EFFICIENT 3DGS-BASED AFFORDANCE REASONING VIA SEMANTICALLY STRUCTURED SHAPE TOKENS

Anonymous authors
Paper under double-blind review

ABSTRACT

3D affordance reasoning requires accurately mapping high-level interaction intent to fine-grained local functional regions. However, existing 3DGS-based affordance methods typically perform cross-modal interaction directly between spatially organized low-level Gaussian features and textual intent, without effectively modeling geometric relationships through hierarchical semantic structures. This can lead to incomplete localization and local fragmentation in object structures, while also introducing substantial computational redundancy. Meanwhile, reliance on low-level spatial patterns limits the abstraction of transferable functional structures and cross-category generalization. To address this issue, we propose ST-Afford, an efficient framework for 3DGS-based affordance reasoning that organizes dense Gaussian geometry into compact semantically structured representations and further dynamically selects task-relevant geometric information based on interaction intent. Specifically, we introduce a Semantically Structured Geometry Extractor, which compresses and encodes high-dimensional 3DGS features into compact shape-prior tokens. Based on these tokens, an Intent-Guided Dynamic Fusion module uses interaction text as a condition to dynamically select task-relevant geometric semantic information and fuse it with local Gaussian features, enabling accurate localization of fine-grained functional regions from high-level interaction intent. Experimental results show that ST-Afford substantially reduces inference overhead while generating more complete and accurate affordance reasoning, and achieves stronger zero-shot cross-category generalization on unseen categories.

1 INTRODUCTION

Affordance reasoning aims to ground high-level interaction intent into fine-grained functional regions on objects, providing actionable spatial cues for embodied agents to perform grasping, manipulation, and action planning (Zitkovich et al., 2023). Early studies mainly addressed affordance localization from 2D images or videos. Sawatzky *et al.* formulated affordance detection as pixel-wise multi-label segmentation with sparse keypoint supervision (Sawatzky et al., 2017), while Luddecke and Worgotter predicted dense affordance maps from single RGB images (Lüddecke & Wörgötter, 2017). However, 2D observations are inherently limited by perspective projection, occlusion, and the lack of explicit depth, making it difficult to faithfully capture the 3D geometry and spatial relationships of functional regions.

To overcome these limitations, later studies extended affordance reasoning to explicit 3D representations, with point clouds becoming a common choice (Qi et al., 2017). 3D AffordanceNet performed point-wise affordance estimation directly on point clouds (Deng et al., 2021), while LASO further incorporated natural language for interaction-conditioned functional-region localization (Li et al., 2024). Nevertheless, point clouds represent surfaces through discrete and irregular samples, limiting their ability to preserve continuous geometry and fine-grained local structure (Mescheder et al., 2019). 3D Gaussian Splatting (Kerbl et al., 2023) provides a denser representation through anisotropic Gaussian primitives and was recently introduced into affordance reasoning by 3DAffordSplat (Wei et al., 2025). Despite its richer geometry, existing 3DGS-based methods still reason largely over dense low-level Gaussian features without first organizing them into compact semantic structures. Consequently, high-level interaction intent repeatedly interacts with substantial task-irrelevant geometry, increasing computational overhead and encouraging reliance on instance-

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.

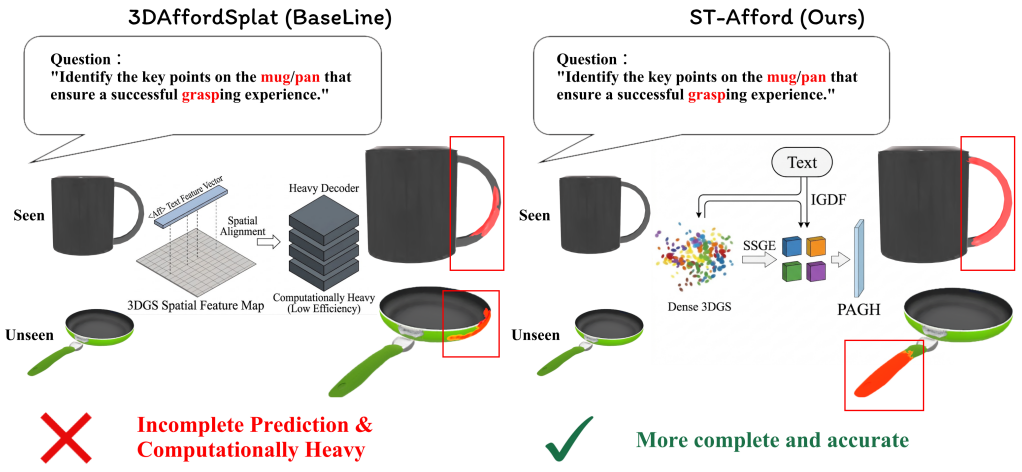

Figure 1: **Comparison between ST-Afford and 3DAffordSplat.** ST-Afford replaces dense Gaussian-level reasoning with compact semantic priors and intent-guided selection, producing more complete and coherent affordance predictions.

specific spatial patterns, which can hinder both fine-grained localization and cross-category generalization.

To address these issues, we propose **ST-Afford**, an efficient framework for 3DGS-based affordance reasoning. Instead of directly performing cross-modal interaction over dense Gaussian primitives, ST-Afford first organizes the geometry into a compact semantic-geometric representation and then selectively retrieves information relevant to the current interaction intent. Specifically, the Semantically Structured Geometry Extractor compresses high-dimensional 3DGS features into a fixed set of structured *shape-prior tokens* that capture higher-level geometric semantics. The Intent-Guided Dynamic Fusion module then uses the interaction text to select task-relevant shape priors and fuse them back into local Gaussian features, combining structured global context with fine-grained spatial information. This design reduces exhaustive cross-modal reasoning over dense primitives while facilitating more transferable functional representations. Experiments show that ST-Afford improves affordance localization accuracy and spatial coherence, substantially increases inference efficiency, and achieves stronger zero-shot generalization to unseen object categories.

The main contributions of this work are as follows:

- We propose an efficient 3DGS-based affordance reasoning framework that organizes dense Gaussian geometry into a compact semantically structured representation, enabling effective grounding from high-level interaction intent to fine-grained local functional regions.
- We introduce ordered semantic-geometric priors into 3DGS affordance reasoning by adapting causal register modeling and semantic prefix supervision, and further design an intent-conditioned fusion mechanism to retrieve task-relevant structural information for fine-grained localization.
- Extensive experiments demonstrate that ST-Afford substantially reduces inference overhead while producing more accurate, complete, and spatially coherent affordance predictions, and further exhibits stronger zero-shot cross-category generalization on unseen object categories.

