Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33732

AD-SRE: An RGB-D Referring Expression Grounding Method for Complex Spatial Relation Understanding

Shuhui Zhao

School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China
1033230110@stu.jiangnan.edu.cn

Abstract. Referring Expression Grounding aims to localize a target object in an image according to a natural language description. Existing methods mainly perform cross-modal modeling based on RGB images and text, and they have achieved promising results in attribute matching and category recognition. However, in scenes involving complex spatial relations, these methods still face limitations in exploiting geometric information and modeling spatial semantics. To address this problem, this paper proposes AD-SRE, an RGB-D Referring Expression Grounding method based on adaptive depth modulation and spatial relation enhancement. The proposed method introduces a Query-Adaptive Depth Modulation module to adjust the contribution of depth features according to query semantics, and a Spatial Relation Enhancement module to strengthen visual responses related to spatial cues. Experiments on the SUN-Spot dataset show that AD-SRE outperforms both the RGB-Text baseline and the simple RGB-D-Text fusion baseline. Further ablation and category-wise results demonstrate that the proposed modules are especially effective for spatial relation expressions and compound expressions. These results indicate that query-aware depth modeling can improve spatial semantic understanding in RGB-D visual grounding.

Keywords: Referring expression grounding, RGB-D, Transformer, Spatial relation modeling, Multimodal fusion.

1. Introduction

Referring Expression Grounding aims to localize the target region in an image according to a natural language description. It is an important task at the intersection of computer vision and natural language processing [1]. Unlike traditional object detection, referring expression grounding requires the model not only to recognize object categories, but also to understand visual attributes, semantic constraints, and relations between the target object and surrounding objects [1, 2]. In recent years, with the rapid development of vision-language modeling and Transformer-based architectures, related studies have gradually shifted from proposal-based multi-stage paradigms to end-to-end cross-modal reasoning frameworks [2, 3]. TransVG is a representative method in this line of research. It formulates visual grounding as joint reasoning over image features and textual features, and directly regresses the target bounding box, thereby avoiding the complex proposal generation and post-processing procedures used in traditional methods [3].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33732

Although existing RGB-Text methods have achieved promising performance, they still face clear limitations in scenes involving complex spatial relations, crowded layouts, and visually similar objects. For expressions mainly defined by appearance attributes, such as color, texture, or category, RGB images usually provide strong discriminative information. However, appearance features alone are often insufficient for expressions involving front-back relations, relative distance, occlusion, or positional constraints. In such cases, the model needs to capture spatial constraints between the target object and reference objects, rather than relying only on local appearance matching. This challenge becomes more significant when multiple objects appear in the same scene, when the reference object is ambiguous, or when the description contains several relational constraints at the same time [1, 4]. Therefore, spatial semantic grounding remains a key challenge in referring expression grounding.

Depth information provides a natural complement to RGB appearance features for addressing these limitations. Unlike RGB images, which mainly describe appearance patterns, depth maps explicitly provide distance distribution, geometric structure, and relative spatial layout. These properties are useful for modeling front-back, near-far, and hierarchical relations [5]. The SUN-Spot dataset further highlights the importance of this problem by providing RGB-D indoor scenes with rich spatial referring expressions and crowded object arrangements [4]. Compared with conventional referring expression datasets, SUN-Spot contains more spatial prepositions and more complex indoor layouts, making it suitable for evaluating spatial semantic understanding in RGB-D scenes [4]. Therefore, introducing depth information is meaningful for improving spatial reasoning in referring expression grounding.

However, simply introducing depth information is not sufficient. Existing RGB-D fusion methods often treat depth as an additional channel or a parallel visual branch and fuse it with RGB features in a fixed manner [6-8]. Such strategies do not fully consider that different language expressions require different degrees of geometric reasoning. Attribute-oriented expressions mainly depend on RGB appearance cues, whereas spatial relation expressions require stronger depth-based geometric support. A fixed fusion strategy may fail to fully exploit the depth modality when spatial reasoning is needed, and may also introduce irrelevant geometric noise when the query mainly depends on appearance attributes. In addition, many cross-modal fusion methods mainly align visual and linguistic features at a global semantic level, while explicit spatial relation words in the query are not sufficiently modeled. This limits the model's ability to selectively activate geometric information according to spatial semantics.

To address these issues, this paper proposes AD-SRE, an RGB-D referring expression grounding method based on adaptive depth modulation and spatial relation enhancement. The core idea is to avoid uniform depth fusion and instead selectively modulate geometric cues according to the semantic characteristics of the input expression. Specifically, a Query-Adaptive Depth Modulation module is introduced to dynamically adjust the contribution of depth features, so that depth information can be strengthened for spatial relation expressions and suppressed when it is less relevant to attribute-oriented descriptions. In addition, a Spatial Relation Enhancement module is designed to exploit relation-sensitive language cues and enhance visual responses associated with spatial semantics. Based on this framework, experiments are conducted on the SUN-Spot dataset to compare the proposed method with an RGB-Text baseline and a simple RGB-D-Text fusion baseline. Ablation studies and category-wise experiments are further designed to analyze the roles of depth information, adaptive modulation, and spatial relation enhancement in different types of referring expressions.

