Syntactic Structure-Guided Visual Grounding with Subject-Centric Feature Enhancement and Verification

Jiepeng Cai[1], **Zhen Xu**[1], **Tiesong Zhao**[2], **Hau-San Wong**[3] and **Si Wu**[1*]
[1]South China University of Technology
[2]Fuzhou University
[3]City University of Hong Kong
{cscaijp, csxuzhen}@mail.scut.edu.cn, t.zhao@fzu.edu.cn, cshswong@cityu.edu.hk, cswusi@scut.edu.cn

Abstract

Visual grounding aims to localize target objects based on natural language descriptions, and the core challenge lies in the cross-modal gap, which is partly caused by the significant differences in semantic structure between language and vision. Existing methods typically rely on holistic sentence-level semantic representations to modulate visual features, while overlooking the inherent structure of textual prompts. In this work, we propose a Syntactic Structure-guided Visual Grounding framework, referred to as SSVG. Specifically, to inject syntactic priors into unstructured visual representations, we design a semantic structure-based feature refinement module to adaptively modulate subject-centric and contextual visual features. To perform cross-modal alignment, we further incorporate a visual semantic consistency verification module, which leverages a subject-aware contrastive learning strategy to constrain and verify the semantic correspondence between the visual prediction and subject-level textual representation, thereby enhancing the model's robustness against semantically similar distractors. Extensive experiments demonstrate that our approach consistently outperforms state-of-the-art methods, and detailed analyses verify the effectiveness of each component.

1 Introduction

Visual grounding focuses on identifying and localizing the image regions that correspond to a given natural language description. In the realm of embodied intelligence and robotic systems, visual grounding plays a pivotal role by allowing agents to anchor language instructions in visual observations, thereby enabling precise object localization for downstream interaction and manipulation. Unlike generic object detection [Redmon and Farhadi, 2018], visual grounding requires cross-modal modeling [Radford *et al.*, 2021; Li *et al.*, 2022; Xue *et al.*, 2025; Wei *et al.*, 2023; Jiang *et al.*, 2024; Xu *et al.*, 2024] over heterogeneous structures, as linguistic descriptions are inherently hierarchical and compositional,

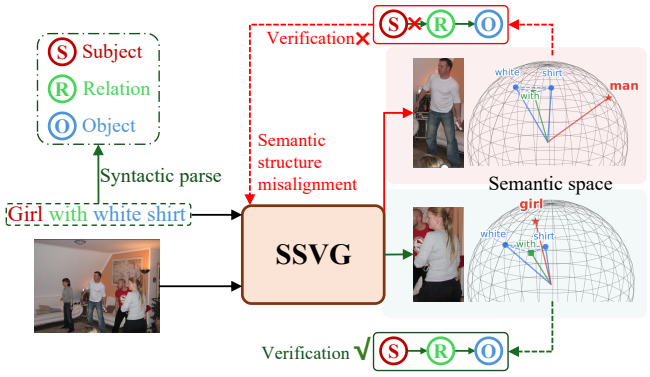

Figure 1: Illustration of the main idea behind the proposed SSVG framework. By parsing textual prompts, we facilitate visual grounding through syntactic structure-guided prediction-and-verification.

whereas visual features are typically encoded as flat aggregations of image patches, leading to a pronounced cross-modal gap.

To alleviate the cross-modal gap between language and vision, TransVG [Deng *et al.*, 2021] feeds textual features and visual features extracted by two independent encoders into a transformer encoder for joint fusion. Going a step further, QRNet [Ye *et al.*, 2022] modulates visual features using the global representation of the entire sentence, enabling the model to adaptively adjust intermediate visual representations under linguistic guidance. Meanwhile, VG-LAW [Su *et al.*, 2023] proposes an active perception modeling paradigm in which the visual backbone actively learns task-adaptive representations and dynamically adjusts the feature extraction process under the guidance of sentence-level semantics. To enable interaction between the visual and textual backbones, SegVG [Kang *et al.*, 2024] designs the triple alignment module, which is iteratively embedded into the intermediate layers of the backbones, utilizing a self-attention mechanism to map vision, text, and query features into a unified multi-modal space. TCRT [Chen *et al.*, 2025] further leverages task-aware prior embeddings generated by the Large Language Model (LLM) to guide visual feature refinement during the cross-modal feature fusion process, while enhancing information interaction between the Referring Expression Comprehension (REC) and Segmentation (RES) tasks. Exist-

*Corresponding author.

ing methods typically treat textual semantics as unstructured global signals, without explicitly accounting for the mismatch between the structure of language and the structure-agnostic nature of visual features, thereby struggling to separate target objects from contextual distractors.