2 Related Work

2.1 3D Affordance Reasoning

Early 3D affordance reasoning mainly relied on point-cloud representations to localize functional regions and model part-level interactions. 3D AffordanceNet (Deng et al., 2021) established dense

affordance prediction on 3D objects, while PartAfford (Xu et al., 2022) further explored part-level affordance discovery across object categories. With the introduction of multimodal cues, IAG-Net (Yang et al., 2023) incorporated visual semantics into 3D affordance grounding, and LASO (Li et al., 2024) further formulated language-guided affordance localization by grounding textual interaction descriptions to local functional regions. Recent studies have increasingly emphasized more generalizable and interaction-aware reasoning, including cross-category transfer and dynamic interaction modeling (Lu et al., 2025; Mao et al., 2026; Wang et al., 2026). More recently, 3DAffordSplat (Wei et al., 2025) introduced 3D Gaussian representations into affordance reasoning and performed cross-modal alignment between Gaussian features and textual intent. While 3DGS provides richer local geometry, existing reasoning still largely operates over dense low-level Gaussian features. ST-Afford instead introduces an intermediate structured geometric representation before interaction-conditioned localization.

2.2 Multimodal Semantic Representation in 3D

Another closely related direction transfers high-level visual and language semantics into 3D representations. Early vision-language 3D methods align geometric features with image- or language-level semantic spaces to support open-vocabulary scene understanding (Hong et al., 2023; Peng et al., 2023; Kerr et al., 2023; Chen et al., 2020). With the emergence of 3DGS, LangSplat (Qin et al., 2024), Feature 3DGS (Zhou et al., 2024), and Gaussian Grouping (Ye et al., 2024) further distill visual or language features into Gaussian primitives, enabling semantic querying, segmentation, and grouping while preserving dense spatial correspondence. Related works also incorporate interaction cues or task-specific information into Gaussian fields for segmentation and robotic manipulation (Choi et al., 2024; Jain et al., 2024; Ji et al., 2025; Shorinwa et al., 2025). These studies demonstrate that high-level semantics can effectively enrich low-level 3D geometry. In contrast, our focus is not merely to attach semantic attributes to individual primitives, but to further organize dense Gaussian features into structured geometric priors that can be selectively queried by interaction intent.

2.3 Structured 3D Representation and Semantic Tokenization

A complementary line of research investigates compact and structured representations for complex 3D geometry. Point-BERT (Yu et al., 2022) tokenizes local point-cloud patches and models their relationships with Transformers, demonstrating the effectiveness of token-based abstraction for 3D representation learning, while register-based representations (Darcet et al., 2024) introduce learnable latent tokens to aggregate information from dense visual features. More recently, LoST (Dutt et al., 2026) explores semantic tokenization for 3D shapes, organizing shape information into an ordered token sequence in which compact prefixes retain dominant semantic content. These approaches motivate moving beyond dense primitive-level features toward more compact semantic abstractions. Building on this direction, ST-Afford adapts structured tokenization to 3DGS-based affordance reasoning and further couples the resulting shape priors with interaction-conditioned selection and local Gaussian localization.

3 Methodology

We propose **ST-Afford**, an efficient framework for 3DGS-based affordance reasoning. As shown in Fig. 2, ST-Afford organizes dense Gaussian features into compact structured priors and selectively retrieves interaction-relevant geometry for fine-grained affordance localization. Specifically, the Semantically Structured Geometry Extractor constructs ordered shape-prior tokens, while the Intent-Guided Dynamic Fusion dynamically selects and associates relevant priors with individual Gaussian primitives. A lightweight Point-wise Affordance Grounding Head then produces point-wise predictions. The corresponding training objectives are described in Sec. 3.3.

3.1 Semantically Structured Geometry Extractor

Dense Gaussian representations preserve fine-grained local geometry, yet their primitive-level organization lacks a compact structural abstraction for interaction with high-level semantic intent. To

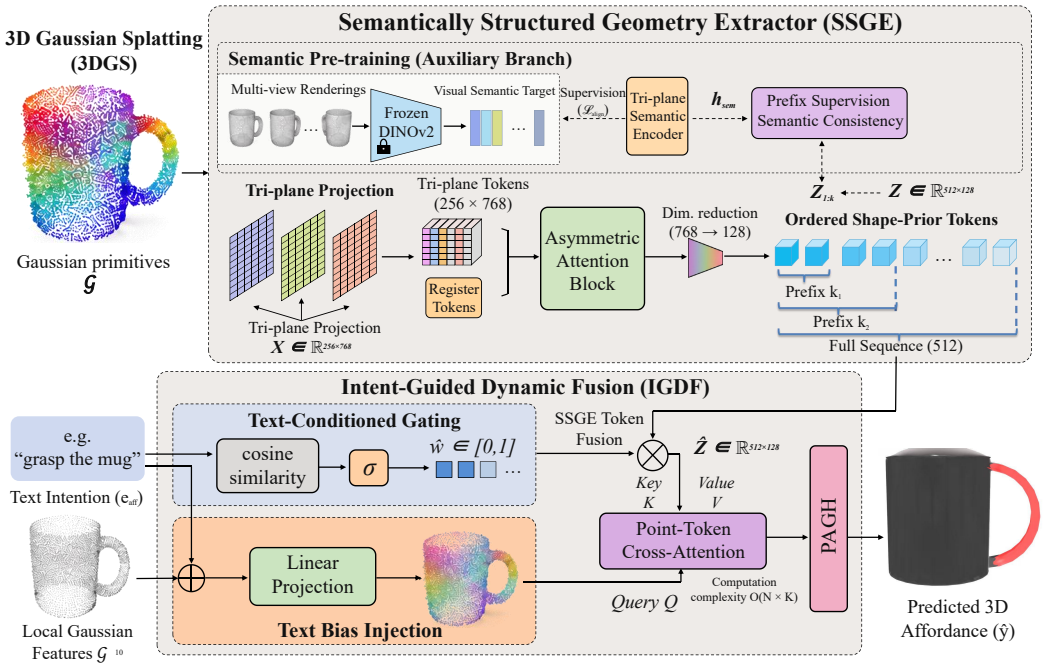

Figure 2: **Overall architecture of ST-Afford.** SSGE organizes dense 3D Gaussian features into compact, semantically structured shape-prior tokens. IGDF then selects intent-relevant priors and fuses them with local Gaussian geometry through point-to-token interaction, followed by lightweight point-wise affordance prediction.

address this issue, we introduce the SSGE. We first project and aggregate the dense Gaussian features onto three orthogonal feature planes to obtain a compact tri-plane representation X. Based on this representation, SSGE first establishes a semantic embedding space for the underlying geometry and then organizes the geometric information into an ordered sequence of shape-prior tokens.