DOI: 10.54254/2755-2721/2026.GU33732

2. Related work

2.1. Referring expression grounding

Referring Expression Grounding, also known as Referring Expression Comprehension, aims to localize the target object referred to by a natural language expression. Different from traditional object detection, the input query is not a fixed category label but a free-form expression, which requires the model to understand object categories, visual attributes, positions, and relations between objects [1, 2]. Existing methods have gradually evolved from proposal-based and modular reasoning frameworks to end-to-end cross-modal alignment methods. Early methods usually relied on candidate region proposals and selected the best-matched region through image-text matching. For example, MAttNet decomposes expressions into different semantic components and models them with modular attention [2]. Although such methods provide interpretable reasoning, their performance is often affected by proposal quality and multi-stage pipeline complexity.

With the development of Transformer-based cross-modal modeling, end-to-end grounding frameworks have become increasingly popular. TransVG jointly encodes visual and textual features with a Transformer and directly predicts the target bounding box, which avoids complex proposal generation and post-processing [3]. However, most existing methods still mainly focus on RGB-text semantic alignment. For expressions involving front-back relations, distance, occlusion, and relative position, RGB appearance features alone may be insufficient to capture geometric structure and spatial relations. Therefore, how to introduce depth information and use it effectively remains important for spatial referring expression grounding.

2.2. RGB-D fusion

RGB-D data contains both appearance information and geometric information, and has been widely used in scene understanding, semantic segmentation, salient object detection, and robotic vision. Compared with RGB images, depth maps explicitly describe distance distribution, surface structure, and spatial layout, which can complement RGB features in scenes with occlusion, similar objects, or complex geometry [5, 6]. Existing RGB-D fusion strategies are usually divided into early fusion, middle fusion, and late fusion according to the fusion stage. Early fusion directly combines RGB and depth at the input or shallow feature level, middle fusion performs interaction in the intermediate feature space, and late fusion combines predictions at the decision stage.

Several representative works have explored multi-level feature injection, residual fusion, encoder-decoder fusion, and attention-based weighting to improve RGB-D representation learning [6-10]. These methods show that depth information can provide useful geometric cues for visual understanding. However, most of them are designed for purely visual tasks, such as semantic segmentation and scene understanding. Their fusion decisions are mainly driven by image content, while the role of language queries is rarely considered. In text-guided localization, different expressions require different degrees of geometric reasoning. Attribute-oriented expressions may rely more on RGB appearance cues, whereas spatial relation expressions require stronger depth-based geometric support. Therefore, query-aware depth fusion remains underexplored in RGB-D referring expression grounding.

DOI: 10.54254/2755-2721/2026.GU33732

2.3. Vision-language pretraining

The development of vision-language pretraining has greatly advanced tasks such as image-text matching, visual question answering, text-guided detection, and referring expression understanding. These methods learn transferable cross-modal representations from large-scale image-text pairs, enabling models to establish semantic correspondence between visual content and natural language at a general level rather than relying entirely on limited downstream annotations. As a result, pretrained vision-language models usually provide stronger semantic representation and transfer ability for downstream vision-language tasks.

Among existing vision-language pretraining methods, CLIP learns a shared vision-language embedding space through large-scale image-text contrastive learning and has shown strong open-vocabulary generalization ability [11]. Since its text encoder can capture category information, visual attributes, and scene semantics, CLIP has been widely used in downstream tasks such as detection, retrieval, and grounding. For the RGB-D text-guided object localization task studied in this paper, CLIP provides a stable way to extract textual representations, especially under relatively small-scale dataset settings.

However, CLIP and other general vision-language pretrained models mainly learn global semantic consistency between images and text. Although they are effective for category recognition and attribute matching, they still have limitations in fine-grained spatial relation modeling. For expressions such as left of, behind, closest to, and in front of, the model needs to reason about relative position, reference objects, and scene geometry. This indicates that general image-text alignment alone is not sufficient for spatial referring expression grounding. In RGB-D scenarios, it is therefore necessary to combine language semantics with geometric cues and to use query-aware mechanisms to control the contribution of depth information.

3. Methodology

3.1. Problem formulation

This paper studies the RGB-D text-guided object localization task in 2D scenes. Given an RGB image, a spatially aligned depth image, and a natural language description, the model aims to predict the 2D bounding box of the unique target referred to by the description. Compared with existing visual grounding methods that use only RGB images and text descriptions, this work further introduces the depth modality to enhance the model's ability to capture spatial relations, geometric hierarchy, and front-back structure.

Formally, let the input RGB image be I^{rgb}, the depth image be I^{d}, and the text description be T. The model outputs the target bounding box as

$$\hat{b} = \left(\hat{x}, \hat{y}, \hat{w}, \hat{h}\right) \tag{1}$$

where $\left(\hat{x}, \hat{y}\right)$ denotes the coordinates of the top-left corner of the predicted box, and $\left(\hat{w}, \hat{h}\right)$ denotes its width and height. During training, the ground-truth bounding box is given by

DOI: 10.54254/2755-2721/2026.GU33732

$$b = (x, y, w, h) \tag{2}$$

The model parameters are optimized jointly with a bounding box regression loss and an overlap loss.

