CIVG: CAUSAL INFERENCE FOR ACCURATE AND ROBUST GENERALIZED VISUAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalized visual grounding, encompassing generalized referring expression comprehension and generalized referring expression segmentation, extends classical visual grounding to multi-target and no-target scenarios. However, existing methods suffer from a critical limitation: they tend to learn superficial co-occurrence patterns induced by dataset biases, and they often fail to reliably reject no-target queries due to spurious correlations. Causal inference offers a principled solution by disentangling dataset-induced biases and enabling counterfactual reasoning about target existence. To this end, we propose CIVG, a novel framework that explicitly integrates causal inference into generalized visual grounding. CIVG introduces a Causality-Aware Deconfounding Module (CADM) that performs front-door adjustment to mitigate confounding biases from both visual and textual modalities simultaneously. We further design a Counterfactual Relevance Regularization (CFR) branch that leverages counterfactual contrastive learning to enhance the model's ability to reject no-target queries. In addition, we present CIVG-Union, a dual-branch prediction strategy that harnesses the complementary strengths of models at different scales. Extensive experiments on two tasks and four benchmarks demonstrate that CIVG achieves state-of-the-art performance, with substantial gains in both accuracy and robustness, particularly in no-target scenarios. Our results validate the effectiveness of integrating causal inference into generalized visual grounding.

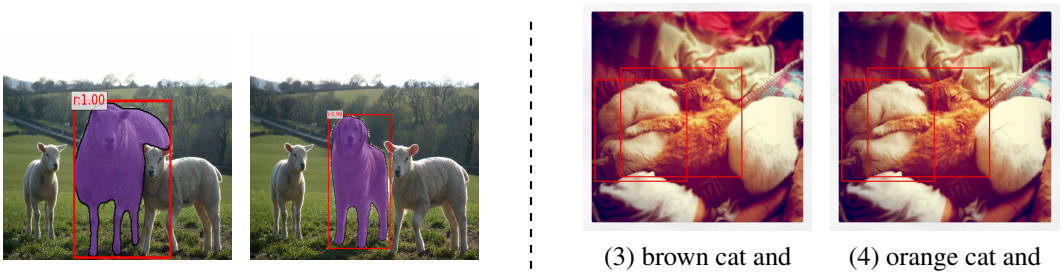
1 INTRODUCTION

Visual grounding aims to localize the specific region(s) within an image based on a given natural language expression (Xiao et al., 2026). Classic tasks include Referring Expression Comprehension (REC) (Wang et al., 2019; Yang et al., 2020) and Referring Expression Segmentation (RES) (Shang et al., 2024; Dai et al., 2024), both of which assume a one-to-one correspondence between each expression and a single target object. However, this assumption does not cover many real-world scenarios, where expressions may refer to multiple objects or even have no valid referent. Models trained under this paradigm tend to produce arbitrary outputs for no-target queries, severely undermining their reliability.

To overcome the one-to-one assumption, generalized visual grounding has been introduced, including Generalized Referring Expression Comprehension (GREC) (He et al., 2023) and Generalized Referring Expression Segmentation (GRES) (Liu et al., 2023a), which support expressions referring to an arbitrary number of objects. The gRefCOCO dataset (Ding et al., 2026) provides the first large-scale resource with multi-target, no-target, and single-target annotations. These developments mark a significant step toward making visual grounding more reflective of real-world demands.

Methodologically, visual grounding has evolved from two-stage proposal matching (Yu et al., 2018; Hong et al., 2022) to one-stage (Yang et al., 2020; 2019) and Transformer-based end-to-end frameworks (Kamath et al., 2021; Deng et al., 2021; Zhu et al., 2022), and more recently to vision-language pre-training (Radford et al., 2021; Wang et al., 2023; Huang et al., 2026b) and multimodal large language models (Huang et al., 2026a; Yuan et al., 2026; Wu et al., 2026). Despite these advances, a critical challenge persists: datasets inherently contain biases, such as stylistic variations in visual appearance and linguistic phrasing (Chen et al., 2025; Qiao et al., 2020). These biases can lead models to overfit to dataset-specific cues—for instance, correlating background styles or fre-

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.



(1) biggest sheep (2) biggest sheep

(a) Background Bias in Images

(3) brown cat and left cat (4) orange cat and left cat

(b) Description Bias in Text

Figure 1: Illustration of Confounding Bias Within Each Modality. (a) illustrates background bias in images, where target objects frequently co-occur with others of the same category, causing InstanceVG (Dai et al., 2026b) to incorrectly identify a dog as the "biggest sheep" in (2). (b) depicts bias in textual descriptions, where the central cat is inconsistently described as either brown or orange, introducing linguistic bias.

quent wordings with particular target objects—rather than learning genuine semantic associations. This spurious reliance ultimately compromises the robustness and generalization ability of visual grounding systems. Figure 1 illustrates the biases present in each modality.

To address the confounding bias, we resort to causal inference, which explicitly models the underlying data-generating process to disentangle spurious correlations and enable counterfactual reasoning, allowing models to focus on genuine semantic causality and assess whether an expression truly refers to an object in the image. Causal inference (Pearl & Mackenzie, 2018) has shown promise across machine learning tasks, including image recognition (Chen et al., 2023), visual question answering (Liu et al., 2023c), and video temporal grounding (Wang et al., 2026a). These methods typically construct a structural causal model (Pearl, 2009) to represent the relationships among variables, and apply causal tools such as the *do-operator* to estimate causal effects beyond observational correlations. In addition, techniques like front-door and back-door adjustments are used to control for latent confounders—either by leveraging mediating variables that transmit causal influence, or by blocking spurious dependencies introduced by common causes. These mechanisms enable the model to better capture true causal relationships, rather than being misled by stylistic artifacts or co-occurrence patterns. Motivated by these successful applications, we investigate how causal principles can be leveraged in generalized visual grounding, particularly under visual and linguistic biases as well as no-target scenarios, where conventional methods struggle.

In this paper, we first construct a structural causal model for visual grounding that explicitly incorporates confounders from both image and text modalities. Based on this model, we propose a novel framework for generalized visual grounding that jointly considers these confounding factors. By simultaneously mitigating visual and linguistic biases, we aim to achieve more accurate, robust, and generalizable grounding performance in open-world scenarios where expressions may refer to any number of objects—or none at all.

