks may provide imperfect geometric cues. Exploring uncertainty-aware prior estimation or more flexible bidirectional interaction between localization and segmentation could further improve the robustness of cross-task coupling. Finally, parameter-efficient adaptation substantially reduces the number of trainable parameters, but inference still requires forwarding through the frozen multimodal encoder, which may remain computationally demanding for very high-resolution imagery. Future work could investigate token pruning, dynamic-resolution processing, or lightweight multimodal encoders to further reduce inference cost.

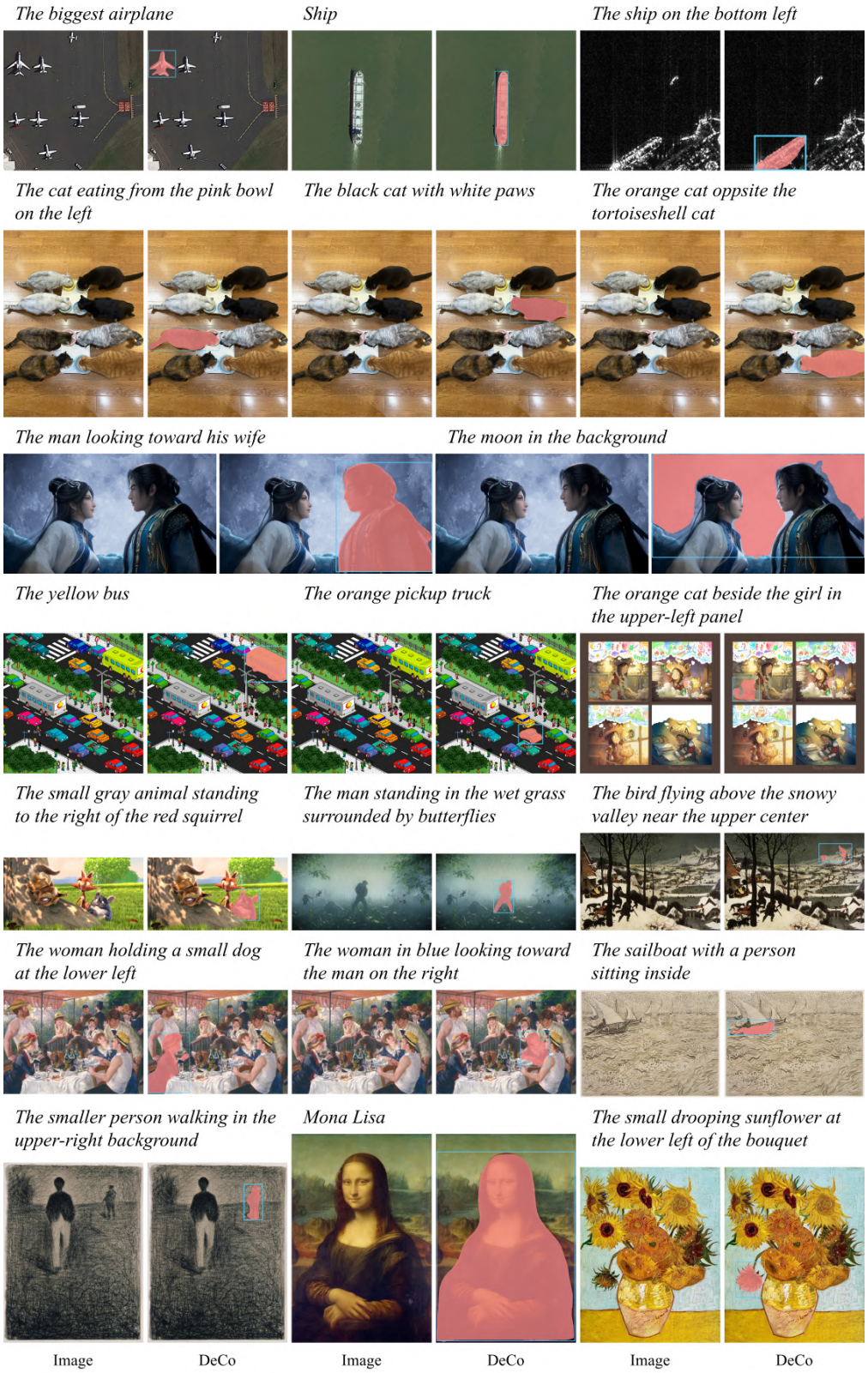

Figure 9: Open-world qualitative results of DeCo with BEiT3-L trained under the mixed setting.