3.2. Overall framework

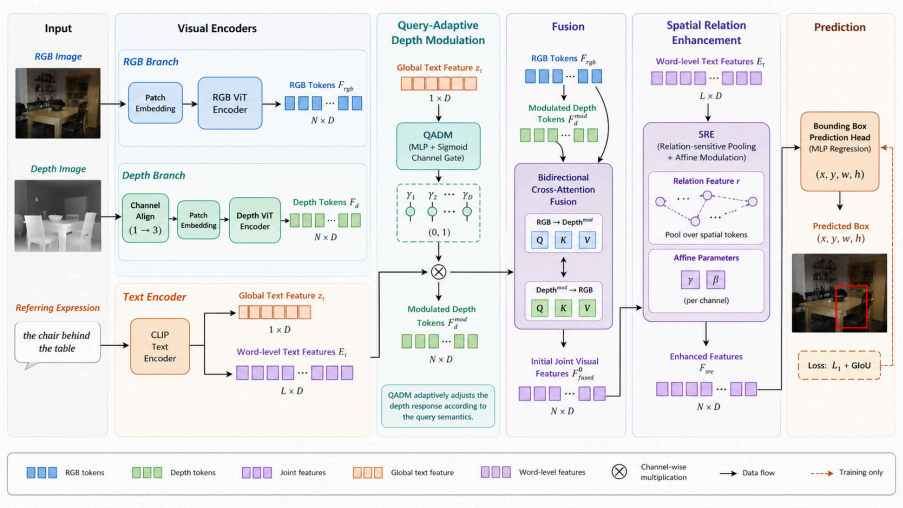


Figure 1. Overall framework of the proposed AD-SRE model

The overall architecture of the proposed AD-SRE model is shown in Figure 1. It consists of an RGB encoding branch, a depth encoding branch, a text encoding branch, a Query-Adaptive Depth Modulation (QADM) module, a Spatial Relation Enhancement (SRE) module, and a bounding box regression head. Different from conventional RGB-D fusion strategies with fixed feature interaction, the proposed framework dynamically modulates depth information according to the semantic content of the input query, especially spatial relation cues.

In the feature extraction stage, the RGB image and the depth image are mapped into visual token sequences, respectively. The text branch encodes the natural language description and produces both a global query representation and word-level representations. The global query representation is used to guide depth modulation, while the word-level features provide relation-sensitive cues for spatial enhancement.

Based on these features, the QADM module adjusts the response strength of the depth branch according to query semantics. The modulated depth features are then fused with RGB features through a cross-modal attention mechanism to obtain a joint visual representation. Finally, the SRE module further enhances the fused features using spatial linguistic cues, and the enhanced multimodal representation is fed into the bounding box regression head to produce the final localization result.

3.3. Visual grounding pipeline

Image Encoder. The image encoder takes the RGB image and the depth image as inputs, respectively. To ensure that the two visual modalities share a consistent representation form for later

DOI: 10.54254/2755-2721/2026.GU33732

fusion, this paper adopts a unified Transformer-based visual encoding framework to encode the RGB image and the depth image separately [12].

For the input RGB image, patch embedding is first applied to divide it into image patches of fixed size, which are then mapped into a visual token sequence:

$$F^{rgb}=\{f_1^{rgb},f_2^{rgb},\ldots,f_N^{rgb}\}\in R^{N\times C} \tag{3}$$

For the depth image, its single-channel nature is first taken into account. A lightweight alignment strategy is then used to expand it into a three-channel form while preserving the same spatial resolution as the RGB image. After that, the depth image is passed through a patch embedding module with the same structure as the RGB branch to obtain depth tokens:

$$F^{d}=\{f_1^{d},f_2^{d},\ldots,f_N^{d}\}\in R^{N\times C} \tag{4}$$

where N denotes the number of visual tokens, and C denotes the feature dimension.

Text Semantic Encoder. The text branch uses a pretrained CLIP text encoder to encode the input natural language description T. This produces a global text feature $z_t\in R^C$, as well as word-level text features

$$E_t=\{e_1,e_2,\ldots,e_L\}\in R^{L\times C} \tag{5}$$

where L denotes the number of text tokens. The global text feature is mainly used for adaptive modulation of the depth modality, while the word-level text features are used for subsequent spatial relation modeling.

Query-Adaptive Depth Modulation (QADM). To model the different levels of dependence on geometric information across language queries, a QADM module is introduced to perform text-driven channel-wise modulation on depth features. A channel gating vector is generated from the global text feature z_t as

$$g=\sigma(MLP(z_t))\in R^C \tag{6}$$

where MLP(·) denotes a two-layer multilayer perceptron, and σ(·) denotes the Sigmoid activation function.

Then, the channel response of each depth token is rescaled to obtain the modulated depth features:

$$\tilde{f}_n^{d}=g\odot f_n^{d},\ n=1,2,\ldots,N \tag{7}$$

DOI: 10.54254/2755-2721/2026.GU33732

This process adaptively controls the response strength of the depth modality across different channels based on the query semantics. When the expression mainly relies on appearance attributes such as color, category, or texture, the model suppresses geometric responses in the depth branch that are less relevant to localization. When spatial relations such as front-back relations, relative distance, left side relations, or proximity are involved, the model strengthens the depth responses that encode geometric structure. In this way, more discriminative spatial cues are provided for subsequent fusion.

