Supplementary Material for ORD: Object-Relation Decoupling for Generalized 3D Visual Grounding

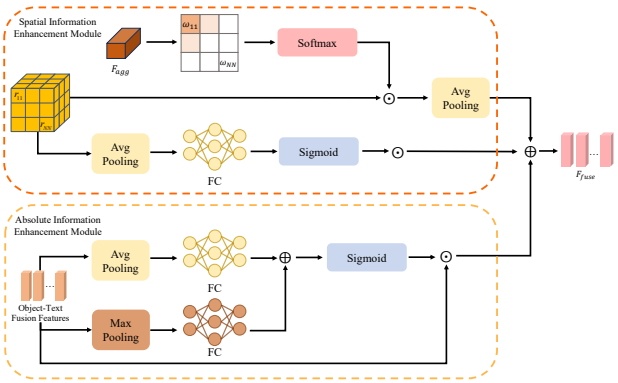

Figure 1. Architecture of the proposed dual-branch information enhancement module. The spatial-information enhancement branch (top) computes spatial-information attention W^R by interacting local relative spatial features with the global aggregated feature F_{agg} and reweighting spatial relations, while the absolute-information enhancement branch (bottom) derives channel-wise absolute-information attention W^O from object–text fused features via average/max pooling and shared FC layers. The two attentions are applied to recalibrate spatial and absolute cues and are fused to produce the enhanced representation F_{fuse}.

1. Method

1.1. Absolute Information Enhancement Module and Spatial Information Enhancement Module

In 3D scenes, spatial relations among objects encompass not only relative positions but also their absolute coordinates within the environment. In practice, however, such structural relations can be sparse, posing a challenge for model capacity. To effectively capture these sparse relations and strengthen spatial reasoning, we design absolute information enhancement module and spatial information enhancement module, as illustrated in Fig. 1. The module comprises a spatial-information enhancement branch and an absolute-information enhancement branch. By learning adaptive weights, it aggregates relative spatial features in a weighted manner, thereby amplifying spatial cues most relevant to the query. We employ a dual-weighting scheme that combines W^R (spatial-information attention) and W^O (absolute-information attention): W^R enables fine-grained modeling of multi-level dependencies among objects in space, while W^O enhances absolute positional cues of objects. This explicit modeling of dependencies between the target and the anchor improves the expressivity of the fused representation.

We first fuse linguistic features with 3D object features to obtain cross-modal fused embeddings $F_{\text{fused}} \in \mathbb{R}^{N_{\text{obj}} \times d_{\text{in}}}$. To capture global context, we prepend an aggregation token T_1 to the input sequence and feed $[T_1; F_{\text{fused}}]$ into a Transformer-based encoder for global interaction, yielding the aggregated feature F_{agg}:

$$F_{\text{agg}} = \mathrm{T}\big([T_1; F_{\text{fused}}]\big)[0] \in \mathbb{R}^{d_{\text{in}}}. \tag{1}$$

Qu et al. [2] showed that feature-product operations can strengthen semantically similar features, enabling the model to learn higher-order latent patterns. Inspired by this, to more precisely capture relative spatial relations between the target and the anchor, we apply a feature-product interaction between local spatial features and the global aggregated feature to obtain a preliminary spatial-information attention. We then combine this with fused object features via a self-enhancement branch that dynamically models both spatial relations and absolute positional cues. Concretely, we compute the spatial interaction weights

$$w_{ij}^R = F_{i,j} \cdot F_{\text{agg}}, \tag{2}$$

and normalize them with a softmax to obtain the spatial-information attention W^R:

$$W_{ij}^R = \mathrm{softmax}(w_{ij}^R). \tag{3}$$

Finally, we reweight the original spatial features using W^R and optimize the alignment with a predefined attention loss, yielding the spatially weighted features

$$F_{i,j}^{\text{w}} = W_{ij}^R \cdot F_{i,j}. \tag{4}$$

To highlight the crucial role of absolute information during interaction, we further introduce an absolute-information attention mechanism. Specifically, we obtain

channel-wise attention from F_{fused} via pooling at the channel level and use it to adaptively recalibrate each feature channel. In this way, salient attribute channels are emphasized while irrelevant or redundant responses are suppressed, forming the absolute-information attention W^O that complements W^R in absolute information enhancement module and spatial information enhancement module.

Concretely, we compute channel-wise descriptors from the fused features F_{fused} using max pooling (MP) and average pooling (AP). MP takes the maximum over all spatial points within each channel to obtain a channel descriptor, whereas AP takes the mean within each channel. The two descriptors are then passed through a lightweight two-layer perceptron (shared across channels) to produce per-channel confidences. Finally, we fuse the MP/AP branches to obtain the absolute-information attention W^O:

$$W^O = \sigma\Big(FC_2\big(\delta\big(FC_1(\text{AP}(\mathbf{F}_{\text{fused}}))\big)\big) + FC_2\big(\delta\big(FC_1(\text{MP}(\mathbf{F}_{\text{fused}}))\big)\big)\Big). \quad (5)$$

where σ is the sigmoid activation and δ denotes ReLU. The projection layers are parameterized as $FC_1 \in \mathbb{R}^{\frac{d}{2}\times d}$ and $FC_2 \in \mathbb{R}^{d\times\frac{d}{2}}$, with d the channel dimensionality.