3.1.1 Tri-plane Semantic Encoding

To provide the subsequent structured tokens with stable object-level semantics, we first learn a semantic encoder over the tri-plane representation. Following the semantic alignment strategy of LoST (Dutt et al., 2026), we train an auxiliary tri-plane semantic encoder with a RIDA-inspired objective to map the tri-plane representation into a semantic space supervised by multi-view visual observations.

Given a tri-plane feature X, the encoder performs local plane-wise feature extraction, patch embedding, Transformer contextualization, and attention pooling. We denote the overall mapping as

$$h_{\text{sem}} = f_\theta(X), \tag{1}$$

where $f_\theta(\cdot)$ denotes the auxiliary tri-plane semantic encoder and h_{sem} is the resulting global semantic latent.

During training, a frozen DINOv2 encoder (Oquab et al., 2024) extracts teacher features from multi-view renderings of the same object, providing semantic supervision for learning the mapping from 3D tri-plane geometry to the visual semantic space. In addition to object-level semantic correspondence, the training preserves inter-object semantic relations and local spatial structure, allowing h_{sem} to serve as a stable semantic target for subsequent shape-prior tokenization. Once trained, the semantic encoder is frozen.

3.1.2 ORDERED SHAPE-PRIOR TOKENIZATION

Given the tri-plane representation X, we obtain tri-plane tokens T through patch embedding and introduce learnable register tokens R, which are jointly processed by a Transformer with asymmetric causal attention. Tri-plane tokens interact only within their own stream, while register tokens can attend to the full tri-plane representation and, within the register stream, only to themselves and preceding tokens. This asymmetric information flow enables the register tokens to aggregate global geometric context without feeding compressed information back into the low-level feature stream, while the causal dependency imposes an explicit ordering in which earlier tokens provide the basis for subsequent structural representations.

After Transformer encoding, we retain only the register outputs and compress their feature dimension through a lightweight bottleneck, yielding the shape-prior sequence $Z = \{z_1, z_2, \ldots, z_K\}$.

Causal dependency alone, however, does not guarantee that the resulting token order carries meaningful semantic structure. We therefore introduce random-prefix supervision during semantic pre-training. For each training sample, we randomly retain the first k tokens,

$$Z_{1:k} = \{z_1, z_2, \ldots, z_k\}, \tag{2}$$

and average the retained tokens before mapping them through a lightweight semantic projection head $D(\cdot)$ to obtain the prefix representation $\tilde{h}_k$.

The projected representation $\tilde{h}_k$ is constrained to preserve the frozen semantic latent $h_{\text{sem}} = f_\theta(X)$, while representations from different prefix lengths of the same object are encouraged to remain semantically consistent. Combined with causal dependency, this random-prefix supervision encourages earlier tokens to capture salient structural semantics, while later tokens progressively complement them with additional geometric information.

Since the downstream task requires structured geometric priors rather than 3D generation, we use a lightweight two-layer MLP as $D(\cdot)$ to predict the frozen semantic latent. The projection head is used only during semantic pre-training and removed afterwards. The resulting ordered shape-prior tokens provide a compact semantic-geometric representation for subsequent interaction-conditioned reasoning.

3.2 INTENT-GUIDED DYNAMIC FUSION

The shape-prior tokens extracted by SSGE characterize the structural properties of an object independently of a specific interaction. However, different interaction intents may correspond to different functional structures on the same object. We therefore introduce IGDF to dynamically select task-relevant structural priors according to the interaction instruction and associate them with local Gaussian primitives.

3.2.1 SPATIALLY CONDITIONED INTENT QUERY

Let e_{aff} denote the interaction-text embedding and G denote the physical attributes of the Gaussian primitives, including their position, scale, and rotation. A global text embedding alone cannot distinguish the spatial states of individual Gaussian primitives. We therefore inject local Gaussian geometry as a spatial condition into the textual representation and construct point-specific queries as

$$Q = W_{\text{cond}} e_{\text{aff}} + W_{\text{geo}} G. \tag{3}$$

The textual branch provides the global interaction intent, while the Gaussian branch preserves local spatial variation across primitives. Consequently, the resulting queries encode both the semantic condition of the current interaction and the local geometric state of each candidate region.

3.2.2 INTENT-CONDITIONED SHAPE-PRIOR SELECTION

Although SSGE compresses dense geometry into a compact shape-prior representation, not all priors are equally relevant to a given interaction. Before feature fusion, we therefore dynamically modulate each shape-prior token according to its semantic relevance to the interaction text.

For the k-th shape token z_k, its relevance weight is computed as

$$a_k = \sigma\left(\cos\left(W_Z z_k,\, W_{\text{text}} e_{\text{aff}}\right)\right), \qquad \hat{z}_k = a_k z_k, \tag{4}$$

where W_Z and W_{text} project the geometric and textual features into a shared latent space, respectively, and $\sigma(\cdot)$ denotes the Sigmoid function. The resulting gated priors are denoted as $\hat{Z} = \{\hat{z}_k\}_{k=1}^{K}$.

This text-conditioned modulation allows the effective structural representation of an object to vary with the interaction instruction, suppressing less relevant geometric information while emphasizing structures associated with the current functional intent.

3.2.3 POINT-TO-TOKEN FEATURE FUSION AND AFFORDANCE LOCALIZATION

Given the spatially conditioned queries Q and the intent-gated shape priors $\hat{Z}$, we employ point-to-token cross-attention to associate the selected structural information with individual Gaussian primitives. Specifically, Gaussian queries serve as the Query, while the gated shape-prior tokens provide the Key and Value. This allows different Gaussian primitives to retrieve different combinations of structural information according to both their local geometry and the current interaction intent.