Cross Fusion Layer. After depth modulation, a cross-modal attention mechanism is adopted to fuse RGB features with the modulated depth features [13]. Compared with direct addition or concatenation, cross-modal attention explicitly models the correspondence between the two visual modalities, enabling fine-grained alignment between appearance information and geometric structure.

Specifically, RGB features and the modulated depth features are projected into queries, keys, and values, respectively:

$$\mathrm{Q}^{\mathrm{rgb}}=\mathrm{F}^{\mathrm{rgb}}\mathrm{W}_{\mathrm{Q}}^{\mathrm{rgb}},\ \mathrm{K}^{\mathrm{rgb}}=\mathrm{F}^{\mathrm{rgb}}\mathrm{W}_{\mathrm{K}}^{\mathrm{rgb}},\ \mathrm{V}^{\mathrm{rgb}}=\mathrm{F}^{\mathrm{rgb}}\mathrm{W}_{\mathrm{V}}^{\mathrm{rgb}} \tag{8}$$

$$\mathrm{Q}^{\mathrm{d}}=\widetilde{\mathrm{F}}^{\mathrm{d}}\mathrm{W}_{\mathrm{Q}}^{\mathrm{d}},\ \mathrm{K}^{\mathrm{d}}=\widetilde{\mathrm{F}}^{\mathrm{d}}\mathrm{W}_{\mathrm{K}}^{\mathrm{d}},\ \mathrm{V}^{\mathrm{d}}=\widetilde{\mathrm{F}}^{\mathrm{d}}\mathrm{W}_{\mathrm{V}}^{\mathrm{d}} \tag{9}$$

where $\mathrm{W}_{\mathrm{Q}}^{\mathrm{rgb}},\mathrm{W}_{\mathrm{K}}^{\mathrm{rgb}},\mathrm{W}_{\mathrm{V}}^{\mathrm{rgb}},\mathrm{W}_{\mathrm{Q}}^{\mathrm{d}},\mathrm{W}_{\mathrm{K}}^{\mathrm{d}}$ and $\mathrm{W}_{\mathrm{V}}^{\mathrm{d}}$ are learnable parameter matrices.

Then, with RGB features as queries and depth features as keys and values, the depth-guided enhancement of RGB features is computed as

$$\mathrm{F}^{\mathrm{rgb}\leftarrow\mathrm{d}}=\mathrm{softmax}\left(\frac{\mathrm{Q}^{\mathrm{rgb}}(\mathrm{K}^{\mathrm{d}})^{\top}}{\sqrt{\mathrm{C}}}\right)\mathrm{V}^{\mathrm{d}} \tag{10}$$

Similarly, with depth features as queries and RGB features as keys and values, the RGB-guided enhancement of depth features is computed as

$$\mathrm{F}^{\mathrm{d}\leftarrow\mathrm{rgb}}=\mathrm{softmax}\left(\frac{\mathrm{Q}^{\mathrm{d}}(\mathrm{K}^{\mathrm{rgb}})^{\top}}{\sqrt{\mathrm{C}}}\right)\mathrm{V}^{\mathrm{rgb}} \tag{11}$$

Finally, the bidirectional interaction results are fused to obtain the initial joint visual features:

$$\mathrm{F}_{0}^{\mathrm{fused}}=\mathrm{Concat}\left(\mathrm{F}^{\mathrm{rgb}\leftarrow\mathrm{d}},\mathrm{F}^{\mathrm{d}\leftarrow\mathrm{rgb}}\right)\mathrm{W}_{\mathrm{f}} \tag{12}$$

where W_{f} is a linear projection matrix used to map the concatenated features back to a unified dimension.

DOI: 10.54254/2755-2721/2026.GU33732

Through this process, complementary modeling between RGB and depth can be achieved at the visual feature level. As a result, appearance attributes and geometric structure can be more fully integrated into a joint representation.

Spatial Relation Enhancement (SRE). After obtaining the joint visual features, a SRE module is further introduced to explicitly model spatial relation semantics in the text. Since not all words in the text directly contribute to spatial reasoning, relation-sensitive semantics are extracted from the word-level text features and encoded into a relation feature r. For simplicity, the text token features that contain spatial relation cues are pooled in a weighted manner to obtain

$$\mathrm{r}=\mathrm{Pool}(\mathrm{E_t})\in\mathrm{R^C} \tag{13}$$

where $\mathrm{Pool}(\cdot)$ denotes the relation-sensitive pooling operation.

Based on the relation feature r, a relation-aware scaling term and a bias term are further generated as

$$\gamma=\sigma(\mathrm{W}_\gamma\mathrm{r}),\ \beta=\mathrm{W}_\beta\mathrm{r} \tag{14}$$

where W_γ and W_β are learnable parameters.