We then apply the absolute-information attention to F_{fused} to yield the channel-recalibrated features F^{O}:

$$F^{\text{O}} = W^O \odot F_{\text{fused}}, \quad (6)$$

where $\odot$ denotes channel-wise (Hadamard) multiplication. Finally, we fuse the absolute-information branch with the spatially weighted features F^{w} to obtain the aggregated representation F_{fuse}:

$$F_{\text{fuse}} = F^{\text{O}} + F^{\text{w}}. \quad (7)$$

In summary, the dual-weighting scheme introduces complementary attentional weights at both the spatial-relation level and the absolute-position (channel) level. This jointly emphasizes salient, modality-relevant cues while suppressing irrelevant responses, mitigating the impact of sparse and complex spatial signals and thereby strengthening the multimodal fused representation.

1.2. Loss

Suppose the original sentence contains k anchor placeholders; concatenating these k anchors with one target yields a label sequence of length $S = k + 1$. After textual aggregation, the classifier produces $\mathbf{P}_{\text{sent}} \in \mathbb{R}^{S\times C}$ over C classes, with ground-truth indices $\mathbf{T}_{\text{sent}} \in \mathbb{R}^{S}$. The sentence loss is

$$\mathcal{L}_{\text{Sent}} = -\frac{1}{S}\sum_{i=1}^{S}\log\frac{\exp\big(\mathbf{P}_{\text{sent}}[\,i,\,T_{\text{sent},i}\,]\big)}{\sum_{j=1}^{C}\exp\big(\mathbf{P}_{\text{sent}}[\,i,\,j\,]\big)}. \quad (8)$$

For each object pair (i, j), we guide the model to learn a sparse distribution over inter-object spatial relations. Concretely, starting from the spatial-attention tensor $W^R \in \mathbb{R}^{N^x_{\text{obj}}\times N^y_{\text{obj}}\times d}$ in Eq. (3), we first project the d-dimensional features to scalars via a fully connected layer and normalize them with a softmax to obtain a relation probability map P^R:

$$P^R = \text{softmax}\big(FC(W^R)\big), \qquad P^R \in \mathbb{R}^{N^x_{\text{obj}}\times N^y_{\text{obj}}}. \quad (9)$$

Here, P^R_{ij} is the normalized probability that object i (target) forms a relation with object j (anchor). To supervise this distribution, we use a binary relation matrix $Y \in \{0,1\}^{N^x_{\text{obj}}\times N^y_{\text{obj}}}$, where $Y_{ij} = 1$ iff i is the target and j is its corresponding anchor; otherwise $Y_{ij} = 0$. The relation regression loss is

$$\mathcal{L}_r = \frac{1}{(N_{\text{obj}})^2}\sum_i\sum_{j\neq i}\text{BCE}\big(P^R_{ij},\,Y_{ij}\big). \quad (10)$$

Let $P_{\text{obj}} = F' \cdot L$ be the logits from object features F' and class embeddings L, and let T_{obj} be the one-hot labels; then

$$\mathcal{L}_{\text{Object}} = -\sum_{j=1}^{N_{\text{obj}}}\sum_{c=1}^{C}T_{\text{obj}}(j,c)\,\log\left(\frac{\exp\big(P_{\text{obj}}(j,c)\big)}{\sum_{c'=1}^{C}\exp\big(P_{\text{obj}}(j,c')\big)}\right). \quad (11)$$

2. Experiment

2.1. Evaluation Metrics

For the ReferIt3D [1] dataset, the candidate set is derived from ground-truth annotations and represented as object identities rather than conventional bounding boxes. Performance is measured by accuracy (**acc**), i.e., the proportion of predictions whose object identity matches the ground-truth label. In addition, the dataset is further stratified by difficulty and viewpoint into *easy/hard*, *view-dependent/view-independent*, and an *overall* category to provide a comprehensive assessment of model performance.

2.2. Qualitative Results

As illustrated in Fig. 3, we visualize our method's predictions of target and anchor positions under complex textual descriptions. Colored boxes denote different positional information: blue indicates the model-predicted target location, green marks the Ground Truth, and yellow denotes the predicted anchor. In each example, the panels from left to right are: our predicted target, the overall prediction of our method, the Ground Truth location, and the predicted anchor. Evidently, when confronted with viewpoint-dependent phrasing or multi-anchor structures, the proposed approach accurately captures the spatial relation between the target and the correct anchor, thereby avoiding

Figure 2. Qualitative comparison of grounding results with existing methods on 3D visual grounding benchmarks.

target shifts caused by incorrect referents. For instance, given the text "the lamp to the right when looking from the fireplace," methods that fail to identify "fireplace" as the viewpoint reference misinterpret "right" (e.g., predict-

Figure 3. Visualization of text-conditioned predictions of target and anchor positions.

ing the lamp to the right of the bed), whereas our method grounds the target at the semantically correct location under the intended viewpoint. Overall, these visualizations substantiate the advantage of centering target–anchor spatial relations: they enhance spatial perception and improve the robustness of cross-modal alignment.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision – ECCV 2020*, pages 422–440. Springer,

2020. 2

[2] Yanru Qu, Han Cai, Kan Ren, Weinan Zhang, Yong Yu, Ying Wen, and Jun Wang. Product-based neural networks for user response prediction. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1149–1154. IEEE, 2016. 1