REASONING COMPLEX RELATIONS: OBJECT-LEVEL COGNITIVE HYPERGRAPH FOR 3D INTENTION GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding and grounding human intentions is an important direction in embodied intelligence. Recently, 3D Intention Grounding (3D-IG) has been proposed to localize target objects in 3D scenes according to abstract, ambiguous, and non-descriptive natural-language intentions. Unlike conventional 3D Visual Grounding, which aligns 3D objects with explicit referring descriptions, 3D-IG requires the model to infer latent functional semantics, such as object attributes and affordances, from human intentions. Therefore, 3D-IG requires reasoning beyond simple cross-modal matching, involving intention understanding, object relations, and scene-level constraints. However, existing methods still largely rely on implicit intent-object alignment, without explicitly modeling multi-object functional relations and scene-level context. To address these challenges, we propose a Cognitive Hypergraph framework that, inspired by the coarse-to-specific process of human cognition, reformulates 3D-IG from local matching into structured reasoning over intentions, functional object groups, and scene context. Specifically, candidate objects are represented as graph nodes, while higher-order hyperedges organize them into intention-conditioned functional groups. By introducing global intention and scene-level priors, the framework jointly models target, intention, functional groups, and scene context, and performs structured reasoning over the cognitive hypergraph to obtain intention-aware candidate representations for final grounding. Experiments on 3D Intention Grounding and 3D Visual Grounding benchmarks demonstrate consistent performance gains across different grounding settings.

1 INTRODUCTION

With the rapid development of 3D visual perception (Li et al., 2025a; Corsetti et al., 2025; Zhang et al., 2025) and multimodal understanding (Ning et al., 2025), grounding natural language in 3D scenes has attracted increasing attention (Shi et al., 2024; Guo et al., 2025). Unlike conventional 3D Visual Grounding (Zhu et al., 2024; Kang et al., 2025a), which mainly relies on explicit descriptions such as color, position, or geometric attributes, 3D Intention Grounding (3D-IG) requires the model to infer functional requirements from abstract, ambiguous, and non-descriptive intents, and then localize the object that satisfies the intended requirement in a 3D scene. For example, in real-world Human-Computer Interaction (Majumdar et al., 2024) and Embodied AI scenarios (Lin et al., 2025; Wang et al., 2024a), users are more likely to say 'I need something that can support my back' or 'I need an object that can hold water,' rather than directly describing the appearance of the target. Therefore, 3D-IG requires reasoning beyond simple cross-modal matching (Li et al., 2025c), involving intention understanding, functional relations among multiple objects, and scene context.

Despite recent progress in intention modeling, current 3D-IG methods still face a fundamental challenge: the cognitive gap between abstract human intentions and 3D physical objects. Most existing methods follow a cross-modal matching paradigm, directly learning implicit correspondences between language expressions and candidate objects. However, such implicit alignment overlooks the human cognitive process of understanding intentions and interpreting scenes, where abstract needs are typically resolved through intermediate reasoning rather than direct object matching. As illustrated on the left side of Fig. 1(a), existing methods (Kang et al., 2025b) mainly rely on local intent cues and align them with individual candidate objects, reducing 3D-IG to local text-object matching. Although this strategy can capture shallow semantic correlations, it struggles to reason over functional relations among objects and scene context, resulting in reasoning gaps in complex environments and limited

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.

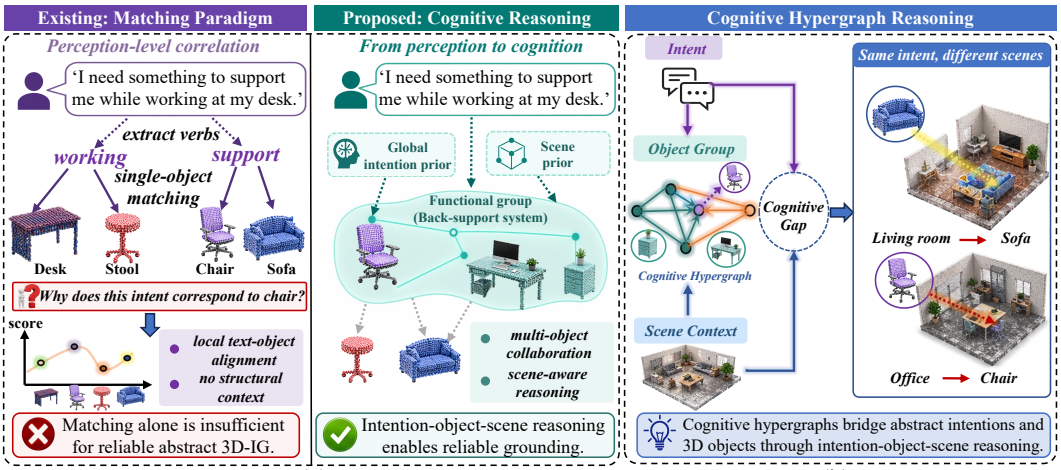


Figure 1: (a) Existing 3D-IG methods rely on local intent-object matching, overlooking functional relations among objects and scene context. Our method reformulates 3D-IG as cognitive reasoning by organizing candidates into intention-conditioned functional groups guided by a global intention prior and an intention-aware scene prior. (b) Overview of cognitive hypergraph reasoning, which jointly models intention, functional groups, and scene context to bridge the gap between abstract intentions and 3D objects. The right example shows the same intent may indicate different targets across scenes.

interpretability. A closer analysis reveals two major limitations. First, existing methods mainly match intention queries with individual candidate objects, overlooking intention-conditioned functional relations among objects. In realistic scenes, target selection often depends on functional relations among multiple objects rather than isolated object semantics, requiring reasoning over how these objects jointly support a specific intention. For example, an 'office-work' intention is more naturally associated with a functional group consisting of a desk, chair, and monitor. Second, existing methods lack explicit modeling of global scene context, even though the same intent may correspond to different targets across scenes. As shown in Fig. 1(b), the target for back support may be a sofa in a living room but a chair in an office. These limitations suggest that local intent-object alignment alone is insufficient for robust intention grounding.

Based on these observations, we argue that 3D intention grounding should be formulated as a cognitive reasoning problem over multi-object structures and scene context, rather than as simple text-object matching. As illustrated on the right side of Fig. 1(a), inspired by the coarse-to-specific process of human cognition (Bayne et al., 2019), we model reasoning from holistic scene understanding, through latent functional group formation, to object-level target identification. Different from conventional hypergraphs that mainly encode higher-order structural or semantic dependencies, our cognitive hypergraph organizes candidates into intention-conditioned functional groups, with construction guided by intention and reasoning further informed by scene context (Fujita, 2025).

To this end, we propose a Cognitive Hypergraph framework for 3D intention reasoning. The core idea is to instantiate the above cognitive reasoning process as a cognitive hypergraph, where candidate objects are represented as graph nodes and higher-order hyperedges encode intention-conditioned functional groups, as shown in Fig. 1(b). Different from conventional hypergraphs that mainly capture higher-order correlations among nodes, our cognitive hypergraph further interprets such relations under the current intention and scene context, modeling not only which objects are related but also how their functional relations depend on the current intention. Specifically, we introduce a global intention prior and an intention-aware scene prior to jointly guide hypergraph reasoning, modulating object relations by intention semantics and scene context. The reasoning process aggregates object nodes into context-modulated hyperedges and propagates group-level information back to candidate nodes, yielding intention-conditioned representations for grounding. In this way, target localization is achieved through joint reasoning over target, intention, functional groups, and scene context, thereby bridging the cognitive gap between abstract intentions and physical objects.

Experiments on 3D Intention Grounding and 3D Visual Grounding benchmarks demonstrate the effectiveness of our method, with consistent performance gains across different grounding settings.

To summarize, our contributions are as follows:

• We propose a cognitive-hypergraph-based framework for 3D intention reasoning, which reformulates 3D-IG from a cross-modal matching paradigm into a cognitive reasoning paradigm.

• We develop a unified target-intention-scene modeling mechanism that captures intention-conditioned functional relations through higher-order hyperedges and integrates a global intention prior with an intention-aware scene prior for multi-object reasoning.

• We design a prior-guided cognitive hypergraph reasoning process that propagates functional group information to object nodes for target grounding.

2 Related Work

2.1 3D Grounding and Intention Grounding

3D Visual Grounding localizes target objects in 3D scenes from natural-language descriptions. Existing methods have improved grounding through proposal-language alignment (Chen et al., 2020; Achlioptas et al., 2020), graph-based relation modeling (Huang et al., 2021), multi-view or point-level reasoning (Huang et al., 2022; Luo et al., 2022), dense text-object alignment (Wu et al., 2023), and 3D vision-language pretraining (Zhu et al., 2023; Jia et al., 2024; Zhu et al., 2024). However, these methods mainly handle descriptive expressions containing explicit object categories, attributes, or spatial cues. 3D Intention Grounding (3D-IG) further requires inferring latent functional needs from abstract intentions. IntentNet (Kang et al., 2025b) introduces this task, but existing methods still largely rely on implicit intent-object matching, making it difficult to explicitly model functional semantics, multi-object collaboration, and scene-level constraints. In contrast, our method formulates 3D-IG as cognitive hypergraph reasoning over intention-oriented functional groups. More related works on 3D Intention Grounding, relation modeling, and functional 3D scene understanding are provided in Appendix A.1 and Appendix A.2.

2.2 Hypergraph-based higher-order Reasoning

Hypergraphs model higher-order dependencies by connecting multiple nodes with one hyperedge (Bretto, 2013), and have been widely studied in hypergraph neural networks and representation learning (Feng et al., 2019; Bai et al., 2021; Antelmi et al., 2023). Recent cognitive hypergraph studies further suggest the potential of hyperedges for representing richer cognitive associations beyond pairwise relations (Citraro et al., 2023; Fujita, 2025). However, existing hypergraph-based methods mainly capture structural or semantic correlations, answering which nodes are related, but not why they should be grouped under a specific intention or how the grouping should be constrained by 3D scene context. Our cognitive hypergraph instantiates nodes as physical object proposals and interprets hyperedges as intention-oriented functional groups guided by global intention and scene priors. Further discussions on cognitive hypergraphs are provided in Appendix A.3.

3 Method

As illustrated in Fig. 2, we propose a Cognitive Hypergraph Reasoning framework for 3D Intention Grounding. Given a natural-language intention query and a 3D scene point cloud, the framework first derives a global intention prior and an intention-aware scene prior. The intention prior then guides the construction of a cognitive hypergraph, where candidate objects are represented as nodes and higher-order hyperedges organize them into intention-conditioned potential functional groups. Based on the constructed hypergraph, the global cognitive priors further guide group-level reasoning and selective evidence propagation, producing enhanced candidate representations for final target grounding. Further details on the conceptual definition, functional interpretation, and task-specific instantiation of cognitive hypergraphs for 3D intention grounding are provided in Appendix B.1.

Problem Definition. In 3D Intention Grounding, given a 3D scene point cloud $X = \{p_i\}_{i=1}^{N}$ and a natural-language intention query $Q = \{w_t\}_{t=1}^{T}$, where $p_i \in \mathbb{R}^3$ denotes the coordinate of the i-th point and w_t denotes the t-th token, the goal is to localize the target object that best satisfies the intention. Specifically, from a candidate box set $\mathcal{B} = \{b_m\}_{m=1}^{M}$, the prediction is obtained by:

$$b^* = \arg\max_{b_m \in \mathcal{B}} s(b_m, Q), \tag{1}$$

where M is the number of candidate objects and b^* denotes the highest-scoring candidate. Unlike conventional 3D Visual Grounding, where language descriptions usually contain explicit visual

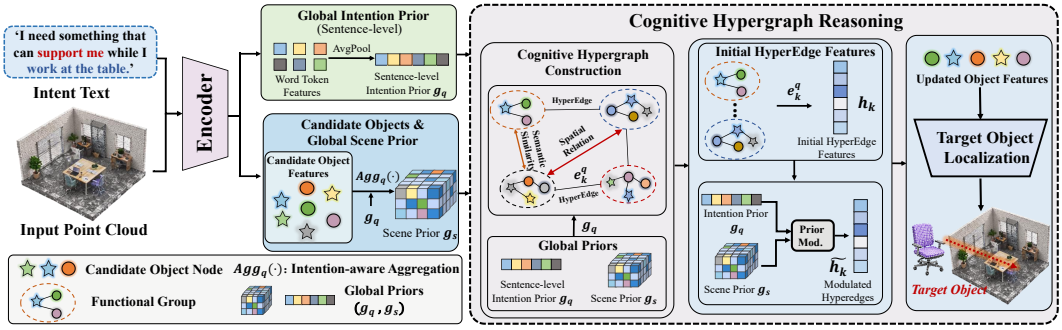


Figure 2: **The overall architecture**. The intent text and point cloud are first encoded to obtain the global intention and scene priors. Candidate objects serve as nodes in the cognitive hypergraph, while intention-conditioned functional relevance, semantic similarity, and spatial relations jointly determine higher-order functional hyperedges. Guided by the global cognitive priors, group-level information is further contextualized and selectively propagated back to object nodes, producing enhanced candidate representations for final target localization.

attributes such as color, shape, or spatial position, 3D-IG queries often express implicit functional demands, usage purposes, or scene-oriented behaviors. Therefore, this task requires structured reasoning from abstract intention to physical objects in the 3D scene, rather than shallow local text-object matching. For queries with multiple annotated targets, all annotated objects are treated as valid targets. Accordingly, Eq. (1) is used for Top-1 evaluation, while ranking-based metrics such as AP are computed using the grounding scores of all candidate objects.