Then, affine modulation is applied to the joint visual features to obtain the enhanced visual features:

$$\mathrm{F^{sre}}=(1+\gamma)\odot\mathrm{F_0^{fused}}+\beta \tag{15}$$

In this formulation, γ controls the enhancement strength of different feature dimensions, while β provides an offset associated with spatial relations.This module is essentially a feature enhancement mechanism driven by spatial semantics. Its role is not to reconstruct the basic fusion between visual modalities. Instead, it further highlights discriminative responses related to relative position, distance relations, and scene layout based on the existing joint visual features. In this way, the model can better understand complex spatial expressions.

3.4. Bounding box prediction head and loss functions

After obtaining the final enhanced multimodal features $\mathrm{F^{sre}}$, a bounding box prediction head is used to output the target box coordinates. Specifically, the fused features are first aggregated and then mapped into the bounding box parameter space through a multilayer perceptron. In this way, the target 2D bounding box $\widehat{\mathrm{b}}$ is directly regressed [3].

During training, the model is optimized with a combination of bounding box regression loss and overlap loss. The regression branch uses the L1 loss:

$$\mathrm{L_{bbox}}=\|\widehat{\mathrm{b}}-\mathrm{b}\|_1 \tag{16}$$

DOI: 10.54254/2755-2721/2026.GU33732

To further constrain the overlap between the predicted box and the ground-truth box, the GIoU loss is introduced [14]:

$$L_{giou}=1\text{-}GIoU(\hat{b},b) \tag{17}$$

The total loss is defined as

$$L=\lambda_1 L_{bbox}+\lambda_2 L_{giou} \tag{18}$$

where λ_1 and λ_2 are the loss weighting coefficients.

With this design, the model can make full use of RGB appearance information, depth-based geometric information, and textual semantic constraints within an end-to-end framework, and directly output the target location corresponding to the input description.

4. Experiments

4.1. Dataset and experimental settings

Dataset. Experiments are conducted on the SUN-Spot dataset. SUN-Spot is an RGB-D dataset designed for spatial Referring Expression Grounding. It contains 1,948 RGB-D images and 7,987 referring expressions [4]. Compared with conventional referring expression datasets, it includes richer spatial prepositions and more crowded scenes. Therefore, it is particularly suitable for evaluating the model's ability in spatial semantic understanding and geometric reasoning.

Experimental settings. According to the task setting in this paper, each sample consists of an RGB image, its aligned depth image, a text description, and a target 2D bounding box. At the input stage, both RGB and depth images are resized to 640×640. The depth image is first aligned in size and then expanded into three channels by channel duplication, so that it is compatible with the visual backbone. On the text side, the original single-sentence input is preserved, with a maximum length of 40 tokens. Tokens beyond this length are truncated, while shorter sequences are padded. A masking mechanism is used to ignore padded positions during computation.

Implementation details. The visual and text branches use a ViT encoder and a CLIP text encoder, respectively. All models are implemented in PyTorch and trained and evaluated on an NVIDIA GeForce RTX 3090 GPU with 24 GB memory. During training, the AdamW optimizer is used [15]. The batch size is set to 16, and the model is trained for 40 epochs. The learning rate for the visual and text backbones is set to 1×10^{-5}, while the learning rate for newly introduced components, including the fusion modules and prediction head, is set to 1×10^{-4}. The weight decay is 1×10^{-4}. To improve training stability, the dropout rate in Transformer-related modules is set to 0.1, and the learning rate is decayed to one tenth of its original value after the 30th epoch. The loss function consists of the L1 loss and the GIoU loss, which jointly constrain the localization error and the overall overlap quality of the predicted box. For a fair comparison, all compared methods are trained under the same number of epochs, input resolution, and optimization strategy.

Evaluation metrics. IoU, Acc@0.5, and Acc@0.75 are adopted as the evaluation metrics. IoU measures the spatial overlap between the predicted box and the ground-truth box, and is defined as:

DOI: 10.54254/2755-2721/2026.GU33732

$$\mathrm{IoU}=\frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|} \tag{19}$$

where B_p denotes the predicted bounding box, B_{gt} denotes the ground-truth bounding box, and $|\cdot|$ denotes the area of a region. In addition, Acc@0.5 and Acc@0.75 are also reported. A sample is considered correctly localized when the IoU between the predicted box and the ground-truth box is greater than 0.5 or 0.75, respectively. These two metrics are used to evaluate model performance under different localization accuracy requirements.

4.2. Compared methods

To verify the effectiveness of the proposed method, three representative methods are compared: TransVG, Naive RGB-D + Text, and AD-SRE.

TransVG. As a classical RGB-Text grounding baseline, TransVG uses only the RGB image and the text description for target localization. It adopts an end-to-end Transformer-based framework and directly regresses the target bounding box [3]. This baseline is used to evaluate the performance of RGB-Text grounding without depth information.

Naive RGB-D + Text. Naive RGB-D + Text is constructed as a simple tri-modal baseline. It introduces the depth modality into the RGB-Text framework, but does not include the Query-Adaptive Depth Modulation module or the Spatial Relation Enhancement module. This baseline is used to examine whether simply adding depth information can improve spatial grounding performance.