In contrast to existing methods, which rely on the holistic semantics of the entire sentence as the modulation signal and overlook the inherent semantic structure of language, we propose a Syntactic Structure-guided Visual Grounding (SSVG) framework that exploits the semantics of textual expressions by explicitly parsing subject-centric and contextual semantics (as shown in Figure 1), and further leveraging the resulting semantic structure to guide visual feature enhancement. To learn subject-centric visual representations, we design semantic structure-guided feature refinement module, which leverages the parsed structural relations in the textual description to reorganize flattened visual features into subject-centric and context-aware representations. This allows the visual features to explicitly align with the structural information of the text. We further propose a visual semantic consistency verification module to assess the semantic reliability of predicted visual regions. The module reconstructs subject-related representations from the predictions and applies subject-aware contrastive learning to constrain the proximity between the reconstructed visual representations and the associated textual semantics. This design further reinforces the guiding role of syntactic structure in aligning visual region features. Extensive experiments on multiple benchmark datasets demonstrate the effectiveness and robustness of the proposed approach. The main contributions of this work can be summarized as follows:

- Different from existing methods that typically modulate visual features using holistic sentence-level representations, we propose a novel visual grounding framework that incorporates linguistic syntactic structures into both prediction and verification processes to significantly improve visual grounding performance.
- By parsing the syntactic structure of the textual prompts, we leverage the resulting dependency relations to modulate the distribution of visual features, highlighting the region aligned with the principal subject.
- We further perform subject-aware contrastive learning to constrain and verify the semantic correspondence between the visual prediction and subject-level textual representation, enhancing the model's discriminability regarding semantically similar distractors.

2 Related Work

2.1 Detection-and-Selection Methods

Constrained by the limitations of early object detection technologies [Girshick, 2015; Ren *et al.*, 2015]—specifically regarding key modules such as Non-Maximum Suppression (NMS) and Region of Interest (RoI)—early mainstream visual grounding methods predominantly adopted a two-stage strategy [Wang *et al.*, 2018; Hong *et al.*, 2019a; Yang *et al.*, 2020a; Yu *et al.*, 2018].The methods first generated a set of candidate regions and then identified the target region with the highest confidence by matching visual regions with the corresponding textual descriptions. During this period, a large number of studies based on the two-stage paradigm were proposed. However, the inherent limitations of detection-and-selection methods gradually became evident. First, the need to generate a large number of dense region proposals in the initial stage introduces substantial computational overhead, thereby reducing overall efficiency. Second, the final localization performance is highly dependent on the quality of the candidate regions; once the initial proposals are suboptimal, subsequent region-text matching is often unable to compensate for the shortcomings.

2.2 One-step Methods

As detection frameworks continued to evolve [Redmon and Farhadi, 2018], the research focus gradually shifted from traditional detection-and-selection paradigms to more compact single-stage designs [Yang *et al.*, 2019; Luo *et al.*, 2020; Liao *et al.*, 2020]. With the emergence of single-stage detectors, subsequent visual grounding studies began to adopt methods that eliminate the need for explicit region proposal generation. Under this paradigm, linguistic semantics are directly integrated into the intermediate feature representations of the detector. The model performs joint reasoning over a set of predefined dense anchors and directly predicts the target bounding box with the highest confidence, enabling a more efficient and unified localization process. As a pioneering work in this direction, FAOA [Yang *et al.*, 2019] proposed to fuse visual and textual representations at the feature level and feed the resulting multimodal features directly into the prediction module to accomplish target localization and detection. A series of subsequent works have continuously improved the performance of one-step visual grounding methods on complex descriptions through techniques such as correlation filtering modeling, recursive sub-query construction, and enhanced reasoning capabilities.

Motivated by the breakthrough success of Transformers [Vaswani *et al.*, 2017] in both vision and language tasks, TransVG [Deng *et al.*, 2021] introduced the Transformer architecture into the visual grounding task. By leveraging self-attention mechanisms, it enables deep cross-modal interactions between visual and textual features, and employs a learnable semantic token to directly regress the target bounding box coordinates. To address the limited degree of interaction between the visual backbone and linguistic information in TransVG, QRNet [Ye *et al.*, 2022] further proposed a query-aware dynamic attention mechanism along with a multi-scale feature fusion strategy. This formulation enables the model to adaptively modulate intermediate visual representations under textual guidance, thereby producing visual features that are better aligned with semantic constraints. To alleviate the supervision sparsity problem in visual grounding tasks, SegVG [Kang *et al.*, 2024] converts target bounding boxes that were originally used only for regression into pixel-level segmentation supervision signals. To fully exploit linguistic information, SSRVG [Zheng *et al.*, 2025] achieves a reasoning process that performs coarse localization using contextual information and then refines the localization with

