QUERY-CONDITIONED RECONSTRUCTION REFINEMENT FOR 3D AFFORDANCE GROUNDING UNDER OCCLUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

3D affordance grounding aims to localize affordance regions specified by natural-language instructions. However, partial observations and severe occlusion often result in missing functional surfaces, while reconstructed geometry may contain unreliable local structures that adversely affect affordance grounding. To address this issue, we propose a query-conditioned reconstruction refinement framework for 3D affordance grounding under occlusion. Given partial multi-view RGB-D observations, we first perform amodal geometry completion to recover the complete object structure. We then introduce query-conditioned localization reliability estimation to assess the reliability of reconstructed local regions with respect to the target affordance, and use the estimated reliability to guide adaptive local detail refinement. Furthermore, we develop query-guided spatial-functional reasoning to jointly exploit geometric spatial relations and query-conditioned functional relations among local regions. By selectively refining reconstructed geometry according to its reliability for the target affordance, our framework enables robust affordance grounding under occlusion. Experimental results on the public dataset demonstrate the effectiveness of the proposed framework, achieving 24.0 IoU on the 3D affordance grounding task under occlusion.

1 INTRODUCTION

Understanding where an object can be functionally interacted with is important for robotic manipulation. Language-guided 3D functional region grounding aims to identify functional regions from natural-language descriptions, providing a direct connection between human instructions and actionable locations in 3D environments Thomason et al. (2022); Roh et al. (2022); Nguyen et al. (2023); Li et al. (2024); Toan et al. (2024). Such grounding is particularly important in realistic environments, where the desired interaction region is often determined not only by object geometry but also by the functional intent expressed in language. However, practical robotic systems typically operate with incomplete observations, making reliable localization challenging when relevant surfaces are occluded or absent from the captured views.

As illustrated in Fig. 1, given the same depth views and functional query, existing approaches directly perform affordance grounding on reconstructed geometry, whereas our framework further refines the reconstructed regions according to their usefulness for the target query before localization.

Existing methods can associate language descriptions with 3D geometry for functional-region localization Jain et al. (2022); Luo et al. (2022); Chen et al. (2022); Zhang et al. (2023); Nguyen et al. (2023); Li et al. (2024); Toan et al. (2024); Tabib et al. (2024), while 3D reconstruction and point cloud completion methods recover geometric structures beyond the observed regions Yan et al. (2022); Zhou et al. (2022); Li et al. (2023). Affostruction Park et al. (2026) further combines complete geometry with language-guided functional localization, enabling grounding beyond the directly observed surfaces. However, for localization, *complete does not necessarily mean directly usable*. Reconstruction errors can vary substantially across regions, while functional localization focuses on regions relevant to the current query. Consequently, the same reconstruction error can have very different effects depending on its location and the target function: an error

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.

in an irrelevant region may have little influence, whereas a small geometric deviation around a functional region can alter its boundary, local structure, or final localization. Thus, reconstruction quality alone does not indicate whether a reconstructed region is useful for the target localization task. Moreover, a reconstruction method optimized for overall geometric fidelity is not necessarily optimized for preserving the local structures that are critical to a specific affordance. This distinction becomes more important under occlusion, where reconstructed regions contain different amounts of observed and inferred geometric evidence. This raises a key question: *how should reconstructed geometry be selectively refined according to the current affordance query?*

To address this issue, we introduce *query-conditioned reconstruction refinement*, which evaluates the localization reliability of reconstructed local regions with respect to the current language query and uses it to guide adaptive local detail refinement. Rather than treating all reconstructed geometry as equally useful, the proposed approach selectively enhances regions that provide more reliable evidence for the target affordance while suppressing the influence of potentially inaccurate reconstructed structures. In this way, reconstruction is not treated as the final geometric representation for grounding, but as an intermediate representation that is further adapted to the downstream localization task. This design allows the model to distinguish between geometric information that is merely available after completion and information that is actually useful for identifying the target functional region. By conditioning the refinement process on the affordance query, the same reconstructed object can consequently be processed differently for different functional requirements.

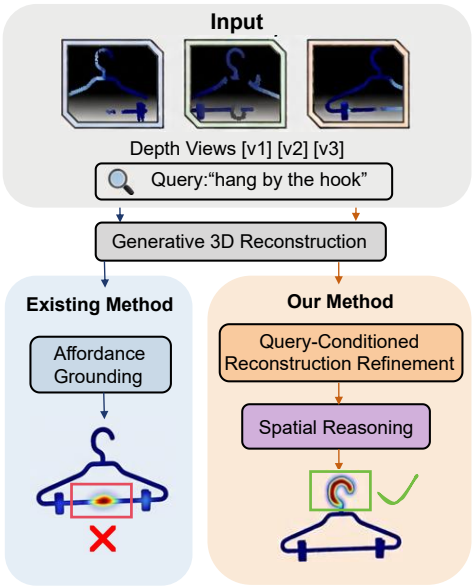

Figure 1: Comparison of 3D affordance grounding with and without query-conditioned reconstruction refinement.