AD-SRE. AD-SRE is the proposed full model. Compared with Naive RGB-D + Text, it further introduces query-adaptive depth modulation and spatial relation enhancement, so that depth information can be dynamically used according to query semantics and spatial relation cues can be explicitly strengthened.

Based on these methods, the experiments mainly focus on the following questions:

• whether pure RGB-Text grounding methods show limitations in spatial referring expression scenarios;

• whether introducing the depth modality can improve localization performance over RGB-Text grounding;

• whether the proposed spatial relation enhancement and adaptive depth modulation mechanisms can bring further gains beyond simple RGB-D fusion.

Table 1. Overall performance comparison of different methods on the SUN-Spot dataset

methods	Input Modalities	IoU	Acc@0.5	Acc@0.75
TransVG	RGB+Text	0.48	0.64	0.35
Naive RGB-D + Text	RGB+Depth+Text	0.50	0.66	0.36
AD-SRE	RGB+Depth+Text	0.52	0.68	0.39

As shown in Table 1, TransVG achieves relatively stable baseline performance when only RGB images and text descriptions are used as inputs. This indicates that end-to-end vision-language modeling is effective for Referring Expression Grounding. However, the SUN-Spot dataset contains many samples with spatial relation descriptions, object occlusion, and interference from similar

DOI: 10.54254/2755-2721/2026.GU33732

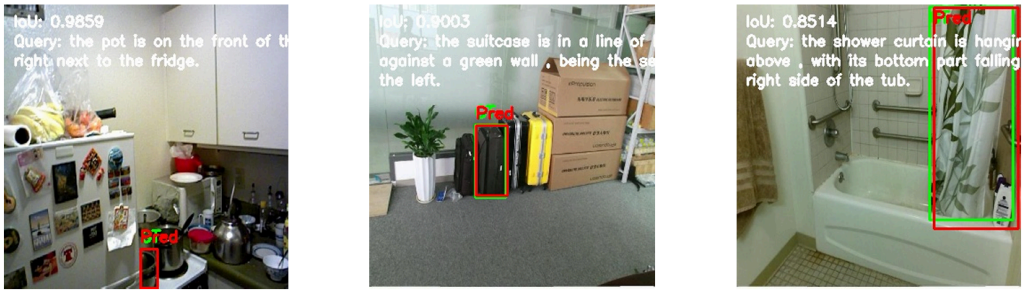
objects. As a result, pure RGB-Text methods still have clear limitations in modeling geometric structure and reasoning about relative position, which restricts their overall performance.

After introducing the depth modality, Naive RGB-D + Text achieves consistent improvements over TransVG on all three metrics, with absolute gains of 0.02 in IoU, 0.02 in Acc@0.5, and 0.01 in Acc@0.75. These results suggest that depth information plays a positive role in spatial Referring Expression Grounding. It can compensate, to some extent, for the limitations of 2D appearance features in handling front-back relations, relative distance, and local occlusion.

The proposed AD-SRE further achieves the best results on all three metrics. Compared with Naive RGB-D + Text, AD-SRE brings additional absolute gains of 0.02 in IoU, 0.02 in Acc@0.5, and 0.03 in Acc@0.75. Among them, the improvement on Acc@0.75 is the most evident. This indicates that the proposed method not only improves coarse localization performance, but also achieves better bounding box regression under stricter localization requirements.

These comparison results show that simply introducing the depth modality can bring certain performance gains, but the improvement remains limited. Only by further incorporating spatial semantic information from the text and adaptively adjusting the contribution of depth features can the advantages of RGB-D-Text tri-modal modeling be more fully exploited. This also verifies the effectiveness and rationality of the proposed AD-SRE model from the perspective of overall performance.

To further provide an intuitive view of the localization behavior of the proposed method in typical scenarios, some visualization results of AD-SRE on the SUN-Spot dataset are presented.



(a) (b) (c)

Figure 2. Visualization results of the proposed AD-SRE in typical localization scenarios

As shown in Figure 2, the proposed method can localize the target object accurately in scenes involving front-back relations, proximity relations, and complex background interference. In particular, when there are many similar objects or the spatial relations are more complicated, the model can still predict the target box according to the key spatial semantics in the text description. This demonstrates that the proposed method has good robustness and interpretability in complex scenes.

4.3. Ablation study

To further analyze the contribution of each proposed module, ablation experiments are conducted based on the full AD-SRE model. Four model settings are compared. Naive RGB-D + Text removes both QADM and SRE, retaining only the basic RGB-D-Text fusion structure. w/o Modulation removes QADM while keeping SRE, and w/o Relation removes SRE while keeping QADM. The full AD-SRE model retains both modules. The results are shown in Table 2.

DOI: 10.54254/2755-2721/2026.GU33732

Table 2. Ablation results of different modules in AD-SRE

methods	QADM	SRE	IoU	Acc@0.5	Acc@0.75
Naive RGB-D + Text			0.50	0.66	0.36
w/o Modulation		✓	0.51	0.67	0.38
w/o Relation	✓		0.51	0.66	0.37
AD-SRE	✓	✓	0.52	0.68	0.39

As shown in Table 2, both QADM and SRE contribute positively to RGB-D text-guided localization. Compared with the Naive RGB-D + Text baseline, w/o Modulation achieves higher performance on all three metrics, especially on Acc@0.75. This indicates that explicit spatial relation enhancement helps the model better capture relative position semantics and improves localization quality under stricter evaluation criteria.