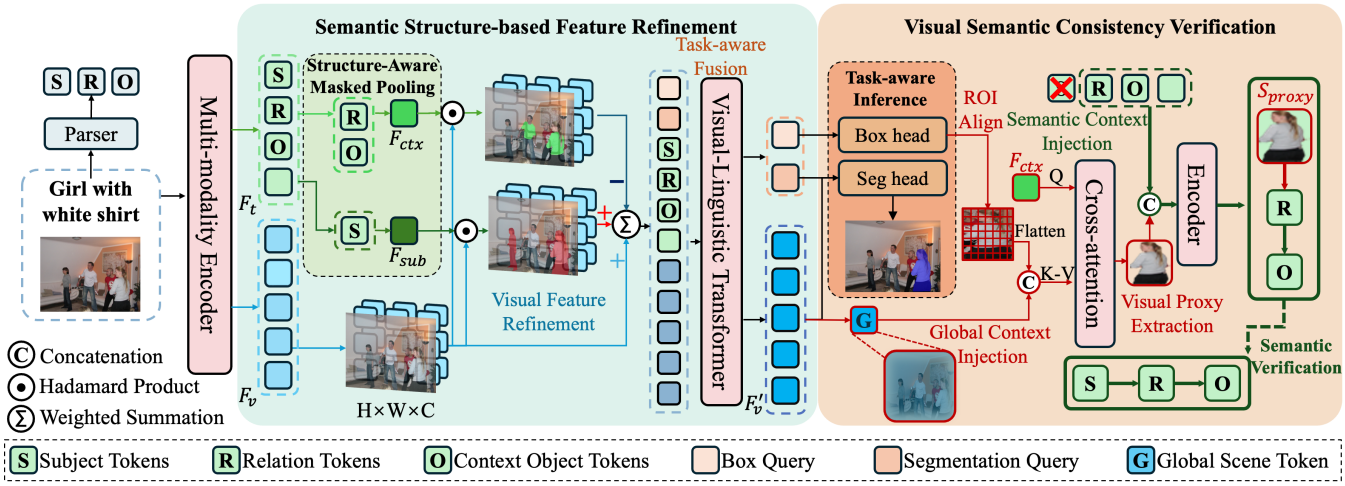

Figure 2: Overview of the proposed framework. The model first performs syntactic parsing on the input text to explicitly extract syntactic relational structures, which are then leveraged to refine visual features using structured linguistic information. Subsequently, two task tokens for the REC and RES tasks, together with the refined multimodal features, are fed into a visual–linguistic transformer for feature fusion, and the target bounding box and segmentation mask are produced by the corresponding prediction heads. After obtaining the predictions, the framework further introduces a visual semantic consistency verification module, which reconstructs subject-related semantic representations from the predicted visual regions and uses the representations to verify the semantic consistency between the visual predictions and the input language, providing additional constraints on the model predictions.

the subject by softly splitting the expression into subject and context. However, the serial design of SSRVG leads to a significant dependency of the subject localization stage on the output of the context stage.

The proposed work differs from existing methods in the following aspects: (1) Unlike TransVG [Deng *et al.*, 2021] and SegVG [Kang *et al.*, 2024], which primarily emphasize cross-modal feature fusion architectures, we focus on the representational mismatch between linguistic descriptions and structure-agnostic visual representations. (2) In contrast to methods like QRNet [Ye *et al.*, 2022] and VGLaw [Su *et al.*, 2023], which treat the entire textual expression as a holistic feature to modulate visual representations, our approach explicitly decomposes the text into distinct semantic components and applies them separately to guide the modulation of visual features. (3) Different from SSRVG [Zheng *et al.*, 2025], which also performs syntactic parsing of the text, SSRVG uses the context as an auxiliary query for cascade reasoning during the decoding stage. In contrast, our SSVG explicitly restructures the visual feature distribution during the encoding stage to reduce the interference of the context on the subject's features.

3 Method

3.1 Overview

As illustrated in Figure 2, we propose a syntactic structure-guided visual grounding framework. The framework first employs an off-the-shelf dependency parser to decompose the input expression into a principal subject representation and structured contextual constraints, thereby establishing a subject-centric semantic prior. Guided by this decomposition, the Semantic Structure-based Feature Refinement (SSFR) module adaptively modulates visual features by emphasizing regions relevant to the subject while suppressing interference induced by contextual information. The refined representations are then leveraged to perform deep cross-modal reasoning for joint referring expression comprehension and segmentation. Finally, rather than simply accepting the forward predictions, the framework incorporates a Visual Semantic Consistency Verification (VSCV) module. This module projects the predicted visual regions back into the semantic space to measure their affinity with the principal subject. By leveraging subject-aware contrastive learning, it acts as a posterior constraint to penalize the cross-modal semantic gap, thereby ensuring robust semantic consistency between the visual predictions and the intended linguistic subject.

To inject syntactic priors into unstructured visual representations, we first extract the core triplet structure from the caption T. Utilizing a dependency parser, we identify the core components $\mathcal{G} = \{S, R, O\}$. Specifically, S corresponds to the syntactic root or nominal subject (nsubj), while the contextual constraints (R, O) are derived from connected head verbs or prepositional phrases (e.g., dobj, pobj). Based on the parsed structure, we further construct binary structure masks $M^k_{\text{struct}} \in \{0,1\}^L$ for the subject and context components (where $k \in \{\text{sub}, \text{ctx}\}$). These masks indicate the token positions associated with each component, serving as explicit guidance for subsequent semantic structure-aware interaction.

3.2 Semantic Structure-Guided Feature Refinement

We employ a unified multi-modal encoder to extract joint representations. Given an input image I and a referring expression T, the encoder outputs visual features $F_v \in \mathbb{R}^{N\times D}$ and linguistic features $F_t \in \mathbb{R}^{L\times D}$, where $F_{t,i}$ denotes the feature

vector corresponding to the i-th token. To alleviate the inherent misalignment between unstructured visual patches and structured linguistic logic, we propose the semantic structure-guided feature refinement module. This module leverages the explicit structural priors derived in Section 3.1 to functionally decouple subject-centric components from contextual constraints, thereby guiding the visual encoding process.