Functional localization further requires understanding how local regions relate to their surrounding context. A region that is ambiguous from its local geometry alone can become identifiable through its spatial configuration and functional relationships with neighboring regions. Recent 3D grounding studies have explored spatial relations, contextual information, and language-conditioned reasoning to improve region-level correspondence Chen et al. (2022); Chang et al. (2024); Xu et al. (2024); Wang et al. (2024b); Shi et al. (2024); Yuan et al. (2024); Feng et al. (2024); Li et al. (2025b); Liu et al. (2025). Motivated by these observations, we develop *query-guided spatial-functional reasoning* to jointly model spatial context and query-dependent functional relationships among local regions. This complements the proposed reconstruction refinement by further exploiting the contextual information among refined regions for final affordance localization. In particular, spatial context can provide complementary evidence when the geometry of an individual region is incomplete or ambiguous, while functional relations help determine whether neighboring regions jointly support the queried interaction.

Our focus is on improving the utilization of the multi-view RGB-D observations already available to the system. By combining query-conditioned reconstruction refinement with query-guided spatial-functional reasoning, our framework selectively processes reconstructed geometry according to its usefulness for the target functional region and performs context-aware localization without iterative viewpoint acquisition. This formulation therefore separates geometric completion from affordance-aware processing, allowing the reconstructed representation to be further adapted to the specific requirements of functional localization.

Our contributions are summarized as follows:

- We introduce a *query-conditioned reconstruction refinement framework* for language-guided 3D affordance grounding under occlusion, which adapts reconstructed geometry to the requirements of the target affordance rather than treating all reconstructed regions uniformly.
- We propose query-conditioned localization reliability estimation and reliability-guided adaptive local refinement, which assess reconstructed local regions with respect to the target query and selectively enhance their local geometric details.
- We develop query-guided spatial-functional reasoning to jointly model spatial context and query-dependent functional relationships among local regions, enabling more effective affordance localization on the refined reconstructed geometry.

2 Related Work

2.1 3D Geometry Completion and Reconstruction

3D point cloud completion aims to recover complete object geometry from partial observations. Recent methods improve structural recovery under severe incompleteness through viewpoint-aware, self-supervised, and local-structure modeling. VAPCNet Fu et al. (2023) incorporates scanning viewpoints, while P2C Cui et al. (2023) explores self-supervised completion from partial observations. SVDFormer Zhu et al. (2023) exploits self-view augmentation and self-structure information, and HyperCD Lin et al. (2023) improves completion optimization with a hyperbolic Chamfer distance. AnchorFormer Chen et al. (2023) further uses regional anchors to capture diverse local structures.

Recent approaches incorporate multimodal and generative priors. CSDN Zhu et al. (2024) integrates image information for coarse-to-fine completion, PointAttN Wang et al. (2024a) models structural dependencies with self- and cross-attention, and PointSea Zhu et al. (2025b) performs global-to-local completion. PCDreamer Wei et al. (2025) introduces multi-view diffusion priors and shape consolidation, while GenPC Li et al. (2025a) leverages pretrained 3D generative priors for zero-shot completion. RealPC Pathak et al. (2025) further highlights the challenges of applying completion methods to realistic incomplete scans.

These methods primarily focus on recovering geometrically plausible complete shapes, where reliability is associated with reconstruction quality. For functional-region localization, however, geometric reliability may not directly reflect localization reliability. Our work therefore evaluates reconstructed local regions with respect to the downstream language-guided localization task and uses the resulting reliability to guide local feature refinement.

2.2 Language-guided 3D Functional Localization

Recent studies investigate 3D affordance and functional-region localization by combining geometry with language. Grounding 3D Object Affordance from 2D Interactions Yang et al. (2023) transfers interaction information from images to 3D representations, while LASO Li et al. (2024) formulates language-guided affordance segmentation as dense 3D prediction. Language-Conditioned Affordance-Pose Detection Toan et al. (2024) explores open-vocabulary affordance detection with language-conditioned manipulation poses, and AffordanceLLM Qian et al. (2024) investigates vision-language models for affordance grounding and unseen object-action combinations.

More recent methods consider multimodal observations and complex functional understanding. LMAffordance3D Zhu et al. (2025a) performs affordance grounding from language, visual observations, and interaction information under partial observations, while SeqAfford Yu et al. (2025) extends affordance segmentation to sequential instructions. Affostruction Park et al. (2026) combines generative reconstruction with language-guided affordance grounding to infer functional regions on unobserved object parts, and further employs affordance-driven active view selection.