3.1 Global Cognitive Prior Modeling

To provide global cognitive guidance for subsequent functional relation modeling and reasoning, we introduce two global cognitive priors: an intention prior and an intention-aware scene prior. For the language modality, we encode the intention query $Q = \{w_t\}_{t=1}^{T}$ with a text encoder and obtain token-level features $\{e_t\}_{t=1}^{T}$, where $e_t \in \mathbb{R}^d$. A sentence-level intention prior is computed as:

$$g_q = \mathrm{AvgPool}(\{e_t\}_{t=1}^{T}), \tag{2}$$

where $g_q \in \mathbb{R}^d$ captures the overall functional semantics of the current user intention.

For the scene modality, given M candidate objects, we obtain their object-level representations as the basis for modeling scene-level contextual information:

$$o_m \in \mathbb{R}^d, \quad m = 1, 2, \ldots, M, \tag{3}$$

where o_m denotes the feature representation of the m-th candidate object. These representations encode semantic, geometric, and spatial contextual information, providing the basis for subsequent relation modeling and reasoning. To capture scene context relevant to the current intention, we further aggregate candidate objects according to their intention-conditioned relevance:

$$g_s = \sum_{m=1}^{M} \eta_m^q o_m, \qquad \eta_m^q = \frac{\exp(f_s(o_m, g_q))}{\sum_{j=1}^{M} \exp(f_s(o_j, g_q))}, \tag{4}$$

where $f_s(\cdot)$ is a lightweight scoring function and η_m^q measures the relevance of each candidate to the current intention. Therefore, $g_s \in \mathbb{R}^d$ summarizes the scene context that is most relevant to the current user need, rather than treating all candidate objects equally. The intention prior g_q is used to guide functional relation construction, while g_q and g_s jointly provide global cognitive guidance for subsequent hypergraph reasoning. More details are provided in Appendix B.2.

3.2 Cognitive Hypergraph Construction

To model intention-conditioned functional relations among candidate objects, we construct a cognitive hypergraph whose relational structure is dynamically guided by the current user intention. Given the object representations $\{o_m\}_{m=1}^{M}$ obtained above, we represent each candidate object as a node v_m, forming the node set $\mathcal{V} = \{v_m\}_{m=1}^{M}$.

Algorithm 1: Cognitive Hypergraph Reasoning for 3D Intention Grounding

Input: Point cloud X, intention query Q, candidate box set $\mathcal{B}$.
Output: Matched target box b^*.
Step 1: Global Cognitive Prior Modeling
Encode Q to obtain token features $\{e_t\}_{t=1}^{T}$ and compute the global intention prior g_q.
Extract candidate object features $\{o_m\}_{m=1}^{M}$ from X and compute the scene prior g_s.
Step 2: Cognitive Hypergraph Construction
Instantiate candidate objects as nodes $\mathcal{V} = \{v_m\}_{m=1}^{M}$.
Compute intention-specific functional relevance a_{ij}^q and relation scores r_{ij}^q.
Construct functional hyperedges $\mathcal{E}^q = (e_k^q)_{k=1}^{K}$ and form the cognitive hypergraph $\mathcal{H}^q = (\mathcal{V}, \mathcal{E}^q)$.
Step 3: Cognitive Hypergraph Reasoning
for each hyperedge $e_k^q \in \mathcal{E}^q$ **do**
    Aggregate member node features to obtain h_k.
    Modulate the hyperedge representation to obtain $\tilde{h}_k$ using g_q and g_s.
end for
for each node $v_m \in \mathcal{V}$ **do**
    Aggregate intention-aware hyperedge messages to obtain u_m.
    Update the node feature to o'_m with gated residual reasoning using u_m, g_q, and g_s.
end for
Step 4: Final Target Grounding
for each updated candidate object o'_m **do**
    Compute grounding score s_m with the grounding head.
end for
Select the final target: $b^* = \arg\max_{b_m \in \mathcal{B}} s_m$.
Return b^*

To construct functional relations conditioned on the current intention, we first estimate the intention-specific functional relevance between two candidate objects o_i and o_j as:

$$a_{ij}^q = \sigma\left(f_q(o_i, o_j, g_q)\right), \tag{5}$$

where $f_q(\cdot)$ is a lightweight two-layer MLP that jointly encodes the two object features and the global intention prior. The resulting $a_{ij}^q \in (0, 1)$ measures how relevant the relation between o_i and o_j is to the current intention. We then define the intention-conditioned relation score as:

$$r_{ij}^q = [\alpha \cdot \mathrm{sim}(o_i, o_j) + (1 - \alpha) \cdot \phi(\Delta p_{ij})] \cdot a_{ij}^q, \tag{6}$$

where $\mathrm{sim}(\cdot, \cdot) \in [0, 1]$ denotes the normalized cosine similarity, Δp_{ij} denotes the relative displacement between the corresponding box centers, and $\phi(\Delta p_{ij})$ is a Gaussian-style spatial affinity function. The hyperparameter α is set to 0.6 based on validation performance to balance semantic similarity and spatial proximity. In this way, the same object pair can exhibit different relation strengths under different user intentions. Further analysis of α is provided in Section 4.4.

Based on the intention-conditioned relation matrix $R^q = [r_{ij}^q]$, each node serves as a center to form a fixed-size functional hyperedge with its top-ranked neighbors. Specifically, for node v_k, we construct:

$$e_k^q = \{v_k\} \cup \mathrm{Top}_{|e|-1}\left(\{r_{kj}^q\}_{j \neq k}\right), \tag{7}$$

where $|e|$ denotes the hyperedge size. Since each candidate node is used as a center node, the number of constructed hyperedges is $K = M$, and K is not an additional tunable hyperparameter. In our implementation, we set $|e| = 4$ by default according to the validation performance, with its sensitivity analyzed in Appendix C.4. Each hyperedge therefore captures a potential functional object group dynamically conditioned on the current intention. The resulting cognitive hypergraph is defined as:

$$\mathcal{H}^q = (\mathcal{V}, \mathcal{E}^q), \tag{8}$$

where $\mathcal{E}^q = (e_k^q)_{k=1}^{K}$ denotes the constructed hyperedges. Different from conventional hypergraphs that mainly encode relatively fixed higher-order correlations among nodes, $\mathcal{H}^q$ dynamically reorganizes object relations according to the current user intention. Therefore, the same 3D scene can induce different functional group structures under different intentions, providing an intention-specific relational basis for subsequent cognitive reasoning. More details are provided in Appendix B.3.

3.3 COGNITIVE HYPERGRAPH REASONING

After constructing the cognitive hypergraph and obtaining the global cognitive priors, we perform structured reasoning through two stages: node-to-hyperedge aggregation and hyperedge-to-node propagation. For the k-th intention-conditioned hyperedge e_k^q, we first aggregate its member nodes to obtain the initial hyperedge representation:

$$h_k = \mathrm{Agg}(\{o_m \mid v_m \in e_k^q\}), \tag{9}$$

where $\mathrm{Agg}(\cdot)$ denotes relation-weighted aggregation over the selected member nodes using their intention-conditioned relation scores, with the detailed formulation provided in Appendix B.3. To further incorporate the current intention and scene context, we inject the global intention prior g_q and the intention-aware scene prior g_s into the hyperedge update:

$$\tilde{h}_k = \psi_h([h_k \,\|\, g_q \,\|\, g_s]), \tag{10}$$

where $[\cdot\|\cdot]$ denotes feature concatenation and $\psi_h(\cdot)$ is a lightweight two-layer MLP for feature projection. The resulting $\tilde{h}_k$ therefore represents a context-aware hyperedge that integrates group-level object relations with intention semantics and scene context.

After hyperedge modulation, we selectively propagate group-level evidence back to object nodes according to their relevance. Let $\mathcal{N}(m) = \{k \mid v_m \in e_k^q\}$ denote the set of hyperedges incident to node v_m. The aggregated hyperedge message is computed as:

$$u_m = \sum_{k \in \mathcal{N}(m)} \beta_{mk}^q \tilde{h}_k, \qquad \beta_{mk}^q = \mathrm{softmax}_{k \in \mathcal{N}(m)}\, f_\beta(o_m, \tilde{h}_k, g_q), \tag{11}$$

where β_{mk}^q measures the relevance of each functional group to candidate node v_m under the current intention and scene context, and $f_\beta(\cdot)$ is implemented as a two-layer MLP with ReLU activation that outputs a scalar relevance score. We then update the candidate representation by combining the original node feature, the aggregated hyperedge message, and the global cognitive priors:

$$o'_m = o_m + \boldsymbol{\gamma}_m \odot \psi_o([o_m \,\|\, u_m \,\|\, g_q \,\|\, g_s]), \tag{12}$$

where $\boldsymbol{\gamma}_m \in (0,1)^d$ is a context-adaptive gating vector generated from the current node, hyperedge message, and global priors. The gated residual update preserves the original object representation while selectively incorporating functional evidence relevant to the current intention and scene context.

Overall, the reasoning process progressively refines candidate objects through intention-conditioned functional groups and context-aware evidence propagation, providing enhanced representations for subsequent target grounding. More details are provided in Appendix B.4.

3.4 INTENTION-AWARE GROUNDING AND OPTIMIZATION

After cognitive hypergraph reasoning, we obtain enhanced candidate representations $\mathcal{O}' = \{o'_m\}_{m=1}^M$, where each o'_m integrates the original object-level feature with intention-conditioned functional information and scene context, yielding a richer representation for intention-aware target grounding. Based on these representations, we compute the grounding score for each candidate by:

$$s_m = s(b_m, Q) = \mathrm{MLP}([o'_m \,\|\, g_q]), \tag{13}$$

where g_q provides global intention semantics for evaluating the compatibility between each reasoning-enhanced candidate and the current intention. The final prediction is obtained as:

$$b^* = \arg\max_{b_m \in \mathcal{B}} s_m, \tag{14}$$

where b^* denotes the predicted target box, and $\mathcal{B}$ is the set of candidate boxes. The highest-scoring candidate is used for Top-1 evaluation, while AP is computed over all candidate grounding scores.

The training objective consists of the detection loss, intention grounding loss, and cognitive consistency regularization:

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{ig} + \lambda_c \mathcal{L}_{cog}, \tag{15}$$

where $\mathcal{L}_{det}$ supervises candidate object generation and $\mathcal{L}_{ig}$ optimizes the grounding scores for intention-aware target selection. The consistency regularization $\mathcal{L}_{cog}$ encourages the updated object representations to remain aligned with their associated context-aware hyperedge representations, strengthening node–hyperedge semantic consistency during training. We set $\lambda_c = 0.1$ according to the ablation study in Section 4.4. Appendix B.5 details the cognitive consistency loss, while Appendix B.6 specifies multi-target proposal assignment, grounding supervision, and inference.

4 EXPERIMENTS

For the 3D Intention Grounding (3D-IG) task, we follow IntentNet (Kang et al., 2025b) to evaluate cross-modal reasoning between natural-language intentions and 3D object grounding. We report results on both validation and test sets and compare our method with representative baselines under the same setting. We further conduct experiments on 3D Visual Grounding to assess the general effectiveness of our method in localization accuracy and robustness.