To initiate the visual feature refinement, we implement a joint semantic localization mechanism that successively aggregates structural prototypes and identifies relevant visual regions. Utilizing the binary structure masks M^k_{struct} derived from the parsing stage (where $k \in \{\text{sub}, \text{ctx}\}$), we first obtain the decoupled prototypes t_k, which then serve as structural bases to generate attention maps M_k. Specifically, to exploit the syntactic dependencies for precise visual grounding, we project the visual features into distinct semantic subspaces corresponding to the subject and context components:

$$t_k = \frac{\sum_{i=1}^{L} M^k_{\text{struct}}(i) \cdot F_{t,i}}{\sum_{i=1}^{L} M^k_{\text{struct}}(i)}, \qquad M_k = \sigma\left((F_v W_{v,k}) t_k^T\right), \tag{1}$$

where $M^{\text{ctx}}_{\text{struct}}$ represents the logical union of relation and object tokens, and $W_{v,k}$ denotes the learnable projection matrix tailored for the specific syntactic component. This formulation ensures that M_{sub} highlights regions aligned with the principal subject, whereas M_{ctx} captures the supporting contextual environment. With the semantic roles explicitly localized, we finally employ a semantic structure-aware modulation strategy to rectify the visual feature distribution:

$$F_v^{\text{refined}} = F_v + \alpha \cdot (M_{\text{sub}} \odot F_v) - \beta \cdot (M_{\text{ctx}} \odot F_v), \tag{2}$$

where $\odot$ represents the Hadamard product. α and β are learnable scaling factors initialized to control the modulation intensity. In this formulation, the term $\alpha \cdot (M_{\text{sub}} \odot F_v)$ acts as a positive reinforcement to boost the subject response, while the subtraction term $\beta \cdot (M_{\text{ctx}} \odot F_v)$ functions as a semantic dampener to mitigate contextual interference. To achieve holistic multi-granularity perception, we establish a unified multi-task query generation mechanism that bridges object-level localization and pixel-level segmentation. We define learnable task queries $Q_{\text{task}} = \{q_{\text{rec}}, q_{\text{res}}\}$ to act as distinct decoding initiators. These queries interact with the linguistic features F_t to encode task-specific semantic priors, generating the text-conditioned queries Q'_{task}.

3.3 Visual Semantic Consistency Verification

To generate high-quality candidate regions for subsequent verification, we first perform a unified cross-modal encoding. We construct a heterogeneous sequence by concatenating the updated task queries, textual tokens, and flattened visual patches. This sequence is processed by the visual-linguistic transformer $\mathcal{T}$ to facilitate deep multimodal fusion:

$$H_{\text{out}} = \mathcal{T}(\mathcal{C}(Q'_{\text{task}}, F_t, F_v) + P_{\text{pos}}), \tag{3}$$

where $\mathcal{C}(\cdot)$ denotes the concatenation operator, and P_{pos} denotes the learnable positional embeddings. The output sequence H_{out} is then parsed into task-specific embeddings. We employ a regression MLP and a specialized segmentation head to decode the preliminary bounding box $\hat{b}$ and segmentation mask M_{seg} directly from the unified representations.Unlike conventional approaches that blindly accept the forward predictions, our framework introduces a novel visual semantic consistency verification mechanism. This module operates on the premise that a correct prediction should naturally map back to the semantic space of the principal subject.

To explicitly validate the semantic consistency of predicted regions via inverse reasoning, we construct a hybrid visual set V_{hybrid} by concatenating fine-grained local features $V_{\text{roi}} \in \mathbb{R}^{N \times D}$ (via RoI Align on $\hat{b}$) with a global scene token v_{global} (via pooling on F'_v). Augmented with distinct learnable positional embeddings to preserve local geometry, this hybrid representation provides corrective supervision to rectify structural misalignment, ensuring the model's focus remains strictly aligned with the principal subject. Subsequently, we introduce a context-guided visual aggregation mechanism. Mirroring linguistic dependencies where context constrains the subject, we utilize the contextual embedding F_{ctx} as a query to distill a visual proxy s_{proxy} from V_{hybrid}:

$$s_{\text{proxy}} = \mathcal{A}(F_{\text{ctx}}, V_{\text{hybrid}}, V_{\text{hybrid}}), \tag{4}$$

where $\mathcal{A}(Q, K, V)$ denotes the standard Multi-Head Attention operation. F_{ctx} serves as the query, while V_{hybrid} serves as both the key and value. This operation aggregates visual cues strictly corresponding to the linguistic context, providing a compact representation for the final consistency verification.

We implement this structural verification by employing a hybrid sequence modeling mechanism. Specifically, we first filter out the original subject embeddings from the textual sequence to isolate the remaining linguistic context F_{rest}. Subsequently, we inject the extracted visual proxy s_{proxy} into this context to form the verification input. This process is formulated as:

$$F_{\text{rest}} = \{f \in F_t \mid f \neq t_{\text{sub}}\}, \qquad h_{\text{ver}} = \mathcal{E}(\mathcal{C}(F_{\text{rest}}; s_{\text{proxy}})), \tag{5}$$

where $\mathcal{C}(\cdot)$ denotes the concatenation operator, and $\mathcal{E}$ represents the transformer encoder. This effectively projects the visual features into the structural semantic space to verify their compatibility with the surrounding context.

To bridge the inherent domain gap between structured linguistic dependencies and flattened visual tokens, we propose a structure-aware contrastive Loss. The core motivation is to enforce that the visual representation derives its validity strictly from its semantic dependency on the coupled object, preventing it from functioning as an isolated instance. Formally, we minimize the contrastive objective over the data distribution $\mathcal{D}$:

$$\mathcal{L}_{\text{ver}} = \mathbb{E}_{i \sim \mathcal{D}}\left[-\log \frac{\exp(s(h_{\text{ver},i}, \text{sg}(t_{\text{sub},i}))/\tau)}{Z_i}\right], \tag{6}$$

where i indexes the training sample, $s(\cdot,\cdot)$ denotes cosine similarity, τ is the temperature parameter, and $\text{sg}(\cdot)$ is the stop-gradient operation. The latter ensures that textual targets act as fixed references, allowing gradients to propagate exclusively through the visual pathway, thereby preserving the aligned semantic space.