The main contributions of this work are as follows:

- We revisit generalized visual grounding from a causal inference perspective and construct a structural causal model that incorporates both visual and textual confounders, thereby providing a theoretical foundation for causal inference in this task.
- We propose CIVG, a unified framework that jointly handles GREC and GRES. Its core is the Causality-Aware Deconfounding Module(CADM), which applies front-door adjustment to mitigate confounding bias in the representations of images and text separately. In addition, we employ counterfactual contrastive learning to enhance the model's ability to identify no-target scenarios, and adopt a dual-branch architecture to further improve model performance.
- Extensive experiments on four widely used benchmark datasets demonstrate that CIVG achieves state-of-the-art performance and exhibits superior robustness in discriminating no-target scenarios, validating the effectiveness of our causal modeling approach.

2 RELATED WORK

2.1 VISUAL GROUNDING

Early REC methods followed a two-stage pipeline (Yu et al., 2018; Hong et al., 2022): first generating object proposals using off-the-shelf detectors then ranking these proposals against the language query via cross-modal matching. Later, one-stage approaches unified detection and reasoning within a single framework, achieving real-time performance by fusing language features directly into the visual feature map (Yang et al., 2020; 2019). More recently, Transformer-based architectures have become dominant (Dai et al., 2024), enabling end-to-end cross-modal modeling through self-attention mechanisms. Concurrently, RES has evolved from CNN-based cross-modal fusion methods (Huang et al., 2020) to Transformer-based approaches (Yang et al., 2022; Ding et al., 2021) that leverage global contextual interactions. A notable trend is the joint training of REC and RES within a multi-task learning framework. Representative works (Luo et al., 2020; Xiao et al., 2024) show that joint training yields consistent performance gains on both REC and RES benchmarks.

To overcome the one-to-one assumption of classic visual grounding, generalized visual grounding has been introduced, encompassing GREC and GRES. Unlike classic REC/RES, GREC and GRES support expressions that refer to an arbitrary number of targets, including multi-target and no-target cases. The gRefCOCO dataset (Ding et al., 2026) provides the first large-scale benchmark for these generalized tasks. DMMI (Hu et al., 2023) proposed a new benchmark for beyond-single-target segmentation, while RefSegformer (Wu et al., 2024) enhanced transformer-based models with no-target discrimination. In addition, Dai et al. (2026b); Nguyen et al. (2026) and related works handle generalized visual grounding tasks through instance-aware capabilities.

However, most existing methods rely on cross-modal correlations, which can be compromised by dataset biases, undermining model generalization and robustness. For instance, RefCOCOg dataset exhibits insensitivity to linguistic structure and statistical biases (Chen et al., 2025; Qiao et al., 2020), while gRefCOCO suffers from a long-tailed distribution of no-target samples (Dai et al., 2025a).

2.2 CAUSAL INFERENCE FOR VISUAL LEARNING

Causal inference has recently gained increasing attention in vision-language research as a principled approach to mitigating spurious correlations and improving model robustness. The core idea is to model the underlying data-generation process via structural causal models and to perform interventions—such as back-door or front-door adjustment—to isolate true causal effects from confounding biases (Jiao et al., 2024). Causal inference has been applied to a variety of vision-language tasks, including visual question answering (Liu et al., 2023c), weakly-supervised video object grounding (Wang et al., 2022a), and video temporal grounding (Wang et al., 2026a). In the context of 2D visual grounding, Huang et al. (2022) identified confounding bias between language and location and addressed text-side confounders, while leaving visual confounders largely unaddressed. Conversely, in remote-sensing visual grounding, Wang et al. (2026b) explored image-side confounders but did not jointly account for textual biases.

Despite these advances, two limitations persist. Existing methods address image-side or text-side confounders in isolation, and causal intervention remains confined to classic REC/RES under the one-to-one assumption, leaving generalized visual grounding largely unexplored. This gap motivates our approach, which extends causal deconfounding to generalized visual grounding by jointly modeling multimodal confounding factors.

3 STRUCTURAL CAUSAL MODEL OF VISUAL GROUNDING

To formally characterize the confounding effect inherent in visual grounding, we construct a structural causal model (SCM) (Pearl & Mackenzie, 2018) that explicitly distinguishes the roles of visual and linguistic inputs. As illustrated in Figure 2, we define the input image as I, the referring text as T, and the final localization result as Y, with the full input set denoted as $X = \{I, T\}$. In this directed acyclic graph, directed edges indicate causal influences, where the joint input X causes the prediction Y.

Most existing grounding methods rely on the observational likelihood $P(Y \mid X)$, which implicitly captures spurious associations arising from back-door paths $X \leftarrow Z \rightarrow Y$. The confounders Z in our task are bimodal: on the visual side, they encompass global image styles, background contexts, and co-occurrence regularities between objects and their surroundings; on the linguistic side, they include stylistic traits such as wording preferences, syntactic variations, and idiosyncratic phrasing habits. These factors jointly shape the observed data distribution and can induce misleading correlations between seemingly predictive features and the ground-truth target, thereby undermining the model's robustness and cross-domain adaptability.

Because the true confounders are often high-dimensional and lack explicit semantic definitions, we introduce an intermediate variable M to serve as a front-door proxy. This constructs an explicit causal pathway $X \rightarrow M \rightarrow Y$, enabling tractable causal intervention even when the confounders remain unobserved. The implementation details of the front-door adjustment are elaborated in the subsequent sections.

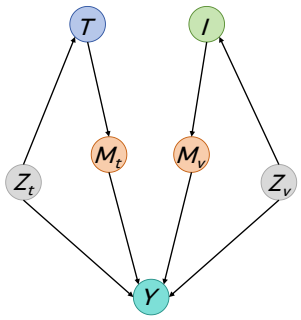


Figure 2: Illustration of the structural causal model of Visual Grounding. I, T, and Y denote the image, text, and prediction. M_v and M_t denote the visual and textual mediators. Z_v and Z_t denote the visual and textual confounders.