4.1 EXPERIMENTAL SETUP

Dataset. For 3D Intention Grounding, we follow the official evaluation protocol of Intent3D (Kang et al., 2025b), built upon ScanNet (Dai et al., 2017). We further evaluate the generality of our method on three 3D Visual Grounding benchmarks: Nr3D (Achlioptas et al., 2020), Sr3D (Achlioptas et al., 2020), and ScanRefer (Chen et al., 2020). More dataset details are provided in Appendix C.1.

Evaluation Metric. Following standard 3D grounding protocols, we evaluate localization performance based on the correspondence between predicted and ground-truth objects. For Intent3D, we report Top-1 Accuracy and Average Precision (AP) under IoU thresholds of 0.25 and 0.5. For the 3D Visual Grounding benchmarks, we follow their standard grounding accuracy protocols.

4.2 IMPLEMENTATION DETAILS

We follow IntentNet (Kang et al., 2025b) and use PointNet++ (Qi et al., 2017) and RoBERTa (Liu et al., 2019) as the point-cloud and text encoders, respectively. The model is trained for 120 epochs on four RTX 3090 GPUs with a batch size of 16. We use 1024 point-cloud tokens, set $|e| = 4$, $\alpha = 0.6$, and $\lambda_c = 0.1$. More details are provided in Appendix C.2.

4.3 COMPARISON WITH THE STATE OF THE ART

Validation on 3D Intention Grounding. Table 1 reports the results on the Intent3D validation and test sets. Following the official protocol, we compare with expert models, foundation models, LLM-based methods, and the task-specific IntentNet (Kang et al., 2025b). Our method achieves the best performance in both Top-1 Accuracy and Average Precision. In particular, at IoU@0.5, it improves the validation Top-1 Accuracy from 40.83% to 45.78% over IntentNet. The gains mainly come from the proposed cognitive hypergraph reasoning. By organizing candidate objects as nodes and constructing higher-order functional hyperedges, our method explicitly models intention-oriented object collaboration. The global intention and scene priors further guide hyperedge modulation and node updating, enabling the model to infer targets under both functional intention semantics and scene-level constraints. This improves representation discriminability for 3D intention grounding.

Table 1: 3D Intention Grounding results on Intent3D validation and test sets. Acc denotes Top1-Acc. ^{0} indicates zero-shot results. The best results are in **bold**, and the second-best results are underlined.

Method	Detector	Val Set				Test Set			
		Acc@0.25	Acc@0.5	AP@0.25	AP@0.5	Acc@0.25	Acc@0.5	AP@0.25	AP@0.5
BUTD-DETR (Jain et al., 2022)	GroupFree (Liu et al., 2021)	47.12	24.56	31.05	13.05	47.86	25.74	31.41	13.46
EDA (Wu et al., 2023)	GroupFree (Liu et al., 2021)	43.11	18.91	14.02	5.00	44.00	19.62	14.56	5.18
3D-ViSTA (Zhu et al., 2023)	Mask3D (Schult et al., 2022)	42.76	30.37	36.10	19.93	43.88	31.44	37.29	22.00
Chat-3D-v2^0 (Huang et al., 2023)	PointGroup (Jiang et al., 2020)	5.86	5.24	0.15	0.13	5.63	4.93	0.14	0.11
Chat-3D-v2 (Huang et al., 2023)	PointGroup (Jiang et al., 2020)	36.71	32.78	3.23	2.58	33.46	29.32	2.67	2.10
IntentNet (Kang et al., 2025b)	GroupFree (Liu et al., 2021)	58.34	40.83	41.90	25.36	58.92	42.28	44.01	27.60
Ours	GroupFree (Liu et al., 2021)	**64.18**	**45.78**	**44.91**	**29.26**	**65.39**	**46.46**	**47.71**	**32.01**

Visualization Analysis. Figure 3 presents qualitative comparisons between our method and IntentNet on the Intent3D dataset. Our method achieves more accurate intention-aware localization, especially when the query requires functional reasoning beyond direct text-object matching. In the first row, the intent 'rest my feet while sitting on a couch' implies that the target should support the feet and be spatially and functionally related to the couch. While IntentNet only localizes partial candidates, our method better captures all objects consistent with the ground truth by modeling the functional relation between the couch and nearby support-related objects. In the second row, IntentNet mistakenly selects a suitcase, likely due to its coarse transport-related semantics. In contrast, our method correctly localizes the laundry basket by jointly considering intention semantics, object functionality, and scene context. These results show that cognitive hypergraph reasoning effectively captures intention-oriented object collaboration and reduces ambiguity caused by shallow functional similarity. Further visual and failure case analyses are provided in Appendix D.1 and D.2.

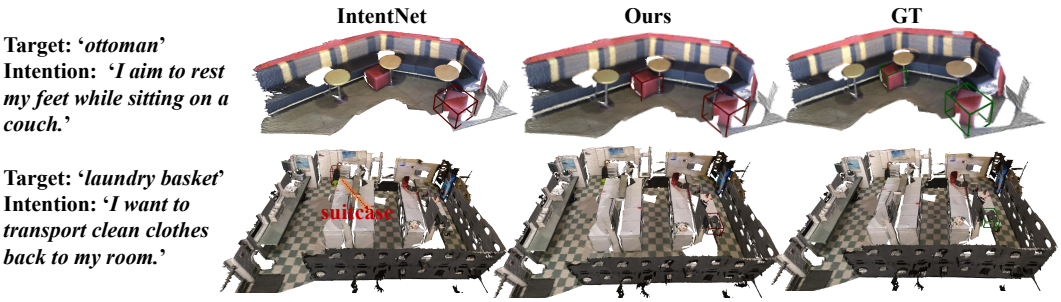


Figure 3: **Qualitative comparison on Intent3D.** Our method localizes intention-relevant objects more accurately by modeling object collaboration and scene context.

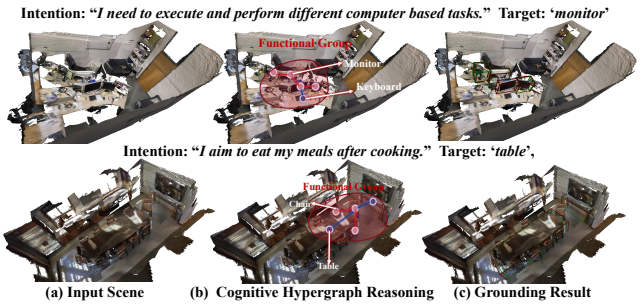


Figure 4: Hypergraph reasoning visualization analysis.

Hypergraph Reasoning Visualization Analysis. Figure 4 visualizes the hypergraph reasoning process for two representative examples. Given an intention query, the model identifies candidate object nodes and constructs intention-oriented functional groups for target grounding. In the first case, the functional group for computer-based tasks mainly contains the monitor and keyboard, which provide core support for task execution. In the second case, the model forms a dining-related group centered on the table and surrounding supportive objects, explaining the post-cooking dining intention. These results show that hypergraph reasoning provides interpretable functional evidence for target localization.

Validation on 3D Visual Grounding. To further evaluate the generality of our method, we conduct experiments on 3D Visual Grounding. Although this task mainly relies on explicit referring expressions, challenging cases still require reasoning over object relations and scene context. Our cognitive hypergraph addresses this by modeling higher-order object dependencies and refining candidate representations with global intention and scene priors. As shown in Table 2, our method achieves the best performance on both Nr3D and Sr3D. Compared with PQ3D (Zhu et al., 2024), the overall accuracy improves from 66.7 to 69.1 on Nr3D and from 79.7 to 83.5 on Sr3D, with consistent gains on the hard splits. These results demonstrate the generality of cognitive hypergraph reasoning beyond 3D intention grounding. Additional results on ScanRefer (Chen et al., 2020) are provided in Appendix C.3.

Table 2: Performance Comparison on 3D Visual Grounding for Nr3D and Sr3D Datasets.

Methods	Nr3D		Sr3D	
	Overall	Hard	Overall	Hard
ReferIt3D (Achlioptas et al., 2020)	35.6	27.9	40.8	31.5
TGNN (Huang et al., 2021)	37.3	30.6	45.0	36.9
3D-SPS (Luo et al., 2022)	51.5	45.1	62.6	65.4
EDA (Wu et al., 2023)	52.1	46.1	68.1	62.9
MVT (Huang et al., 2022)	55.4	49.1	64.5	58.8
G^3-LQ (Wang et al., 2024b)	58.4	50.7	73.1	66.3
MiKASA (Chang et al., 2024)	64.4	59.4	75.2	67.3
GPS (Jia et al., 2024)	64.9	57.8	77.5	71.6
MA^2TransVG (Xu et al., 2024)	65.2	57.6	73.9	69.3
PQ3D (Zhu et al., 2024)	66.7	58.7	79.7	72.8
SeeGround (Li et al., 2025a)	46.1	38.3	–	–
TSP3D (Guo et al., 2025)	48.7	–	57.1	–
Ours	**69.1**	**61.3**	**83.5**	**76.3**

4.4 Ablation Study

Component Analysis. Table 3 presents component ablations starting from a full-sentence matching baseline without IntentNet's verb–object alignment; variant definitions and controlled comparisons of relational structure and scene pooling are provided in Appendices B.7 and C.5, respectively. CHC improves AP@0.5 from 20.36% to 23.37%, demonstrating the benefit of intention-conditioned higher-order relation modeling, while GCP alone reaches 22.60%,

Table 3: Component ablation on the Intent3D validation set. CHC, GCP, and CHR denote Cognitive Hypergraph Construction, Global Cognitive Prior Injection, and Cognitive Hypergraph Reasoning, respectively. $\mathcal{L}_{\text{cog}}$ denotes consistency regularization.

Methods				Metrics			
CHC	GCP	CHR	$\mathcal{L}_{\text{cog}}$	Top1-Acc@0.25	Top1-Acc@0.5	AP@0.25	AP@0.5
				52.34	36.83	34.60	20.36
✓				55.82	39.64	37.96	23.37
	✓			54.15	38.23	37.64	22.60
✓		✓		58.15	41.67	39.76	24.37
✓	✓			60.75	42.18	40.50	25.19
✓	✓	✓		62.33	43.27	42.07	26.34
✓	✓	✓	✓	**64.18**	**45.78**	**44.91**	**29.26**

confirming the usefulness of global intention and scene context. Combining CHC with CHR further improves AP@0.5 to 24.37%, while CHC+GCP achieves 25.19%, showing the complementarity of relational structure and global cognitive priors. Integrating CHC, GCP, and CHR further increases AP@0.5 to 26.34%, demonstrating the benefit of prior-guided adaptive hypergraph reasoning. Finally, adding cognitive consistency regularization yields the best AP@0.5 of 29.26%, further validating the importance of node–hyperedge consistency for robust and effective structured reasoning.

Efficiency Analysis. Table 4 reports the computational efficiency of different model variants, where inference latency is measured as the average time per sample. Compared with the Baseline, adding CHC slightly increases latency from 26.3 ms to 27.4 ms and parameters from 148.2M to 148.6M, due to the additional intention-conditioned relation modeling and group aggregation. Introducing GCP further increases latency to 28.0 ms and parameters to 148.8M. With the complete cognitive hypergraph reasoning module, the full model requires 30.3 ms per sample and 149.6M parameters. Compared with IntentNet, our model incurs only moderate additional computation while maintaining the same peak memory usage of 6.1 GB, demonstrating a favorable accuracy–efficiency trade-off.

Table 4: Efficiency comparison in terms of training time, inference latency, peak memory usage, and model parameters.

Method	Training Time (hours)	Inference (ms)	Memory (GB)	Params (M)
IntentNet (Kang et al., 2025b)	32	27.1	6.1	148.6
Baseline	28	26.3	6.1	148.2
Baseline + CHC	29	27.4	6.1	148.6
Baseline + CHC + GCP	30	28.0	6.1	148.8
Full Model (Ours)	34	30.3	6.1	149.6

Analysis of the Balance Coefficient α in Cognitive Hypergraph Construction. In cognitive hypergraph construction, the balance coefficient α controls the relative contribution of semantic similarity and spatial affinity in Eq. (6). A smaller α overemphasizes spatial proximity and may connect nearby but functionally irrelevant objects, while a larger α weakens spatial constraints and may introduce unreasonable semantic associations. As shown in Figure 5, the model achieves the best performance on the Intent3D validation set when $\alpha = 0.6$, indicating that semantic similarity should be slightly emphasized while spatial relations remain necessary for constructing scene-consistent functional groups. Additional analyses of the hyperedge size $|e|$ are provided in Appendix C.4.

