SPATIAL RELATIONS ARE NOT GEOMETRY ALONE

Anonymous authors
Paper under double-blind review

ABSTRACT

3D spatial reasoning requires effective representations of inter-object spatial relations. Existing methods typically construct spatial relations from predefined geometric schemes, while incorporating query semantics and reference frames at stages outside relation formation. This separation leads to two fundamental limitations: geometric measurements are fixed before the model knows what spatial evidence the query requires, while directional semantics are not defined directly under the relevant reference frame. Consequently, query-critical geometric cues can be irreversibly omitted, while the mapping from reference frames to directional semantics is poorly constrained and must be learned implicitly. Downstream reasoning is forced to operate on incomplete and ambiguous spatial representations, directly degrading reasoning performance. We therefore propose FROG, a unified relation-centric framework for 3D spatial reasoning, built on the principle that spatial relations should be formed jointly from reFerence fRames, geOmetry, and lanGuage. An order-aware geometry controller captures the relational semantics of the query and drives language-adaptive spatial measurement, adapting measurement directions and relative weights across spatial scales according to the query. In parallel, a relation-centric reference-frame mechanism uses observer orientation to define frame-dependent directional semantics directly during object–object relation formation, without modifying the underlying object or scene representations. By jointly integrating geometry, language, and reference frames into relation formation, FROG effectively mitigates the omission of query-critical geometric cues and the ambiguity of directional semantics, providing downstream reasoning with more complete and context-consistent spatial relation representations. Extensive experiments across multiple 3D LLM backbones, object settings, and spatial reasoning benchmarks demonstrate performance gains and validate the complementary roles of language-adaptive relation modeling and lightweight reference-frame integration within a unified relation-formation framework.

1 INTRODUCTION

3D spatial reasoning (Chen et al., 2024; Huang et al., 2024a) refers to the process of deriving task-relevant spatial information from scenes and is fundamental to 3D vision-language understanding (Zhu et al., 2025; Ma et al., 2025; Hu et al., 2026), underpinning downstream tasks such as visual grounding (Park et al., 2026; Huang et al., 2025) and visual question answering (Ma et al., 2026; Linghu et al., 2024). Spatial reasoning requires not only recognizing individual objects but also understanding how they are spatially related (Fig. 1, left), making effective representations of inter-object spatial relations essential (Yu et al., 2025; Baik et al., 2025; Zhou et al., 2026).

Existing methods (Zhou et al., 2026; Huang et al., 2026; Yu et al., 2025) continue to stack sophisticated reasoning mechanisms on top of a flawed assumption: effective relation representations can be determined in advance by a fixed geometric scheme. As illustrated in Fig. 1 (middle), this assumption gives rise to two fundamental problems. First, geometric measurement is independent of language. Different queries require different spatial evidence (Shang et al., 2026; Park et al., 2026), yet a predefined geometric scheme determines what spatial information is encoded before the model knows what the query requires. Once query-critical cues are omitted at this stage, downstream reasoning is forced to operate on incomplete representations, leading directly to erroneous spatial judgments. Second, directional semantics are independent of the reference frame. The same pairwise displacement may indicate different or even opposite directions under different reference

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

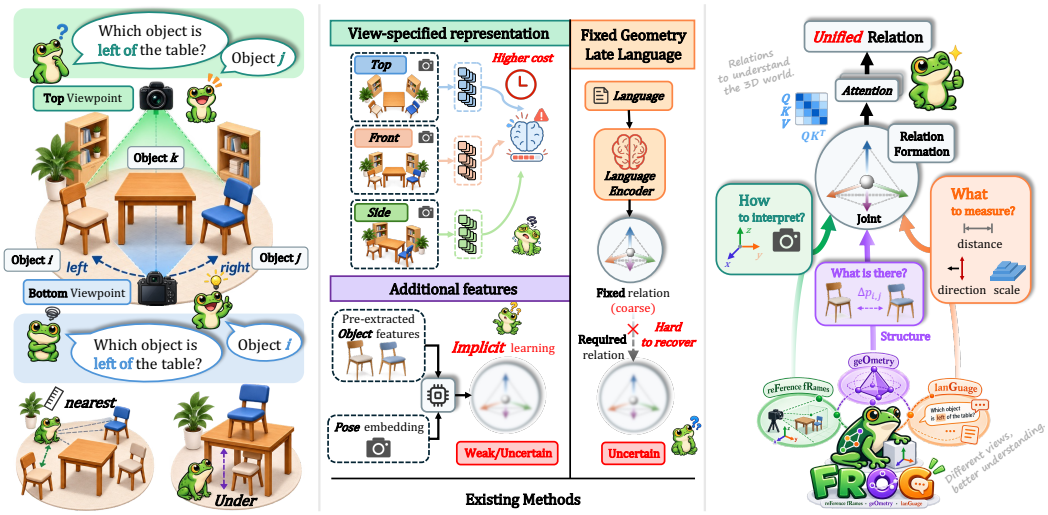


Figure 1: Illustration of the problem setting, limitations of existing methods, and the FROG framework, which jointly integrates geometry, language, and reference frames into relation formation.

frames (Liu et al., 2026; Song et al., 2025; Huang et al., 2025), making the reference frame indispensable for defining direction. Existing methods (Yuan et al., 2025; Zhao et al., 2026; Zhang et al., 2026) either transform object or scene representations according to the viewpoint before relation formation, or inject camera or pose features only afterward. Both treat a fundamentally relation-level problem outside relation formation through complex auxiliary mechanisms. This leaves the mapping from reference frames to directional semantics poorly constrained and forces it to be learned implicitly, thereby resulting in degraded and unstable performance across viewpoints. Ultimately, recent research has increasingly focused on sophisticated reasoning over spatial relations while neglecting the more basic problem of how those relations should be represented in the first place.

This motivates a more fundamental question: how should representations of inter-object spatial relations be formed in the first place? Their formation cannot be determined by geometry alone: language determines which geometric cues are required by the given query and therefore what spatial information should be measured, while the reference frame determines how directional measurements should be interpreted. The same object pair has different spatial relation semantics under different queries or reference frames, without any change to the underlying scene geometry itself. Spatial relation formation is therefore inherently context-dependent, with geometry, language, and the reference frame jointly determining the resulting relation representation.

Building on this perspective, we propose a unified relation-centric framework for 3D spatial reasoning, in which relation representations are jointly determined by reFerence fRames, geOmetry, and lanGuage (FROG), as shown in Fig. 1 (right). To make relation formation sensitive to the relational semantics of the linguistic query, we introduce an order-aware geometry controller, which contextualizes query tokens while preserving their order and produces query-specific control signals for geometric reasoning. Conditioned on these signals, language-adaptive spatial measurement dynamically adapts measurement directions and reweights spatial scales according to the query, allowing language to determine not only how spatial relations are used, but also how the underlying geometry is measured. Meanwhile, the relation-centric reference frame directly defines reference-frame-dependent directional semantics during object–object relation formation within LLM attention using only observer orientation, without modifying object or scene representations or requiring camera position or full 6-DoF pose. Together, these designs mitigate the omission of query-critical geometric cues and the ambiguity of frame-dependent directional semantics, allowing downstream reasoning to operate on more complete and context-consistent spatial relation representations. Extensive experiments across different 3D LLM backbones, object settings, and spatial reasoning benchmarks demonstrate improved spatial reasoning performance, validating the complementary roles of language-adaptive relation modeling and lightweight reference-frame integration.

In summary, the main contributions of this work are as follows:

- We propose FROG, a unified relation-centric framework built around a new perspective on spatial relation formation: relations should be jointly determined by geometry, language, and reference frames rather than treated as fixed geometric quantities.
- We develop three complementary modules: an order-aware geometry controller that captures relational semantics in queries, a language-adaptive spatial measurement module that adapts measurements to the query, and a relation-centric reference frame module that incorporates observer orientation into relation formation to define directional semantics.
- Extensive experiments across diverse backbones, object settings, and benchmarks show substantial gains in 3D spatial reasoning. Controlled ablations and reference-frame interventions further validate the effectiveness and complementary roles of our designs.

2 Related Work

Language Conditioning in Spatial Relation Modeling. ViL3DRel (Chen et al., 2022) encodes pairwise distances and orientations as spatial relations and uses language to condition spatial attention. CORE-3DVG (Yang et al., 2023), MiKASA (Chang et al., 2024), and CoT3DRef (Abdelrahman et al., 2024) exploit contextual relations, multiple anchors, and anchor chains for compositional grounding, while D-LISA (Zhang et al., 2024) introduces language-informed spatial attention. More recently, ORD (Huang et al., 2026) aligns relational predicates with geometric relations, and QuatRoPE (Zhou et al., 2026) forms pairwise spatial relations directly within LLM attention. Despite different designs, these methods largely follow a common pattern: geometric measurements form spatial relations, while language determines how those relations are attended to, aligned, or composed (Chen et al., 2022; Huang et al., 2026; Zhou et al., 2026; Hsu et al., 2023; Chang et al., 2024). FROG instead makes relation formation itself language-dependent, allowing the query to directly control inter-object geometric measurement according to its relational semantics.

Reference Frames in 3D Spatial Reasoning. MVT (Huang et al., 2022), ViewRefer (Guo et al., 2023), and ViewSRD (Huang et al., 2025) address viewpoint dependence through multi-view representations or structured multi-view reasoning. VPP-Net (Shi et al., 2024) explicitly predicts the speaker viewpoint, while 3D Situation Awareness (Yuan et al., 2025) models observer position and orientation. More recent methods align 3D representations with recovered ego poses, as in Direction-aware 3D LMM (Liu et al., 2026), or condition multimodal features on camera representations, as in SpaceMind (Zhao et al., 2026). These methods incorporate viewpoint information through representation construction, scene alignment, or feature conditioning. For directional relations, however, the reference frame directly determines how pairwise geometry should be interpreted. FROG therefore incorporates observer orientation directly into relation formation, defining frame-dependent directional semantics without modifying object or scene representations.

Taken together, prior work uses language and reference-frame information at different stages, but generally not within relation formation. FROG instead integrates both into relation formation: language controls geometric measurement, while the reference frame determines directional semantics.