The shape-prior tokens therefore provide compact object-level structural context, while the point queries preserve the local spatial identity of individual Gaussian primitives. The resulting fused representation F_{fused} remains spatially aligned with the original Gaussian primitives.

Since this point-wise correspondence is preserved throughout the fusion process, no additional spatial decoder is required. We directly employ a lightweight PAGH to produce the final affordance prediction:

$$\hat{y} = \sigma\left(W_{\text{pred}} F_{\text{fused}} + b_{\text{pred}}\right) \odot M, \tag{5}$$

where M denotes the valid Gaussian mask and $\hat{y}$ contains the predicted affordance scores of the Gaussian primitives.

Through this process, IGDF first identifies interaction-relevant structures in the compact shape-prior space and then associates the selected structural information with local Gaussian primitives, establishing the correspondence between high-level interaction semantics and fine-grained functional regions.

3.3 TRAINING OBJECTIVES

ST-Afford is trained in two stages. During semantic pre-training, RIDA aligns the tri-plane representation with multi-view visual semantics using

$$\mathcal{L}_{\text{RIDA}} = \mathcal{L}_{\text{global}} + \lambda_{\text{rank}} \mathcal{L}_{\text{rank}} + \lambda_{\text{spatial}} \mathcal{L}_{\text{spatial}}, \tag{6}$$

where the three terms preserve global 3D–visual correspondence, inter-object semantic relations, and local spatial structure, respectively. After freezing the semantic encoder, SSGE is optimized by

$$\mathcal{L}_{\text{SSGE}} = \mathcal{L}_{\text{sem}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{prefix}} \mathcal{L}_{\text{prefix}} + \lambda_{\text{ord}} \mathcal{L}_{\text{ord}}, \tag{7}$$

where $\mathcal{L}_{\text{sem}}$ and $\mathcal{L}_{\text{rec}}$ align prefix representations with the frozen semantic target, while $\mathcal{L}_{\text{prefix}}$ maintains semantic consistency across prefix lengths. To further encourage progressive organization of the ordered shape-prior tokens, we introduce

$$\mathcal{L}_{\text{ord}} = \frac{1}{K-1} \sum_{k=1}^{K-1} \left[d(\tilde{h}_{k+1}, h_{\text{sem}}) - d(\tilde{h}_k, h_{\text{sem}}) \right]_+, \tag{8}$$

where $d(\cdot, \cdot)$ denotes cosine distance and $[\cdot]_+$ is the hinge operator. This objective discourages semantic degradation as additional shape-prior tokens are introduced.

For downstream affordance reasoning, the model is optimized with

$$\mathcal{L}_{\text{aff}} = \mathcal{L}_{\text{focal}} + \mathcal{L}_{\text{dice}} + \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{ctr}}\mathcal{L}_{\text{ctr}}, \tag{9}$$

where $\mathcal{L}_{\text{focal}}$ and $\mathcal{L}_{\text{dice}}$ supervise functional-region localization, $\mathcal{L}_{\text{text}}$ aligns localized geometry with the interaction description, and $\mathcal{L}_{\text{cons}}$ promotes consistent shape priors across objects sharing the same interaction intent. We further introduce an intent-conditioned contrastive objective

$$\mathcal{L}_{\text{ctr}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\text{sim}(g_i, t_i)/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(g_i, t_j)/\tau)}, \tag{10}$$

where g_i denotes the pooled geometric representation of the ground-truth affordance region, t_i is the corresponding interaction embedding, and τ is a temperature parameter. This objective improves the discriminability of functional representations under different interaction intents.

4 Experiments

4.1 Experimental Settings and Evaluation Metrics

We evaluate ST-Afford under both seen- and unseen-category settings. For seen categories, we follow the dataset and split of 3DAffordSplat (Wei et al., 2025), which contains 21 daily-object categories, 18 affordance types, and 52 object-affordance pairs divided into training, validation, and test sets. To assess zero-shot cross-category generalization, we construct an unseen-category test set with 175 instances whose categories are excluded from training. This evaluates whether functional structures learned from seen categories can transfer to objects with different global shapes. Since our study focuses on manipulation-oriented interactions, the unseen set mainly covers grasping and wrap-grasping affordances.

All experiments are conducted on two NVIDIA A100 GPUs. We use MAE, SIM, AUC, and IoU as the main metrics. MAE measures the discrepancy between predicted affordance scores and ground-truth annotations, SIM measures the similarity between their spatial distributions, AUC evaluates the distinction between affordance and non-affordance regions, and IoU measures their spatial overlap. Together, these metrics assess point-wise prediction quality and the spatial completeness of localized affordance regions.

4.2 Comparison on Seen Categories

As shown in Table 1, ST-Afford achieves the best performance across all metrics on the seen-category test set, with clear gains in SIM and IoU, indicating more complete and spatially coherent functional-region localization. Figure 3 shows that IAGNet (Yang et al., 2023) produces weak responses on fine-grained regions, while LASO (Li et al., 2024) exhibits region confusion and over-expansion. 3DAffordSplat (Wei et al., 2025) performs better with 3DGS features but still shows local shifts and incomplete coverage on complex structures. In contrast, ST-Afford better matches the ground truth by using SSGE to construct compact structured priors and IGDF to retrieve interaction-relevant geometry while preserving Gaussian-level spatial correspondence. Some discrepancies remain for very small or topologically complex regions due to the difficulty of preserving fine-grained geometry.

ST-Afford also reaches approximately 11.74 FPS. SSGE compresses dense Gaussian features into compact shape priors before cross-modal reasoning, while IGDF produces point-aligned fused features, avoiding the heavy spatial decoder used in prior 3DGS-based approaches. The lightweight

Table 1: Quantitative comparison on the seen-category dataset.