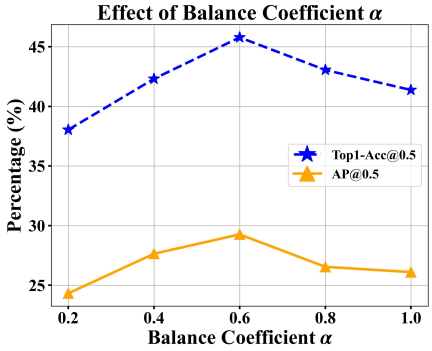


Figure 5: Effect of Balance Coefficient α in Cognitive Hypergraph Construction.

Analysis of the Cognitive Consistency Weight λ_c. To evaluate the effect of cognitive consistency regularization, we conduct an ablation study on the validation set by varying λ_c in the training objective. As shown in Table 5, removing the consistency term ($\lambda_c = 0$) results in lower performance, with AP@0.5 and Top1-Acc@0.5 decreasing to 26.34 and 43.27, respectively. Increasing λ_c to 0.1 improves semantic consistency between candidates and functional groups, achieving the best AP@0.5 of 29.26 and Top1-Acc@0.5 of 45.78. Performance declines at $\lambda_c = 0.2$ and 0.5, suggesting that excessive alignment with hyperedge representations weakens object-level discrimination. A moderate weight therefore balances node–hyperedge consistency and discriminability.

Table 5: Ablation study on the cognitive consistency weight λ_c.

Weight (λ_c)	0	0.05	0.1	0.2	0.5
AP@0.5	26.34	<u>28.64</u>	**29.26**	28.02	27.31
Top1-Acc@0.5	43.27	<u>44.97</u>	**45.78**	43.88	41.96

5 Conclusion

In this paper, we propose a Cognitive Hypergraph framework for 3D Intention Grounding, aiming to bridge abstract human intentions and physical objects in 3D scenes through structured cognitive reasoning. Unlike existing methods that mainly rely on implicit intent-object matching, our method models candidate objects and their higher-order functional relations within a cognitive hypergraph, further enhanced by a global intention prior and an intention-aware scene prior. This design enables intention-aware reasoning over object, intention, and scene semantics for more reliable grounding. Extensive experiments on 3D Intention Grounding and 3D Visual Grounding benchmarks demonstrate the effectiveness, interpretability, and generalization capability of our method. We hope this work provides a new perspective for advancing 3D grounding from local cross-modal matching toward cognition-driven structured reasoning in embodied intelligence.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and manuscript polishing, including sentence rephrasing, grammar checking, and readability improvement. They were not used to develop the research methodology, design experiments, implement the proposed method, process experimental data, or interpret the results. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. It involves no human-subject or animal experiments and introduces no new collection of personally identifiable information. All experiments are conducted on publicly available benchmark datasets in accordance with their applicable usage terms. We are not aware of additional privacy, security, or ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, training objectives, and model configurations in Sections 3 and 4, with additional implementation details, component definitions, and experimental settings provided in Appendices B and C. All datasets used in this work are publicly available. We will release the source code and training instructions to facilitate reproduction of the reported results.

REFERENCES

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.

Alessia Antelmi, Gennaro Cordasco, Mirko Polato, Vittorio Scarano, Carmine Spagnuolo, and Dingqi Yang. A survey on hypergraph representation learning. *ACM Computing Surveys*, 56(1):1–38, 2023.

Song Bai, Feihu Zhang, and Philip HS Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637, 2021.

Tim Bayne, David Brainard, Richard W Byrne, Lars Chittka, Nicky Clayton, Cecilia Heyes, Jennifer Mather, Bence Ölveczky, Michael Shadlen, Thomas Suddendorf, et al. What is cognition? *Current biology*, 29(13):R608–R615, 2019.

Alain Bretto. Hypergraph theory. *An introduction. Mathematical Engineering. Cham: Springer*, 1, 2013.

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

Salvatore Citraro, Simon De Deyne, Massimo Stella, and Giulio Rossetti. Towards hypergraph cognitive networks as feature-rich models of knowledge. *EPJ Data Science*, 12(1):31, 2023.

Jaime Corsetti, Francesco Giuliari, Alice Fasoli, Davide Boscaini, and Fabio Poiesi. Functionality understanding and segmentation in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24550–24559, 2025.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7395–7408, 2025.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3558–3565, 2019.

Yifan Feng, Yifan Zhang, Shihui Ying, Shaoyi Du, and Yue Gao. Kernelized hypergraph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):8938–8954, 2025.

Takaaki Fujita. Cognitive hypergraphs and superhypergraphs: A novel framework for complex relational modeling. *Neutrosophic Computing & Machine Learning*, 39, 2025.

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 Computer Vision and Pattern Recognition Conference*, pp. 3666–3675, 2025.

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. *CoRR*, 2023.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems*, 37:113991–114017, 2024.

Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1610–1618, 2021.

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*, pp. 15524–15533, 2022.

Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pp. 417–433. Springer, 2022.

Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, pp. 289–310. Springer, 2024.

Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pp. 4867–4876, 2020.

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*, pp. 3905–3915, 2025a.

Weitai Kang, Mengxue Qu, Jyoti Kini, Yunchao Wei, Mubarak Shah, and Yan Yan. Intent3d: 3d object detection in rgb-d scans based on human intention. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, and Junwei Liang. Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3707–3717, 2025a.

Yang Li, Aming Wu, Zihao Zhang, and Yahong Han. Novel class discovery for point cloud segmentation via joint learning of causal representation and reasoning. *Advances in Neural Information Processing Systems*, 2025b.

Yixuan Li, Zan Wang, and Wei Liang. R2g: Reasoning to ground in 3d scenes. *Pattern Recognition*, 168:111728, 2025c.

Xuewu Lin, Tianwei Lin, Lichao Huang, Hongyu Xie, and Zhizhong Su. Bip3d: Bridging 2d images and 3d perception for embodied intelligence. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9007–9016, 2025.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2949–2958, 2021.

Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16454–16463, 2022.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16488–16498, 2024.

Zhenhua Ning, Zhuotao Tian, Shaoshuai Shi, Guangming Lu, Daojing He, Wenjie Pei, and Li Jiang. Enhancing spatial reasoning in multimodal large language models through reasoning-based segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7851–7860, 2025.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

Zhipeng Qian, Yiwei Ma, Zhekai Lin, Jiayi Ji, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Multi-branch collaborative learning network for 3d visual grounding. In *European Conference on Computer Vision*, pp. 381–398. Springer, 2024.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*, 2022.

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

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*, pp. 196–213. Springer, 2024.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19757–19767, 2024a.

Yuan Wang, Yali Li, and Shengjin Wang. Gˆ 3-lq: Marrying hyperbolic alignment with explicit semantic-geometric modeling for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13917–13926, 2024b.

Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19231–19242, 2023.

Can Xu, Yuehui Han, Rui Xu, Le Hui, Jin Xie, and Jian Yang. Multi-attribute interactions matter for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17253–17262, 2024.

Chenyangguang Zhang, Alexandros Delitzas, Fangjinhua Wang, Ruida Zhang, Xiangyang Ji, Marc Pollefeys, and Francis Engelmann. Open-vocabulary functional 3d scene graphs for real-world indoor spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19401–19413, 2025.

Haomeng Zhang, Chiao-An Yang, and Raymond A Yeh. Multi-object 3d grounding with dynamic modules and language-informed spatial attention. *Advances in Neural Information Processing Systems*, 37:123237–123260, 2024.

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*, pp. 15225–15236, 2023.

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*, pp. 2911–2921, 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.

Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, et al. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8120–8132, 2025.

Appendix

A Related Work . . . 14
A.1 3D Visual Grounding and 3D Intention Grounding . . . 14
A.2 Relation Modeling and Functional Reasoning in 3D Scenes . . . 15
A.3 Hypergraph-based Higher-order Reasoning . . . 15
B Supplementary Method and Implementation Details . . . 16
B.1 Definition and Interpretation of Cognitive Hypergraphs . . . 16
B.2 Feature Interfaces and Global Cognitive Priors . . . 16
B.3 Details of Cognitive Hypergraph Construction . . . 17
B.4 Details of Cognitive Hypergraph Reasoning . . . 18
B.5 Details of Cognitive Consistency Regularization . . . 19
B.6 Grounding Supervision and Inference . . . 19
B.7 Configuration and Controlled Comparisons . . . 20
C Supplementary Experiments . . . 22
C.1 Datasets . . . 22
C.2 Implementation Details . . . 22
C.3 Validation on 3D Visual Grounding . . . 23
C.4 Additional Hyperparameter Analysis . . . 24
C.5 Controlled Design Comparisons . . . 24
D Supplementary Visualizations and Failure Analysis . . . 26
D.1 Additional Visualization Analysis . . . 26
D.2 Failure Cases and Model Limitations . . . 27

A Related Work

A.1 3D Visual Grounding and 3D Intention Grounding

3D Visual Grounding aims to localize target objects in 3D scenes according to natural-language descriptions. Early works, such as ScanRefer (Chen et al., 2020) and ReferIt3D (Achlioptas et al., 2020), establish the task by aligning textual descriptions with 3D object proposals, laying the foundation for language-guided 3D object localization. Subsequent methods improve cross-modal alignment from different perspectives. For example, TGNN (Huang et al., 2021) introduces graph-based relation modeling to capture contextual object dependencies, MVT (Huang et al., 2022) exploits multi-view feature fusion for richer visual representation, 3D-SPS (Luo et al., 2022) progressively selects referred points for single-stage grounding, EDA (Wu et al., 2023) decouples textual expressions for dense alignment, and 3D-ViSTA (Zhu et al., 2023) leverages pre-trained 3D vision-language alignment to enhance multimodal representation learning. More recent methods further advance grounded 3D understanding through explicit semantic-geometric modeling (Wang et al., 2024b), scene-aware transformers (Chang et al., 2024), large-scale 3D vision-language pretraining (Jia et al., 2024), unified promptable queries (Zhu et al., 2024), zero-shot open-vocabulary grounding (Li et al., 2025a), and efficient text-guided voxel pruning (Guo et al., 2025). These works have significantly promoted 3D language grounding, especially when the input language contains explicit object categories, attributes, or spatial descriptions. However, they still mainly focus on matching descriptive language to visible object properties, which limits their ability to handle abstract human intentions that imply latent functional needs.

Nevertheless, conventional 3D Visual Grounding is not fully suited for 3D Intention Grounding (3D-IG), where the input language is usually abstract, functional, and non-descriptive. Instead of directly describing the target appearance, 3D-IG requires the model to infer the object that can satisfy a latent human need. IntentNet (Kang et al., 2025b) introduces this task and explores verb-object alignment for intention-driven object localization. Nevertheless, existing 3D-IG methods still largely rely on implicit intent-object matching or prior functional annotations, making it difficult to explicitly reason about functional semantics, multi-object collaboration, and scene-level constraints. In contrast, our method formulates 3D-IG as cognitive hypergraph reasoning, where candidate objects are organized into intention-oriented functional groups and jointly reasoned with global intention and scene priors.

A.2 Relation Modeling and Functional Reasoning in 3D Scenes

Relation and context modeling are essential for robust 3D scene understanding, since target plausibility is often determined not only by individual object semantics, but also by object-object relations and global scene context. Existing 3D grounding methods have investigated different forms of relational and contextual reasoning. TGNN (Huang et al., 2021) models text-guided object relations with graph neural networks, while MA^{2}TransVG (Xu et al., 2024) explores multi-attribute interactions for 3D visual grounding. Other works further introduce explicit semantic-geometric modeling (Wang et al., 2024b), scene-aware transformers (Chang et al., 2024), and contextual object perception (Shi et al., 2024) to improve object disambiguation in cluttered 3D scenes. Multi3DRefer (Zhang et al., 2023) extends grounding from a single target to multiple objects, and D-LISA (Zhang et al., 2024) introduces dynamic modules and language-informed spatial attention for multi-object grounding. Recent reasoning-based grounding methods also suggest that explicit reasoning is beneficial for bridging language expressions and 3D scene structures (Li et al., 2025c). These methods show that object relations and contextual cues are important for disambiguating targets in complex 3D scenes.