Methods	Multi-task	Venue	RefCOCO			RefCOCO+			RefCOCOg	
			val	testA	testB	val	testA	testB	val	test
Detection-and-Selection Methods										
MAttNet [Yu *et al.*, 2018]	×	CVPR18	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
RvG-Tree [Hong *et al.*, 2019b]	×	TPAMI19	75.06	78.61	69.85	63.51	67.45	56.66	66.95	66.51
NMTree [Liu *et al.*, 2019]	×	ICCV19	76.41	81.21	70.09	66.46	72.02	57.52	65.87	66.44
One-step Methods										
TransVG [Deng *et al.*, 2021]	×	ICCV21	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73
QRNet [Ye *et al.*, 2022]	×	CVPR22	84.01	85.85	82.34	72.94	76.17	63.81	71.89	73.03
VG-LAW [Su *et al.*, 2023]	✓	CVPR23	86.62	89.32	83.16	76.37	81.04	67.50	76.90	76.96
SimVG [Dai *et al.*, 2024]	×	NeurIPS24	87.63	90.22	84.04	78.65	83.36	71.82	80.37	80.51
EEVG [Chen *et al.*, 2024]	✓	ECCV24	88.08	90.33	85.50	77.97	82.44	69.15	79.60	80.24
OneRef [Xiao *et al.*, 2024]	✓	NeurIPS24	91.89	94.31	88.58	86.38	90.38	79.47	86.82	87.32
SSRVG [Zheng *et al.*, 2025]	×	AAAI25	92.73	94.04	90.75	85.12	88.84	78.97	88.30	88.51
SSVG	✓	IJCAI26	**93.27**	**94.57**	**90.94**	**86.74**	**91.11**	**79.95**	**88.46**	**88.93**

Table 1: Comparison with existing visual grounding methods on RefCOCO, RefCOCO+, and RefCOCOg for the REC task. Multi-task indicates that the model is capable of performing both REC and RES tasks.

Methods	Multi-task	Venue	RefCOCO			RefCOCO+			RefCOCOg	
			val	testA	testB	val	testA	testB	val	test
RefTR [Li and Sigal, 2021]	✓	NeurIPS21	70.56	73.49	66.57	61.08	64.65	52.73	58.73	58.51
SeqTR [Zhu *et al.*, 2022]	✓	ECCV22	67.26	69.79	64.12	54.14	58.93	48.19	55.67	55.64
VG-LAW [Su *et al.*, 2023]	✓	CVPR23	75.62	77.51	72.89	66.63	70.38	58.89	65.63	66.08
RefFormer [Wang *et al.*, 2024]	✓	NeurIPS24	74.47	77.92	72.30	66.70	72.28	60.43	65.61	66.26
PVD [Cheng *et al.*, 2024]	✓	AAAI24	75.07	77.29	70.13	64.39	69.15	57.19	63.22	63.89
Potamoi [Guan *et al.*, 2025]	✓	T-ITS25	-	76.32	69.77	-	71.18	57.74	-	65.20
EEVG [Chen *et al.*, 2024]	✓	ECCV24	78.23	79.27	76.58	69.04	72.65	62.33	69.15	70.01
OneRef [Xiao *et al.*, 2024]	✓	NeurIPS24	79.83	81.86	76.99	74.68	77.90	69.58	74.06	74.92
SSVG	✓	IJCAI26	**81.57**	**82.55**	**78.76**	**75.56**	**79.17**	**69.60**	**75.50**	**76.04**

Table 2: Comparison with existing visual grounding methods on RefCOCO, RefCOCO+, and RefCOCOg for the RES task.

Crucially, the normalization term Z_i is designed to filter semantic collisions and incorporate structural constraints. It is computed within the sampled mini-batch $\mathcal{B}$ as:

$$\begin{aligned} Z_i = &\exp(s(h_{\text{ver},i}, \text{sg}(t_{\text{sub},i}))/\tau) \\ &+ \sum_{j \in \mathcal{N}(i)} \exp(s(h_{\text{ver},i}, \text{sg}(t_{\text{sub},j}))/\tau) \\ &+ \mathbb{I}_i^{\text{obj}} \cdot \exp(s(h_{\text{ver},i}, \text{sg}(t_{\text{obj},i}))/\tau), \end{aligned} \tag{7}$$

where $\mathcal{N}(i) = \{j \in \mathcal{B} \setminus \{i\} \mid \text{id}(j) \neq \text{id}(i)\}$ is the set of negative samples. This strict definition explicitly excludes in-batch samples sharing the same semantic identity (i.e., token ID) as i, thereby preventing false negatives. Furthermore, $t_{\text{obj},i}$ represents the structural hard negative (the object feature), and the indicator $\mathbb{I}_i^{\text{obj}}$ ensures this constraint is applied only when a valid object component exists.