Method	FPS ↑	MAE ↓	SIM ↑	AUC ↑	IoU ↑
IAGNet	∼0.39	0.2639	0.1713	0.4894	0.0357
LASO	∼0.72	0.2254	0.4094	0.7619	0.2944
3DAffordSplat	∼0.61	0.2136	0.4354	0.8391	0.3084
ST-Afford (Ours)	**∼11.74**	**0.1377**	**0.5675**	**0.8938**	**0.4511**

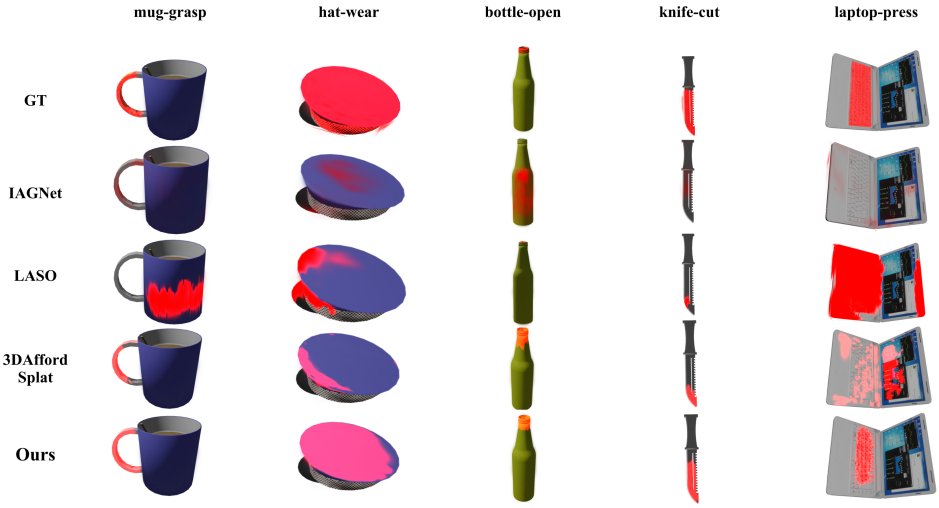

Figure 3: **Qualitative comparison of affordance reasoning on the seen categories.**

PAGH then performs point-wise localization, improving both localization quality and inference efficiency.

4.3 Generalization to Unseen Categories

Unseen-category objects require the model to learn functional structures that transfer across categories, such as recognizing handles with different shapes as graspable regions. To ensure a fair comparison, we use 3DAffordSplat (Wei et al., 2025) as the main baseline, as it also takes 3DGS as input. Other methods mainly operate on point clouds, whereas our unseen-category test set contains only 3DGS annotations and no corresponding point-cloud data. Converting 3DGS to point clouds may introduce geometric information loss and additional noise. Therefore, we compare only with 3DAffordSplat, which uses the same 3DGS representation (Fig. 4).

Table 2: Quantitative comparison on the unseen-category dataset.

Method	MAE ↓	SIM ↑	AUC ↑	IoU ↑
3DAffordSplat	0.2155	0.4502	0.7947	0.2424
ST-Afford (Ours)	**0.1918**	**0.5154**	**0.8237**	**0.3793**

As shown in Table 2, ST-Afford outperforms the baseline on all four metrics, with a clear improvement in IoU. SSGE converts Gaussian features into a fixed number of structured tokens, while IGDF selects relevant geometric information according to the textual intent. Therefore, the model can utilize local functional structures shared across categories rather than relying entirely on the overall shapes of the training categories.

4.4 Ablation Study and Analysis

The ablation results in Table 3 validate the contributions of both SSGE and IGDF. Removing DiffGS together with SSGE causes substantial degradation, indicating that direct reasoning over low-level Gaussian features is insufficient for bridging interaction semantics and local functional structures,

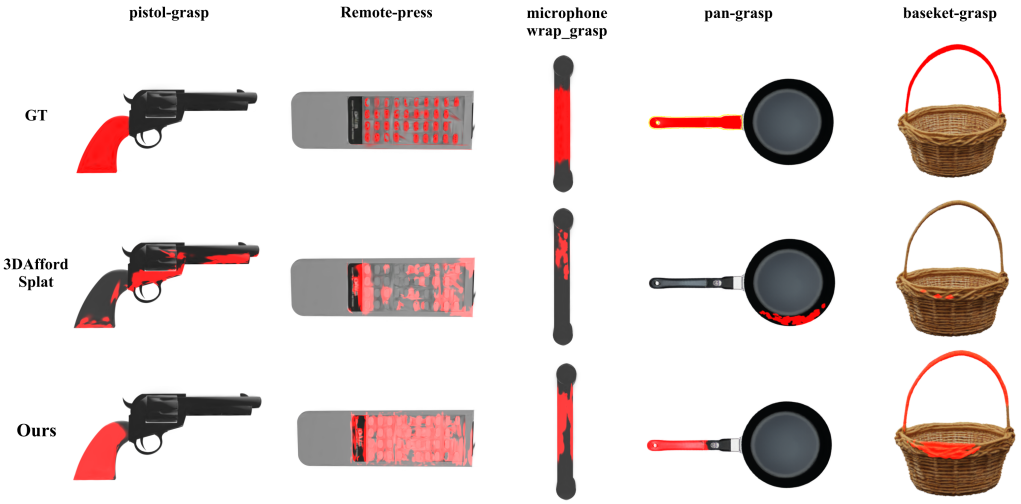

Figure 4: **Qualitative comparison of affordance reasoning on the unseen categories.**

Table 3: Ablation results of the main components in ST-Afford.

Variant	Modification	MAE ↓	SIM ↑	AUC ↑	IoU ↑
A1	No DiffGS + SSGE	0.4493	0.2281	0.5374	0.0918
A2	No level of semantic	0.1694	0.5272	0.8814	0.3926
B1	No IGDF	0.3167	0.2720	0.6542	0.1036
B2	No Gate	0.1958	0.4993	0.8755	0.2690
B3	No Text Bias	0.1925	0.3991	0.8478	0.2626
A0	**Full ST-Afford**	**0.1377**	**0.5675**	**0.8938**	**0.4511**

while removing semantic-level organization also reduces performance, showing that token compression alone is inadequate. These results highlight the importance of semantic distillation, causal organization, and prefix-based supervision in constructing discriminative semantic-geometric priors. For IGDF, removing the full fusion module consistently degrades localization, while the token-gating and spatial-conditioning ablations show complementary roles: the former selects interaction-relevant shape priors and suppresses irrelevant structures, whereas the latter preserves Gaussian-level geometric differences for local prediction. Together, SSGE and IGDF enable effective grounding from high-level interaction intent to fine-grained functional regions.