Beyond spatial or semantic relations, recent works have begun to emphasize functional understanding in 3D environments. Functionality understanding and segmentation (Corsetti et al., 2025) studies how 3D objects support functional uses, while open-vocabulary functional 3D scene graphs (Zhang et al., 2025) explicitly model functional relations among objects in real-world indoor spaces. Large-scale 3D vision-language pretraining and LLM-based 3D scene understanding further promote grounded scene understanding and language-driven interaction (Jia et al., 2024; Huang et al., 2024; Kang et al., 2025a). Multimodal and embodied scene understanding methods (Daxberger et al., 2025; Ning et al., 2025; Majumdar et al., 2024; Wang et al., 2024a; Zhu et al., 2025) further indicate that 3D agents need to reason over spatial layouts, object affordances, and task contexts rather than merely align language with object categories. However, most existing methods still focus on pairwise relations, explicit visual descriptions, implicit LLM-based reasoning, or general spatial understanding. They do not explicitly model how multiple objects form intention-oriented functional groups under a specific human intent. Our method addresses this gap by constructing intention-conditioned cognitive hyperedges that organize candidate objects into higher-order functional groups, then modulating the resulting group representations with global intention and scene priors.

A.3 Hypergraph-based Higher-order Reasoning

Conventional graphs model pairwise relations between nodes, which may be insufficient for complex scenes where multiple objects jointly determine the functional meaning of a target. Hypergraphs provide a natural tool for modeling higher-order dependencies by connecting multiple nodes with a single hyperedge (Bretto, 2013). Hypergraph neural networks (Feng et al., 2019), hypergraph convolution and attention (Bai et al., 2021), and recent surveys on hypergraph representation learning (Antelmi et al., 2023) have demonstrated the effectiveness of hypergraphs in capturing higher-order structural relations. Beyond generic hypergraph representation learning, recent studies have also explored cognitive hypergraph structures for knowledge and relational modeling. Citraro et al. (2023) propose hypergraph cognitive networks as feature-rich models of knowledge, showing that hyperedges can better represent higher-order associations among concepts than ordinary pairwise cognitive networks. Fujita (2025) further discusses cognitive hypergraphs and superhypergraphs as a framework for complex relational modeling. These studies suggest the potential of hypergraph structures for representing cognitive relations beyond pairwise links. In 3D perception, higher-order reasoning has also been explored in tasks such as point cloud representation and novel class discovery (Li et al., 2025b),

demonstrating the potential of hypergraph-based structures to capture complex dependencies beyond pairwise relations.

Despite these advances, existing hypergraph-based methods mainly focus on structural or semantic correlations among nodes, or discuss cognitive hypergraphs from a general knowledge modeling perspective. They typically answer which nodes are related, but do not explicitly explain why these nodes should be grouped together under a particular intention, nor how such groups should be constrained by the current 3D scene context. This distinction is crucial for 3D-IG, where a target object is selected according to whether it satisfies an abstract functional need in a specific scene. Different from generic hypergraphs and general cognitive hypergraph formulations, our cognitive hypergraph is instantiated over physical object proposals in 3D scenes and interprets higher-order object relations under the guidance of global intention and scene priors. As a result, hyperedges are not merely structural connections, but intention-oriented functional groups that support cognitive reasoning over target, intention, object collaboration, and scene context.

B Supplementary Method and Implementation Details

B.1 Definition and Interpretation of Cognitive Hypergraphs

A hypergraph represents higher-order relations by connecting multiple nodes through a single hyperedge (Bretto, 2013; Bai et al., 2021). In 3D intention grounding, a query often specifies a desired use rather than the target's category or appearance (Kang et al., 2025b). Target selection can therefore depend on object affordances and on how other objects in the scene provide functional context (Corsetti et al., 2025; Zhang et al., 2025). We use *cognitive* to denote a goal-directed, coarse-to-specific organization of this evidence: the intention specifies the functional goal, object groups provide intermediate hypotheses, and group-level evidence informs the final choice of a target. Earlier cognitive-hypergraph studies use higher-order associations to represent concepts and knowledge (Citraro et al., 2023; Fujita, 2025). Our formulation instantiates this idea on physical object proposals for intention-conditioned grounding in 3D scenes.

Our task-specific contribution lies in the integration of intention-conditioned hyperedge construction, context-guided group representation, and selective group-to-object propagation for target grounding. For each candidate object, the intention representation g_q conditions the functional relevance between object pairs, which is combined with semantic similarity and relative box positions to compute relation scores; the top-ranked neighbors then form an intention-conditioned hyperedge. Thus, the hypergraph structure is dynamically recomputed for each scene–query pair. A hyperedge represents a potential functional-context hypothesis rather than an annotated functional group: its members may correspond to alternative valid targets or different objects that provide complementary evidence, and not every member is necessarily required for the intended activity. After construction, the member features are aggregated into a group representation, which is further modulated by g_q and the intention-aware scene prior g_s. The resulting group-level evidence is selectively propagated to incident candidate objects before target scoring. Notably, the scene prior guides this subsequent reasoning stage rather than directly entering the pairwise relation score, while scene information is also reflected during hypergraph construction through object features and relative box positions. Here, prior refers to an input-derived contextual feature rather than a separately specified probability distribution. Appendix C.5 examines these distinctions empirically. Table 10 evaluates query-conditioned construction and subsequent context-guided reasoning against an adapted generic hypergraph baseline. Table 11 compares group-level hyperedge propagation with pairwise graph propagation under the same Top-3 neighbor selection and $\lambda_c = 0$. These results support the grounding utility of the proposed design under the evaluated settings, while the learned hyperedges remain hypotheses about useful object groups rather than verified functional relations.

B.2 Feature Interfaces and Global Cognitive Priors

Object and language features. Each candidate object is represented by a box b_m, its center $p_m \in \mathbb{R}^3$, and an object-level feature $o_m \in \mathbb{R}^d$. We project the object features and RoBERTa token features into a common embedding space with $d = 256$ using separate learned linear layers. For detector-based settings, the object features are taken from the final detection feature layer before box prediction, while protocols with ground-truth candidates use the corresponding instance-level

features. The grounding branch operates on these representations without introducing additional category labels or ground-truth scene information.

For detector-based settings such as Intent3D (Kang et al., 2025b), we retain up to $M = 256$ proposals, ranked by objectness when more candidates are available, without applying query-dependent filtering before hypergraph construction. Padded entries are excluded from all subsequent pooling, ranking, and loss computations. In the following, M denotes the number of valid candidate objects in the current scene. The candidate count is distinct from both the number of raw input points and the number of encoded point-cloud tokens.

Masked intention pooling. The sentence-level pooling in Section 3.1 averages the valid lexical-token features. Let $\mu_t \in \{0, 1\}$ indicate whether token t is included in the pooling operation. Padding and sequence-boundary tokens are excluded. The pooled intention representation is computed as:

$$g_q = \frac{\sum_{t=1}^{T_{\max}} \mu_t e_t}{\sum_{t=1}^{T_{\max}} \mu_t}, \tag{16}$$

where $T_{\max} = 256$, and each valid query contains at least one lexical token. The mask is applied after contextual token encoding, ensuring that the pooled representation preserves information from the complete encoded query.

Intention-aware scene pooling. The scene prior follows the intention-conditioned pooling in Section 3.1. The scoring network f_s takes $[o_m \| g_q]$ and produces one scalar per candidate. Its softmax is normalized across valid candidates within the same scene–query pair. The resulting weights are used to aggregate object features into $g_s \in \mathbb{R}^d$. This summary is computed once from the input object features and reused during hyperedge modulation and node updating. Uniform average pooling is a separate ablation obtained by replacing all candidate weights with $1/M$.

B.3 Details of Cognitive Hypergraph Construction

Semantic and spatial affinity. To specify the affinity functions used in Section 3.2, we normalize the object features as $\bar{o}_m = o_m / \max(\|o_m\|_2, \epsilon)$ and define:

$$\mathrm{sim}(o_i, o_j) = \frac{1 + \bar{o}_i^\top \bar{o}_j}{2}, \qquad \phi(\Delta p_{ij}) = \exp\left(-\frac{\|p_i - p_j\|_2^2}{2\sigma_p^2}\right). \tag{17}$$

We use $\epsilon = 10^{-6}$ and a spatial bandwidth $\sigma_p = 1.0\,\mathrm{m}$. Box centers are expressed in metric scene coordinates; no scene-size rescaling is applied to this distance. The numerical cosine value is clipped to $[-1, 1]$ before rescaling. Only the features used for computing cosine affinity are normalized in this way; the reasoning branch retains the projected object features.

The semantic–spatial balance is $\alpha = 0.6$. The intention-specific scorer f_q operates on the ordered concatenation $[o_i \| o_j \| g_q]$. No symmetrization is applied: r_{ij}^q measures the suitability of object j for a group centered on object i, and need not equal r_{ji}^q. Since the intention gate lies in $(0, 1)$, it modulates relative relation strengths while retaining the base semantic–spatial affinity as an upper bound. Semantic similarity and proximity therefore serve as grouping biases, rather than explicit labels of functional collaboration.

Membership selection. Each valid node serves as a center once. For center v_k, we exclude k from neighbor ranking, select the top $\min(|e| - 1, M - 1)$ candidate indices according to r_{kj}^q, and include the center itself. The default hyperedge size is $|e| = 4$. Ties are resolved by candidate index, and scenes with fewer than four valid candidates use all available candidates.

We retain one hyperedge entry per center, including entries with identical memberships. Consequently, $K = M$ refers to an indexed family of hyperedges, rather than a deduplicated set. Overlapping groups are allowed. The binary incidence matrix is:

$$H_{mk}^q = \begin{cases} 1, & v_m \in e_k^q, \\ 0, & \text{otherwise}, \end{cases} \qquad H^q \in \{0, 1\}^{M \times K}. \tag{18}$$

Rows index nodes and columns index center-associated groups. Every node belongs to at least its own hyperedge. Membership is constructed once for each scene–query pair and is not iteratively

Table 6: Network interfaces. Widths specify the input, hidden, and output dimensions; brackets denote feature concatenation.

Network	Input	Widths
f_s	$[o_m \| g_q]$	$2d \to d \to 1$
f_q	$[o_i \| o_j \| g_q]$	$3d \to d \to 1$
ψ_h	$[h_k \| g_q \| g_s]$	$3d \to d \to d$
f_β	$[o_m \| \tilde{h}_k \| g_q]$	$3d \to d \to 1$
ψ_o	$[o_m \| u_m \| g_q \| g_s]$	$4d \to d \to d$
f_g	$[o_m \| u_m \| g_q \| g_s]$	$4d \to d \to d$
Grounding head	$[o'_m \| g_q]$	$2d \to d \to 1$

recomputed within the reasoning block. A scene with no valid proposal returns an empty prediction and does not enter hypergraph reasoning; its grounding and consistency losses are set to zero while the detection objective remains applicable.

Differentiable weighting within selected groups. Hard top-k selection determines discrete hyperedge memberships, so gradients do not propagate through the selected neighbor indices. To retain a differentiable path for the selected relations, we use their continuous relation scores to weight node-to-hyperedge aggregation. For each selected node–hyperedge incidence, including the center node, we define:

$$\omega^q_{mk} = \frac{H^q_{mk}(r^q_{km} + \epsilon)}{\sum_{j=1}^{M} H^q_{jk}(r^q_{kj} + \epsilon)}, \qquad h_k = \sum_{m=1}^{M} \omega^q_{mk} o_m. \tag{19}$$

The self-relation r^q_{kk} is computed using the same scoring function, although the center is excluded from neighbor ranking. The weights are nonnegative and sum to one within each hyperedge. Equal member scores recover mean aggregation, while unequal scores preserve the relative relevance estimated during construction.

During backpropagation, selected indices are treated as fixed, but the selected scores and member features are not detached. Consequently, the grounding and consistency objectives can update f_q through ω^q_{mk}. This provides a continuous training path for selected relations without claiming that the discrete selection itself is differentiable. Unselected relations receive no direct gradient through this aggregation path, although they share scorer parameters with selected relations. Training and inference use the same membership and weighting operations.