4 METHODOLOGY

In this section, we introduce the CIVG model, whose overall architecture is illustrated in Figure 3. The multimodal encoder extracts visual and textual features $\mathbf{F}_v$ and $\mathbf{F}_t$. The model is then divided into three key components. First, Causality-Aware Deconfounding Module (CADM) performs front-door adjustment to mitigate confounding bias, yielding causality-aware feature representations $\mathbf{F}'_v$ and $\mathbf{F}'_t$. These features are subsequently passed to the instance-aware branch and the Counterfactual Relevance Regularization(CFR) branch. The counterfactual branch constructs counterfactuals to suppress spurious correlations. The instance-aware branch jointly performs global segmentation, instance segmentation, and detection, inheriting the design of InstanceVG (Dai et al., 2026b). For the multimodal encoder, we adopt BEiT-3 (Wang et al., 2023) and instantiate three variants with different scales: CIVG-B, CIVG-L, and CIVG-Union.

4.1 CAUSALITY-AWARE DECONFOUNDING MODULE

As analyzed in Sec. 3, observational learning $P(Y \mid X)$ is vulnerable to back-door confounding via $X \leftarrow Z \rightarrow Y$. To address this, we introduce a front-door mediator M and estimate the causal effect $X \rightarrow M \rightarrow Y$ by front-door adjustment.

Front-door Causal Formulation. According to the front-door criterion, the interventional distribution is

$$\begin{aligned} P(Y \mid do(X=x)) &= \sum_m P(Y \mid do(M=m))\, P(M=m \mid do(X=x)) \\ &= \sum_m P(M=m \mid X=x) \sum_{x'} P(Y \mid X=x', M=m) P(X=x') \\ &= \mathbb{E}_{x' \sim P(X)}\, \mathbb{E}_{m \sim P(M \mid X=x)} \left[P(Y \mid X=x', M=m) \right], \end{aligned} \tag{1}$$

where x' denotes samples from the empirical training distribution and m is the mediator derived from the current input x. Since computing expectations over the entire training distribution is infeasible, we approximate them using modality-specific prototype memory banks and attention-based aggregation.

To approximate the front-door adjustment, following prior works (Wang et al., 2024; 2025; 2026a), we adopt a linear mapping model and define a causal feature function as a linear combination of sampled inputs and mediators:

$$F(x', m) = \mathbb{E}_{x' \sim P(X)}\, \mathbb{E}_{m \sim P(M \mid X=x)}\, f(x', m), \quad f(x', m) = h(x') + g(m), \tag{2}$$

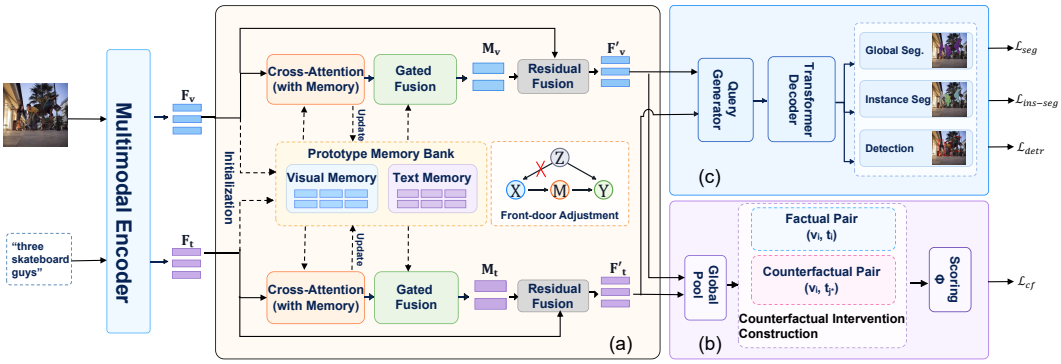


Figure 3: Overview of the CIVG framework. (a) The Causality-Aware Deconfounding Module, which performs front-door adjustment to mitigate confounding bias. (b) The Counterfactual Relevance Regularization branch, which constructs counterfactuals to suppress spurious correlations. (c) The instance-aware branch, which performs joint multi-task perception.

which yields the separable form:

$$F(x', m) = \mathbb{E}_{x' \sim P(X)}[h(x')] + \mathbb{E}_{m \sim P(M|X=x)}[g(m)]. \tag{3}$$

As it is intractable to get a closed-form solution of expectations involving the complex representation space, the estimation is achieved by the query mechanism. Let x'_i and m_j be sampled input and mediator representations from the input and mediator distributions, respectively. The expectations are computed as:

$$\mathbb{E}_{x' \sim P(X)}[h(x')] = \sum_i \alpha_i h(x'_i), \quad \alpha_i = \frac{\exp(p^\top x'_i)}{\sum_k \exp(p^\top x'_k)}, \tag{4}$$

$$\mathbb{E}_{m \sim P(M|X=x)}[g(m)] = \sum_j \beta_j g(m_j), \quad \beta_j = \frac{\exp(q^\top m_j)}{\sum_k \exp(q^\top m_k)}, \tag{5}$$

where $p = p(x)$ and $q = q(x)$ are attention queries derived from the current input to guide the aggregation of sampled features and mediators. This is efficiently implemented via multi-head attention, enabling seamless integration of causal adjustments into existing transformer-based frameworks with minimal modifications

Mediator Construction and Prototype Memory. Prototype memories are initialized from the training set by applying K-means clustering to globally pooled visual and textual encoder features. The resulting cluster centers are projected into the hidden dimension and stored as modality-specific memory banks. During training, these memories are updated via cosine-similarity-based soft assignment and exponential moving average, while remaining non-trainable buffers. They thus serve as a compact empirical approximation of the training distribution $P(X)$, and their attended representations approximate the expectations in equation 4. For each modality, the mediator is constructed from the current input features and the corresponding prototype memory: cross-attention between local features and the prototype memory yields attended features, which serve as samples for the expectation over $P(M \mid X = x)$ in equation 5. These attended features are then fused with the original local features via gated feature fusion, producing the final deconfounded mediator representations M_v and M_t.

Residual Fusion. To preserve task-relevant information while mitigating confounding biases, we fuse the original features with the causally disentangled representations via learnable gates. The gating parameters are computed from the global representations:

$$\lambda_v = \sigma\left(\text{MLP}_v\left(\text{Pool}(\mathbf{F}_v)\right)\right), \quad \lambda_t = \sigma\left(\text{MLP}_t\left(\text{Pool}(\mathbf{F}_t)\right)\right), \tag{6}$$

$$\begin{aligned} \mathbf{F}'_v &= \lambda_v \cdot \mathbf{F}_v + (1 - \lambda_v) \cdot M_v, \\ \mathbf{F}'_t &= \lambda_t \cdot \mathbf{F}_t + (1 - \lambda_t) \cdot M_t. \end{aligned} \tag{7}$$

Where σ is the sigmoid function. This residual design retains beneficial non-confounded cues while suppressing spurious correlations. Through this modality-specific front-door adjustment, we obtain causally disentangled representations for both visual and language inputs, enabling robust grounding.