3 Methods

3.1 Context-Dependent Spatial Relation Formation

Given a language query $\mathcal{Q}$ and two objects with centers $\mathbf{c}_i, \mathbf{c}_j \in \mathbb{R}^3$, let $\Delta\mathbf{c}_{ij} = \mathbf{c}_j - \mathbf{c}_i \in \mathbb{R}^3$ denote their pairwise displacement. Rather than forming spatial relations using a geometric scheme fixed independently of the query and reference frame, FROG makes relation formation context-dependent. For decoder layer ℓ and attention head $h \in \{1, \dots, H\}$ where FROG is applied, with H denoting the number of attention heads, we define

$$\mathcal{K}_{ij}^{\ell h}(\mathcal{Q}, \psi) = \mathcal{M}^{\ell h}\Big(\mathcal{T}_{\psi}(\Delta\mathbf{c}_{ij}), \mathbf{\Theta}^{\ell h}(\mathcal{Q})\Big) \in \mathbb{R}^{d_{\text{spec}} \times d_{\text{spec}}}, \tag{1}$$

where $\mathcal{K}_{ij}^{\ell h}$ is the geometric relation operator, $\mathcal{M}^{\ell h}$ is the spatial measurement operator, and $\mathcal{T}_{\psi}$ expresses pairwise displacement under the reference frame determined by observer yaw ψ. The query-dependent controls $\mathbf{\Theta}^{\ell h}(\mathcal{Q}) = \{\boldsymbol{\delta}^{\ell h}(\mathcal{Q}), \boldsymbol{\rho}^{\ell h}(\mathcal{Q}), \beta^{\ell h}(\mathcal{Q})\}$ determine how the transformed

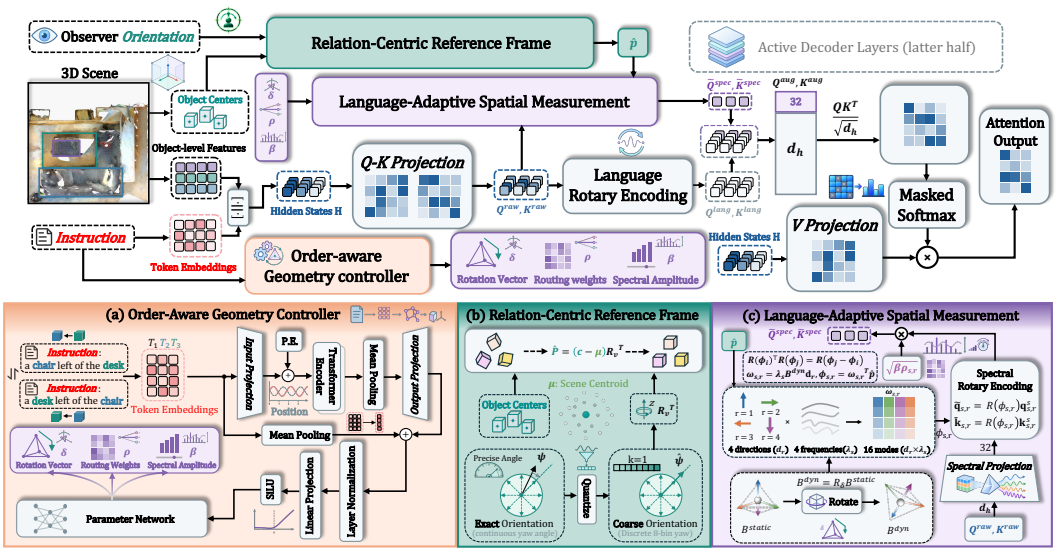


Figure 2: Overview of FROG. Language determines how inter-object geometry is measured, while the reference frame determines how directional geometry is interpreted. Together, they define context-dependent spatial relations that are realized inside self-attention.

geometry is measured. Specifically, $\boldsymbol{\delta}^{\ell h}(\mathcal{Q}) \in \mathbb{R}^3$ controls the measurement basis, $\boldsymbol{\rho}^{\ell h}(\mathcal{Q}) = \{\rho_{sr}^{\ell h}(\mathcal{Q})\}_{s,r=1}^{4} \in \mathbb{R}^{4\times 4}$ allocates importance across direction–frequency modes, and $\beta^{\ell h}(\mathcal{Q}) > 0$ controls the overall geometric contribution. Thus, language specifies the measurement of geometry, while the reference frame specifies its directional interpretation.

3.2 Language-Adaptive Spatial Measurement

Different queries require different geometric evidence, which fixed measurements cannot capture. FROG therefore makes spatial measurement adaptive to the query semantics.

Order-Aware Geometry Controller. Relational semantics can change with the ordering of entities in the query, making order awareness essential for query-adaptive measurement. We therefore introduce the Order-Aware Geometry Controller (OAGC), which captures order-sensitive query semantics and produces layer- and head-specific geometric control signals. Let $\mathcal{Q} = (w_1, \ldots, w_M)$ contain M valid query tokens with backbone embeddings $\mathbf{e}_t \in \mathbb{R}^d$. We project them into a d_c-dimensional controller space, add sinusoidal positional encodings, and process the sequence with a two-layer bidirectional Transformer to obtain contextualized features $\mathbf{z}_t \in \mathbb{R}^{d_c}$. With $\bar{\mathbf{e}} = M^{-1}\sum_{t=1}^{M} \mathbf{e}_t$ and $\bar{\mathbf{z}} = M^{-1}\sum_{t=1}^{M} \mathbf{z}_t$, the order-aware query representation is

$$\mathbf{g} = \bar{\mathbf{e}} + W_{\text{out}}\bar{\mathbf{z}} + \mathbf{b}_{\text{out}} \in \mathbb{R}^d, \qquad W_{\text{out}} \in \mathbb{R}^{d\times d_c}, \quad \mathbf{b}_{\text{out}} \in \mathbb{R}^d. \tag{2}$$

Because pooling follows positional contextualization, $\mathbf{g}$ captures query content together with order-sensitive relational semantics.

A shared parameter network maps $\mathbf{g}$ to routing logits $A^{\ell h} \in \mathbb{R}^{4\times 4}$, rotation logits $\mathbf{t}^{\ell h} \in \mathbb{R}^3$, and an amplitude logit $b^{\ell h} \in \mathbb{R}$:

$$\rho_{sr}^{\ell h}(\mathcal{Q}) = 16\frac{\exp(A_{sr}^{\ell h}/\tau)}{\sum_{s'=1}^{4}\sum_{r'=1}^{4}\exp(A_{s'r'}^{\ell h}/\tau)}, \quad \boldsymbol{\delta}^{\ell h}(\mathcal{Q}) = \delta_{\max}\tanh(\mathbf{t}^{\ell h}), \quad \beta^{\ell h}(\mathcal{Q}) = f_\beta(b^{\ell h}). \tag{3}$$

Here $s, r \in \{1, \ldots, 4\}$ index spatial frequency and direction, $\tau > 0$ is the routing temperature, $\delta_{\max} > 0$ bounds each component of the rotation vector, and $f_\beta : \mathbb{R} \to \mathbb{R}_{>0}$ is a clipped softplus mapping. The normalization gives $\sum_{s,r} \rho_{sr}^{\ell h}(\mathcal{Q}) = 16$, so the average mode weight is one while $\boldsymbol{\rho}^{\ell h}(\mathcal{Q})$ redistributes relative importance across modes and $\beta^{\ell h}(\mathcal{Q})$ controls their overall strength.

Language-Adaptive Spatial Measurement. As shown in Fig. 2(c), the Language-Adaptive Spatial Measurement (LASM) module turns these controls into query-adaptive spatial measurement modes. For each layer-head pair, we maintain a learnable right-handed orthonormal frame $B^{\ell h}_{\text{static}} \in SO(3)$ obtained by Gram–Schmidt orthogonalization. Let $\{\mathbf{d}_r \in \mathbb{R}^3\}_{r=1}^4$ be tetrahedral unit directions satisfying $\sum_{r=1}^4 \mathbf{d}_r \mathbf{d}_r^\top = \frac{4}{3} I_3$, and let $\{\lambda_s > 0\}_{s=1}^4$ denote four fixed spatial frequencies. The tetrahedral directions provide symmetric directional coverage in 3D, while the frequencies span multiple spatial scales. The query-adaptive frame and measurement modes are

$$B^{\ell h}_{\text{dyn}}(\mathcal{Q}) = \exp\Big([\boldsymbol{\delta}^{\ell h}(\mathcal{Q})]_\times\Big) B^{\ell h}_{\text{static}}, \quad \boldsymbol{\omega}^{\ell h}_{sr}(\mathcal{Q}) = \lambda_s B^{\ell h}_{\text{dyn}}(\mathcal{Q}) \mathbf{d}_r \in \mathbb{R}^3. \tag{4}$$

Here $[\cdot]_\times$ maps a vector in $\mathbb{R}^3$ to its 3×3 skew-symmetric matrix. Since $\exp([\boldsymbol{\delta}^{\ell h}(\mathcal{Q})]_\times) \in SO(3)$, the controller rotates the measurement frame while preserving its orthonormal structure. Together, $\boldsymbol{\delta}^{\ell h}(\mathcal{Q})$ steers measurement directions, $\boldsymbol{\rho}^{\ell h}(\mathcal{Q})$ reallocates importance across direction–frequency modes, and $\beta^{\ell h}(\mathcal{Q})$ controls the overall geometric strength. The fixed frequencies and tetrahedral coordinates are detailed in Appendix A.3.