B.4 Details of Cognitive Hypergraph Reasoning

Network architecture. Table 6 summarizes the inputs and dimensions of the networks used in prior modeling, hypergraph reasoning, and grounding. Each network consists of two linear layers with a ReLU activation between them, a hidden width of $d = 256$, and biases in both layers. No additional dropout or normalization is applied inside these MLPs. The outputs of f_q and f_g are passed through a sigmoid; f_s and f_β produce unnormalized scalar scores, while ψ_h and ψ_o produce linearly projected features. The grounding head outputs an unnormalized logit for each candidate. Each network shares its parameters across the nodes or hyperedges to which it is applied, whereas different networks have separate parameters.

Prior modulation and message normalization. The relation-weighted group representation h_k in Eq. (19) is combined with g_q and g_s by ψ_h to obtain the modulated hyperedge feature $\tilde{h}_k$, as described in Section 3.3. For each candidate node v_m, f_β scores only its incident hyperedges, and the resulting scores are normalized by a softmax over those hyperedges. Nonincident hyperedges are excluded from the normalization and contribute no message to v_m. As shown by the network inputs in Table 6, g_s is not a direct input to f_β; it influences message relevance through the modulated feature $\tilde{h}_k$. This node-wise normalization allows each candidate to weight its associated group contexts according to the current intention.

Gated residual update. Let $z_m = [o_m \| u_m \| g_q \| g_s] \in \mathbb{R}^{4d}$ concatenate the original candidate feature, its aggregated hyperedge message, and the global intention and scene priors. The context-adaptive gate introduced in Section 3.3 is defined as:

$$\boldsymbol{\gamma}_m = \sigma(f_g(z_m)), \qquad \boldsymbol{\gamma}_m \in (0,1)^d. \tag{20}$$

The candidate is updated as $o'_m = o_m + \boldsymbol{\gamma}_m \odot \psi_o(z_m)$. The gate scales the projected update independently in each feature dimension, allowing each candidate to incorporate relevant group-level and global contextual evidence while retaining a direct residual path from its original representation. We apply a single reasoning block consisting of node-to-hyperedge aggregation, prior-guided hyperedge modulation, hyperedge-to-node propagation, and this gated update. The resulting candidate features are passed to the grounding head; the block does not iteratively reconstruct the hypergraph or introduce a separate box-refinement head.

Computational structure. Dense relation construction evaluates M^2 ordered pairs, including self-relations used for aggregation. After selection, the incidence structure contains at most $M|e|$ nonzero entries. Weighted member pooling and message summation each require $O(M|e|d)$ arithmetic operations when evaluated over valid incidences. These costs exclude pair-scoring MLPs, neighbor selection, feature projections, and the detection and language backbones. Thus, sparse propagation does not remove the quadratic cost of dense relation construction.

B.5 Details of Cognitive Consistency Regularization

The consistency objective encourages compatibility between updated candidate features and their associated group contexts. Using the incidence matrix, it is written as:

$$\mathcal{L}_{\text{cog}} = \frac{\sum_{m=1}^{M} \sum_{k=1}^{K} H^q_{mk} \left[1 - \cos_\epsilon(o'_m, \tilde{h}_k)\right]}{\sum_{m=1}^{M} \sum_{k=1}^{K} H^q_{mk}}, \tag{21}$$

where $\cos_\epsilon(x,y) = x^\top y / [\max(\|x\|_2, \epsilon) \max(\|y\|_2, \epsilon)]$ uses $\epsilon = 10^{-6}$. Both representations receive gradients; no stop-gradient operation is applied.

The loss is first averaged over valid incidence pairs for each scene–query pair and then averaged across the batch. For $M \geq |e|$, its within-sample denominator is $M|e|$. A node belonging to more hyperedges contributes more alignment terms, so this normalization is not equivalent to weighting all nodes equally.

This objective regularizes representation consistency rather than requiring identical object categories or proving functional collaboration. Member features also contribute to their group representations, so the aligned features share information by construction. The grounding objective remains responsible for discriminating targets from other candidates. Excessively strong alignment may reduce this distinction, especially for complementary objects with different semantic roles. We retain $\lambda_c = 0.1$ as the default weight.

B.6 Grounding Supervision and Inference

Multiple-target proposal assignment. For a scene–query pair, let $\mathcal{B}_Q^{\text{gt}}$ denote the annotated target-box set. Valid targets are determined by these annotations rather than by an assumption that every object with a similar affordance is labeled positive. For each candidate, compute:

$$\begin{aligned} \rho_m &= \max_{b \in \mathcal{B}_Q^{\text{gt}}} \text{IoU}_{3D}(b_m, b), \\ \mathcal{P} &= \{m : \rho_m \geq \tau_+\}, \qquad \mathcal{N}_- = \{m : \rho_m < \tau_-\}, \end{aligned} \tag{22}$$

with $\tau_+ = 0.5$ and $\tau_- = 0.25$. Candidates between the thresholds are ignored by the grounding loss. Assignment is recomputed from the current predicted boxes without differentiating through the resulting labels. Multiple proposals may be positive for the same annotated target, and different annotated targets can contribute positives simultaneously. If the target set is empty, $\mathcal{P}$ is empty and all valid proposals belong to $\mathcal{N}_-$.

Table 7: Default configuration of the proposed method.

Item	Configuration
Feature and hidden dimension	$d = 256$
Maximum candidate count	256
Maximum query length	256 tokens
Semantic–spatial balance	$\alpha = 0.6$
Spatial bandwidth	$\sigma_p = 1.0\,\mathrm{m}$
Hyperedge size	$\min(4, M)$, including the center
Hyperedge count	$K = M$, without deduplication
Member aggregation	Relation-weighted pooling
Reasoning blocks	1
Consistency weight	$\lambda_c = 0.1$
Numerical stabilization	$\epsilon = 10^{-6}$
Positive / negative IoU thresholds	0.5 / 0.25
Visualization probability / NMS thresholds	0.5 / 0.25

Multi-label grounding loss. The grounding head produces a logit s_m for each candidate, yielding a matching probability $\hat{y}_m = \sigma(s_m)$. We supervise labeled proposals independently with binary targets, allowing multiple candidates to receive high scores when a query has multiple annotated targets. To balance the contributions of positive and negative proposals, we define:

$$\mathcal{L}_{\text{ig}} = -\frac{\sum_{m\in\mathcal{P}} \log \hat{y}_m}{\max(1, |\mathcal{P}|)} - \frac{\sum_{m\in\mathcal{N}_-} \log(1-\hat{y}_m)}{\max(1, |\mathcal{N}_-|)}. \tag{23}$$

Empty sums contribute zero. The loss is evaluated with a numerically stable binary-cross-entropy-with-logits implementation and averaged over scene–query pairs. When no proposal reaches the positive threshold, no poorly localized candidate is forced to become a grounding positive; candidate localization remains supervised by the detection objective.

Joint optimization. The overall objective retains the decomposition in Section 3.4. The detection branch uses its native detection loss and assignment rule, including its original component weights. The grounding branch adds the loss in Eq. (23) and the consistency regularizer. Proposal features are not detached before reasoning, allowing the added objectives to update the shared visual features. The RoBERTa-base language encoder is kept frozen, while its learned projection layer is optimized through the intention-dependent computations. The final argmax is used only during inference and is not part of the training gradient path.

Inference and evaluation. Top-1 prediction selects the candidate with the largest s_m, equivalently the largest $\hat{y}_m$. For AP, all valid candidate boxes and their probabilities are supplied to the official Intent3D evaluator, retaining its matching, post-processing, and aggregation conventions at IoU thresholds 0.25 and 0.5. No additional multiplication by detector objectness or custom score threshold is introduced before this evaluator. For visualization only, we retain candidates with $\hat{y}_m \geq 0.5$ and apply class-agnostic 3D NMS with IoU threshold 0.25. These visualization settings do not replace the benchmark evaluation protocol.

For protocols with ground-truth object candidates, candidate identity provides the supervision label directly, and the detection loss is omitted. The distinction between ground-truth and detected candidates must be maintained when comparing datasets or baselines.

B.7 Configuration and Controlled Comparisons

Table 7 summarizes the default configuration of the proposed method. The raw point sampling, detection-backbone settings, and optimization schedule are specified separately in the experimental implementation details. Parameter selection and checkpoint selection are performed on the validation set, while the test set is reserved for final evaluation.

Component variants. Our full-sentence matching baseline represents the complete intention query with the sentence-level feature g_q and directly scores each candidate as $s_m = \mathrm{MLP}([o_m \| g_q])$. It

does not perform the verb–object alignment used by IntentNet (Kang et al., 2025b), nor does it update candidate features through cognitive hypergraph reasoning or scene-prior fusion. This baseline serves as the reference for our component ablations and is not a reproduction of IntentNet.

For the CHC-only variant, we retain cognitive hypergraph construction and relation-weighted member aggregation, while removing global-prior modulation and adaptive hyperedge-to-node reasoning. The incident hyperedge representations are uniformly aggregated as:

$$u_m^{\mathrm{uni}} = \frac{1}{|\mathcal{N}(m)|} \sum_{k \in \mathcal{N}(m)} h_k, \qquad o'_m = o_m + \psi_b([o_m \| u_m^{\mathrm{uni}}]), \tag{24}$$

where ψ_b has widths $2d \to d \to d$. This provides a basic propagation path through which the constructed hyperedges can affect the final candidate representations.

For the GCP-only variant, no hypergraph is constructed, and the global intention and scene priors are directly fused with each candidate:

$$o'_m = o_m + \psi_p([o_m \| g_q \| g_s]), \tag{25}$$

where ψ_p has widths $3d \to d \to d$.

For the CHC+GCP variant, we retain the cognitive hypergraph and relation-weighted member aggregation, but omit the adaptive reasoning operations in CHR. The incident hyperedge representations are first aggregated using u_m^{uni} in Eq. (24), and the global priors are then directly incorporated into the candidate update:

$$o'_m = o_m + \psi_{bp}([o_m \| u_m^{\mathrm{uni}} \| g_q \| g_s]), \tag{26}$$

where ψ_{bp} has widths $4d \to d \to d$. Thus, this variant retains both hypergraph structure and global context, but does not employ the adaptive hyperedge modulation, message weighting, or gated residual update of CHR.

For the CHC+CHR variant, the full hypergraph reasoning pipeline is retained while explicit global-prior injection into the reasoning block is removed. Specifically, the g_q and g_s inputs to hyperedge modulation, message propagation, and node updating are set to zero. The query representation remains available to the intention-conditioned relation construction and the final grounding head, thereby isolating global-prior injection from intention-conditioned hypergraph topology.

Here, GCP refers specifically to the explicit injection of global cognitive priors into candidate reasoning and updating. Therefore, when CHC is enabled, g_q is still used to construct intention-conditioned relations even when GCP is not included as an independent component.

For the non-full component variants in Table 3, cognitive consistency regularization is disabled by setting $\lambda_c = 0$, so that the effects of the architectural components can be examined without additional consistency constraints. The complete model uses the default setting $\lambda_c = 0.1$, which is selected on the validation set and further analyzed in Table 5. Therefore, the final row of Table 3 corresponds to the complete model with CHC, GCP, CHR, and cognitive consistency regularization.

Controlled comparison variants. Within each comparison, we use the same candidate inputs, backbone, training schedule, and evaluation protocol. For the hyperedge-size study, we vary $|e|$ from 2 to 4 while retaining the same relation scoring and reasoning pipeline, with $\lambda_c = 0.1$. When $|e| = 2$, each hyperedge contains a center node and one selected neighbor; thus, this comparison changes both the group size and the number of available neighbors.

To control for neighbor count, we further compare a query-conditioned pairwise graph with the cognitive hypergraph using the same Top-3 neighbors for each center node. The graph represents the three center–neighbor relations as separate pairwise edges, whereas the hypergraph represents the center and its three neighbors as a single four-node hyperedge. Both variants apply the same relation-weighted member aggregation, global-prior modulation, attention-based edge-to-node aggregation, and gated residual node update, using the same network architectures and hidden dimensions. Cognitive consistency regularization is disabled in both variants ($\lambda_c = 0$). This comparison therefore controls neighbor selection and the parameterized reasoning operations while varying whether the selected objects are represented through separate pairs or a shared higher-order group.