4.2 Counterfactual Relevance Regularization

Although the multimodal encoder can align visual and linguistic features, the model may still overfit spurious correlations in the training data—for example, incorrectly associating background textures or object co-occurrence patterns with the target expression. To enhance the model's causal sensitivity to the input expression, we introduce a lightweight counterfactual relevance branch inspired by counterfactual reasoning.

Problem Formulation. Given a batch of N image–text pairs $\{(I_i, T_i)\}_{i=1}^{N}$, let $\mathbf{v}_i \in \mathbb{R}^{D_v}$ and $\mathbf{t}_i \in \mathbb{R}^{D_t}$ be the global pooled visual and textual features after causal deconfounding. The branch judges whether a given pair is factually aligned.

Relevance Scoring Function. We map the joint representation of a factual pair to a relevance score via a binary classifier:

$$r(\mathbf{v}_i, \mathbf{t}_i) = \Phi([\mathbf{v}_i; \mathbf{t}_i]) \in \mathbb{R}, \tag{8}$$

where $[\cdot;\cdot]$ denotes concatenation, and Φ is the factual pair scoring function. A higher score indicates higher image–expression matching confidence.

Counterfactual Intervention Construction. To simulate "how the model would respond if the current image were paired with an incorrect expression," we intervene on text features across samples. For the i-th sample, its counterfactual pair is $(\mathbf{v}_i, \mathbf{t}_j)$ with $j \neq i$. To obtain high-quality negative samples, we adopt hard-negative mining:

$$j^* = \begin{cases} \arg\max_{j \in \mathcal{C}_i} \mathrm{sim}(\mathbf{v}_i, \mathbf{t}_j), & \mathcal{C}_i \neq \emptyset \\ (i \bmod N) + 1, & \mathcal{C}_i = \emptyset, \end{cases} \tag{9}$$

where $\mathcal{C}_i = \{j \neq i \mid \mathrm{ID}(i) \neq \mathrm{ID}(j)\}$, sim is cosine similarity, and $\mathrm{ID}(\cdot)$ denotes the image source identifier to avoid pairing different augmentations of the same original image. For N=1, the counterfactual loss is not computed. This design retains the same visual content while altering only the linguistic condition, forcing the model to focus on causal effects of the language description.

Training Objective. The counterfactual regularization loss $\mathcal{L}_{\text{cf}}$ consists of two terms that jointly enlarge the gap between factual and counterfactual scores: (1) Binary cross-entropy term, where factual pairs are treated as positive (label 1) and counterfactual pairs as negative (label 0).(2) Ranking margin term that enforces the factual score to exceed the counterfactual score by at least a margin.

$$\mathcal{L}_{\text{bce}} = -\frac{1}{2N} \sum_{i=1}^{N} \left[\log \sigma(r_i) + \log \left(1 - \sigma(r_{j^*})\right) \right], \tag{10}$$

$$\mathcal{L}_{\text{rank}} = \frac{1}{N} \sum_{i=1}^{N} \max\left(0,\ \gamma - \sigma(r_i) + \sigma(r_{j^*})\right), \tag{11}$$

where $r_i = r(\mathbf{v}_i, \mathbf{t}_i)$, $r_{j^*} = r(\mathbf{v}_i, \mathbf{t}_{j^*})$, $\sigma(\cdot)$ is the sigmoid function and $\gamma = 0.2$. The use of sigmoid probabilities improves numerical stability and consistency with the BCE term. The final counterfactual loss is

$$\mathcal{L}_{\text{cf}} = \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{rank}}, \tag{12}$$

4.3 Dual-Branch Prediction Framework

Through empirical analysis of model scaling, we observe that different model capacities exhibit complementary strengths. The smaller variant, CIVG-B, excels at target existence discrimination, i.e., judging whether the referred object actually exists in the input image. In contrast, the larger

variant, CIVG-L, achieves higher precision in predicting the spatial extent of the target, particularly in generating accurate instance masks and bounding boxes.

Motivated by this observation, we propose CIVG-UNION, a dual-branch visual grounding architecture that harnesses the complementary capabilities of both variants through a two-stage inference strategy. In the first stage, CIVG-B directly determines whether the referred target is present in the image. If it predicts absence, the model outputs an empty prediction, thereby suppressing false positives for negative samples. Otherwise, in the second stage, CIVG-L generates precise bounding boxes and instance masks as the final grounding result.

Formally, let $f_{\mathrm{B}}(I,T) \in \{0,1\}$ denote the existence prediction of CIVG-B, and $g_{\mathrm{L}}(I,T)$ denote the prediction function of CIVG-L. The final output $\mathcal{Y}$ of CIVG-UNION is given by

$$\mathcal{Y} = \begin{cases} g_{\mathrm{L}}(I,T), & \text{if } f_{\mathrm{B}}(I,T) = 1, \\ \varnothing, & \text{otherwise.} \end{cases} \tag{13}$$

This design ensures that CIVG-L, optimized for precise localization, is only invoked when the target is predicted to exist, thereby improving both precision and efficiency for negative samples.

4.4 TRAINING OBJECTIVES

The overall training objective comprises four components: detection, global semantic segmentation, instance-level segmentation, target existence discrimination, and counterfactual relevance regularization. For detection, following DETR (Carion et al., 2020), the loss $\mathcal{L}_{\text{detr}}$ combines L1, cross-entropy, and GIoU losses for bounding box supervision. For global semantic segmentation, we adopt binary cross-entropy (BCE) and Dice loss (Cheng et al., 2022) between the ground-truth global mask $\mathbf{M}_{\text{gt}}$ and the predicted global mask $\mathbf{S}_{\text{global}}$, as in prior work Liu et al. (2023b), yielding $\mathcal{L}_{\text{seg}}$. For instance-level segmentation, the loss $\mathcal{L}_{\text{ins-seg}}$ also employs BCE and Dice losses, where Hungarian matching establishes one-to-one assignment between semantic queries and instance targets (Dai et al., 2026b). Positive pairs are supervised directly, while negative pairs are penalized with a balancing weight:

$$\mathcal{L}_{\text{ins-seg}} = \frac{1}{N_{\text{pos}}} \sum_{i} \mathcal{L}_{\text{seg}}^{\text{pos}(i)} + \frac{\lambda_{\text{neg}}}{N_{\text{neg}}} \sum_{j} \mathcal{L}_{\text{seg}}^{\text{neg}(j)}, \tag{14}$$

where N_{pos}, N_{neg} are the numbers of positive and negative masks, and $\lambda_{\text{neg}} = 0.2$ by default. Finally, the counterfactual relevance loss $\mathcal{L}_{\text{cf}}$, as defined in Sec. 4.2, encourages factual pairs to score higher than counterfactual interventions.