The w/o Relation setting also improves IoU and Acc@0.75 compared with the baseline, although the gain is relatively smaller. This suggests that query-adaptive depth modulation can reduce the influence of irrelevant geometric information and enhance useful depth cues according to the semantic characteristics of the input expression. Since QADM mainly adjusts the contribution of the depth modality from a global query perspective, its effect is more reflected in improving bounding box quality rather than producing a large increase in coarse localization accuracy.

The full AD-SRE model achieves the best performance on all metrics, indicating that the two modules are complementary rather than redundant. QADM focuses on query-aware control of depth features, while SRE further strengthens visual responses related to spatial relations after multimodal fusion. Their combination enables the model to make better use of RGB appearance, depth geometry, and textual semantics, which verifies the effectiveness of the proposed design.

4.4. Category-wise experiments

Since the proposed method is mainly designed to improve spatial semantic understanding, overall performance alone is not sufficient to fully evaluate its effectiveness. Therefore, category-wise experiments are conducted according to the type of text description. The expressions are divided into three categories: attribute expressions, spatial relation expressions, and compound expressions. Attribute expressions mainly rely on visual cues such as color, appearance, or category. Spatial relation expressions depend on relation words such as left of, behind, closest, and between. Compound expressions contain both attribute constraints and spatial relation constraints, and therefore require the model to perform both semantic alignment and geometric reasoning. The results are shown in Table 3.

Table 3. Localization performance under different types of text descriptions

type	Naive RGB-D + Text			AD-SRE		
	IoU	Acc@0.5	Acc@0.75	IoU	Acc@0.5	Acc@0.75
Attribute	0.54	0.70	0.41	0.55	0.71	0.41
Spatial relation	0.52	0.69	0.39	0.54	0.72	0.42
Compound	0.45	0.59	0.30	0.47	0.61	0.32

DOI: 10.54254/2755-2721/2026.GU33732

As shown in Table 3, different types of expressions present different levels of difficulty. Attribute expressions achieve the best overall performance, indicating that explicit appearance cues are relatively easier for the model to capture. Spatial relation expressions are more challenging because the model needs to reason about the relative positions between objects. Compound expressions obtain the lowest performance, suggesting that combining attribute recognition with spatial relation reasoning imposes higher requirements on multimodal alignment and scene understanding.

Compared with Naive RGB-D + Text, AD-SRE improves performance across all three expression types, but the degree of improvement varies. For attribute expressions, the improvement is relatively limited. This is consistent with the nature of such expressions, because they mainly depend on RGB appearance information and require less geometric reasoning. In this case, the additional benefit of adaptive depth modulation and spatial relation enhancement is naturally smaller.

For spatial relation expressions, AD-SRE brings more evident gains, especially on Acc@0.5 and Acc@0.75. This shows that the proposed method can better exploit depth information and spatial linguistic cues when the query requires relative position reasoning. The result also supports the motivation of introducing SRE, since explicit relation-aware enhancement helps the model focus on visual regions that are more relevant to spatial semantics.

For compound expressions, AD-SRE also improves all metrics over the baseline. Although this category remains the most difficult, the performance gain indicates that the proposed method can better integrate appearance cues, geometric cues, and textual constraints when multiple types of semantic information are involved. Overall, the category-wise results are consistent with the design goal of AD-SRE. The advantage of the proposed method is relatively small on attribute-dominant expressions, but becomes more evident on expressions that require spatial relation understanding or joint attribute-spatial reasoning.

To further complement the category-wise analysis, visualization results of AD-SRE in spatial relation and compound expression scenarios are presented in Figure 3.