The proposed framework is trained in an end-to-end manner using a unified multi-task objective function. We employ L1 loss $\mathcal{L}_{\text{L1}}$ [Girshick, 2015] and Generalized IoU (GIoU) loss $\mathcal{L}_{\text{GIoU}}$ [Rezatofighi *et al.*, 2019] for box regression. For mask segmentation, we utilize binary cross-entropy loss $\mathcal{L}_{\text{focal}}$ [Lin *et al.*, 2017] and dice loss $\mathcal{L}_{\text{dice}}$ [Milletari *et al.*, 2016]. Incorporating the proposed structure-aware consistency verification, the overall optimization formulation is defined as:

$$\min_{\theta} \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}} + \lambda_{\text{ver}} \mathcal{L}_{\text{ver}}, \tag{8}$$

where $\mathcal{L}_{\text{box}} = \mathcal{L}_{\text{L1}} + \mathcal{L}_{\text{GIoU}}$ and $\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{focal}} + \mathcal{L}_{\text{dice}}$. λ_{ver} serves as a scaling factor specifically introduced to balance the auxiliary verification signal with the primary grounding objectives. By jointly optimizing these terms, the model learns to accurately localize and segment the referent while explicitly verifying semantic consistency through the structural verification mechanism.

4 Experiment

4.1 Experimental Setup

Implementation Details. All experiments were implemented using the PyTorch framework and conducted on an NVIDIA A800 GPU cluster for both training and evaluation. During preprocessing, all input images were uniformly resized to a fixed resolution of (448×448). The parameters of the vision–language encoder were initialized with pretrained BEiT-3 weights. The model was trained end to end using the AdamW optimizer for a total of 30 epochs. The initial learning rate was set to (1.5×10^{-5}) and decayed to (1.5×10^{-6}) after the 20th epoch. A batch size of 32 was used throughout training, with a dropout rate of 0.1. The data augmentation strategy followed the settings adopted in prior related work [Yang *et al.*, 2020b]. Experimental evaluation was carried out on three widely used benchmarks for referring expression grounding, namely RefCOCO [Yu *et al.*, 2016], RefCOCO+ [Yu *et al.*, 2016], and RefCOCOg [Nagaraja *et al.*, 2016].

Evaluation Metrics. Model performance was assessed using commonly adopted evaluation metrics in the field. For the referring expression comprehension (REC) task, Preci-

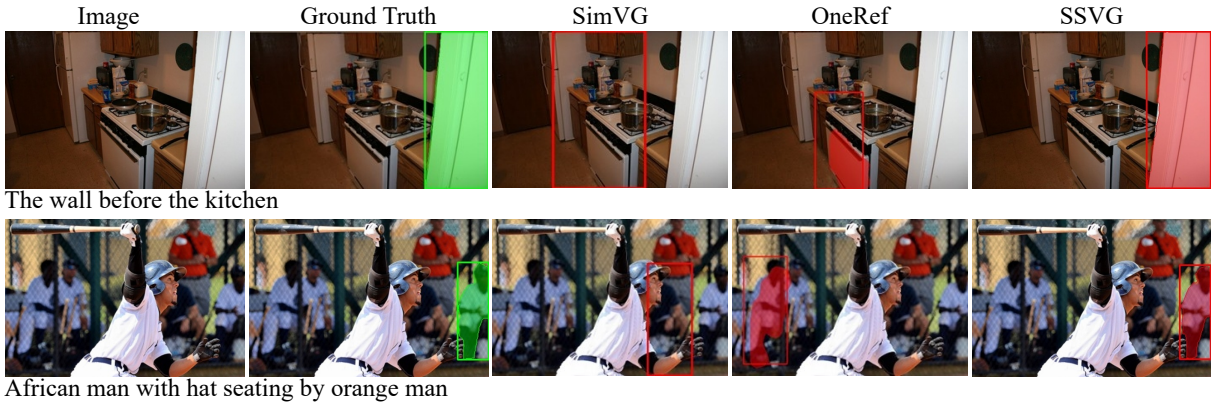

Figure 3: Qualitative comparison of different methods (SimVG, OneRef, and SSVG) on the REC task and RES task. Note that SimVG does not support the RES task. The comparison illustrates the superior performance of our approach in accurately localizing target regions.

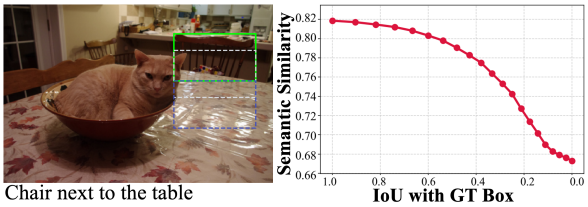

Figure 4: Validation of the visual semantic consistency verification module. The semantic similarity between the reconstructed subject representation and the ground-truth subject embedding decreases monotonically as the IoU between the perturbed bounding box and the ground-truth box decreases.

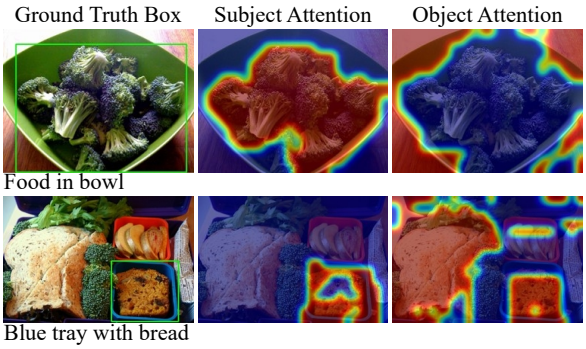

Figure 5: Visualization of subject-aware and context-aware attention maps, demonstrating how the model allocates attention to subject-relevant regions and contextual elements.

sion@0.5 (Prec@0.5) was employed as the evaluation criterion, where a prediction is considered correct if the Intersection over Union (IoU) between the predicted bounding box and the ground-truth box exceeds 0.5. For the referring expression segmentation (RES) task, the mean Intersection over Union (mIoU) was used to measure segmentation performance by averaging the IoU between the predicted and ground-truth masks.