For the scene-prior comparison, we replace intention-aware weighted pooling with uniform average pooling while keeping the remaining architecture unchanged and setting $\lambda_c = 0.1$. The corresponding results are reported in Appendix C.5.

C SUPPLEMENTARY EXPERIMENTS

C.1 DATASETS

Our experiments consist of two parts: 3D Intention Grounding and 3D Visual Grounding. The former evaluates the effectiveness of our method on intention-driven target localization, while the latter further examines its generality on conventional language-guided 3D grounding.

3D Intention Grounding (3D-IG). For the 3D-IG task, we follow the official evaluation protocol of Intent3D (Kang et al., 2025b). Intent3D is built upon ScanNet (Dai et al., 2017) and contains 44,990 intention descriptions across 1,042 indoor 3D scenes. On average, each scene contains 61 object instances and 43 intention descriptions, and each scene-object pair is associated with 6 intention texts. The dataset covers 568 distinct verbs and 2,894 distinct nouns, providing diverse functional and intention-oriented language expressions. Since the language queries are usually abstract and non-descriptive, Intent3D serves as a suitable benchmark for evaluating cross-modal reasoning between human intention understanding and 3D object grounding. For fair comparison, we strictly follow the official data split, including 35,850 training samples, 2,285 validation samples, and 6,855 test samples, with no scene overlap across different splits.

3D Visual Grounding. To further evaluate the generality of our framework, we conduct experiments on three widely used 3D visual grounding benchmarks, including Nr3D, Sr3D, and ScanRefer, all built upon ScanNet. Nr3D (Achlioptas et al., 2020) contains 41,503 human-annotated referring expressions across 707 indoor scenes, where each expression describes a unique target object among distractors. Sr3D (Achlioptas et al., 2020) contains 83,572 template-generated expressions, with an emphasis on spatial relationships between objects. ScanRefer (Chen et al., 2020) contains 51,583 natural-language descriptions referring to 11,046 objects across 800 indoor scenes, and evaluates object localization under both Unique and Multiple settings. Following the standard protocols of each benchmark, we report grounding accuracy under their corresponding evaluation settings.

C.2 IMPLEMENTATION DETAILS

To ensure a fair comparison with existing 3D Intention Grounding methods, we follow the implementation setting of IntentNet (Kang et al., 2025b). Specifically, PointNet++ (Qi et al., 2017) and a frozen RoBERTa-base encoder (Liu et al., 2019) encode the 3D scene and intention query, respectively. For each scene, we randomly sample 50,000 XYZ-RGB points using a fixed random seed of 1184, following the preprocessing protocol of IntentNet (Kang et al., 2025b). PointNet++ (Qi et al., 2017) produces 1,024 point-level tokens from these input points, and the maximum query length is 256 tokens. The PointNet++ backbone (Qi et al., 2017) is initialized from a pretrained Group-Free detector checkpoint (Liu et al., 2021). Following IntentNet, candidate objects for Intent3D are obtained from Group-Free detector proposals. For the 3D Visual Grounding benchmarks, we follow the official PQ3D protocol (Zhu et al., 2024): Nr3D and Sr3D use ground-truth object instances, while ScanRefer uses ground-truth instances during training and predicted object instances during evaluation. The corresponding object-level features are used as object representations for cognitive hypergraph reasoning. We train the model with AdamW using a weight decay of 5×10^{-4}. The initial learning rates are 10^{-3} for the PointNet++ backbone and 10^{-4} for the remaining trainable modules. We use a step learning-rate schedule with a configured milestone of 80 epochs, a decay factor of 0.1, and no warm-up. The full model is trained for 120 epochs on four NVIDIA RTX 3090 GPUs with a total batch size of 16, requiring approximately 34 hours. Unless otherwise specified, all compared variants use the same backbone, input configuration, and training settings.

For cognitive hypergraph construction, each candidate object is instantiated as a graph node. We compute intention-conditioned relation scores by jointly considering functional relevance, semantic similarity, and spatial affinity. The semantic–spatial balance coefficient is set to $\alpha = 0.6$ according to validation performance. Each node then serves as a center node and selects its top-ranked neighbors to form a hyperedge. We use $|e| = 4$ by default, which achieves the best validation performance while providing a compact higher-order functional group.

For global cognitive priors, the global intention prior g_q is obtained by average pooling over valid token-level RoBERTa features, while the intention-aware scene prior g_s is computed by intention-conditioned weighted pooling over candidate object features, as detailed in Appendix B.2. During cognitive hypergraph reasoning, member-node features are first aggregated by relation-weighted

Table 8: Performance comparison on 3D Visual Grounding for the ScanRefer dataset.

Method	Venue	Unique		Multiple		Overall	
		0.25	0.5	0.25	0.5	0.25	0.5
ScanRefer (Chen et al., 2020)	ECCV20	67.6	46.2	32.1	21.3	39.0	26.1
EDA (Wu et al., 2023)	CVPR23	85.8	68.6	49.1	37.6	54.6	42.3
VPP-Net (Shi et al., 2024)	CVPR24	86.1	67.1	50.3	39.0	55.7	43.3
M3DRef-CLIP (Zhang et al., 2023)	ICCV23	-	77.2	-	36.8	-	44.7
G3-LQ (Wang et al., 2024b)	CVPR24	88.6	73.3	50.2	39.7	56.0	44.7
MCLN (Qian et al., 2024)	ECCV24	86.9	72.7	52.0	40.8	57.2	45.5
MA^{2}TransVG (Xu et al., 2024)	CVPR24	86.3	74.1	53.8	41.4	57.9	45.7
ConcreteNet (Unal et al., 2024)	ECCV24	86.4	82.1	42.4	38.4	50.6	46.5
D-LISA (Zhang et al., 2024)	NeurIPS24	-	75.5	-	40.0	-	46.9
GPS (Jia et al., 2024)	ECCV24	-	77.9	-	42.7	-	48.1
Chat-Scene (Huang et al., 2024)	NeurIPS24	89.6	**82.5**	47.8	42.9	55.5	50.2
PQ3D (Zhu et al., 2024)	ECCV24	86.7	78.3	51.5	46.2	57.0	51.2
Robin3D (Kang et al., 2025a)	ICCV25	-	-	-	-	**60.8**	**55.1**
Ours	-	**90.8**	81.3	**54.7**	**48.6**	60.3	53.7

pooling to obtain the initial hyperedge representation. The resulting hyperedge is then modulated by g_q and g_s. Group-level evidence is selectively propagated back to incident candidate nodes, where g_q further conditions message relevance and the scene prior influences propagation through the modulated hyperedge representations. Finally, a context-adaptive gated residual update integrates the aggregated group evidence with the original candidate representation and global priors.

The overall training objective follows Section 3.4:

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{ig} + \lambda_c \mathcal{L}_{cog}, \tag{27}$$

where $\mathcal{L}_{det}$ supervises candidate object generation, $\mathcal{L}_{ig}$ supervises intention grounding, and $\mathcal{L}_{cog}$ regularizes the consistency between updated candidate representations and their associated context-aware hyperedge representations. We set $\lambda_c = 0.1$ according to validation performance. For protocols using ground-truth object candidates, the detection loss is omitted, as described in Appendix B.6. During inference, the model constructs the intention-conditioned cognitive hypergraph from the candidate objects, performs prior-guided hypergraph reasoning, and ranks candidates according to their grounding scores.

C.3 Validation on 3D Visual Grounding

Table 8 reports the 3D visual grounding results on ScanRefer (Chen et al., 2020) under the Unique, Multiple, and Overall settings. The Multiple subset accounts for the majority of samples and is more challenging than the Unique subset, since multiple objects with similar categories or appearances may coexist in cluttered indoor scenes. This setting therefore requires stronger object disambiguation, contextual reasoning, and scene-level understanding. Our method achieves strong and competitive performance across all settings. On the Unique subset, our method achieves the best result at IoU@0.25 with 90.8, while obtaining a competitive 81.3 at IoU@0.5. On the more challenging Multiple subset, our method achieves the best results of 54.7 and 48.6 at IoU@0.25 and IoU@0.5, respectively, among methods with reported split-wise results. Compared with PQ3D (Zhu et al., 2024), our method improves Multiple@0.25 from 51.5 to 54.7 and Multiple@0.5 from 46.2 to 48.6, corresponding to gains of 3.2 and 2.4 points. These results demonstrate the effectiveness of cognitive hypergraph reasoning in resolving ambiguity among multiple competing candidates through contextual object relations. In terms of overall performance, our method achieves 60.3 at IoU@0.25 and 53.7 at IoU@0.5. Compared with PQ3D (Zhu et al., 2024), our method improves Overall@0.25 by 3.3 points and Overall@0.5 by 2.5 points. Compared with MA^{2}TransVG (Xu et al., 2024), the improvement is particularly pronounced at the stricter IoU threshold, increasing Overall@0.5 from 45.7 to 53.7. These gains show that our framework is effective not only for target retrieval but also for more precise localization. The improvement mainly benefits from the proposed target–intention–scene reasoning mechanism, where candidate objects are organized as nodes, higher-order hyperedges capture object relations, and global intention and scene priors jointly guide the reasoning process. This allows candidate representations to be refined under both referring

semantics and scene context, which is particularly beneficial in complex scenes with visually similar distractors. Although Robin3D (Kang et al., 2025a) reports slightly higher Overall scores of 60.8 and 55.1, it is an LLM-based framework relying on large-scale instruction-following models, which may introduce additional inference overhead. In contrast, our method provides a task-specific structured reasoning framework based on cognitive hypergraphs, explicitly modeling target–intention–scene relations without relying on complex large-language-model decoding.

Overall, the ScanRefer results demonstrate that our framework generalizes beyond intent-centric grounding and remains effective for conventional 3D visual grounding, particularly in challenging scenes involving multiple competing objects.

C.4 Additional Hyperparameter Analysis

Analysis of the Hyperedge Size $|e|$. To evaluate the influence of higher-order functional group scale, we conduct an ablation study on the validation set by varying the hyperedge size $|e|$, which denotes the number of candidate objects contained in each hyperedge. As shown in Table 9, both overly small and overly large hyperedge sizes lead to inferior performance. When $|e| = 2$, the hypergraph is close to pairwise relation modeling, which limits its ability to capture higher-order object collaboration under user intentions. As $|e|$ increases, more candidate objects are grouped into each functional hyperedge, allowing the model to capture richer contextual and collaborative cues. The best performance is achieved when $|e| = 4$, reaching the highest AP@0.5 and Top1-Acc@0.5. However, further increasing $|e|$ to 5 or 6 introduces more irrelevant objects into the functional groups, which weakens the intention-oriented semantics of hyperedges and harms grounding accuracy. These results show that a moderate hyperedge size is essential for effective cognitive hypergraph reasoning.

Table 9: Ablation study on the hyperedge size $|e|$.

Hyperedge Size ($\|e\|$)	2	3	4	5	6
AP@0.5	25.34	<u>27.21</u>	**29.26**	26.87	25.46
Top1-Acc@0.5	41.52	<u>43.26</u>	**45.78**	42.41	40.32

C.5 Controlled Design Comparisons

To examine the key design choices of our framework, we conduct controlled comparisons on the Intent3D validation set. The backbone, candidate inputs, and training schedule are matched within each comparison, while the specified design factor is varied. We examine four aspects: generic versus cognitive hypergraph modeling, the effect of hyperedge size, hyperedge versus pairwise graph propagation under the same Top-3 neighbor selection, and intention-aware versus uniform scene pooling. The consistency weight used for each comparison is specified in the corresponding table.

Generic vs. Cognitive Hypergraph Modeling. We investigate the contribution of the proposed intention-conditioned cognitive design using an adapted KHGNN (Feng et al., 2025) as the starting point. We adapt KHGNN to the same backbone, candidate object representations, and hyperedge size used in our framework. In this baseline, candidate objects are modeled through generic higher-order hypergraph relations, while the complete intention sentence is represented as a global textual input.

Starting from this adapted KHGNN baseline, we progressively modify the hypergraph construction and reasoning components to obtain our cognitive hypergraph framework. We first introduce query-conditioned hypergraph construction (CHC), then incorporate global cognitive priors (GCP), and finally introduce adaptive cognitive hypergraph reasoning (CHR). The resulting variants therefore represent successive modifications of the same starting framework. Cognitive consistency regularization is disabled throughout this comparison ($\lambda_c = 0$), and all variants use the same backbone, candidate inputs, hyperedge size, and training protocol.