3.3 Relation-Centric Reference Frame

Directional semantics are inherently reference-frame dependent, as the same pairwise displacement can be interpreted as different directions under different viewpoints. To avoid ambiguous directional interpretation, we introduce a relation-centric reference frame that directly defines frame-dependent semantics during relation formation. For any yaw angle α, let $R_v(\alpha) \in SO(3)$ denote the world-to-observer rotation. We use the convention that $\alpha = 0$ faces the world $+x$ direction and positive yaw rotates the forward axis toward $+y$. Spatial language is naturally organized around coarse directional semantics. We therefore quantize observer yaw ψ to the nearest of eight uniformly spaced orientations, $\widehat{\psi} = Q_8(\psi)$, where adjacent bins differ by 45°. This coarse quantization reduces sensitivity to small orientation perturbations, yielding a more robust directional reference frame. Pairwise displacement is invariant to any shared translation $\mathbf{a} \in \mathbb{R}^3$, since $R_v(\widehat{\psi})[(\mathbf{c}_j - \mathbf{a}) - (\mathbf{c}_i - \mathbf{a})] = R_v(\widehat{\psi})(\mathbf{c}_j - \mathbf{c}_i)$. Thus, observer orientation directly determines the directional interpretation of object–object displacement. We therefore define the reference-frame transform as

$$R_v(\widehat{\psi}) = \begin{bmatrix} \sin\widehat{\psi} & -\cos\widehat{\psi} & 0 \\ \cos\widehat{\psi} & \sin\widehat{\psi} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathcal{T}_\psi(\Delta\mathbf{c}_{ij}) = \begin{cases} R_v(\widehat{\psi})\Delta\mathbf{c}_{ij}, & \text{if orientation is available,} \\ \Delta\mathbf{c}_{ij}, & \text{otherwise.} \end{cases} \tag{5}$$

The rows of $R_v(\widehat{\psi})$ correspond to the observer's right, forward, and up axes, so $\mathcal{T}_\psi$ directly expresses pairwise displacement under the directional reference frame.

For token-wise computation, let $\boldsymbol{\mu} \in \mathbb{R}^3$ denote a shared scene center. We use $\widehat{\mathbf{p}}_i = R_v(\widehat{\psi})(\mathbf{c}_i - \boldsymbol{\mu})$ when orientation is available and $\widehat{\mathbf{p}}_i = \mathbf{c}_i$ otherwise. Then $\widehat{\mathbf{p}}_j - \widehat{\mathbf{p}}_i = \mathcal{T}_\psi(\Delta\mathbf{c}_{ij})$, allowing the reference frame to act on relation geometry while leaving object and scene representations unchanged.

3.4 Joint Relation Formation in Attention

FROG realizes the context-dependent relation operator directly inside self-attention, allowing query-adaptive and reference-frame-aware relations to directly influence object interactions. We apply the relation branch to the latter half of the decoder layers, denoted by $\mathcal{L}_{\text{act}}$, while earlier layers retain the original attention computation. For each token i, we define a geometric-role indicator $m_i \in \{0, 1\}$: $m_i = 1$ for object-level tokens associated with a 3D center $\mathbf{c}_i$, and $m_i = 0$ otherwise. Multiple tokens originating from the same object share the same center.

At an active layer $\ell \in \mathcal{L}_{\text{act}}$ and head h, let $\mathbf{q}_i^{\text{raw},\ell h}, \mathbf{k}_i^{\text{raw},\ell h} \in \mathbb{R}^{d_h}$ denote the per-head query and key before the backbone's one-dimensional RoPE, and let $\mathbf{q}_i^{\text{lang},\ell h}, \mathbf{k}_i^{\text{lang},\ell h} \in \mathbb{R}^{d_h}$ denote their standard RoPE-transformed counterparts. We instantiate the spectral dimension introduced in Eq. 1 as $d_{\text{spec}} = 32$. Layer-specific projections shared across heads produce

$$\mathbf{q}_i^{\text{spec},\ell h} = P_Q^\ell \mathbf{q}_i^{\text{raw},\ell h}, \qquad \mathbf{k}_i^{\text{spec},\ell h} = P_K^\ell \mathbf{k}_i^{\text{raw},\ell h}, \qquad P_Q^\ell, P_K^\ell \in \mathbb{R}^{32 \times d_h}. \tag{6}$$

Each spectral feature is partitioned into sixteen two-dimensional blocks $\mathbf{q}_{i,sr}^{\text{spec},\ell h}, \mathbf{k}_{i,sr}^{\text{spec},\ell h} \in \mathbb{R}^2$. For each mode (s, r), we assign each geometric-role token i the spatial phase $\phi_{i,sr}^{\ell h}(\mathcal{Q}, \psi) =$

$\big(\boldsymbol{\omega}_{sr}^{\ell h}(\mathcal{Q})\big)^\top \widehat{\mathbf{p}}_i$. Let $\mathcal{R}(\phi) \in SO(2)$ denote planar rotation by angle ϕ. Using $\mathcal{R}(\phi_i)^\top \mathcal{R}(\phi_j) = \mathcal{R}(\phi_j - \phi_i)$, the relative phase satisfies

$$\begin{aligned} \phi_{j,sr}^{\ell h}(\mathcal{Q},\psi) - \phi_{i,sr}^{\ell h}(\mathcal{Q},\psi) &= \big(\boldsymbol{\omega}_{sr}^{\ell h}(\mathcal{Q})\big)^\top (\widehat{\mathbf{p}}_j - \widehat{\mathbf{p}}_i) \\ &= \big(\boldsymbol{\omega}_{sr}^{\ell h}(\mathcal{Q})\big)^\top \mathcal{T}_\psi(\Delta \mathbf{c}_{ij}). \end{aligned} \tag{7}$$

Thus, token-wise rotary phases reduce exactly to the query-adaptive, reference-frame-aware pairwise displacement used for relation formation.

Combining the language-adaptive measurements with the reference-frame transform gives the concrete relation operator

$$\mathcal{K}_{ij}^{\ell h}(\mathcal{Q},\psi) = \beta^{\ell h}(\mathcal{Q})\,\mathrm{BlkDiag}_{s,r}\left[\rho_{sr}^{\ell h}(\mathcal{Q})\mathcal{R}\Big(\big(\boldsymbol{\omega}_{sr}^{\ell h}(\mathcal{Q})\big)^\top \mathcal{T}_\psi(\Delta \mathbf{c}_{ij})\Big)\right] \in \mathbb{R}^{32\times 32}, \tag{8}$$

where $s, r \in \{1,\dots,4\}$ and $\mathrm{BlkDiag}_{s,r}[\cdot]$ places the sixteen 2×2 mode operators along the block diagonal. Equation 8 instantiates Eq. 1: the reference frame specifies displacement, while the query controls measurement directions, mode weights, and geometric strength.

The corresponding geometric interaction is

$$G_{ij}^{\ell h} = m_i m_j \big(\mathbf{q}_i^{\mathrm{spec},\ell h}\big)^\top \mathcal{K}_{ij}^{\ell h}(\mathcal{Q},\psi)\mathbf{k}_j^{\mathrm{spec},\ell h} \in \mathbb{R}. \tag{9}$$

The geometric term is then added to the backbone content interaction:

$$S_{ij}^{\ell h} = \frac{\big(\mathbf{q}_i^{\mathrm{lang},\ell h}\big)^\top \mathbf{k}_j^{\mathrm{lang},\ell h} + G_{ij}^{\ell h}}{\sqrt{d_h}} + M_{ij}, \tag{10}$$

where $S_{ij}^{\ell h} \in \mathbb{R}$ is the pre-softmax attention score and $M_{ij} \in \mathbb{R} \cup \{-\infty\}$ is the additive attention mask. The value branch, softmax, value aggregation, and output projection follow the backbone.

4 Experiments

4.1 Experimental Setup

Datasets. We jointly train our models on seven 3D vision-language tasks. Specifically, we use ScanRefer (Chen et al., 2020) for single-object visual grounding, Multi3DRefer (Zhang et al., 2023) for multi-object visual grounding, Scan2Cap (Chen et al., 2021) and Nr3D-caption (Achlioptas et al., 2020) for dense captioning, ScanQA (Azuma et al., 2022) and SQA3D (Ma et al., 2023) for visual question answering, and ObjAlign (Huang et al., 2024a) for object alignment. We use ScanRefer, Multi3DRefer, and SQA3D as our standard evaluation benchmarks. To more comprehensively assess 3D spatial reasoning capabilities, we additionally evaluate on ASR (Zhou et al., 2026) and Real-3DQA (Ma et al., 2026). ASR and Real-3DQA are used only for evaluation.

Evaluation Metrics. For ScanRefer, we report Acc@0.25 and Acc@0.50, measuring grounding accuracy at 3D IoU thresholds of 0.25 and 0.50. We additionally report the same metrics on the Multiple subset, where multiple objects of the same semantic class as the target are present. Multi3DRefer is evaluated using F1@0.25 and F1@0.50 at the corresponding IoU thresholds. For SQA3D, we report top-1 exact match (EM@1) for answer accuracy. ASR follows the 3D grounding protocol and is evaluated with Acc@0.25 and Acc@0.50. For Real-3DQA, we report Exact Match (EM), Relaxed Match (EM-R), and Viewpoint Rotation Score (VRS), where VRS measures consistency across viewpoint rotations by averaging success rates over one to four viewpoints.

4.2 Comparison with Existing Methods

Overall Performance across Backbones and Tasks. As shown in Table 1, FROG yields consistent improvements across different backbones and object settings. On ScanRefer, FROG improves Acc@0.50 by 14.8/9.5 points with GT objects and by 7.0/5.6 points with Mask3D for Chat-Scene and 3DGraphLLM, respectively. Gains are more pronounced on the Multiple subset, reaching up to 16.4 points, where distinguishing same-category instances requires finer-grained relational reasoning. FROG also consistently improves Multi3DRefer, including an 8.6-point gain in F1@0.50 for Chat-Scene. Since Multi3DRefer provides no viewpoint annotations, these gains highlight language-adaptive relation formation without reference-frame information. FROG also remains competitive on SQA3D, showing that its gains extend beyond grounding to broader 3D question answering.

Table 1: Comparison on ScanRefer, Multi3DRefer, and SQA3D. Our method uses the officially provided viewpoint annotations for ScanRefer and SQA3D, while no viewpoint information is used for Multi3DRefer because such annotations are not officially available for this dataset.