These studies demonstrate the value of combining geometry and language for functional localization. In contrast, our framework focuses on the reliability of reconstructed geometry for the current localization task. We estimate query-conditioned localization reliability for reconstructed local re-

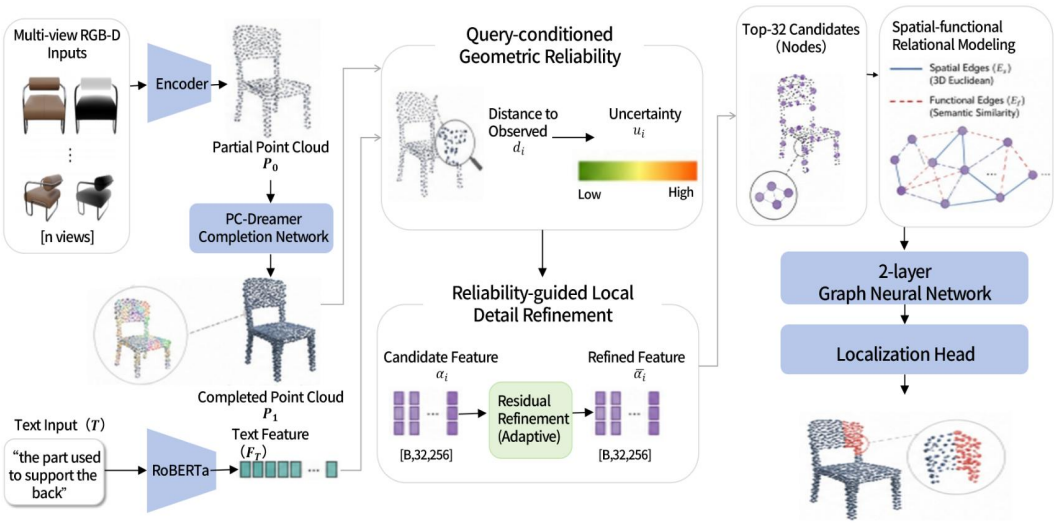

Figure 2: Overview of the proposed framework. Given partial multi-view RGB-D observations and a language query, the framework first reconstructs the amodal object geometry, then evaluates the localization reliability of local reconstructed regions and uses it to guide local feature refinement. Query-guided spatial-functional reasoning is subsequently employed to localize the target functional region.

gions, use it to regulate local detail refinement, and further employ query-guided spatial-functional reasoning to selectively exploit spatial and functional relationships for the target affordance.

3 Methodology

3.1 Problem Formulation and Framework Overview

We study language-guided 3D affordance grounding from partial multi-view RGB-D observations. Given a set of RGB-D observations $\mathcal{V} = \{(I_k, D_k, K_k, T_k)\}_{k=1}^{N}$ and a natural-language affordance query q, the goal is to localize the functional region of the object described by the query.

As illustrated in Fig. 2, the framework takes multi-view RGB-D observations and a text query as input. The partial point cloud is first completed to obtain the amodal geometry, followed by query-conditioned reliability estimation and reliability-guided local feature refinement. The refined local features are then used to construct spatial-functional relations and predict the target affordance region.

3.2 Amodal Geometry Completion

The RGB-D observations obtained from limited viewpoints only cover part of the object surface. Direct affordance grounding on partial geometry may fail when the functional region is occluded or unobserved. We therefore employ PCDreamer as the amodal point-cloud completion backbone to recover the complete object structure.

The RGB-D observations are first transformed into a common coordinate system and fused into a partial point cloud P_0. PCDreamer then generates the completed point cloud:

$$P_1 = \mathcal{G}_{\mathrm{PCD}}(P_0), \tag{1}$$

where P_1 contains both observed and reconstructed geometry. The completed point cloud is subsequently partitioned into local regions, and each region is encoded as a local feature α_i. The language query is encoded into a semantic representation F_T for subsequent reliability estimation and functional reasoning.

3.3 RELIABILITY-GUIDED 3D AFFORDANCE GROUNDING

The completed point cloud provides additional geometric evidence for occluded regions, but reconstructed geometry may contain locally unreliable structures under severe occlusion. Treating all reconstructed regions equally can introduce misleading geometric evidence into affordance grounding. We therefore estimate the reliability of each local region conditioned on the target affordance and use it to guide subsequent feature refinement.The completed geometry is subsequently represented by a set of local regions, each retaining both geometric information and its observation/completion status. These local representations, together with the language representation, are then fed into the reliability-guided grounding module for query-conditioned processing.

Query-conditioned Geometric Reliability. For each local region, we extract its geometric feature α_i, while the language query is encoded into a semantic representation F_T. Together with the completion confidence c_i^{pcd} and observation state m_i, these features are used to estimate the localization reliability:

$$u_i = \sigma\left(\psi\left(\alpha_i, c_i^{\mathrm{pcd}}, m_i, F_T\right)\right), \tag{2}$$

where $u_i \in [0,1]$ denotes the reliability of the i-th local region for the current affordance query. This query-conditioned formulation allows the usefulness of reconstructed geometry to vary with the target affordance.