The total training loss is a weighted sum of all four terms:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{detr}}\mathcal{L}_{\text{detr}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{instance}}\mathcal{L}_{\text{ins-seg}} + \lambda_{\text{cf}}\mathcal{L}_{\text{cf}}, \tag{15}$$

with hyper-parameters $\lambda_{\text{detr}} = 0.1$, $\lambda_{\text{seg}} = 1.0$, $\lambda_{\text{instance}} = 1.0$, and $\lambda_{\text{cf}} = 0.1$.

5 EXPERIMENTS

5.1 DATASETS & EVALUATION METRICS

For the generalized referring expression segmentation (GRES) task, we adopt the gRefCOCO (Liu et al., 2023a), Ref-ZOM (Hu et al., 2023), and R-RefCOCO/+/g datasets (Wu et al., 2024). For the generalized referring expression comprehension (GREC) task, we employ the gRefCOCO dataset (Ding et al., 2026) for both training and testing. Due to space limitations, we report the main results on gRefCOCO in the main paper, while the complete results on all datasets are provided in Appendix A.

Following prior works (Dai et al., 2026b; Luo et al., 2025), we use the following evaluation metrics. For GRES, we compute gIoU, cIoU, and N-acc on gRefCOCO; oIoU, mIoU, and acc on Ref-ZOM; and mIoU, mRR, and rIoU on R-RefCOCO. For GREC, we report Precision@(F1=1, IoU≥ 0.5) and N-acc. (no-target accuracy).Precision@(F1=1, IoU≥ 0.5) is abbreviated as F1score. gIoU averages instance-wise IoUs per image, assigning 1 for correct no-target cases and 0 for false positives. The

Table 1: Comparison with the state-of-the-art methods on gRefCOCO Liu et al. (2023a) dataset for GRES task. † Reproduced by us using the official codebase under identical settings.

Method	Val			TestA			TestB		
	gIoU	cIoU	N-acc.	gIoU	cIoU	N-acc.	gIoU	cIoU	N-acc.
MLLM Methods									
LISA-7B (Lai et al., 2024)	61.63	61.76	54.67	66.27	68.50	50.01	58.84	60.63	51.91
GSVA-7B (Xia et al., 2024)	66.47	63.29	62.43	71.08	69.93	65.31	62.23	60.47	60.56
SAM4MLLM-QW-7B (Chen et al., 2024)	68.96	66.33	62.96	70.54	70.13	65.82	63.98	63.21	61.61
Specialist Methods									
MattNet (Yu et al., 2018)	48.24	47.51	41.15	59.30	58.66	44.04	46.14	45.33	41.32
LTS (Jing et al., 2021)	52.70	52.30	–	62.64	61.87	–	50.42	49.96	–
VLT (Ding et al., 2021)	52.00	52.51	47.17	63.20	62.19	48.74	50.88	50.52	47.82
CRIS (Wang et al., 2022b)	56.27	55.34	–	63.42	63.82	–	51.79	51.04	–
LAVT (Yang et al., 2022)	58.40	57.64	49.32	65.90	65.32	49.25	55.83	55.04	48.46
ReLA (Ding et al., 2026)	63.98	62.91	56.29	70.12	69.43	58.96	61.29	60.15	58.59
CoHD (Luo et al., 2025)	68.42	65.17	63.68	72.67	71.85	64.00	63.60	62.63	60.37
GC^{3}VG (Dai et al., 2026a)	73.28	69.52	70.66	74.10	72.62	69.18	66.28	64.46	64.33
PropVG (Dai et al., 2025b)	73.29	69.23	72.83	74.43	74.20	69.87	65.87	64.76	64.97
InstanceVG† (Dai et al., 2026b)	72.74	68.69	71.23	74.11	73.30	71.09	65.76	64.62	65.18
CIVG-B(ours)	74.01	69.66	<u>74.75</u>	75.31	74.65	<u>72.14</u>	66.68	65.5	<u>66.31</u>
CIVG-L(ours)	<u>74.98</u>	<u>70.67</u>	72.87	<u>76.59</u>	<u>75.76</u>	69.75	<u>68.54</u>	<u>67.87</u>	62.82
CIVG-Union(ours)	**77.39**	**72.11**	**80.53**	**77.08**	**76.16**	**75.87**	**69.11**	**68.32**	**68.98**

Table 2: Comparison with the state-of-the-art methods on gRefCOCO He et al. (2023) dataset for GREC tasks. † Reproduced by us using the official codebase under identical settings.

Methods	Val		TestA		TestB	
	F1score	N-acc.	F1score	N-acc.	F1score	N-acc.
MCN (Luo et al., 2020)	28.0	30.6	32.3	32.0	26.8	30.3
VLT (Ding et al., 2021)	36.6	35.2	40.2	34.1	30.2	32.5
MDETR (Kamath et al., 2021)	42.7	36.3	50.0	34.5	36.5	31.0
UNINEXT (Yan et al., 2023)	58.2	50.6	46.4	49.3	42.9	48.2
SimVG (Dai et al., 2024)	62.1	54.7	64.6	57.2	54.8	57.2
PropVG (Dai et al., 2025b)	72.2	72.8	68.8	69.9	59.0	65.0
InstanceVG† (Dai et al., 2026b)	72.0	72.1	69.4	71.1	59.2	64.8
CIVG-B(ours)	74.1	<u>74.7</u>	70.3	<u>72.1</u>	60.8	<u>66.3</u>
CIVG-L(ours)	<u>75.2</u>	72.9	<u>71.6</u>	69.8	<u>62.9</u>	62.8
CIVG-Union(ours)	**77.7**	**80.3**	**72.3**	**75.9**	**63.6**	**70.0**

cIoU metric evaluates intersection versus union pixels. In Ref-ZOM, mIoU and oIoU represent the average IoU for referred objects and cIoU, respectively. For R-RefCOCO, rIoU measures segmentation quality by including negative sentences in the mIoU calculation. N-acc. in gRefCOCO and Acc. in Ref-ZOM represent the ratio of correctly classified no-target expressions. mRR in R-RefCOCO computes the recognition rate for no-target expressions.