As shown in Table 10, the adapted KHGNN baseline achieves 36.79% Top1-Acc@0.5 and 21.38% AP@0.5. Introducing query-conditioned construction improves these metrics to 39.64% and 23.37%, respectively, supporting the benefit of incorporating intention information into hypergraph construction. Adding global cognitive priors further improves performance to 42.18% and 25.19%, indicating the value of intention and scene context beyond the constructed relational structure. With adaptive CHR, performance reaches 43.27% Top1-Acc@0.5 and 26.34% AP@0.5.

Together, these modifications improve Top1-Acc@0.5 and AP@0.5 over the adapted KHGNN baseline by 6.48 and 4.96 percentage points, respectively, without cognitive consistency regulariza-

Table 10: Progressive modifications of an adapted KHGNN baseline on the Intent3D validation set. Each subsequent row incorporates the preceding modifications. All variants use the same backbone, candidate inputs, hyperedge size, and training protocol, with $\lambda_c = 0$.

Variant	Query-Cond.	Global Prior	Adaptive CHR	Top1-Acc@0.5	AP@0.5
Adapted KHGNN (Feng et al., 2025)				36.79	21.38
+ Query-conditioned Construction	✓			39.64	23.37
+ Global Cognitive Priors	✓	✓		42.18	25.19
+ Adaptive CHR	✓	✓	✓	**43.27**	**26.34**

tion. The results support the contribution of intention-conditioned construction and context-guided reasoning when extending a generic hypergraph framework to 3D intention grounding.

Table 11: Controlled comparisons on the Intent3D validation set. Settings are matched within each block. The graph and hypergraph structure comparison uses the same Top-3 neighbors and $\lambda_c = 0$.

Variant	Top1-Acc@0.5	AP@0.5
Hyperedge size ($\lambda_c = 0.1$)		
Cognitive Hypergraph ($\|e\| = 2$)	41.52	25.34
Cognitive Hypergraph ($\|e\| = 4$)	**45.78**	**29.26**
Matched-neighbor structure ($\lambda_c = 0$)		
Query-conditioned Graph (Top-3)	38.71	23.42
Cognitive Hypergraph ($\|e\| = 4$)	**43.27**	**26.34**
Scene-prior aggregation ($\lambda_c = 0.1$)		
Uniform Average Pooling	44.75	26.92
Intention-aware Weighted Pooling	**45.78**	**29.26**

Effect of hyperedge size. We first examine how the number of objects in each hyperedge affects intention grounding. When $|e| = 2$, a hyperedge contains its center and one selected neighbor; increasing $|e|$ to 4 allows the center to aggregate information from three selected neighbors while retaining the same relation scoring and reasoning pipeline. As shown in Table 11, this change improves Top1-Acc@0.5 from 41.52 to 45.78 and AP@0.5 from 25.34 to 29.26, gains of 4.26 and 3.92 points, respectively. The result supports the use of a larger intention-conditioned group, through which a candidate can incorporate evidence from several relevant objects. However, increasing $|e|$ also increases the number of selected neighbors, so this comparison alone does not isolate the effect of group-level propagation.

Query-conditioned graph vs. cognitive hypergraph. To control for neighbor count, we further compare a pairwise graph with our cognitive hypergraph using the same query-conditioned Top-3 neighbors for each center. The graph models the selected relations through pairwise edges, whereas the hypergraph forms a group representation from the center and its neighbors before propagating group-level information to candidate objects. We set $\lambda_c = 0$ for both variants to exclude the effect of cognitive consistency regularization. The graph variant's construction and reasoning operations are detailed in Appendix B.7. As shown in Table 11, the hypergraph achieves 43.27 Top1-Acc@0.5 and 26.34 AP@0.5, compared with 38.71 and 23.42 for the graph, yielding gains of 4.56 and 2.92 points. Under matched neighbor selection, this result supports the benefit of group-level hyperedge aggregation and propagation over the evaluated pairwise graph variant. It also indicates that the improvement from enlarging hyperedges cannot be explained solely by access to more neighbors.

Uniform vs. intention-aware scene pooling. We next assess the contribution of intention-conditioned scene aggregation. We replace the weighted pooling in Eq. (4) with uniform average pooling, $g_s^{\text{uni}} = \frac{1}{M}\sum_{m=1}^{M} o_m$, while keeping the remaining architecture and training settings unchanged. Uniform pooling treats all candidates equally, whereas intention-aware pooling gives greater weight to candidates relevant to the current query. As shown in Table 11, intention-aware pooling improves Top1-Acc@0.5 from 44.75 to 45.78 and AP@0.5 from 26.92 to 29.26, gains of 1.03 and 2.34 points, respectively. The larger gain in AP@0.5 is consistent with improved ranking across

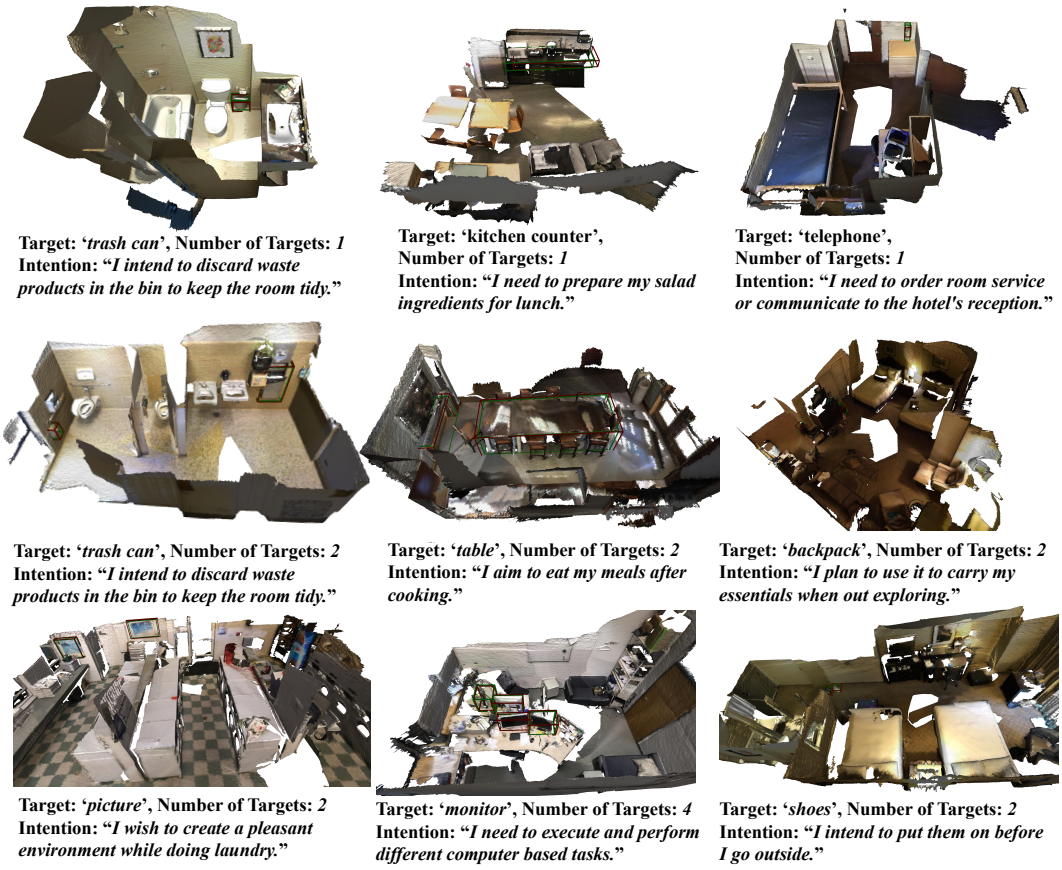


Figure 6: **Additional visualization results of our method on the Intent3D dataset.** Green boxes indicate ground-truth annotations, while red boxes denote our detection results.

candidates, beyond the highest-scoring prediction. These results support using a query-dependent scene prior to guide hyperedge modulation and node updating.

Overall, the three comparisons in Table 11 address different design choices. The hyperedge-size study identifies the benefit of incorporating multiple relevant neighbors; the matched-neighbor comparison shows an advantage for the evaluated group-level propagation over pairwise graph propagation; and the pooling comparison supports an intention-aware scene prior. Together, they provide evidence for combining local object groups with query-dependent scene context in 3D intention grounding.

D Supplementary Visualizations and Failure Analysis

D.1 Additional Visualization Analysis

To further evaluate the effectiveness of our method, we conduct additional visualization experiments on Intent3D, as shown in Figure 6, where green boxes denote ground-truth objects and red boxes denote the predictions of our model. The visualization results cover diverse indoor scenes, including bathrooms, kitchens, living rooms, bedrooms, and hotel-like rooms, involving different functional intentions such as discarding waste, preparing food, communicating, carrying items, and performing computer-based tasks. As shown in Figure 6, our method can accurately ground target objects under both single-target and multi-target settings. For example, when the intention is to discard waste products, the model correctly identifies trash cans in different bathroom scenes, demonstrating its ability to infer object functionality from abstract intention expressions. In kitchen and dining-related cases, such as locating a kitchen counter for preparing salad ingredients or identifying tables for eating meals after cooking, our method effectively leverages scene context and object functionality

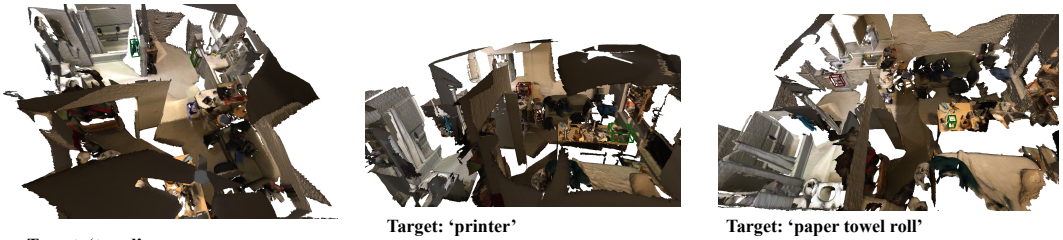


Figure 7: **Failure cases on Intent3D.** Our method may fail under three challenging conditions: missing proposals, geometric ambiguity, and functional confusion. Specifically, the target towel is missed by the detector, the printer is confused with geometrically similar desktop clutter, and the paper towel roll is mistaken for a towel due to similar wiping affordance.

to identify the intended targets. Moreover, in multi-target scenarios involving trash cans, tables, backpacks, monitors, and shoes, the model assigns high grounding scores to multiple functionally relevant candidates corresponding to annotated targets, rather than focusing only on a single visually salient object. This behavior is consistent with the multi-target evaluation setting, where all annotated target objects are treated as valid targets.

These visualization results further demonstrate the effectiveness of the proposed cognitive hypergraph reasoning framework in modeling intention-oriented functional relations and scene-level constraints. By organizing candidate objects into functional groups and incorporating global intention and scene priors, our method achieves robust grounding across diverse scenes, different object scales, and varying numbers of target instances. These qualitative results provide intuitive evidence of the effectiveness of our framework for complex 3D Intention Grounding scenarios.

D.2 Failure Cases and Model Limitations

Figure 7 shows three representative failure cases. In the first case, the target towel is missed by the detector and therefore cannot be included as a node in the cognitive hypergraph. Although the intention, "I keep it after taking a bath," provides a relevant cue, subsequent reasoning cannot recover a target that is absent from the candidate proposals. This illustrates the method's dependence on proposal recall. The second case concerns geometric ambiguity. For the intention 'I will print documentation received electronically,' the model selects a box-like object in the desktop clutter instead of the printer. The similar geometry and surrounding office context make these candidates difficult to distinguish, indicating that fine-grained object recognition remains important even when the intention provides a functional cue. The third case involves functional ambiguity. For the intention 'I will wipe my wet kitchen counter after washing my veggies,' the model selects a towel rather than the annotated paper towel roll. Both objects can plausibly satisfy the wiping intention, so distinguishing the annotated target requires finer-grained visual or contextual evidence.

Overall, the method remains limited by proposal recall, fine-grained object discrimination, and ambiguity between objects with similar functions. Future work could improve these aspects through stronger candidate generation and more detailed geometric and affordance representations.