Model	Detector / Segmentation	LLM	ScanRefer				Multi3DRefer		SQA3D
			Acc @0.25	Acc @0.50	Multi @0.25	Multi @0.50	F1 @0.25	F1 @0.50	EM @1
3D-VisTA (Zhu et al., 2023)	Mask3D	-	50.6	45.8	43.7	39.1	–	–	48.5
M3DRef-CLIP (Zhang et al., 2023)	PointGroup	-	51.9	44.7	43.8	36.8	42.8	38.4	–
VPP-Net (Shi et al., 2024)	Group-Free	-	55.7	43.3	50.3	39.0	–	–	–
TSP3D (Guo et al., 2025)	N/A	-	56.5	46.7	51.0	42.4	–	–	–
PV-Ground (Shang et al., 2026)	N/A	-	59.3	47.8	54.6	43.4	–	–	–
PQ3D (Zhu et al., 2024)	Joint Opt.	-	57.0	51.2	51.5	46.2	–	50.1	47.1
EG-3DVG (Park et al., 2026)	N/A	-	57.1	50.1	51.1	44.2	–	–	–
LEO (Huang et al., 2024b)	Mask3D	Vicuna-7B-v1.1	–	–	–	–	–	–	50.0
Video-3D LLM (Zheng et al., 2025)	Mask3D	LLaVA-Video-7B	58.1	51.7	50.9	45.3	58.0	52.7	58.6
Robin3D (Kang et al., 2025)	Mask3D	Vicuna-7B-v1.5	60.8	55.1	–	–	64.9	59.7	56.9
S^2-MLLM (Xu et al., 2026)	Mask3D	LLaVA-Video-7B	59.2	52.7	52.4	46.6	–	–	–
QuatRoPE(3DGraphLLM) (Zhou et al., 2026)	Mask3D	Vicuna-7B-v1.5	58.2	52.5	54.3	49.2	60.6	56.0	55.2
LLaVA-3D (Zhu et al., 2025)	N/A	LLaVA-1.5-7B	54.1	42.4	–	–	–	–	55.6
Chat-Scene (Huang et al., 2024a)	GT	Llama-3.2-1B-Instruct	50.7	50.3	42.7	42.3	53.3	52.9	50.7
Chat-Scene + FROG	GT	Llama-3.2-1B-Instruct	**65.3** ↑14.6	**65.1** ↑14.8	**58.8** ↑16.1	**58.7** ↑16.4	**61.6** ↑8.3	**61.5** ↑8.6	**52.8** ↑2.1
3DGraphLLM (Zemskova & Yudin, 2025)	GT	Llama-3.2-1B-Instruct	55.9	55.8	47.9	47.7	58.6	58.4	51.1
3DGraphLLM + FROG	GT	Llama-3.2-1B-Instruct	**65.5** ↑9.6	**65.3** ↑9.5	**58.9** ↑11.0	**58.7** ↑11.0	**62.1** ↑3.5	**61.9** ↑3.5	**53.3** ↑2.2
Chat-Scene (Huang et al., 2024a)	Mask3D	Vicuna-7B-v1.5	55.5	50.2	47.8	42.9	57.1	52.4	**54.6**
Chat-Scene + FROG	Mask3D	Vicuna-7B-v1.5	**63.3** ↑7.8	**57.2** ↑7.0	**57.0** ↑9.2	**51.3** ↑8.4	**60.2** ↑3.1	**55.6** ↑3.2	53.4 ↓1.2
3DGraphLLM (Zemskova & Yudin, 2025)	Mask3D	Vicuna-7B-v1.5	57.0	51.3	50.6	45.5	60.1	55.4	53.1
3DGraphLLM + FROG	Mask3D	Vicuna-7B-v1.5	**62.5** ↑5.5	**56.9** ↑5.6	**56.4** ↑5.8	**50.9** ↑5.4	**60.9** ↑0.8	**56.1** ↑0.7	**54.2** ↑1.1

Table 2: Results on Real-3DQA (Ma et al., 2026) and its viewpoint-rotation evaluation. One–Four report the percentages of question groups correct under at least one to four viewpoint variants. Dist., Dir., Count., and Exist. denote Distance, Direction, Counting, and Existence question types.

Model	Type	Backbone	Real-3DQA		Viewpoint Rotation					Question Types			
			EM	EM-R	**VRS**	One	Two	Three	Four	Dist.	**Dir.**	Count.	Exist.
GPT4Scene (Qi et al., 2026)	VLM	Qwen2-VL-7B	33.1	36.9	18.2	55.5	14.3	2.5	0.5	15.8	25.2	31.7	36.5
3D-LLM (Hong et al., 2023)	3D-LLM	Flan-T5-XL	7.5	10.4	9.6	33.2	4.1	1.1	0.1	5.5	24.6	**24.2**	44.2
Chat-3D v2 (Huang et al., 2023)	3D-LLM	Vicuna-7B	3.4	9.7	6.6	23.2	2.7	0.5	0.0	4.3	14.6	16.7	21.2
LEO (Huang et al., 2024b)	3D-LLM	Vicuna-7B-v1.1	14.3	19.1	**14.3**	**46.9**	8.1	1.6	0.4	**11.8**	23.2	20.0	36.5
Chat-Scene (Huang et al., 2024a)	3D-LLM	Vicuna-7B-v1.5	17.0	22.1	12.9	43.3	7.1	1.2	0.1	10.1	23.7	16.7	40.4
Chat-Scene + FROG	3D-LLM	Vicuna-7B-v1.5	**23.3** ↑6.3	**27.7** ↑5.6	13.7 ↑0.8	42.7 ↓0.6	**10.1** ↑3.0	**1.7** ↑0.5	**0.4** ↑0.3	10.2 ↑0.1	**26.8** ↑3.1	19.2 ↑2.5	**48.1** ↑7.7

Zero-Shot Spatial Reasoning Evaluation on Real-3DQA. Table 2 further evaluates FROG on Real-3DQA (Ma et al., 2026), a challenging zero-shot benchmark that filters out questions answerable from language priors alone and places greater emphasis on genuine 3D spatial relation reasoning. FROG delivers broad improvements across nearly all reported metrics, spanning answer accuracy, viewpoint-rotation robustness, and different spatial question types. In particular, EM and EM-R improve by 6.3 and 5.6 points, respectively, while the stricter Two–Four viewpoint consistency levels all increase. FROG also improves all four spatial question types, with particularly clear gains on Direction (+3.1) and Existence (+7.7). These results demonstrate that the improved relation representations generalize beyond grounding to challenging zero-shot 3D reasoning.

Reference-Frame-Aware Relation Modeling. Table 3 compares different strategies for incorporating viewpoint information into 3D grounding. By integrating the reference frame directly into relation formation, FROG-A achieves 65.1 Acc@0.50 on ScanRefer, substantially outperforming prior multi-view and representation-level approaches. More importantly, FROG can readily leverage externally recovered reference frames when official viewpoint annotations are unavailable: incorporating PoseRecover viewpoints (Liu et al., 2026) on Multi3DRefer improves F1@0.50 from 61.5 to 65.3. FROG-B further achieves the strongest performance on both datasets, reaching 66.4/65.3 on ScanRefer/Multi3DRefer. When official ScanRefer viewpoints are replaced by PoseRecover predictions, FROG-C still achieves 63.8 Acc@0.50, demonstrating that the relation-centric formulation generalizes across different viewpoint sources without modifying representations. Together, these results show that directly defining frame-dependent directional semantics at relation formation provides an effective, robust, and flexible mechanism for reference-frame-aware 3D reasoning.

Table 3: Comparison of viewpoint-aware strategies for 3D grounding. Multi-view, Rep.-level, and Relation-level denote viewpoint integration levels. Official and PoseRecover denote dataset-provided and predicted viewpoints, respectively. Our results use Chat-Scene-1B with GT objects.

Model	View Source		View Integration			ScanRefer Acc@0.50	Multi3DRefer F1@0.50
	ScanRefer	Multi3DRefer	Multi-view	Rep.-level	Relation-level		
MVT (Huang et al., 2022)	Multi-view	–	✓	✗	✗	33.3	–
ViewRefer (Guo et al., 2023)	Multi-view	–	✓	✗	✗	33.7	–
ViewSRD (Huang et al., 2025)	Multi-view	–	✓	✗	✗	36.0	–
VPP-Net (Shi et al., 2024)	Pred. viewpoint	–	✗	✓	✗	43.3	–
ConcreteNet (Unal et al., 2024)	GT camera sup.	–	✓	✓	✗	46.5	–
PoseAlign-T (Liu et al., 2026)	PoseRecover	PoseRecover	✗	✓	✗	58.7	–
PoseAlign-E (Liu et al., 2026)	PoseRecover	PoseRecover	✗	✓	✗	46.9	50.2
FROG-A	**Official**	**None**	✗	✗	✓	65.1	61.5
FROG-B	**Official**	**PoseRecover**	✗	✗	✓	**66.4**	**65.3**
FROG-C	**PoseRecover**	**PoseRecover**	✗	✗	✓	63.8	64.7

Table 4: Performance–cost comparison under the same Chat-Scene-1B backbone.

Model	Model cost		Inference cost				Spatial reasoning performance			
	Extra P. (M) ↓	Param. Δ (%) ↓	Prefill (ms) ↓	Decode (ms/tok) ↓	E2E (×) ↓	Mem. (GB) ↓	ScanRefer Acc@0.50 ↑	Multi3DRefer F1@0.50 ↑	ASR Acc@0.25 ↑	ASR Acc@0.50 ↑
Chat-Scene (Huang et al., 2024a)	0	0	48.73	40.481	1.000	6.388	50.3	52.9	22.92	22.92
+ QuatRoPE (Zhou et al., 2026)	0	0	85.92	43.357	1.095	6.388	55.0 ↑4.7	57.7 ↑4.8	25.89 ↑2.97	25.89 ↑2.97
+ FROG	3.988	0.315	92.53	41.869	1.067	6.404	**65.1** ↑14.8	**61.5** ↑8.6	**32.14** ↑9.22	**31.85** ↑8.93

Efficiency and Transferability of Relation Modeling. Table 4 shows that FROG achieves substantial spatial reasoning gains with a lightweight design. Under the same Chat-Scene-1B backbone, FROG adds only 3.988M parameters (0.315%) and incurs 1.067× end-to-end runtime, while improving ScanRefer Acc@0.50 by 14.8 points and Multi3DRefer F1@0.50 by 8.6 points. On the zero-shot ASR benchmark, FROG improves Acc@0.25/0.50 by 9.22/8.93 points and substantially surpasses QuatRoPE. Since ASR provides no viewpoint input and emphasizes spatial reasoning, these gains further demonstrate the benefits of language-adaptive spatial measurement. Overall, FROG delivers stronger and more transferable spatial reasoning with minimal additional cost.