5.2 Experimental Setup

For generalized tasks (GREC/GRES), the maximum sentence length is capped at 50 words. The model is trained for 10 epochs, and the learning rate is multiplied by 0.1 at the 7th epoch. The multi-modal encoder is frozen, and the initial learning rate for the remaining parameters is set to 5×10^{-4}. We employ the Adam optimizer for full-precision training.Prototype memory banks are constructed via K-means clustering with $K = 256$. All experiments are performed on two RTX 3090 GPUs.

5.3 Main results

Main results on GRES As demonstrated in Table 1, our proposed CIVG-B achieves state-of-the-art performance on all metrics on the gRefCOCO dataset. The improvement in N-acc. is particularly

remarkable, indicating stronger robustness in no-target scenarios and validating the effectiveness of our causal inference framework. In comparison, CIVG-L improves over CIVG-B in both gIoU and cIoU, yet exhibits a notable drop in N-acc. Furthermore, CIVG-Union brings additional performance gains, demonstrating the effectiveness of the dual-branch prediction framework. We also achieve state-of-the-art performance on the Ref-ZOM and R-RefCOCO/+/g datasets, as shown in Table 4 and Table 5, respectively.

Main results on GREC As presented in Tab. 2, CIVG-B achieves state-of-the-art performance on the GREC benchmark, demonstrating both accuracy and robustness. Moreover, CIVG-Union yields further improvements over both CIVG-B and CIVG-L, surpassing CIVG-L by +2.5%, +0.9%, and +0.7% in F1score on the validation, testA, and testB sets and CIVG-B by +5.6%, +3.8%, and +3.7% in N-acc. on the validation testA, and testB sets,respectively.

5.4 Ablation Studies

To comprehensively evaluate each component in CIVG, we conduct a systematic ablation study on the gRefCOCO validation set, decomposing CIVG into the Causality-Aware Deconfounding Module (CADM), the Counterfactual Relevance Regularization (CFR), and the Dual-Branch Prediction (DBP). Starting from a baseline without these modules, we progressively integrate each component individually and jointly, using both BEiT-B and BEiT-L as encoders. As shown in Table 3, both CADM and CFR improve all metrics, with CADM yielding particularly significant gains in N-acc., demonstrating that causal deconfounding and counterfactual contrast enhance accuracy and robustness. DBP brings further improvements, validating dual-branch prediction. We additionally ablate the internal design of CADM and analyze the parameters and computational complexity, with details in Appendix A.3; results confirm the effectiveness of jointly deconfounding both modalities and residual fusion, as well as the efficiency of our modules.

Table 3: Comprehensive ablation study on gRefCOCO validation set.

Branch	Modules			Performance Metrics			
	CADM	CFR	DBP	F1score	gIoU	cIoU	N-acc.
BEiT3-B				71.97	72.74	68.69	72.12
	✓			73.73	73.91	69.54	74.17
		✓		72.32	73.64	69.31	73.49
	✓	✓		**74.06**	**74.01**	**69.66**	**74.75**
BEiT3-L				74.46	74.16	69.66	70.56
	✓			75.02	74.72	70.49	72.35
		✓		74.67	74.56	70.03	71.23
	✓	✓		**75.17**	**74.98**	**70.67**	**72.87**
Union			✓	75.64	75.13	70.87	75.21
	✓		✓	76.91	76.56	71.85	78.94
		✓	✓	76.50	75.82	71.64	77. 56
	✓	✓	✓	**77.74**	**77.39**	**72.11**	**80.53**

6 Conclusion

In this paper, we first construct a structural causal model for visual grounding. Motivated by this SCM, we propose CIVG, a causal framework for generalized visual grounding, aiming to address the limitations of existing correlation-driven methods and to introduce causal inference into generalized visual grounding. We employ front-door adjustment and counterfactual construction to enhance the model's causal awareness. This framework effectively mitigates spurious correlations induced by dataset biases. Experimental results on multiple benchmarks demonstrate the effectiveness of our method in improving grounding accuracy and robustness.

AI USE STATEMENT

In this work, we used generative AI tools for two purposes: (1) retrieval and discovery, specifically to aid in searching for and identifying relevant literature on causal inference and related work; and (2) to polish writing. We have not used generative AI tools for any other tasks with required disclosure, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we did not use generative AI tools for any tasks with recommended disclosure. We have reviewed all AI-assisted work. Specifically, all literature retrieved with the assistance of AI tools was manually verified for relevance, accuracy, and correct citation through a comprehensive literature survey conducted by the authors. All text polished by AI was reviewed and edited by the authors to ensure accuracy and originality. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

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

Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S.-H. Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 513–524, 2025. doi: 10.1109/CVPRW67362.2025.00056.

Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *European Conference on Computer Vision*, pp. 323–340. Springer, 2024.

Zitan Chen, Zhuang Qi, Xiao Cao, Xiangxian Li, Xiangxu Meng, and Lei Meng. Class-level structural relation modeling and smoothing for visual representation learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 2964–2972, 2023.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1280–1289. IEEE, 2022.

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, Sen Yang, Wenxiao Cai, Yanpeng Sun, and Wankou Yang. Deris: Decoupling perception and cognition for enhanced referring image segmentation through loopback synergy. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19936–19946. IEEE, 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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7058–7068, 2025b. doi: 10.1109/ICCV51701.2025.00663.

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, 2026a. doi: 10.1109/TCSVT.2025.3632245.

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(1):448–465, 2026b. doi: 10.1109/TPAMI.2025.3607387.

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

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16301–16310. IEEE, 2021.

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.

Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15501–15510. IEEE, 2021.

Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. Grec: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*, 2023.

Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):684–696, 2022. doi: 10.1109/TPAMI.2019.2911066.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4044–4054. IEEE, 2023.

Jianqiang Huang, Yu Qin, Jiaxin Qi, Qianru Sun, and Hanwang Zhang. Deconfounded visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 998–1006, 2022.