During training, the reliability is supervised using the discrepancy between reconstructed and ground-truth local geometry:

$$u_i^* = \exp\left(-\beta\ \mathrm{CD}(S_i, S_i^{gt})\right), \tag{3}$$

where S_i and S_i^{gt} denote the reconstructed and ground-truth local geometries, respectively. The reliability loss $\mathcal{L}_{\mathrm{rel}}$ is defined between the predicted and target reliability.

Reliability-guided Local Detail Refinement. The predicted reliability is further used to regulate query-relevant local feature refinement. Specifically, the refined feature is computed as

$$\bar{\alpha}_i = \alpha_i + u_i\, \mathcal{R}(\alpha_i, F_T), \tag{4}$$

where $\mathcal{R}(\cdot)$ denotes the query-conditioned refinement function. Reliable regions receive stronger query-relevant feature enhancement, while potentially unreliable reconstructed regions are suppressed. The resulting refined features $\bar{\alpha}_i$ are then passed to the subsequent spatial-functional reasoning module.

3.4 QUERY-GUIDED SPATIAL-FUNCTIONAL REASONING

Spatial-functional Relational Modeling. Local geometric features alone may not distinguish the target functional region from neighboring regions. Functional regions are characterized by their geometry and relationships with surrounding parts, and these relationships may vary with the language query. We therefore jointly model spatial and query-conditioned functional relationships over the refined local regions.

Specifically, we construct a query-guided spatial-functional graph

$$G_{\mathrm{QSF}} = (V, E_s, E_f(q)), \tag{5}$$

where each node corresponds to a local region, E_s represents geometry-based spatial relationships, and $E_f(q)$ represents query-conditioned functional relationships. Spatial relationships are derived from relative geometry, while functional relationships are learned from local representations and query semantics.

The two relation types are propagated separately and subsequently integrated to update each local region representation. This allows each region to incorporate surrounding context while preserving the distinction between spatial proximity and functional relevance.

Functional Region Prediction. After relational reasoning, each local region contains refined geometric and spatial-functional contextual information. These representations are fused with localization reliability and the language representation:

$$F_i = \text{Fusion}\left(\bar{\alpha}_i, h_i^{\text{SF}}, u_i, F_T\right), \tag{6}$$

where h_i^{SF} denotes the output representation of the spatial-functional reasoning module.

A localization head predicts the affordance score for each local region. The region with the highest query-conditioned score is selected as the predicted functional region, while region-level scores can also be projected to the corresponding points to obtain a dense affordance map. The final prediction is generated directly from the given multi-view RGB-D observations without iterative active-view selection.

3.5 Training Objective

The framework is trained by jointly considering geometry completion, localization-aware reliability estimation, and functional-region localization. The completion objective follows the PCDreamer backbone, while the downstream objectives supervise reliability estimation and final localization:

$$\mathcal{L} = \lambda_{\text{comp}}\mathcal{L}_{\text{comp}} + \lambda_{\text{rel}}\mathcal{L}_{\text{rel}} + \lambda_{\text{loc}}\mathcal{L}_{\text{loc}}, \tag{7}$$

where $\mathcal{L}_{\text{comp}}$, $\mathcal{L}_{\text{rel}}$, and $\mathcal{L}_{\text{loc}}$ denote the completion, localization-aware reliability, and functional localization objectives, respectively.

In practice, the completion backbone is first kept fixed while the localization-aware reliability, local refinement, relational reasoning, and localization modules are optimized. Limited joint fine-tuning is then performed with a smaller learning rate for the completion backbone. This preserves the pretrained geometric prior while adapting the reconstructed representation to functional localization.

4 Experiments

4.1 Experimental Setup

Dataset and Protocol. We evaluate our method on the Affogato benchmark for 3D affordance grounding, which provides 3D object representations, natural-language affordance descriptions, and corresponding affordance annotations Lee et al. (2026). We use the training split for optimization and the held-out test split for evaluation. Since our method targets affordance grounding from incomplete observations, the input consists of partial multi-view RGB-D observations rather than ground-truth complete geometry.

Input Construction. The RGB-D observations are transformed into a common object coordinate system using the corresponding camera intrinsics and poses and fused into a partial point cloud P_o. PCDreamer Wei et al. (2025) is then used to recover an amodal point-cloud representation P_c. The completed geometry is subsequently partitioned into local regions for localization-aware geometric reasoning and affordance prediction. The observation state of each local region is retained as an auxiliary cue, allowing the downstream modules to distinguish directly observed geometry from reconstructed geometry.