4.3 Ablations

Table 5: Ablation studies of FROG on ScanRefer (Chat-Scene-1B, GT objects). Left: RB, LA, and RF denote the static relation branch, language adaptation, and relation-centric reference frame. Right: OA, δ, ρ, and β denote order awareness, rotation, mode routing, and amplitude control.

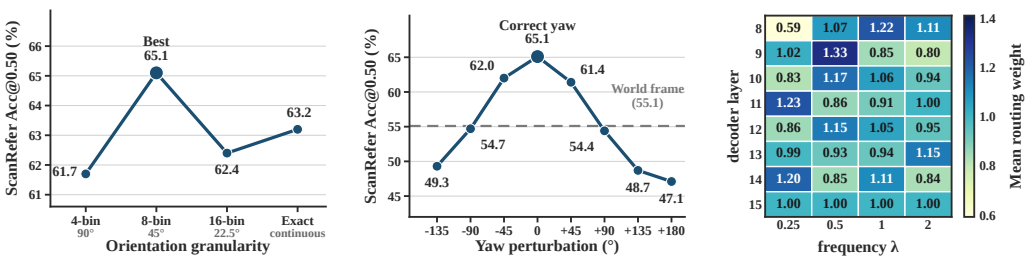
(a) Component Ablation

Model	RB	LA	RF	Overall		Multiple	
				Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50
Chat-Scene	✗	✗	✗	50.7	50.3	42.7	42.3
+ RB	✓	✗	✗	54.0	53.9	45.8	45.7
+ RB + LA	✓	✓	✗	56.0	55.8	48.2	47.9
+ RB + RF	✓	✗	✓	64.0	63.8	57.6	57.3
FROG	✓	✓	✓	**65.3**	**65.1**	**58.8**	**58.7**

(b) OAGC Ablation

Variant	OA	δ	ρ	β	Overall		Multiple	
					Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50
w/o OA	✗	✓	✓	✓	63.7	63.4	57.0	56.6
w/o δ	✓	✗	✓	✓	62.3	62.0	55.4	55.2
w/o ρ	✓	✓	✗	✓	63.1	62.9	56.7	56.4
w/o β	✓	✓	✓	✗	62.5	62.1	55.7	55.2
FROG	✓	✓	✓	✓	**65.3**	**65.1**	**58.8**	**58.7**

Table 5 shows that the main components of FROG provide complementary improvements in spatial reasoning. Starting from the static relation branch, language adaptation improves ScanRefer Acc@0.50 from 53.9 to 55.8, while relation-centric reference-frame modeling raises it to 63.8. Combining both further reaches 65.1 and improves the Multiple subset to 58.7, showing the complementary roles of query-adaptive geometric measurement and reference-frame-aware directional modeling. The OAGC ablation further shows that order awareness, measurement rotation, mode routing, and amplitude control jointly contribute to adapting geometric evidence in direction, spatial scale, and overall strength. Fig. 3 provides a closer view of the resulting behavior. Fig. 3(a) shows that the 8-bin orientation achieves the best performance, supporting the use of coarse directional semantics for reference-frame modeling. Fig. 3(b) shows that performance peaks at the correct yaw and decreases as the reference frame is increasingly perturbed, directly demonstrating the importance of accurate directional interpretation. Finally, Fig. 3(c) reveals different routing preferences across decoder layers and spatial frequencies, indicating that FROG dynamically combines geometric evidence at different scales throughout the reasoning process.



(a) (b) (c)

Figure 3: Further analysis of FROG. (a) Effect of orientation discretization granularity. (b) Robustness to yaw perturbations. (c) Layer-wise mean routing weights over spatial frequencies.

4.4 Qualitative Results

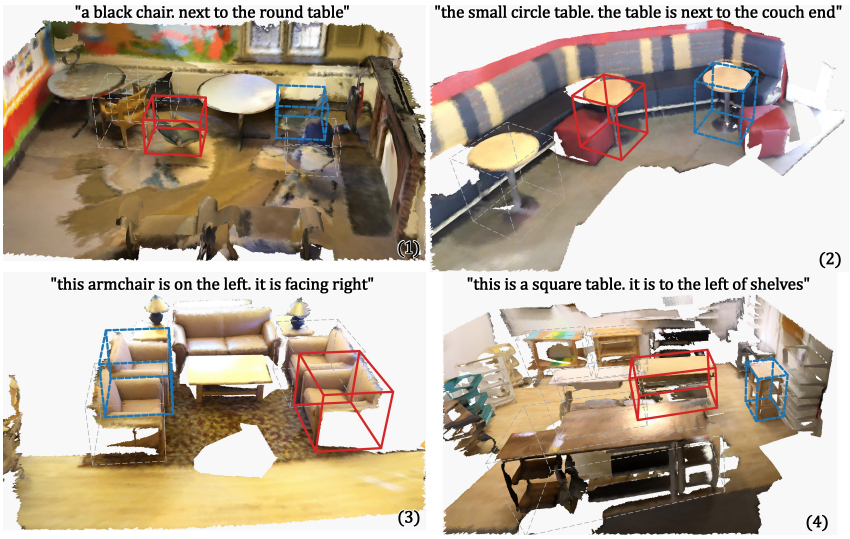


Figure 4: Qualitative grounding on ScanRefer with challenging same-category distractors. Solid red boxes denote Baseline (Huang et al., 2024a), and dashed blue boxes denote FROG.

Figure 4 illustrates four representative cases where spatial relations are critical for distinguishing same-category objects. In the first example, the Baseline selects another chair but fails to satisfy the *next to* relation with the round table, whereas FROG identifies the chair consistent with this proximity cue. In the second, the Baseline chooses a more central table, while FROG correctly follows the *next to the couch end* relation to locate the intended instance. In the third, the Baseline confuses two similar armchairs and selects the one on the opposite side, whereas FROG jointly respects the *left* and *facing right* cues. In the last example, the Baseline selects another table that does not match the specified anchor relation, while FROG correctly grounds the table *left of* the shelves. Overall, these cases show that FROG more faithfully uses explicit relational cues to resolve same-category ambiguities, rather than relying primarily on object appearance or category similarity.

5 Conclusion

This study revisits how spatial relations should be formed for 3D spatial reasoning and proposes a unified relation-centric framework, FROG. By developing an order-aware geometry controller and language-adaptive spatial measurement, FROG allows queries to guide inter-object geometric measurement. Further, by incorporating observer orientation into relation formation, FROG directly defines frame-dependent directional semantics. Together, these designs provide downstream reasoning with more complete and context-consistent spatial relation representations. Extensive experiments validate the effectiveness of FROG in improving spatial reasoning across diverse settings. Future work will explore more flexible relation modeling for broader 3D reasoning scenarios.

AI USE STATEMENT

We did not use generative AI tools for tasks requiring mandatory disclosure under the ICLR 2027 AI Policy, including generating synthetic datasets, developing theoretical or conceptual frameworks, formulating or proving mathematical claims, proposing or refining hypotheses, designing research methodology or experiments, implementing methods, translating content, cleaning or reformatting datasets, conducting qualitative or thematic analysis, or interpreting experimental results. We used generative AI tools for language editing, grammar correction, improving the clarity and readability of the manuscript, and limited assistance with software code editing and debugging. All AI-assisted work was carefully reviewed and verified by the authors. The authors take full responsibility for the final content of this work, including all text, claims, results, and artifacts.

REPRODUCIBILITY STATEMENT

We provide detailed implementation and evaluation specifications to support reproduction of our results. Appendix A describes the implementation of FROG, including geometric-role tokens, the Order-Aware Geometry Controller, Language-Adaptive Spatial Measurement, the Relation-Centric Reference Frame, relation formation inside attention, training and initialization, inference with KV caching, and model dimensions and parameter overhead. Appendix B further specifies the training tasks and data processing, backbone and object settings, reference-frame inputs, evaluation and decoding protocols, ablation settings, and efficiency-profiling procedure. The reported experiments use the same object-level backbone representations and evaluation metrics described in the main paper, while all FROG-specific architectural choices and optimization settings required to reproduce the method are documented in the appendix.

REFERENCES

Eslam Abdelrahman, Mohamed Ayman Mohamed, Mahmoud Ahmed, Habib, and Mohamed Elhoseiny. Cot3dref: Chain-of-thoughts data-efficient 3d visual grounding. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 11871–11896, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/32f9049217da6e718a426b07242dff73-Paper-Conference.pdf.

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, pp. 422–440. Springer, 2020.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19107–19117. IEEE, 2022.

Sangwon Baik, Hyeonwoo Kim, and Hanbyul Joo. Learning 3d object spatial relationships from pre-trained 2d diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8418–8428, October 2025.

Chun-Peng Chang, Shaoxiang Wang, Alain Pagani, and Didier Stricker. Mikasa: Multi-key-anchor & scene-aware transformer for 3d visual grounding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14131–14140. IEEE, 2024.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, June 2024.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pp. 202–221. Springer, 2020.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in neural information processing systems*, 35:20522–20535, 2022.

Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3193–3203, 2021.

Wenxuan Guo, Xiuwei Xu, Ziwei Wang, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Text-guided sparse voxel pruning for efficient 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3666–3675, June 2025.

Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15372–15383, October 2023.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 20482–20494. Curran Associates, Inc., 2023. doi: 10.52202/075280-0900. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/413885e70482b95dcbeeddc1daf39177-Paper-Conference.pdf.

Joy Hsu, Jiayuan Mao, and Jiajun Wu. Ns3d: Neuro-symbolic grounding of 3d objects and relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2614–2623, June 2023.

Wenbo Hu, Jingli Lin, Yilin Long, Yunlong Ran, Lihan Jiang, Yifan Wang, Chenming Zhu, Runsen Xu, Tai Wang, and Jiangmiao Pang. G^2VLM: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9535–9546, June 2026.

Haifeng Huang, Zehan Wang, Rongjie Huang, Luping Liu, Xize Cheng, Yang Zhao, Tao Jin, and Zhou Zhao. Chat-3d v2: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*, 2023.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, and Zhou Zhao. Chat-scene: Bridging 3d scene and large language models with object identifiers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=t3BhmwAzhv.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3D world. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20413–20451. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/huang24ae.html.

Ronggang Huang, Haoxin Yang, Yan Cai, Xuemiao Xu, Huaidong Zhang, and Shengfeng He. Viewsrd: 3d visual grounding via structured multi-view decomposition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9726–9736, October 2025.