Lihong Huang, Sheng-hua Zhong, Zhi Zhang, and Yan Liu. From pixels to logic: A perception-reasoning decomposition framework for open-world referring expression comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5058–5066, 2026a.

Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10485–10494. IEEE, 2020.

Yihong Huang, Fei Ma, Yihua Shao, Jingcai Guo, Zitong YU, Laizhong Cui, and Qi Tian. Nüwa: Mending the spatial integrity torn by vlm token pruning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 100825–100856, 2026b.

Licheng Jiao, Yuhan Wang, Xu Liu, Lingling Li, Fang Liu, Wenping Ma, Yuwei Guo, Puhua Chen, Shuyuan Yang, and Biao Hou. Causal inference meets deep learning: A comprehensive survey. *Research*, 7:0467, 2024.

Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9853–9862. IEEE, 2021.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 1760–1770. IEEE, 2021.

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

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23592–23601. IEEE, 2023a.

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, 2023b.

Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 11624–11641, 2023c.

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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10031–10040. IEEE, 2020.

Zhuoyan Luo, Yinghao Wu, Tianheng Cheng, Yong Liu, Yicheng Xiao, Hongfa Wang, Xiao-Ping Zhang, and Yujiu Yang. Cohd: A counting-aware hierarchical decoding framework for generalized referring expression segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22685–22694. IEEE, 2025.

E-Ro Nguyen, Hieu Le, Dimitris Samaras, and Michael S. Ryoo. Phrase-instance alignment for generalized referring segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 8486–8495, June 2026.

Judea Pearl. Causal inference in statistics: An overview. 2009.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect.* Allen Lane London, 2018.

Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2020.

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

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

Haoyu Wang, Le Wang, Sanping Zhou, Jingyi Tian, Zheng Qin, Yabing Wang, Gang Hua, and Wei Tang. Towards precise embodied dialogue localization via causality guided diffusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13350–13360. IEEE, 2025.

Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu, and Qijun Chen. Vision-and-language navigation via causal learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13139–13150. IEEE, 2024.

Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1960–1968, 2019.

Qiyi Wang, Senda Chen, and Ying Shen. Causalvtg: Towards robust video temporal grounding via causal inference. *Advances in Neural Information Processing Systems*, 38:103609–103633, 2026a.

Wei Wang, Junyu Gao, and Changsheng Xu. Weakly-supervised video object grounding via causal intervention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3933–3948, 2022a.

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

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*, 64:4503815–4503815, 2026b. doi: 10.1109/TGRS.2026.3703028.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11676–11685. IEEE, 2022b.

Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Toward robust referring image segmentation. *IEEE Transactions on Image Processing*, 33:1782–1794, 2024.

Yixuan Wu, Yang Zhang, Jian Wu, Philip Torr, and Jindong Gu. Postalign: Multimodal grounding as a corrective lens for mllms. In *International Conference on Learning Representations*, volume 2026, pp. 78875–78894, 2026.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3858–3869, 2024. doi: 10.1109/CVPR52733.2024.00370.

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, 2024.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Toward visual grounding: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3): 2749–2771, 2026. doi: 10.1109/TPAMI.2025.3630635.

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

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18134–18144. IEEE, 2022.

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4682–4692. IEEE, 2019.

Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *European conference on computer vision*, pp. 387–404. Springer, 2020.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1307–1315. IEEE, 2018.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with mllm for dense grounded understanding of images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2026. doi: 10.1109/TPAMI.2026.3720379.

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. Springer, 2022.

A ADDITIONAL EXPERIMENTS

A.1 DETAILS OF DATASETS

gRefCOCO(Liu et al., 2023a; He et al., 2023) extends REC with multi-target and no-target expressions. It contains 278,232 expressions, including 80,022 multi-target and 32,202 no-target expressions, spanning 60,287 object instances in 19,994 images. We follow the UNC partition with training, validation, testA, and testB splits.

Ref-ZOM(Hu et al., 2023) is derived from COCO and contains 55,078 images with 74,942 annotated objects. The training split includes 43,749 images and 58,356 objects, while the test split includes 11,329 images and 16,586 objects. Its annotations cover one-to-zero, one-to-one, and one-to-many scenarios, corresponding to non-, single-, and multi-target GRES cases.

R-RefCOCO/+/g(Wu et al., 2024) comprises R-RefCOCO, R-RefCOCO+, and R-RefCOCOg, all derived from the classical RES benchmarks. Negative sentences are added to the training set at a 1:1 ratio with positive sentences to improve robustness. For evaluation, only the validation set follows the official UNC partition.

A.2 MORE EXPERIMENTS REUSLTS

Results on GRES As shown in table 4, on Ref-ZOM, CIVG-B achieves performance comparable to the state of the art. Furthermore, CIVG-L and CIVG-Union successively attain higher performance. For the R-RefCOCO dataset, as reported in the table 5, CIVG-B achieves state-of-the-art results. Compared with CIVG-L, CIVG-Union achieves improvements in mRR and rIoU, accompanied by a slight decrease in mIoU, indicating that CIVG-Union attains superior performance in no-target scenarios.

Table 4: Comparison with state-of-the-art methods on the Ref-ZOM (Hu et al., 2023) dataset. † Reproduced by us using the official codebase under identical settings.

Method	Backbone	mIoU	oIoU	Acc.
MLLM Methods				
LISA-7B(Lai et al., 2024)	SAM-ViT-H	65.39	66.41	93.39
GSVA-7B(Xia et al., 2024)	SAM-ViT-H	68.13	68.29	94.59
Specialist Methods				
MCN(Luo et al., 2020)	DarkNet-53	54.70	55.03	75.81
VLT(Ding et al., 2021)	DarkNet-53	60.43	60.21	79.26
LAVT(Yang et al., 2022)	Swin-B	64.78	64.45	83.11
DMMI(Hu et al., 2023)	Swin-B	68.21	68.77	87.02
CoHD(Luo et al., 2025)	Swin-B	69.81	68.99	93.34
InstAlign(Nguyen et al., 2026)	Swin-B	70.81	71.13	94.23
InstanceVG†(Dai et al., 2026b)	BEiT3-B	71.15	70.91	96.72
CIVG-B(ours)	BEiT3-B	71.71	71.47	97.25
CIVG-L(ours)	BEiT3-L	<u>73.87</u>	<u>73.56</u>	<u>97.72</u>
CIVG-Union(ours)	BEiT3-L	**75.04**	**74.81**	**98.74**

Table 5: Comparison with State-of-the-Art Methods on the R-RefCOCO/+/g(Wu et al., 2024) Dataset. † Reproduced by us using the official codebase under identical settings.