Implementation Details. We use PCDreamer as the pretrained amodal point-cloud completion backbone. Local regions and language queries are encoded into 256-dimensional geometric and semantic features, respectively. The downstream modules are optimized using AdamW with a learning rate of 1×10^{-4}, while the PCDreamer backbone is fine-tuned with a learning rate of 1×10^{-5}. We use two message-passing layers for spatial-functional reasoning and a two-layer MLP for affordance prediction. All experiments are conducted on 4 NVIDIA RTX 4090 GPUs with a batch size of 16.

Evaluation Metrics We evaluate 3D affordance grounding using four complementary metrics: average Intersection over Union (aIoU), Area Under the ROC Curve (AUC), Similarity (SIM), and Mean Absolute Error (MAE) Park et al. (2026). aIoU measures the spatial overlap between the predicted and ground-truth affordance regions. AUC evaluates the ranking quality of predicted

Table 1: Quantitative comparison on the 3D affordance grounding task. The best results are highlighted in bold.

Method	aIoU ↑	AUC ↑	SIM ↑	MAE ↓
OpenAD Nguyen et al. (2023)	3.1	64.8	0.329	0.150
PointRefer Li et al. (2024)	10.5	76.1	0.405	0.120
Espresso-3D Lee et al. (2026)	13.6	**79.0**	0.429	**0.111**
Affostruction Park et al. (2026)	19.1	72.0	0.426	0.217
Ours	**24.0**	75.9	**0.451**	0.177

affordance scores over the object surface. SIM measures the similarity between the predicted and ground-truth affordance distributions, while MAE measures the absolute prediction error. Higher aIoU, AUC, and SIM indicate better performance, whereas lower MAE indicates better performance.

4.2 Quantitative Results

Comparison with Existing Methods. We compare our framework with representative 3D affordance grounding methods, including OpenAD Nguyen et al. (2023), PointRefer Li et al. (2024), Espresso-3D Lee et al. (2026), and Affostruction Park et al. (2026). Following the commonly adopted evaluation protocol, we report four metrics for affordance grounding: aIoU, AUC, SIM, and MAE. The quantitative results are summarized in Table 1.

Table 1 summarizes the quantitative comparison across four commonly used evaluation metrics. Our framework achieves 24.0 aIoU, 75.9 AUC, 0.451 SIM, and 0.177 MAE. Among the reported results, our method achieves the highest aIoU and SIM, indicating accurate functional-region overlap and strong agreement between the predicted and ground-truth affordance distributions. The comparison should be interpreted in the context of the input setting. Existing methods are reported according to their respective evaluation protocols, while our framework specifically targets partial multi-view RGB-D observations in which the target object may be partially occluded or incompletely observed. Therefore, the numerical results are primarily used to provide a reference for affordance grounding performance rather than to establish a strict ranking across different observation protocols.

Under partial observations, geometry completion provides additional information for regions that are not directly visible. However, the reconstructed geometry may contain spatially varying errors, which can subsequently affect affordance localization. Our framework therefore does not treat the completed geometry as uniformly reliable. Instead, it estimates localization reliability conditioned on the affordance query and uses this reliability to regulate reconstructed local features before performing query-guided spatial-functional reasoning. The resulting performance demonstrates that explicitly modeling reconstruction reliability can improve affordance grounding under incomplete observations. In particular, the gains in aIoU and SIM suggest that reliability-aware processing helps the model better identify functional regions and maintain consistency with the target affordance distribution, despite the uncertainty introduced by partial observation and geometry completion.

4.3 Qualitative Results

We further compare qualitative affordance predictions under incomplete and occluded observations. The compared methods receive the same partial multi-view RGB-D observations and affordance query.As shown in Fig. 3, incomplete observations can lead to missing or insufficient geometric evidence around the queried affordance region. Geometry completion provides additional structural information for these unobserved areas, but different methods still produce distinct affordance predictions on the reconstructed geometry. Across the shown examples, the proposed method generally generates affordance regions that are more consistent with the corresponding ground-truth regions, including the seat of the chair, the drawer front, the tabletop, the sofa back, and the cabinet opening. In contrast, predictions from existing methods may exhibit spatial deviations or less accurate local responses around the target regions. These differences are particularly noticeable when the queried functional region is partially observed or requires information from reconstructed geometry. The qualitative results indicate that the proposed framework can better exploit the reconstructed object structure while focusing the prediction on the functional region specified by the query. This provides