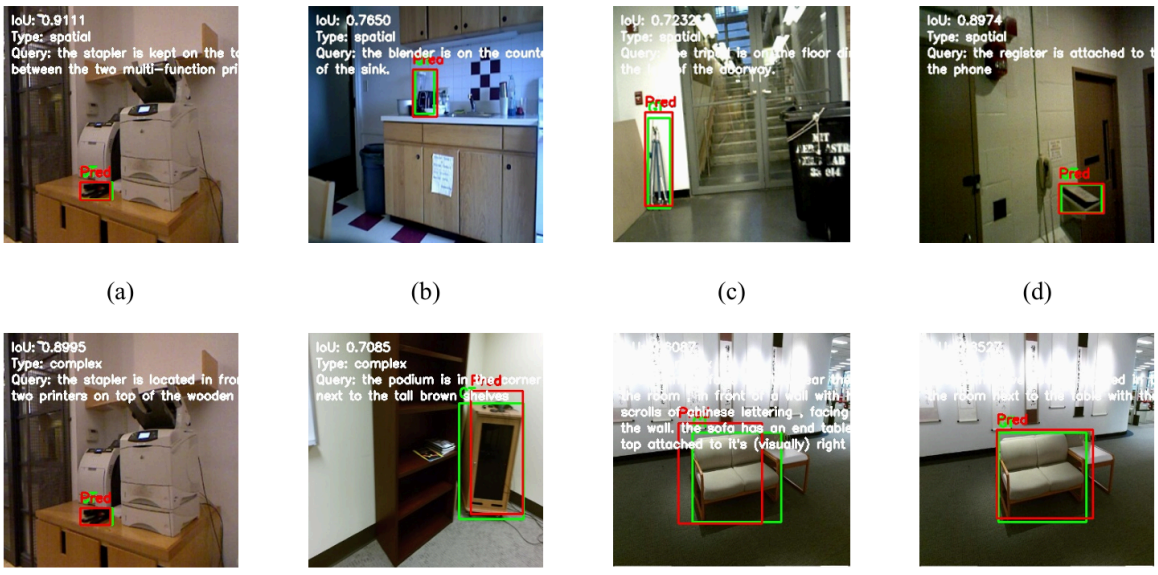


Figure 3. Visualization results of AD-SRE for complex expressions of different types

As shown in Figure 3, AD-SRE can effectively localize the target object in both spatial relation expression scenarios and compound expression scenarios. For the spatial relation examples shown in

DOI: 10.54254/2755-2721/2026.GU33732

subfigures (a)–(d), the model can effectively exploit relative positional information in the scene and complete localization according to relation words such as between, on, and attached to. For the compound expression examples shown in subfigures (e)–(h), the model can still identify the target region when both attribute constraints and spatial constraints are involved. Although some deviation between the predicted box and the target boundary still exists in a few difficult samples, the overall results indicate that the proposed method has good robustness and adaptability in complex spatial semantic understanding tasks.

5. Conclusion

This paper addresses the problems of insufficient spatial relation understanding and inadequate use of depth information in Referring Expression Grounding for RGB-D scenes. A tri-modal localization framework based on spatial semantic enhancement and adaptive depth modulation is proposed. By introducing the Query-Adaptive Depth Modulation module and the Spatial Relation Enhancement module, the proposed method enables text-guided dynamic adjustment of depth information and enhanced modeling of spatial relations. Experimental results show that the proposed method achieves better overall localization performance and performs particularly well in challenging spatial expression scenarios, especially on spatial relation expressions and compound expressions. The method can integrate RGB, Depth, and Text more effectively, and provides useful insights for vision-language grounding research in complex scenes.

References

[1] Yanyuan Qiao, Chaorui Deng, Qi Wu. Referring Expression Comprehension: A Survey of Methods and Datasets. , 2020.
[2] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, Tamara L. Berg. MAttNet: Modular Attention Network for Referring Expression Comprehension. , 2018.
[3] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, Houqiang Li. TransVG: End-to-End Visual Grounding with Transformers. , 2022.
[4] Mauceri, Cecilia, Palmer, Martha, Heckman, Christoffer. SUN-Spot: An RGB-D Dataset With Spatial Referring Expressions. //Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2019.
[5] Song, Shuran, Lichtenberg, Samuel P., Xiao, Jianxiong. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
[6] Hazirbas, Caner, Ma, Lingni, Domokos, Csaba, Cremers, Daniel. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. //Asian Conference on Computer Vision, 2016.
[7] Park, Seonwook J., Hong, Kye-Hyeon, Lee, Sanghoon, Lee, Jinsoo, Cho, Minsu, Kim, Suha Kwak. RDFNet: RGB-D Multi-Level Residual Feature Fusion for Indoor Semantic Segmentation. //Proceedings of the IEEE International Conference on Computer Vision, 2017.
[8] Jiang, Jindong, Zheng, Lunan, Luo, Fei, Zhang, Zhijun. RedNet: Residual Encoder-Decoder Network for Indoor RGB-D Semantic Segmentation. arXiv preprint arXiv: 1806.01054, 2018.
[9] Hu, Xinxin, Yang, Kailun, Fei, Lei, Wang, Kaiwei. ACNet: Attention Based Network to Exploit Complementary Features for RGBD Semantic Segmentation. //2019 IEEE International Conference on Image Processing (ICIP), 2019.
[10] Seichter, Daniel, Köhler. Efficient RGB-D Semantic Segmentation for Indoor Scene Analysis. //2021 IEEE International Conference on Robotics and Automation (ICRA), 2021.
[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. , 2021.
[12] Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, Weissenborn, Dirk, Zhai, Xiaohua, Unterthiner, Thomas, Dehghani, Mostafa, Minderer, Matthias, Heigold, Georg, Gelly, Sylvain, others. An Image is Worth 16x16

DOI: 10.54254/2755-2721/2026.GU33732

Words: Transformers for Image Recognition at Scale. //International Conference on Learning Representations, 2021.

[13] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, Polosukhin, Illia Attention Is All You Need. //Advances in Neural Information Processing Systems, 2017.

[14] Rezatofighi, Hamid, Tsoi, Nathan, Gwak, JunYoung, Sadeghian, Amir, Reid, Ian, Savarese, Silvio. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[15] Loshchilov, Ilya, Hutter, Frank. Decoupled Weight Decay Regularization. //International Conference on Learning Representations, 2019.