4.2 Comparison with State-of-the-Arts

In this section, we conduct a systematic comparison between the proposed method and current state-of-the-art (SOTA) methods on three widely used benchmarks. Table 1 reports the quantitative comparison results on the REC task across the evaluated datasets, while Table 2 presents the corresponding quantitative results for the RES task. In addition, Figure 3 provides qualitative comparisons for both REC and RES tasks, illustrating the visual grounding performance of different methods.

According to the results reported in Table 1, our method consistently outperforms the recent SOTA method SSRVG across all datasets. On the testA, testB, and val subsets of RefCOCO, our method achieves improvements of 0.54%, 0.53%, and 0.19%, respectively, over SSRVG. Similarly, on the testA, testB, and val subsets of RefCOCO+, the performance gains are 1.62%, 2.27%, and 0.98%, respectively. On the val and test subsets of RefCOCOg, our approach also yields consistent improvements of 0.16% and 0.42%. In addition, the proposed method can be seamlessly extended to the RES task without requiring architectural modifications. As shown in Table 2, compared with existing SOTA method OneRef, the proposed method achieves an average improvement of 1.12 percentage points across all evaluated datasets.

From a qualitative analysis perspective, Figure 3 presents visual comparison results of different methods on both the REC and RES tasks. Compared with SimVG and OneRef, our method is able to more accurately focus on target regions that are semantically consistent with the referring expressions, particularly in scenarios involving complex attribute descriptions and strong background distractions. Taking the second-row example in Figure 3 as an illustration, the image contains a complex background with multiple visually similar human instances. The corresponding referring expression, “African man with hat seating by orange man”, involves multiple attribute constraints as well as relational descriptions. Under such conditions, competing methods are more likely to be confused by similar distractor objects, leading to localization shifts or incorrect instance selection. In contrast, our method can effectively distinguish between multiple similar instances and leverage both semantic and relational cues to localize the target instance that precisely matches the refer-

Figure 6: Qualitative ablation study on the RefCOCOg test set.

Method	RefCOCOg	
	val	test
SSVG	**88.46**	**88.93**
w/o SSFR	87.83	88.43
w/o VSCV	87.39	88.04
w/o SSFR & VSCV	87.27	88.28

Table 3: Quantitative ablation study for the REC task on RefCOCOg.

ring expression, demonstrating its strong capability in complex semantic understanding.

5 Analysis of Main Components

Can Visual Semantic Consistency Verification Accurately Reflect Semantic Alignment? To validate the Visual Semantic Consistency Verification (VSCV)module, we conduct a controlled perturbation experiment by gradually disturbing the ground-truth bounding box to generate 20 boxes with IoU decreasing from 1.0 to 0.0. For each perturbed box, we perform semantic reconstruction and measure the cosine similarity between the reconstructed subject representation and the ground-truth subject embedding. As shown in Figure 4, the reconstruction similarity decreases monotonically with IoU, indicating that the reconstruction module provides a reliable semantic signal that correlates with localization accuracy. We also conduct a quantitative ablation analysis on the VSCV module, with the results reported in Table 3. Specifically, on the RefCOCOg dataset, we adopt Precision@0.5 as the evaluation metric and compare model performance on the val and test subsets. When the VSCV module is removed, the performance on the val set drops from 88.46% to 87.39%, while the performance on the test set decreases from 88.93% to 88.04%, showing a consistent and noticeable degradation. This indicates that by introducing the semantic reconstruction and verification, the VSCV module provides additional and reliable semantic supervision, thereby leading to performance improvements.

How Does Structure-Guided Feature Refinement Differentiate Subject and Context? To analyze the effectiveness of the proposed Semantic Structure-Guided Feature Refinement module (SSFR), we visualize the learned subject-aware attention map M_{sub} and context-aware attention map M_{ctx}. The visualizations provide intuitive evidence of how SSFR differentiates subject-relevant and context-related visual cues before feature fusion. As illustrated in Figure 5, M_{sub} consistently concentrates on regions that are semantically aligned with the referred subject, even when such regions are not the most visually salient. In contrast, M_{ctx} primarily activates on surrounding objects and contextual elements that are visually prominent but semantically auxiliary to the referring expression. This complementary behavior indicates that SSFR successfully decomposes the original visual feature into two structurally distinct components corresponding to subject and context. Meanwhile, we also present a qualitative ablation analysis in Figure 6. It can be observed that the full model is able to stably and accurately localize the target object even in the presence of complex backgrounds and challenging referring expressions. Similarly, we conduct an ablation study on the SSFR module. As shown in Table 3, removing SSFR results in a drop in Precision@0.5 on the val / test subsets of RefCOCOg from 88.46% / 88.93% to 87.83% / 88.43%, respectively, validating the effectiveness of this module. This indicates that injecting syntactic structural information from the text into the visual feature modeling process plays a critical role in improving performance.