Ronggang Huang, Fansen Meng, Huaidong Zhang, and Xuemiao Xu. Ord: Object-relation decoupling for generalized 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30964–30973, June 2026.

Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15524–15533, June 2022.

Weitai Kang, Haifeng Huang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Robin3d: Improving 3d large language model via robust instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3905–3915, October 2025.

Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3d scenes. *Advances in Neural Information Processing Systems*, 37:140903–140936, 2024.

Quan Liu, Weihao Xuan, Junjue Wang, Naoto Yokoya, Ling Shao, and Shijian Lu. Direction-aware 3d large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9668–9678, June 2026.

Wufei Ma, Luoxin Ye, Celso M de Melo, Alan Yuille, and Jieneng Chen. Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17249–17260. IEEE, 2025.

Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Prisacariu. Do 3d large language models really understand 3d spatial relationships? In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 32591–32610, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/3739050312e637d158e39a2859fa8b01-Paper-Conference.pdf.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=IDJx97BC38.

GwangWook Park, Hyo-Jun Lee, Jong-Hyeon Baek, Hanul Kim, and Yeong Jun Koh. Eg-3dvg: Expression and geometry aware grounding decoder for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2625–2634, June 2026.

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 157277–157302, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/fef126561bbf9d4467dbb8d27334b8fe-Paper-Conference.pdf.

Junpeng Shang, Feifei Shao, Jun Xiao, Lin Li, Hongwei Wang, and Dongfang Ma. Pv-ground: Text-guided point-voxel interaction for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38657–38667, June 2026.

Xiangxi Shi, Zhonghua Wu, and Stefan Lee. Viewpoint-aware visual grounding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14056–14065, June 2024.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15768–15780. IEEE, 2025.

Ozan Unal, Christos Sakaridis, Suman Saha, and Luc Van Gool. Four ways to improve verbo-visual fusion for dense 3d visual grounding. In *European Conference on Computer Vision (ECCV)*, October 2024.

Beining Xu, Siting Zhu, Zhao Jin, Junxian Li, and Hesheng Wang. S^2-MLLM: Boosting spatial reasoning capability of mllms for 3d visual grounding with structural guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2557–2569, June 2026.

Li Yang, chunfeng yuan, Ziqi Zhang, Zhongang Qi, Yan Xu, Wei Liu, Ying Shan, Bing Li, Weiping Yang, Peng Li, Yan Wang, and Weiming Hu. Exploiting contextual objects and relations for 3d visual grounding. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49542–49554. Curran Associates, Inc., 2023. doi: 10.52202/075280-2156. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9b91ee0da3bcd61905fcd89e770168fc-Paper-Conference.pdf.

Hanxun Yu, Wentong Li, Song Wang, Junbo Chen, and Jianke Zhu. Inst3d-lmm: Instance-aware 3d scene understanding with multi-modal instruction tuning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14147–14157. IEEE, 2025.

Zhihao Yuan, Yibo Peng, Jinke Ren, Yinghong Liao, Yatong Han, Chun-Mei Feng, Hengshuang Zhao, Guanbin Li, Shuguang Cui, and Zhen Li. Empowering large language models with 3d situation awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19435–19445, June 2025.

Tatiana Zemskova and Dmitry Yudin. 3dgraphllm: Combining semantic graphs and large language models for 3d scene understanding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8885–8895. IEEE, 2025.

Gongjie Zhang, Wenhao Li, Quanhao Qian, Jiuniu Wang, Deli Zhao, Shijian Lu, and Ran Xu. On the generalization capacities of MLLMs for spatial intelligence. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DE5ZJtR4bg.

Haomeng Zhang, Chiao-An Yang, and Raymond A. Yeh. Multi-object 3d grounding with dynamic modules and language-informed spatial attention. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 123237–123260. Curran Associates, Inc., 2024. doi: 10.52202/079017-3917. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/df027cf11469e746ef94d583f9f5537f-Paper-Conference.pdf.

Yiming Zhang, ZeMing Gong, and Angel X. Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15225–15236, October 2023.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. Spacemind: Camera-guided modality fusion for spatial reasoning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16811–16822, June 2026.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8995–9006, June 2025.

Shengli Zhou, Minghang Zheng, Feng Zheng, and Yang Liu. Scalable object relation encoding for better 3d spatial reasoning in large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16758–16767, June 2026.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4295–4305. IEEE, 2025.

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2911–2921, October 2023.

Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *European Conference on Computer Vision*, pp. 188–206. Springer, 2024.

Appendix Overview

Section	Content
A	**Implementation and Reproducibility Details**
A.1	Geometric-role tokens and relation scope
A.2	Order-aware Geometry Controller
A.3	Language-adaptive Spatial Measurement
A.4	Relation-centric Reference Frame
A.5	Joint relation formation in attention
A.6	Training and initialization
A.7	Inference with KV caching
A.8	Model dimensions and parameter overhead
B	**Experimental Protocol Details**
B.1	Training tasks and data processing
B.2	Backbone/object settings for Table 1
B.3	Reference-frame inputs for Tables 3 and 2
B.4	Metrics and decoding for Tables 1 and 2
B.5	Settings for Table 5 and Fig. 3
B.6	Profiling protocol for Table 4
C	**Multi-Seed Evaluation** on ScanRefer, Multi3DRefer, and SQA3D
D	**Additional Qualitative Results** extending Fig. 4
E	**Limitations**

A Implementation and Reproducibility Details

This appendix provides implementation and reproducibility details for the method and experiments in Secs. 3 and 4, complementing the architecture in Fig. 2. We follow the terminology and notation of the main paper throughout. FROG augments an object-level 3D language model with a relation branch in which language controls how inter-object geometry is measured and the reference frame determines how directional geometry is interpreted. The backbone scene representation, token organization, value branch, and output projection are retained.

A.1 Geometric-Role Tokens and Relation Scope

Let the decoder input sequence have length T. Each token i is associated with a geometric-role indicator $m_i \in \{0, 1\}$. We set $m_i = 1$ for object-level tokens associated with a 3D object center $\mathbf{c}_i \in \mathbb{R}^3$, and $m_i = 0$ otherwise. Multiple tokens originating from the same object share the same center. Language, answer, padding, and non-object graph-edge tokens therefore do not contribute an explicit geometric relation term.

Chat-Scene represents each object using object-identity, point-cloud, and image-feature tokens. 3DGraphLLM additionally introduces graph-neighbor structure; object-aligned neighbor features inherit the center of the corresponding neighbor object. FROG preserves these backbone-specific sequence organizations and forms explicit relations between pairs of geometric-role tokens. The relation branch uses object centers as its geometric input, while the original backbone features continue to carry appearance and other object information.

OAGC is evaluated separately from the causal text stream and receives the complete question but not the answer. The geometric controls can therefore depend on the full known query without modifying the decoder causal mask.

A.2 Order-Aware Geometry Controller

Question representation. OAGC uses the backbone tokenizer without introducing additional special tokens. Questions are truncated to at most $M = 128$ tokens and padded within each mini-batch.

Let $a_t \in \{0,1\}$ denote the valid-token mask and $\mathbf{e}_t \in \mathbb{R}^d$ the corresponding backbone token embedding. The controller reads a stop-gradient copy of the token embeddings,

$$\mathbf{e}_t = \mathrm{stopgrad}(\mathrm{Embed}(w_t)), \qquad \bar{\mathbf{e}} = \frac{\sum_t a_t \mathbf{e}_t}{\max(1, \sum_t a_t)}. \tag{11}$$

The stop-gradient is confined to the controller input branch.

The normalized embeddings are projected to a 256-dimensional controller space and combined with fixed sinusoidal positional encodings,

$$\mathbf{x}_t = W_{\mathrm{in}} \,\mathrm{LN}(\mathbf{e}_t) + \mathbf{b}_{\mathrm{in}} + \mathrm{PE}(n_t), \tag{12}$$

$$n_t = \max\left(0, \sum_{u \le t} a_u - 1\right), \tag{13}$$

where cumulative indexing over valid tokens makes the positional indices independent of padding side. The positional encoding is

$$\mathrm{PE}(n, 2k) = \sin\left(n\, 10000^{-2k/256}\right), \tag{14}$$

$$\mathrm{PE}(n, 2k+1) = \cos\left(n\, 10000^{-2k/256}\right). \tag{15}$$

A two-layer bidirectional Transformer encoder with width 256, four attention heads, FFN width 512, GELU activations, pre-normalization, and zero dropout produces contextualized features $\mathbf{z}_t$. A final LayerNorm is applied after the encoder. The pooled query representation is

$$\mathbf{g} = \bar{\mathbf{e}} + W_{\mathrm{out}} \frac{\sum_t a_t \mathbf{z}_t}{\max(1, \sum_t a_t)} + \mathbf{b}_{\mathrm{out}}. \tag{16}$$

Pooling after positional contextualization makes the representation sensitive to query word order, while $\bar{\mathbf{e}}$ provides a direct lexical residual.

Layer- and head-specific controls. A shared parameter network maps $\mathbf{g}$ to a 256-dimensional hidden vector,

$$\mathbf{u} = \mathrm{SiLU}(W_c \,\mathrm{LN}(\mathbf{g}) + \mathbf{b}_c), \tag{17}$$

Three linear output heads then predict controls for all active decoder layers and attention heads: routing logits $A^{\ell h} \in \mathbb{R}^{4\times 4}$, rotation logits $\mathbf{t}^{\ell h} \in \mathbb{R}^3$, and amplitude logits $b^{\ell h} \in \mathbb{R}$.

For each active layer ℓ and head h, the 16 direction–frequency logits are normalized jointly,

$$\rho^{\ell h}_{sr} = 16 \frac{\exp(A^{\ell h}_{sr}/\tau)}{\sum_{s'=1}^{4}\sum_{r'=1}^{4} \exp(A^{\ell h}_{s'r'}/\tau)}, \qquad \tau = 1, \tag{18}$$

so that $\sum_{s,r} \rho^{\ell h}_{sr} = 16$ and the average mode weight equals one. The rotation vector is

$$\boldsymbol{\delta}^{\ell h} = 0.35 \tanh(\mathbf{t}^{\ell h}), \qquad R^{\ell h}_{\delta} = \exp([\boldsymbol{\delta}^{\ell h}]_\times), \tag{19}$$

and the geometric amplitude is

$$\beta^{\ell h} = \mathrm{clip}\big(\mathrm{softplus}\big(\mathrm{softplus}^{-1}(0.02) + 4b^{\ell h}\big), 10^{-6}, 1\big). \tag{20}$$

The three output heads are zero-initialized, so optimization starts from uniform routing, zero query-dependent frame rotation, and $\beta = 0.02$.