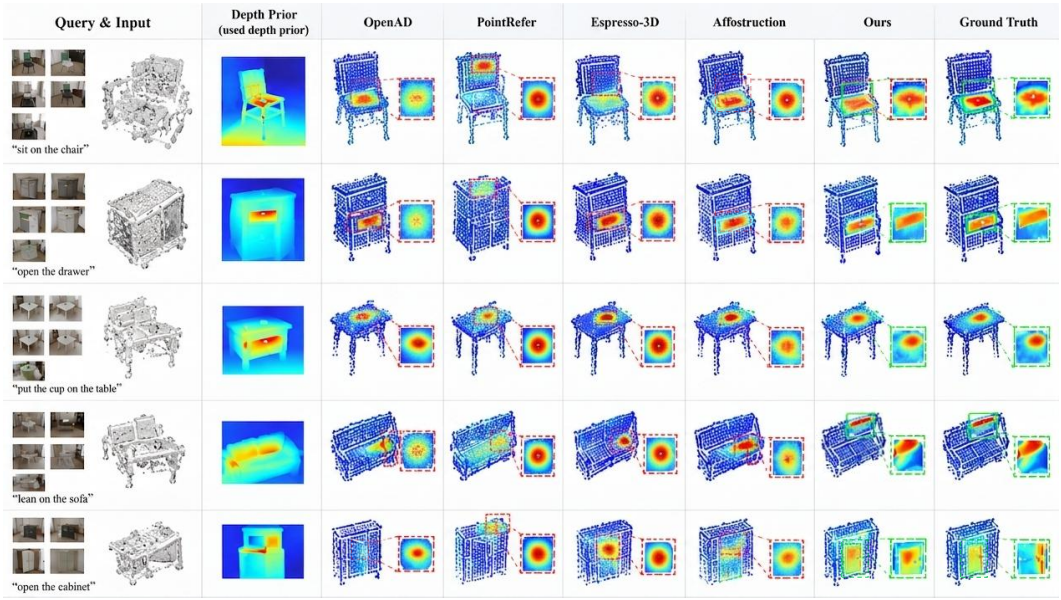

Figure 3: Qualitative comparison of 3D affordance grounding under occlusion. Given the same partial multi-view RGB-D observations and affordance query, different methods produce affordance predictions on the object surface. The last column shows the corresponding ground-truth affordance regions for comparison.

Table 2: Ablation study of the proposed components.

Variant	aIoU	AUC	SIM	MAE
Completion	19.8	71.6	0.425	0.193
+ LGR	21.5	73.4	0.437	0.185
+ RLDR	22.7	74.6	0.444	0.181
+ QSFR	**24.0**	**75.9**	**0.451**	**0.177**

Table 3: Comparison of different local refinement strategies.

Strategy	aIoU	AUC	SIM	MAE
No refinement	21.5	73.4	0.437	0.185
Uniform	22.0	73.8	0.440	0.183
Reliability-guided	**22.7**	**74.6**	**0.444**	**0.181**

Table 4: Performance under different visible-surface ratios.

Visibility	aIoU	AUC	SIM	MAE
20–40%	18.1	69.2	0.412	0.196
40–60%	21.3	72.8	0.434	0.184
60–80%	23.7	75.1	0.448	0.178
80–100%	**26.2**	**78.0**	**0.463**	**0.166**

Table 5: Analysis of localization-aware geometric reliability.

Reliability	Error	aIoU
0.00–0.25	0.182	15.7
0.25–0.50	0.141	18.9
0.50–0.75	0.103	23.1
0.75–1.00	**0.071**	**28.6**

a qualitative illustration of the effectiveness of reliability-guided processing and spatial-functional reasoning for affordance localization under incomplete and occluded observations.

4.4 Ablation Study

We evaluate the contribution of the three main components introduced after amodal geometry completion: localization-aware geometric reliability (LGR), reliability-guided local detail refinement (RLDR), and query-guided spatial-functional reasoning (QSFR). All configurations use the same completed geometry, data split, and training protocol.

Table 2 shows consistent improvements as the proposed components are progressively introduced. The completion-based baseline achieves 19.8 aIoU. Introducing LGR improves aIoU to 21.5 and reduces MAE from 0.193 to 0.185, indicating that query-conditioned reliability provides useful information beyond the completed geometry itself. Adding RLDR further increases aIoU to 22.7 and improves all four evaluation metrics. This result indicates that the estimated reliability can be directly used to regulate local feature refinement, allowing the model to emphasize reconstructed regions that provide more useful evidence for the current affordance query. Finally, QSFR further improves the performance to **24.0** aIoU, **75.9** AUC, **0.451** SIM, and **0.177** MAE. The consistent improvement demonstrates that spatial context and query-dependent functional relationships provide complementary information for affordance localization. Moreover, Table 3 isolates the effect of the refinement strategy. Uniform refinement provides only a limited improvement over the LGR baseline, whereas reliability-guided refinement achieves **22.7** aIoU, **74.6** AUC, **0.444** SIM, and **0.181** MAE. This comparison indicates that the improvement does not simply result from additional local feature processing; rather, regulating the refinement according to localization reliability is important for effectively using reconstructed geometry.

4.5 Further Analysis

We further analyze the behavior of the proposed framework under different observation completeness levels and examine whether the predicted localization reliability is associated with the quality and usefulness of reconstructed local geometry.