5 Conclusion

We present ST-Afford, a 3D affordance reasoning framework for 3D Gaussian Splatting (3DGS). To address the loss of local topological information and limited cross-category generalization that can occur when existing methods directly perform dense cross-modal alignment between unstructured Gaussian points and text, we design the model from two aspects: feature representation and cross-modal fusion. By introducing the SSGE, high-dimensional Gaussian features are converted into a fixed number of compact shape-prior tokens. On this basis, the IGDF module uses text instructions to condition local spatial features and shape priors, respectively, and performs fine-grained affordance reasoning through the lightweight PAGH, without requiring an additional heavy spatial decoder. Experimental results show that ST-Afford reduces the forward inference cost while producing more spatially continuous affordance regions and maintaining the correspondence between textual intent and geometric features. On the unseen-category test set, ST-Afford also demonstrates favorable cross-category generalization. The semantically structured geometric representation and text-conditioned dynamic fusion mechanism developed in this work provide a feasible implementation for affordance reasoning in 3D multimodal understanding and physical interaction.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation during the preparation of the manuscript. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, assist in proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean or reformat datasets, perform qualitative or thematic data analysis, or interpret experimental results.

Additionally, we used generative AI tools to assist with improving the readability of the manuscript and to support information retrieval and discovery, including identifying relevant literature and related work. All AI-assisted content was reviewed and verified by the authors. The authors checked the translated and edited text against the original manuscript and verified the relevant technical content and references. The authors take full responsibility for the final content of this work, including all text, claims, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, training settings, datasets, evaluation protocols, and implementation details in Sections 4.1 and Appendix B.

REFERENCES

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 202–221, 2020. doi: 10.1007/978-3-030-58565-5_13. URL https://arxiv.org/abs/1912.08830.

Seokhun Choi, Hyeonseop Song, Jaechul Kim, Taehyeong Kim, and Hoseok Do. Click-gaussian: Interactive segmentation to any 3d gaussians. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/406_ECCV_2024_paper.php.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1778–1787, 2021. URL https://arxiv.org/abs/2103.16397.

Niladri Shekhar Dutt, Zifan Shi, Paul Guerrero, Chun-Hao Paul Huang, Duygu Ceylan, Niloy J. Mitra, and Xuelin Chen. Lost: Level of semantics tokenization for 3d shapes, 2026. URL https://arxiv.org/abs/2603.17995.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 20482–20494, 2023. doi: 10.52202/075280-0900. URL https://arxiv.org/abs/2307.12981.

Umangi Jain, Ashkan Mirzaei, and Igor Gilitschenski. Gaussiancut: Interactive segmentation via graph cut for 3d gaussian splatting. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2830. URL https://arxiv.org/abs/2411.07555.

Mazeyu Ji, Ri-Zhao Qiu, Xueyan Zou, and Xiaolong Wang. Graspsplats: Efficient manipulation with 3d feature splatting. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 1443–1460. PMLR, 2025. URL https://proceedings.mlr.press/v270/ji25a.html.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. doi: 10.1145/3592433. URL https://arxiv.org/abs/2308.04079.

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19729–19739, 2023. URL https://arxiv.org/abs/2303.09553.

Yicong Li, Na Zhao, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-seng Chua. Laso: Language-guided affordance segmentation on 3d object. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14251–14260, 2024. doi: 10.1109/CVPR52733.2024.01351.

Dongyue Lu, Lingdong Kong, Tianxin Huang, and Gim Hee Lee. Geal: Generalizable 3d affordance learning with cross-modal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1680–1690, 2025. URL https://arxiv.org/abs/2412.09511.

Timo Lüddecke and Florentin Wörgötter. Learning to segment affordances. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 769–776, 2017. doi: 10.1109/ICCVW.2017.96.

Aihua Mao, Kaihang Huang, Yong-Jin Liu, Chee Seng Chan, and Ying He. Vagnet: Grounding 3d affordance from human-object interactions in videos. *arXiv preprint arXiv:2602.20608*, 2026. URL https://arxiv.org/abs/2602.20608.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL https://arxiv.org/abs/1812.03828.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Songyou Peng, Kyle Genova, Chiyu Max Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–824, 2023. URL https://arxiv.org/abs/2211.15654.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 652–660, 2017. URL https://arxiv.org/abs/1612.00593.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20051–20060, 2024. URL https://arxiv.org/abs/2312.16084.

Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5197–5206, 2017. doi: 10.1109/CVPR.2017.552.

Olaolu Shorinwa, Johnathan Tucker, Aliyah Smith, Aiden Swann, Timothy Chen, Roya Firoozi, Monroe David Kennedy, and Mac Schwager. Splat-mover: Multi-stage, open-vocabulary robotic manipulation via editable gaussian splatting. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4748–4770. PMLR, 2025. URL https://proceedings.mlr.press/v270/shorinwa25a.html.

Hanqing Wang, Mingyu Liu, Xiaoyu Chen, Chengwei Ma, Yiming Zhong, Wenti Yin, Yuhao Liu, Zhiqing Cui, Jiahao Yuan, Lu Dai, Zhiyuan Ma, and Hui Xiong. Videoafford: Grounding 3d affordance from human-object-interaction videos via multimodal large language model. *arXiv preprint arXiv:2602.09638*, 2026. URL https://arxiv.org/abs/2602.09638.

Zeming Wei, Junyi Lin, Yang Liu, Weixing Chen, Jingzhou Luo, Guanbin Li, and Liang Lin. 3daffordsplat: Efficient affordance reasoning with 3d gaussians. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2821–2830, 2025. doi: 10.1145/3746027.3754778. URL https://arxiv.org/abs/2504.11218.

Chao Xu, Yixin Chen, He Wang, Song-Chun Zhu, Yixin Zhu, and Siyuan Huang. Partafford: Part-level affordance discovery from 3d objects. *arXiv preprint arXiv:2202.13519*, 2022. URL https://arxiv.org/abs/2202.13519.

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. URL https://arxiv.org/abs/2303.10437.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. doi: 10.1007/978-3-031-73397-0_10. URL https://arxiv.org/abs/2312.00732.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling, 2022. URL https://arxiv.org/abs/2111.14819.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21676–21685, 2024. URL https://arxiv.org/abs/2312.03203.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023. URL https://proceedings.mlr.press/v229/zitkovich23a.html.



Figure 5: **Affordance reasoning across the seen-category dataset.** Representative predictions covering the object categories and affordance types used in the seen-category evaluation.

A Additional Seen-Category Visualization

Figure 5 provides an overview of affordance predictions across the 21 object categories and 18 affordance types in the seen-category dataset. The visualization complements the aggregate quantitative results in the main paper by illustrating the diversity of object geometries and interaction regions covered by the benchmark.