A.3 Language-Adaptive Spatial Measurement

Learnable measurement frame. Each active layer and attention head contains two learnable 3D vectors. Gram–Schmidt orthogonalization defines a right-handed frame,

$$\mathbf{b}_1 = \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|}, \tag{21}$$

$$\mathbf{b}_2 = \frac{\mathbf{a}_2 - (\mathbf{a}_2^\top \mathbf{b}_1)\mathbf{b}_1}{\|\mathbf{a}_2 - (\mathbf{a}_2^\top \mathbf{b}_1)\mathbf{b}_1\|}, \tag{22}$$

$$\mathbf{b}_3 = \mathbf{b}_1 \times \mathbf{b}_2, \qquad B_{\mathrm{static}} = [\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3]. \tag{23}$$

The resulting frame is shared across samples for a given layer-head pair and is adapted to the query through

$$B^{\ell h}_{\mathrm{dyn}}(Q) = \exp([\boldsymbol{\delta}^{\ell h}(Q)]_\times) B^{\ell h}_{\mathrm{static}}. \tag{24}$$

Directions and frequencies. We use four tetrahedral unit directions,

$$\mathbf{d}_1 = (1,1,1)^\top/\sqrt{3}, \qquad \mathbf{d}_2 = (1,-1,-1)^\top/\sqrt{3}, \tag{25}$$

$$\mathbf{d}_3 = (-1,1,-1)^\top/\sqrt{3}, \qquad \mathbf{d}_4 = (-1,-1,1)^\top/\sqrt{3}, \tag{26}$$

and four fixed spatial frequencies,

$$(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (0.25, 0.5, 1, 2). \tag{27}$$

For mode (s, r),

$$\boldsymbol{\omega}_{sr}^{\ell h} = \lambda_s B_{\mathrm{dyn}}^{\ell h} \mathbf{d}_r, \qquad \phi_{i,sr}^{\ell h} = (\boldsymbol{\omega}_{sr}^{\ell h})^\top \widehat{\mathbf{p}}_i. \tag{28}$$

Together, $\boldsymbol{\delta}$, ρ, and β adapt the measurement directions, the relative weighting of the 16 direction–frequency modes, and the overall geometric strength, respectively. Object-center coordinates retain the units used by the backbone data pipeline.

A.4 Relation-Centric Reference Frame

RCRF uses observer yaw around the dataset vertical axis as the reference-frame signal. Let ψ be the continuous yaw. We quantize it to the nearest of eight uniformly spaced orientations. With $\Delta = 2\pi/8$,

$$k = \left\lfloor \frac{(\psi \bmod 2\pi) + \Delta/2}{\Delta} \right\rfloor \bmod 8, \qquad \widehat{\psi} = k\Delta. \tag{29}$$

We use the convention that yaw 0 faces the world $+x$ direction and positive yaw rotates the forward axis toward $+y$. The horizontal forward, right, and vertical up axes are

$$\mathbf{f} = (\cos\widehat{\psi}, \sin\widehat{\psi}, 0)^\top, \qquad \mathbf{r} = \mathbf{f} \times (0,0,1)^\top, \qquad \mathbf{u} = (0,0,1)^\top. \tag{30}$$

The corresponding world-to-observer transform is

$$R_v(\widehat{\psi}) = \begin{bmatrix} \sin\widehat{\psi} & -\cos\widehat{\psi} & 0 \\ \cos\widehat{\psi} & \sin\widehat{\psi} & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{31}$$

Let $v \in \{0,1\}$ indicate whether observer orientation is available for a sample. The coordinates used by the relation branch are

$$\widehat{\mathbf{p}}_i = \begin{cases} R_v(\widehat{\psi})(\mathbf{c}_i - \boldsymbol{\mu}), & v = 1, \\ \mathbf{c}_i, & v = 0, \end{cases} \tag{32}$$

where $\boldsymbol{\mu}$ is a shared scene-centering vector. The shared translation cancels in pairwise displacement,

$$R_v(\widehat{\psi})[(\mathbf{c}_j - \boldsymbol{\mu}) - (\mathbf{c}_i - \boldsymbol{\mu})] = R_v(\widehat{\psi})(\mathbf{c}_j - \mathbf{c}_i), \tag{33}$$

Hence, RCRF changes the directional interpretation of inter-object geometry without modifying the backbone object or scene representations. When orientation is unavailable, the same LASM branch operates on the original world-coordinate centers.

A.5 Joint Relation Formation in Attention

Let $\mathbf{q}_i^{\mathrm{raw}}, \mathbf{k}_i^{\mathrm{raw}} \in \mathbb{R}^{d_h}$ denote the per-head query and key before the backbone one-dimensional RoPE, and let $\mathbf{q}_i^{\mathrm{lang}}, \mathbf{k}_i^{\mathrm{lang}}$ denote their standard RoPE-transformed counterparts. In parallel with this unchanged content path, layer-specific bias-free projections shared across heads produce a 32-dimensional spectral branch,

$$\mathbf{q}_i^{\mathrm{spec}} = P_Q^\ell \mathbf{q}_i^{\mathrm{raw}}, \qquad \mathbf{k}_i^{\mathrm{spec}} = P_K^\ell \mathbf{k}_i^{\mathrm{raw}}, \qquad P_Q^\ell, P_K^\ell \in \mathbb{R}^{32 \times d_h}. \tag{34}$$

For grouped-query attention, keys are expanded to query-head resolution before the head-wise spectral computation. The spectral branch is constructed from pre-RoPE features so that sequence-position RoPE is applied only to the content path.

The spectral features are partitioned into sixteen two-dimensional blocks, one for each direction–frequency mode (s, r). Define

$$\mathcal{R}(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}. \tag{35}$$

For each mode,

$$\widehat{\mathbf{q}}^{\mathrm{spec}}_{i,sr} = m_i \sqrt{\beta^{\ell h} \rho^{\ell h}_{sr}}\, \mathcal{R}(\phi^{\ell h}_{i,sr}) \mathbf{q}^{\mathrm{spec}}_{i,sr}, \tag{36}$$

$$\widehat{\mathbf{k}}^{\mathrm{spec}}_{i,sr} = m_i \sqrt{\beta^{\ell h} \rho^{\ell h}_{sr}}\, \mathcal{R}(\phi^{\ell h}_{i,sr}) \mathbf{k}^{\mathrm{spec}}_{i,sr}. \tag{37}$$

Using $\mathcal{R}(\phi_i)^\top \mathcal{R}(\phi_j) = \mathcal{R}(\phi_j - \phi_i)$, their inner product yields

$$G^{\ell h}_{ij} = m_i m_j \beta^{\ell h} \sum_{s,r} \rho^{\ell h}_{sr} (\mathbf{q}^{\mathrm{spec}}_{i,sr})^\top \mathcal{R}\big((\boldsymbol{\omega}^{\ell h}_{sr})^\top (\widehat{\mathbf{p}}_j - \widehat{\mathbf{p}}_i)\big)\, \mathbf{k}^{\mathrm{spec}}_{j,sr}. \tag{38}$$

This form makes explicit the same three conditioning sources as the relation operator in Sec. 3.4: geometry supplies the object displacement, RCRF determines its directional coordinate system, and OAGC/LASM determine how that displacement is measured.

The resulting pre-softmax attention score is

$$S^{\ell h}_{ij} = \frac{(\mathbf{q}^{\mathrm{lang}}_i)^\top \mathbf{k}^{\mathrm{lang}}_j + G^{\ell h}_{ij}}{\sqrt{d_h}} + M_{ij}, \tag{39}$$

where M_{ij} is the original additive attention mask. The value branch, softmax, value aggregation, and output projection are unchanged. The relation branch is applied to the latter half of the decoder: layers 8–15 for the 1B backbone and layers 16–31 for the 7B backbone.

A.6 Training and Initialization

Objective and trainable parameters. FROG is trained with the backbone autoregressive answer-token objective; prompt, object-sequence, and padding positions are excluded from the answer loss. No additional relation-label or geometry-specific auxiliary supervision is introduced.

Backbone adaptation follows the corresponding base-model training setup, with LoRA rank 16, $\alpha = 16$, and dropout 0.05 applied to the attention query/key/value/output projections and the MLP gate/up/down projections. FROG-specific trainable parameters comprise OAGC, its control heads, the spectral query/key projections, and the learnable static measurement frames; RCRF introduces no learned parameters.

Optimization. The main joint-training runs use an 8-bit Adam optimizer with $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.02, and a learning rate of 1.6×10^{-4}. We use a mini-batch size of 8, cosine learning-rate decay with 0.1 epoch of warm-up, a minimum learning-rate multiplier of 0.01, gradient clipping at 0.01, bfloat16 autocast, and gradient checkpointing. The seven-task training collection is concatenated and shuffled, and the principal joint-training stage is run for three epochs. The 7B Mask3D experiments follow the corresponding two-stage GT-to-Mask3D training procedure.

Initialization. The routing, rotation, and amplitude output heads are zero-initialized, giving $\rho_{sr} = 1$, $\boldsymbol{\delta} = \mathbf{0}$, and $\beta = 0.02$ at the start of optimization. The spectral query/key projections are initialized as truncated identity projections into the 32-dimensional spectral space, and the static measurement frames are initialized independently across active layer-head pairs.

A.7 Inference with KV Caching

During generation, FROG uses the backbone KV-cache interface. The relation-enhanced context is constructed during prefill over the input sequence, and subsequent autoregressive decoding reuses the cached context without changing the backbone generation interface.

A.8 Model Dimensions and Parameter Overhead

The relation branch is applied to the latter half of each decoder, with $d_{\text{spec}} = 32$ in both model scales. The relevant dimensions are summarized below; layer indices are zero-based. Here, Aug. dim. denotes $d_h + d_{\text{spec}}$.