Robustness to Observation Completeness. We divide the test samples according to their visible-surface ratio and evaluate the proposed method under different levels of incomplete observation. The results are reported in Table 4. As the visible-surface ratio increases from 20–40% to 80–100%, aIoU increases from 18.1 to **26.2**, while AUC and SIM increase from 69.2 and 0.412 to **78.0** and **0.463**, respectively. At the same time, MAE decreases from 0.196 to **0.166**. The trend is consistent with the increasing amount of directly observed geometric evidence. Under lower visibility, a larger portion of the object structure must be inferred through completion, making reliable affordance localization more challenging. Nevertheless, the framework maintains effective affordance grounding across the evaluated visibility ranges. This behavior is consistent with the proposed reliability-guided processing: reconstructed geometry can provide additional evidence for occluded regions, while its contribution to local reasoning is regulated according to the estimated reliability.

Localization Reliability Analysis. To examine whether predicted localization reliability is associated with geometric evidence quality and downstream localization performance, we group local regions according to their predicted reliability and report their local reconstruction error and affordance localization quality in Table 5. The results show a clear relationship between reliability and localization quality. As the predicted reliability increases from 0.00–0.25 to 0.75–1.00, the local reconstruction error decreases from 0.182 to **0.071**, while local aIoU increases from 15.7 to **28.6**. This indicates that the reliability prediction captures meaningful differences among reconstructed local regions. Importantly, the reliability is not intended to indicate only whether a point is observed or generated. Instead, it provides a task-oriented estimate of how useful the reconstructed region is for the current affordance localization query. The observed relationship between reliability, reconstruction error, and localization quality supports its use as the weighting factor for adaptive local refinement. Overall, the experiments support the proposed processing strategy: amodal geometry completion provides missing structural evidence, localization-aware reliability evaluates reconstructed regions with respect to the target affordance, reliability-guided refinement selectively enhances useful local details, and query-guided spatial-functional reasoning incorporates contextual information for final affordance grounding.

5 Conclusion

This work addresses language-guided 3D functional region localization under occlusion. We propose a localization-aware framework that combines amodal geometry completion, query-conditioned reliability estimation, and reliability-guided local refinement to reduce the impact of unreliable reconstructed geometry. Query-guided spatial-functional reasoning is further introduced to model contextual relationships and improve functional region localization.

AI USE STATEMENT

We used generative AI tools for language polishing, improving expression, and enhancing the overall writing of the manuscript. We reviewed all AI-assisted content and take responsibility for the final content of this work, including text and arguments produced with the assistance of generative AI.

ETHICS STATEMENT

As 3D affordance grounding may be used to support robotic manipulation, the predicted functional regions should not be treated as fully reliable action decisions; human or system-level verification remains necessary before physical interaction, particularly under severe occlusion or uncertain reconstruction.

REPRODUCIBILITY STATEMENT

The main text describes the model architecture, experimental setup, network configurations, and evaluation metrics.

REFERENCES

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 (CVPR)*, pp. 14131–14140, 2024.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 20522–20535, 2022.

Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. Anchorformer: Point cloud completion from discriminative nodes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13581–13590, 2023.

Ruikai Cui, Shi Qiu, Saeed Anwar, Jiawei Liu, Chaoyue Xing, Jing Zhang, and Nick Barnes. P2c: Self-supervised point cloud completion from single partial clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14305–14314, 2023.

Chun Feng, Joy Hsu, Weiyu Liu, and Jiajun Wu. Naturally supervised 3d visual grounding with language-regularized concept learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13269–13278, 2024.

Zhiheng Fu, Longguang Wang, Lian Xu, Zhiyong Wang, Hamid Laga, Yulan Guo, Farid Boussaid, and Mohammed Bennamoun. Vapcnet: Viewpoint-aware 3d point cloud completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12108–12118, 2023.

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 (ECCV)*, pp. 417–433, 2022.

Junha Lee, Eunha Park, Chunghyun Park, Dahyun Kang, and Minsu Cho. Affogato: Learning open-vocabulary affordance grounding with automated data generation at scale. In *European Conference on Computer Vision (ECCV)*, 2026.

An Li, Zhe Zhu, and Mingqiang Wei. Genpc: Zero-shot point cloud completion via 3d generative priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1308–1318, 2025a.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3707–3717, 2025b.

Shanshan Li, Pan Gao, Xiaoyang Tan, and Mingqiang Wei. Proxyformer: Proxy alignment assisted point cloud completion with missing part sensitive transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9466–9475, 2023.

Yicong Li, Na Zhao, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-Seng Chua. Laso: Language-guided affordance segmentation on 3d object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14251–14260, 2024.

Fangzhou Lin, Yun Yue, Songlin Hou, Xuechu Yu, Yajun Xu, Kazunori D. Yamada, and Ziming Zhang. Hyperbolic chamfer distance for point cloud completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14549–14560, 2023.