B Implementation and Training Details

B.1 Semantic Pre-training Data and Architecture

Semantic pre-training uses pre-computed DiffGS tri-plane features with resolution $768 \times 64 \times 64$. The pre-training corpus contains 23,142 objects from 21 categories. Each object is associated with 19 rendered RGB views at resolution 518×518, from which one view is sampled during training. We use a frozen DINOv2-Small encoder as the visual teacher. Its CLS and patch features provide object-level and spatial semantic supervision, respectively.

For RIDA, the three feature planes are first processed independently by lightweight plane-wise stems and then converted into spatial tokens through patch embedding. A 12-layer Transformer encoder contextualizes these tokens, and a learnable attention-pooling query produces the global semantic representation h_{sem}. For SSGE, the tri-plane representation is patch-embedded into 256 tri-plane tokens and jointly processed with 512 learnable register tokens. The asymmetric causal mask described in the main paper is applied throughout the Transformer. The final register features are compressed to 128 dimensions to form the ordered shape-prior tokens. During SSGE pre-training, the prefix length is randomly sampled from $[16, 512]$.

B.2 Detailed Semantic Pre-training Objectives

RIDA semantic alignment. Let s_i denote the RIDA global feature of the i-th object and t_i the projected DINOv2 global feature from a rendered view of the same object. After ℓ_2 normalization, we use a symmetric InfoNCE objective

$$\mathcal{L}_{\text{global}} = \frac{1}{2}\left[\text{CE}\left(\frac{\bar{S}\bar{T}^\top}{\tau}, \mathbf{y}\right) + \text{CE}\left(\frac{\bar{T}\bar{S}^\top}{\tau}, \mathbf{y}\right)\right], \tag{11}$$

where $\bar{S}$ and $\bar{T}$ stack the normalized student and teacher features in a mini-batch, $\mathbf{y}$ denotes the identity matching labels, and $\tau = 0.07$.

To preserve inter-object semantic relations, we additionally match the normalized pairwise distance matrices:

$$\mathcal{L}_{\text{rank}} = \frac{1}{B^2}\left\|\frac{D_s}{\sigma_s} - \frac{D_t}{\sigma_t}\right\|_F^2, \tag{12}$$

where D_s and D_t are the pairwise Euclidean-distance matrices of the student and teacher features, and σ_s and σ_t denote their corresponding scale factors.

We further preserve local semantic structure through spatial affinity distillation. For spatially aligned student and teacher token sets, their self-affinity matrices are denoted by A_s and A_t, respectively, and

$$\mathcal{L}_{\text{spatial}} = \|A_s - A_t\|_F^2. \tag{13}$$

The complete RIDA objective is

$$\mathcal{L}_{\text{RIDA}} = \mathcal{L}_{\text{global}} + \mathcal{L}_{\text{rank}} + 0.5\,\mathcal{L}_{\text{spatial}}. \tag{14}$$

SSGE ordered-token pre-training. After RIDA is trained, its parameters are frozen and its semantic output h_{sem} is used as the target for SSGE. For a randomly sampled prefix $Z_{1:k}$, we average the retained tokens and map the result through the lightweight two-layer projection head $D(\cdot)$:

$$\tilde{h}_k = D\left(\frac{1}{k}\sum_{i=1}^{k} z_i\right). \tag{15}$$

The semantic alignment and feature reconstruction terms are

$$\mathcal{L}_{\text{sem}} = 1 - \cos\left(\tilde{h}_k, h_{\text{sem}}\right), \qquad \mathcal{L}_{\text{rec}} = \left\|\tilde{h}_k - h_{\text{sem}}\right\|_2^2. \tag{16}$$

To make different prefix lengths preserve consistent semantics, two prefix lengths k_1 and k_2 are independently sampled for each object. Let

$$g_i^{(k)} = \text{Norm}\left(\frac{1}{k}\sum_{m=1}^{k} z_{i,m}\right). \tag{17}$$

The prefix-consistency objective is

$$\mathcal{L}_{\text{prefix}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left({g_i^{(k_1)}}^\top g_i^{(k_2)}/\tau_p\right)}{\sum_{j=1}^{B}\exp\left({g_i^{(k_1)}}^\top g_j^{(k_2)}/\tau_p\right)}, \tag{18}$$

with $\tau_p = 0.1$. The SSGE pre-training objective is

$$\mathcal{L}_{\text{SSGE}} = \mathcal{L}_{\text{sem}} + 0.3\,\mathcal{L}_{\text{rec}} + 0.1\,\mathcal{L}_{\text{prefix}}. \tag{19}$$

Unlike the original generative formulation from which the tokenizer design is adapted, our downstream task only requires the structured encoder. We therefore use a two-layer MLP, $128 \to 768 \to 768$, as the semantic projection head during pre-training and discard it afterwards.

B.3 Downstream Affordance Objectives

For downstream affordance reasoning, the total objective is

$$\mathcal{L}_{\text{aff}} = \mathcal{L}_{\text{focal}} + \mathcal{L}_{\text{dice}} + \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}}, \tag{20}$$

where $\lambda_{\text{text}} = 1$ and $\lambda_{\text{cons}} = 0.1$.

We use a masked Focal Loss to alleviate the foreground–background imbalance:

$$\mathcal{L}_{\text{focal}} = -\frac{1}{|M|}\sum_{i\in M}\left[\alpha(1-\hat{y}_i)^{\gamma} y_i \log(\hat{y}_i+\epsilon) + (1-\alpha)\hat{y}_i^{\gamma}(1-y_i)\log(1-\hat{y}_i+\epsilon)\right], \tag{21}$$

with $\alpha = 0.25$ and $\gamma = 2$. Dice loss is computed for both foreground and background:

$$\mathcal{L}_{\text{dice}} = 2 - \text{Dice}^{+} - \text{Dice}^{-}. \tag{22}$$

For text–geometry alignment, the ground-truth affordance mask is used to pool the point-aligned fused features:

$$f_{\text{geo}} = \frac{\sum_{i\in M} y_i F_{\text{fused},i}}{\sum_{i\in M} y_i + \epsilon}. \tag{23}$$

We then contrast the pooled geometric feature with the corresponding interaction text:

$$\mathcal{L}_{\text{text}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\text{sim}(f^i_{\text{geo}}, e^i_{\text{aff}})/\kappa\right)}{\sum_{j=1}^{B}\exp\left(\text{sim}(f^i_{\text{geo}}, e^j_{\text{aff}})/\kappa\right)}, \tag{24}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and κ is the contrastive temperature.