Backbone	Hidden	Layers	Q/KV heads	Head dim.	Active layers	Aug. dim.
Llama-3.2-1B-Instruct	2048	16	32/8	64	8–15	96
Vicuna-7B-v1.5	4096	32	32/32	128	16–31	160

For the Chat-Scene-1B configuration used in Table 4, FROG adds 3,988,480 trainable parameters (3.988M), corresponding to 0.315% of the full model. The added parameters comprise OAGC and its control network, the spectral query/key projections, and the learnable measurement frames. FROG does not increase sequence length. In active layers, only the query/key representation receives the additional 32-dimensional spectral branch; the value width is unchanged. Relation formation is computed from token-wise spectral features without materializing a dense $T \times T \times 32$ relation tensor.

B Experimental Protocol Details

B.1 Training Tasks and Data Processing

Following Sec. 4.1, joint training uses seven tasks: ScanRefer, Multi3DRefer, Scan2Cap, Nr3D-caption, ScanQA, SQA3D, and ObjAlign. ASR and Real-3DQA are reserved for evaluation. The seven training datasets are concatenated and shuffled without task-specific resampling.

Scenes contain at most 150 objects. In the GT-object setting, centers are computed from aligned 3D instances and use the metric coordinate system of the backbone data pipeline. All tokens associated with the same object share its center. No additional scene-scale normalization is applied before the spatial frequencies in Eq. (27).

B.2 Backbone and Object Settings

The backbone and object configurations corresponding to the main comparison in Table 1 instantiate FROG on both Chat-Scene and 3DGraphLLM while preserving their original object-feature construction. Chat-Scene uses its standard per-object token organization, whereas 3DGraphLLM additionally includes graph-neighbor structure. The 1B setting uses Llama-3.2-1B-Instruct with GT objects, and the 7B setting uses Vicuna-7B-v1.5 with Mask3D objects. In both cases, FROG associates relation coordinates with the existing object-level tokens rather than introducing additional scene tokens.

B.3 Reference-Frame Inputs

For ScanRefer and SQA3D, the viewpoint annotations used in the main experiments are expressed in the aligned scene coordinate frame, converted to the yaw convention in Eq. (30), and quantized according to Eq. (29). The resulting yaw is used only by RCRF; it is not appended to the language sequence or used to modify the backbone scene features.

For tasks without an associated observer orientation, including the standard Multi3DRefer and ASR evaluations, the relation branch follows the world-coordinate path in Eq. (32). Real-3DQA uses the viewpoint rotation associated with each viewpoint variant, converted to the same yaw convention. OAGC and LASM remain unchanged across these settings; orientation, when available, enters FROG only through RCRF. The viewpoint-source variants in Table 3 follow the source combinations reported there.

B.4 Evaluation Protocol

For the results in Table 1, ScanRefer is evaluated with Acc@0.25 and Acc@0.50 using axis-aligned 3D IoU, together with the corresponding metrics on the Multiple subset. Multi3DRefer

uses F1@0.25 and F1@0.50 after matching predicted and target object sets at the respective IoU thresholds. SQA3D reports top-1 exact match (EM@1), and ASR follows the same grounding metrics as ScanRefer. Real-3DQA uses Exact Match (EM), Relaxed Match (EM-R), and Viewpoint Rotation Score (VRS), as reported in Table 2.

Unless an analysis explicitly changes the decoding strategy, generation uses five-beam search with up to 64 newly generated tokens, repetition penalty 3.0, and KV caching. No test-time augmentation is used.

B.5 Ablation and Reference-Frame Analysis

Table 5 follows the component definitions in the main paper. RB denotes the static relation branch, LA enables language-adaptive spatial measurement, and RF enables the relation-centric reference frame. For the OAGC ablations, *w/o OA* replaces the order-aware query representation with order-insensitive pooling; *w/o* δ disables query-dependent measurement-frame rotation; *w/o* ρ uses uniform direction–frequency routing; and *w/o* β replaces query-dependent amplitude control with a static amplitude.

Fig. 3 varies the reference-frame or routing signal while leaving the corresponding model definition unchanged: panel (a) compares orientation granularities, panel (b) perturbs the yaw signal before the standard 8-bin transformation, and panel (c) averages routing weights over evaluation prompts, attention heads, and tetrahedral directions to visualize layer–frequency preferences.

B.6 Efficiency Profiling

The inference-cost measurements in Table 4 use a single NVIDIA GeForce RTX 4090 with batch size 1. We profile 16 uniformly sampled ScanRefer validation examples with greedy decoding, KV caching, and a maximum generation length of 32 new tokens. The first example is warmed up twice before measurement. GPU synchronization is used around timed regions, and peak allocated memory is reset before each measured run. Latency and memory are averaged across the 16 examples using the same profiling pipeline and hardware.

The parameter increase is obtained by enumerating FROG-specific trainable modules at runtime. End-to-end latency measures the complete generation call, including controller and relation-formation overhead.

C Results Across Random Seeds

We further evaluate the stability of FROG by repeating the four principal backbone and object-setting configurations with three independent random seeds ($\{42, 67, 99\}$). Tables 6 and 7 report the results on ScanRefer, Multi3DRefer, and SQA3D. FROG maintains consistently strong performance across random seeds and configurations, with only small variations across repeated runs, demonstrating the stability and reproducibility of the proposed relation modeling.

D Additional Qualitative Results

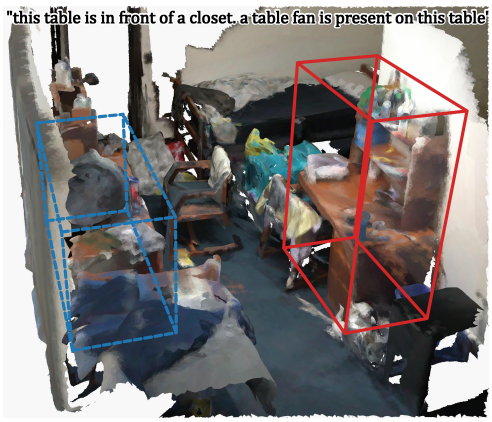
We provide additional representative qualitative examples using the same visualization convention as Fig. 4, covering diverse spatial relations and challenging same-category distractors. Solid red boxes denote Baseline (Huang et al., 2024a), while dashed blue boxes denote FROG.

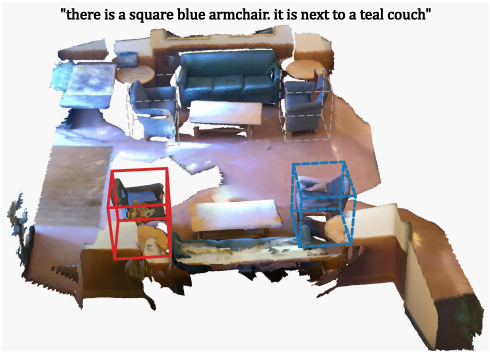
Table 6: Multi-seed results on ScanRefer. We report three independent runs for each principal FROG configuration. Shaded rows report mean ± standard deviation.

Model	Detector / Segmentation	LLM	Seed	ScanRefer			
				Acc @0.25	Acc @0.50	Multi @0.25	Multi @0.50
Chat-Scene + FROG	GT	Llama-3.2-1B-Instruct	42	65.3	65.1	58.8	58.7
			67	64.4	64.3	57.9	57.8
			99	64.5	64.3	57.8	57.6
			Mean	64.73 ± 0.40	64.57 ± 0.38	58.17 ± 0.45	58.03 ± 0.48
3DGraphLLM + FROG	GT	Llama-3.2-1B-Instruct	42	65.0	65.0	58.4	58.3
			67	64.6	64.4	57.9	57.6
			99	65.5	65.3	58.9	58.7
			Mean	65.03 ± 0.37	64.90 ± 0.37	58.40 ± 0.41	58.20 ± 0.45
Chat-Scene + FROG	Mask3D	Vicuna-7B-v1.5	42	62.7	56.5	56.2	50.3
			67	63.3	57.2	57.0	51.3
			99	62.8	56.8	56.4	50.8
			Mean	62.93 ± 0.26	56.83 ± 0.29	56.53 ± 0.34	50.80 ± 0.41
3DGraphLLM + FROG	Mask3D	Vicuna-7B-v1.5	42	62.2	55.9	55.8	49.9
			67	62.1	56.2	56.0	50.3
			99	62.5	56.9	56.4	50.9
			Mean	62.27 ± 0.17	56.33 ± 0.42	56.07 ± 0.25	50.37 ± 0.41

Table 7: Multi-seed results on Multi3DRefer and SQA3D. We report three independent runs for each principal FROG configuration. Shaded rows report mean ± standard deviation.

Model	Detector / Segmentation	LLM	Seed	Multi3DRefer		SQA3D
				F1 @0.25	F1 @0.50	EM @1
Chat-Scene + FROG	GT	Llama-3.2-1B-Instruct	42	61.6	61.5	52.8
			67	62.5	62.4	52.9
			99	61.7	61.5	53.1
			Mean	61.93 ± 0.40	61.80 ± 0.42	52.93 ± 0.12
3DGraphLLM + FROG	GT	Llama-3.2-1B-Instruct	42	61.9	61.8	53.0
			67	61.6	61.4	53.0
			99	62.1	61.9	53.3
			Mean	61.87 ± 0.21	61.70 ± 0.22	53.10 ± 0.14
Chat-Scene + FROG	Mask3D	Vicuna-7B-v1.5	42	60.0	55.3	53.6
			67	60.2	55.6	53.4
			99	59.8	55.3	53.9
			Mean	60.00 ± 0.16	55.40 ± 0.14	53.63 ± 0.21
3DGraphLLM + FROG	Mask3D	Vicuna-7B-v1.5	42	60.3	55.4	54.7
			67	60.1	55.6	54.2
			99	60.9	56.1	54.2
			Mean	60.43 ± 0.34	55.70 ± 0.29	54.37 ± 0.24







E LIMITATIONS

Our evaluation is currently limited to established object-level 3D vision-language benchmarks and predominantly static scene settings. In addition, the current reference-frame formulation focuses primarily on observer orientation for directional spatial reasoning. Future work could extend FROG to more diverse and complex 3D environments, richer reference-frame settings, and dynamic or interactive scenarios with evolving spatial relations. We also plan to explore more general and flexible relation modeling for broader spatial reasoning tasks and real-world applications.