Zhenyang Liu, Yikai Wang, Sixiao Zheng, Tongying Pan, Longfei Liang, Yanwei Fu, and Xiangyang Xue. Reasongrounder: Lvlm-guided hierarchical feature splatting for open-vocabulary 3d visual grounding and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3718–3727, 2025.

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 (CVPR)*, pp. 16454–16463, 2022.

Toan Nguyen, Minh Nhat Vu, An Vuong, Dzung Nguyen, Thieu Vo, Ngan Le, and Anh Nguyen. Open-vocabulary affordance detection in 3d point clouds. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5692–5698, 2023.

Chunghyun Park, Seunghyeon Lee, and Minsu Cho. Affostruction: 3d affordance grounding with generative reconstruction. *arXiv preprint arXiv:2601.09211*, 2026.

Stuti Pathak, Prashant Kumar, Dheeraj Baiju, Nicholus Mboga, Gunther Steenackers, and Rudi Penne. Revisiting point cloud completion: Are we ready for the real-world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25388–25398, 2025.

Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 7587–7597, 2024.

Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *Proceedings of the 5th Conference on Robot Learning*, volume 164, pp. 1046–1056, 2022.

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

Ramesh Ashok Tabib, Dikshit Hegde, and Uma Mudenagudi. Lgafford-net: A local geometry aware affordance detection network for 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5261–5270, 2024.

Jesse Thomason, Mohit Shridhar, Yonatan Bisk, Chris Paxton, and Luke Zettlemoyer. Language grounding with 3d objects. In *Proceedings of the 5th Conference on Robot Learning*, volume 164, pp. 1691–1701, 2022.

Nguyen Toan, Minh Nhat Vu, Baoru Huang, Tuan Van Vo, Vy Truong, Ngan Le, Thieu Vo, Bac Le, and Anh Nguyen. Language-conditioned affordance-pose detection in 3d point clouds. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3071–3078, 2024. doi: 10.1109/ICRA57147.2024.10610008.

Jun Wang, Ying Cui, Dongyan Guo, Junxia Li, Qingshan Liu, and Chunhua Shen. Pointattn: You only need attention for point cloud completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5472–5480, 2024a. doi: 10.1609/aaai.v38i6.28356.

Yuan Wang, Yali Li, and Shengjin Wang. G^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 (CVPR)*, pp. 13917–13926, 2024b.

Guangshun Wei, Yuan Feng, Long Ma, Chen Wang, Yuanfeng Zhou, and Changjian Li. Pcdreamer: Point cloud completion through multi-view diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27243–27253, 2025.

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 (CVPR)*, pp. 17253–17262, 2024.

Xingguang Yan, Liqiang Lin, Niloy J. Mitra, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Shapeformer: Transformer-based shape completion via sparse representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6239–6249, 2022.

Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3d object affordance from 2d interactions in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10905–10915, 2023. doi: 10.1109/ICCV51070.2023.01001.

Chunlin Yu, Hanqing Wang, Ye Shi, Haoyang Luo, Sibei Yang, Jingyi Yu, and Jingya Wang. Seqafford: Sequential 3d affordance reasoning via multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1691–1701, 2025. doi: 10.1109/CVPR52734.2025.00165.

Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, and Zhen Li. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20623–20633, 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 (ICCV)*, pp. 15225–15236, 2023.

Haoran Zhou, Yun Cao, Wenqing Chu, Junwei Zhu, Tong Lu, Ying Tai, and Chengjie Wang. Seedformer: Patch seeds based point cloud completion with upsample transformer. In *European Conference on Computer Vision (ECCV)*, pp. 19–35, 2022.

He Zhu, Quyu Kong, Kechun Xu, Xunlong Xia, Bing Deng, Jieping Ye, Rong Xiong, and Yue Wang. Grounding 3d object affordance with language instructions, visual observations and interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17337–17346, 2025a. doi: 10.1109/CVPR52734.2025.01616.

Zhe Zhu, Honghua Chen, Xing He, Weiming Wang, Jing Qin, and Mingqiang Wei. Svdformer: Complementing point cloud via self-view augmentation and self-structure dual-generator. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14508–14518, 2023.

Zhe Zhu, Liangliang Nan, Haoran Xie, Honghua Chen, Jun Wang, Mingqiang Wei, and Jing Qin. Csdn: Cross-modal shape-transfer dual-refinement network for point cloud completion. *IEEE Transactions on Visualization and Computer Graphics*, 30(7):3545–3563, 2024. doi: 10.1109/TVCG.2023.3236061.

Zhe Zhu, Honghua Chen, Xing He, and Mingqiang Wei. Pointsea: Point cloud completion via self-structure augmentation. *International Journal of Computer Vision*, 133(7):4770–4794, 2025b. doi: 10.1007/s11263-025-02400-y.