6 Conclusion

In this paper, we propose a syntactic structure-guided visual grounding framework that aims to fully exploit linguistic structural information to enhance cross-modal alignment. Specifically, SSVG first leverages the syntactic structure of the input text to perform structured modulation of visual features, enhancing target-related visual representations. Building upon this, we further design a syntactic structure-guided semantic reconstruction and subject-aware contrastive learning strategy to explicitly reduce the semantic gap between target visual representations and the corresponding language descriptions, strengthening the consistency between visual and language representations. Extensive experiments validate the effectiveness of the proposed method in visual grounding tasks.

Acknowledgments

This work was supported in part by TCL Science and Technology Innovation Fund (Project No. 20231752), in part by the GuangDong Basic and Applied Basic Research Foundation (Project No. 2024A1515011437), in part by the Guangdong Provincial Science and Technology Program (Project No. 2025A050508003).

Contribution Statement

Jiepeng Cai and Zhen Xu contributed equally to this work and are recognized as joint first authors.

References

[Chen *et al.*, 2024] 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*, pages 125–141. Springer, 2024.

[Chen *et al.*, 2025] Wenbo Chen, Zhen Xu, Ruotao Xu, Si Wu, and Hau-San Wong. Task-aware cross-modal feature refinement transformer with large language models for visual grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3931–3941, 2025.

[Cheng *et al.*, 2024] 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, pages 1326–1334, 2024.

[Dai *et al.*, 2024] 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.

[Deng *et al.*, 2021] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1769–1779, 2021.

[Girshick, 2015] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[Guan *et al.*, 2025] Runwei Guan, Liye Jia, Shanliang Yao, Fengyufan Yang, Sheng Xu, Erick Purwanto, Xiaohui Zhu, Ka Lok Man, Eng Gee Lim, Jeremy Smith, et al. Watervg: Waterway visual grounding based on text-guided vision and mmwave radar. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[Hong *et al.*, 2019a] 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, 2019.

[Hong *et al.*, 2019b] 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, 2019.

[Jiang *et al.*, 2024] Le Jiang, Yan Huang, Lianxin Xie, Wen Xue, Cheng Liu, Si Wu, and Hau-San Wong. Hunting blemishes: Language-guided high-fidelity face retouching transformer with limited paired data. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5102–5111, 2024.

[Kang *et al.*, 2024] Weitai Kang, Gaowen Liu, Mubarak Shah, and Yan Yan. Segvg: Transferring object bounding box to segmentation for visual grounding. In *Proceedings of the European Conference on Computer Vision*, pages 57–75, 2024.

[Li and Sigal, 2021] 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.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900, 2022.

[Liao *et al.*, 2020] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[Liu *et al.*, 2019] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4673–4682, 2019.

[Luo *et al.*, 2020] 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 Conference on Computer Vision and Pattern Recognition*, pages 10034–10043, 2020.

[Milletari *et al.*, 2016] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision*, pages 565–571, 2016.

[Nagaraja *et al.*, 2016] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Proceedings of the European Conference on Computer Vision*, pages 792–807, 2016.

[Radford *et al.*, 2021] 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*, pages 8748–8763, 2021.

[Redmon and Farhadi, 2018] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

[Rezatofighi *et al.*, 2019] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[Su *et al.*, 2023] 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 Conference on Computer Vision and Pattern Recognition*, pages 10857–10866, 2023.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[Wang *et al.*, 2018] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018.

[Wang *et al.*, 2024] Yabing Wang, Zhuotao Tian, Qingpei Guo, Zheng Qin, Sanping Zhou, Ming Yang, and Le Wang. Referencing where to focus: Improving visual grounding with referential query. *Advances in Neural Information Processing Systems*, 37:47378–47399, 2024.

[Wei *et al.*, 2023] Xiwen Wei, Zhen Xu, Cheng Liu, Si Wu, Zhiwen Yu, and Hau San Wong. Text-guided unsupervised latent transformation for multi-attribute image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19285–19294, 2023.

[Xiao *et al.*, 2024] 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.

[Xu *et al.*, 2024] Feifan Xu, Rui Li, Si Wu, Yong Xu, and Hau San Wong. Text-conditional attribute alignment across latent spaces for 3d controllable face image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9172–9181, 2024.

[Xue *et al.*, 2025] Wen Xue, Chun Ding, Ruotao Xu, Si Wu, Yong Xu, and Hau-San Wong. Retouchgpt: Llm-based interactive high-fidelity face retouching via imperfection prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9059–9067, 2025.

[Yang *et al.*, 2019] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4683–4693, 2019.

[Yang *et al.*, 2020a] Sibei Yang, Guanbin Li, and Yizhou Yu. Relationship-embedded representation learning for grounding referring expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2765–2779, 2020.

[Yang *et al.*, 2020b] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 387–404, 2020.

[Ye *et al.*, 2022] 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 Conference on Computer Vision and Pattern Recognition*, pages 15502–15512, 2022.

[Yu *et al.*, 2016] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision*, pages 69–85, 2016.

[Yu *et al.*, 2018] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.

[Zheng *et al.*, 2025] Shiyi Zheng, Peizhi Zhao, Zhilong Zheng, Peihang He, Haonan Cheng, Yi Cai, and Qingbao Huang. Look around before locating: Considering content and structure information for visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[Zhu *et al.*, 2022] 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 *Proceedings of the European Conference on Computer Vision*, pages 598–615, 2022.