Method	R-RefCOCO			R-RefCOCO+			R-RefCOCOg		
	mIoU	mRR	rIoU	mIoU	mRR	rIoU	mIoU	mRR	rIoU
CRIS(Wang et al., 2022b)	43.58	76.62	29.01	32.13	72.67	21.42	27.82	74.47	14.60
EFN(Feng et al., 2021)	58.33	64.64	32.53	37.74	77.12	24.24	32.53	75.33	19.44
VLT(Ding et al., 2021)	61.66	63.36	34.05	50.15	75.37	34.19	49.67	67.31	31.64
LAVT(Yang et al., 2022)	69.59	58.25	36.20	56.99	73.45	36.98	59.52	61.60	34.91
LAVT+(Yang et al., 2022)	54.70	82.39	40.11	45.99	86.35	39.71	47.22	81.45	35.46
RefSegformer(Wu et al., 2024)	68.78	73.73	46.08	55.82	81.23	42.14	54.99	71.31	37.65
CoHD(Luo et al., 2025)	74.16	84.27	53.61	64.59	87.49	49.07	63.56	82.68	42.16
InstanceVG†(Dai et al., 2026b)	76.15	91.10	62.34	69.73	94.20	59.10	70.08	91.82	53.71
CIVG-B (ours)	76.98	91.96	62.59	70.18	94.40	59.35	70.64	92.14	54.83
CIVG-L (ours)	**81.74**	93.50	68.67	**73.52**	94.77	65.38	**72.26**	93.58	60.82
CIVG-Union (ours)	78.69	**95.97**	**70.39**	72.19	**97.01**	**65.58**	70.71	**95.78**	**60.95**

A.3 ABLATION STUDY

Table 6: Ablation on CADM.

Variant	F1score	gIoU	cIoU	N-acc
w/o VSD	73.46	73.37	69.30	72.95
w/o TSD	73.46	73.07	69.11	72.08
w/o RS	73.65	73.48	69.39	73.34
Full	**74.06**	**74.01**	**69.66**	**74.75**

Table 7: Parameters of different core components.

Method	Params (M)	MACs (G)
BEiT-3	170.89	36.10
+Instance-aware	182.69 (+11.80)	59.42 (+23.32)
+CADM	194.73 (+11.83)	60.52 (+1.10)
+CFR	195.52 (+0.79)	60.54 (+0.02)

ablation study of CADM To verify the effectiveness of our proposed simultaneous deconfounding of the textual and visual modalities, as well as the effectiveness of residual fusion, we conduct ablation studies using CIVG-B on the validation set of the gRefCOCO dataset. We separately remove visual-side deconfounding(VSD), text-side deconfounding(TSD), and residual fusion(RS). The results are presented in the table 6. The results demonstrate that simultaneously deconfounding the visual and textual modalities is effective, and further validate the effectiveness of residual fusion.

Analysis of Parameters and Complexity. We present a detailed incremental breakdown of the parameters and computational complexity of each module in CIVG in Table 7. As can be observed, the combined parameter size of the proposed CADM and CFR modules is approximately 12.62 M, accounting for only 6.4% of the total model parameters. In terms of computational complexity, CADM and CFR contribute 1.9% of the overall computational cost, highlighting the efficiency of the proposed modules.

A.4 VISUALIZATION

In this section, we present qualitative examples, as illustrated in figure 4. Although CIVG achieves satisfactory localization accuracy, it still exhibits limitations on small and occluded objects. However, CIVG demonstrates robust understanding of text-referred objects, maintaining high precision and recall even in complex, multi-target, or no-target scenarios.

B LIMITATION & FUTURE WORK

CIVG brings a novel causal inference perspective to generalized visual grounding. Unlike correlation-based approaches, it formulates multi-target and no-target grounding as a causal problem and introduces a front-door intervention that jointly accounts for visual- and text-side confounders, thereby explicitly targeting spurious correlations such as language–location shortcuts.

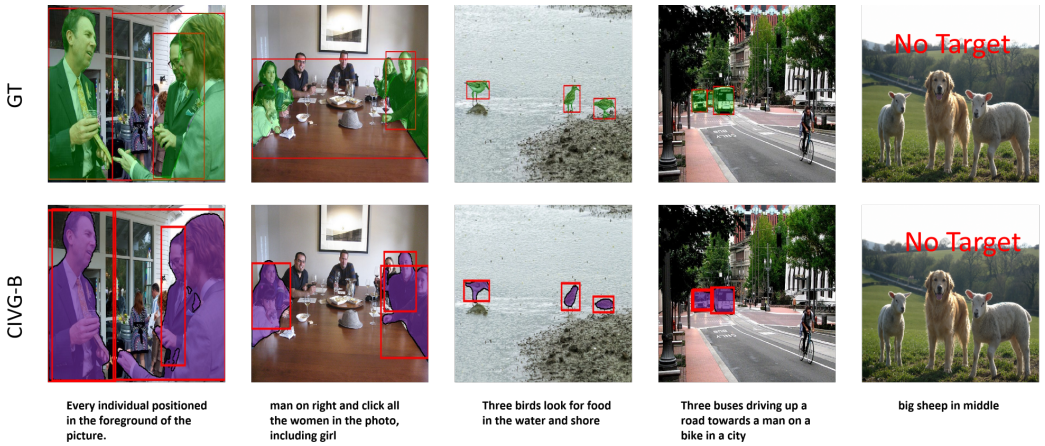


Figure 4: Visualization of complex and multi-referent object situations, as well as no-target scenarios.

However, the current formulation has two main limitations. First, the structural causal model remains insufficiently specified for visual grounding: the causal variables, mediation pathways, and identification assumptions are described only at a high level, and the visual and textual mediators are not modeled with grounding-specific granularity. Second, the front-door adjustment adopts a separable additive approximation following prior work, which may oversimplify potentially nonlinear, interactive, or context-dependent relationships among inputs, mediators, and targets. The causal module should therefore be viewed as a principled approximation rather than a fully identified causal estimator.

Future work will develop a more concrete and fine-grained structural causal model for generalized visual grounding and explore more complex relationships in front-door adjustment. We also plan targeted deconfounding and distribution-shift evaluations to better validate the causal claims.