Finally, for two objects A and B associated with the same interaction intent, cross-object consistency is imposed in the ordered token space:

$$\mathcal{L}_{\text{cons}} = \frac{1}{K}\sum_{k=1}^{K}\left[1 - \cos\left(z^A_k, z^B_k\right)\right]. \tag{25}$$

This regularizer does not assume explicit part-to-part correspondence; it encourages objects sharing an interaction intent to maintain compatible representations in the common ordered token space.

B.4 Optimization and Implementation Settings

RIDA is pre-trained for 100 epochs on one NVIDIA A100 GPU using AdamW with learning rate 10^{-4} and weight decay 0.04. A cosine learning-rate schedule is used, together with mixed-precision training and gradient clipping with maximum norm 1.0. SSGE is then trained for 200 epochs on two NVIDIA A100 GPUs with AdamW, the same base learning rate and weight decay, a 10-epoch warm-up, and cosine annealing. RIDA remains frozen throughout SSGE pre-training.

For downstream affordance training, we use two NVIDIA A100 GPUs. Each scene is truncated to at most 32,768 Gaussian primitives. The training batch size is 16 with two-step gradient accumulation, and the validation batch size is 4. The downstream model is trained for 200 epochs with a base learning rate of 10^{-4} and weight decay 5×10^{-4}. The SSGE encoder is updated with a reduced learning rate of 5×10^{-5} to preserve the pre-trained structural prior. A StepLR schedule with step size 60 and decay factor 0.7 is used. All experiments use random seed 42.

During downstream fine-tuning, LoRA adapters are applied to the query and value projection layers in IGDF. We use rank 32, scaling factor 32, and dropout 0.1. The complete downstream training takes approximately 29.15 hours under this configuration.

Table 4: Main optimization settings for semantic pre-training and downstream affordance training.

Setting	RIDA	SSGE	Downstream
Training epochs	100	200	200
GPUs	1×A100	2×A100	2×A100
Optimizer	AdamW	AdamW	AdamW
Base learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Encoder learning rate	–	–	5×10^{-5}
Weight decay	0.04	0.04	5×10^{-4}
LR schedule	Cosine	Warm-up + Cosine	StepLR
Gradient clipping	1.0	1.0	–
Prefix range	–	$[16, 512]$	–
Gradient accumulation	–	–	2
Max. Gaussian primitives	–	–	32,768

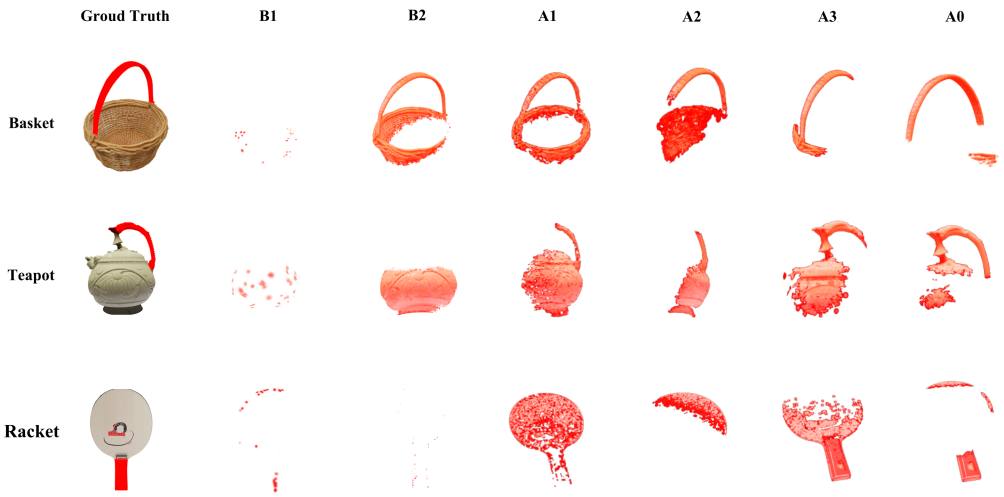

Figure 6: **Qualitative visualization of the ablation study.** The visualization highlights differences in affordance responses among the full model and the ablated variants.

C Additional Ablation Details

Figure 6 provides additional qualitative results for the ablation variants. For cases in which weak affordance responses are difficult to distinguish on top of the rendered object, the object surface is omitted in the visualization to make differences in the predicted response maps easier to inspect.

The ablation variants reported in Table 3 are defined as follows:

- **A0 (Full Model):** the complete ST-Afford model.
- **A1 (w/o DiffGS + SSGE):** removes the structured representation branch and directly fuses low-level Gaussian point features with text through cross-attention, approximating dense primitive-level cross-modal reasoning.
- **A2 (w/o Semantic Ordering):** retains the SSGE architecture but removes random-prefix truncation and prefix-consistency supervision during pre-training. The resulting tokens therefore no longer receive explicit supervision to organize information according to semantic importance.
- **B1 (w/o IGDF):** removes text-conditioned fusion, including token gating and the textual term in point-query construction, and predicts using geometric information only.
- **B2 (w/o Gate):** retains the spatially conditioned query construction but removes text-conditioned gating over the shape-prior tokens.
- **B3 (w/o Text Bias):** retains shape-prior token gating but removes the textual term from point-query construction.

Table 5: Per-affordance comparison on the seen-category test set.

Affordance	MAE ↓				SIM ↑				AUC ↑				IoU ↑			
	IAGNet	LASO	3DAffordSplat	ST-Afford	IAGNet	LASO	3DAffordSplat	ST-Afford	IAGNet	LASO	3DAffordSplat	ST-Afford	IAGNet	LASO	3DAffordSplat	ST-Afford
lay	0.2478	0.3559	**0.2364**	0.2720	0.1491	0.3604	**0.4157**	0.3861	0.5146	0.6914	**0.7610**	0.6813	0.0336	**0.2545**	0.1905	0.2394
sit	0.2001	0.2449	0.2029	**0.1676**	0.1112	**0.4152**	0.4113	0.3999	0.4305	0.8471	**0.8624**	0.8241	0.0162	**0.3698**	0.2968	0.2983
support	0.3787	0.2527	0.2271	**0.1969**	0.0748	0.5281	**0.5388**	0.5178	0.3450	**0.8685**	0.8629	0.8488	0.0201	**0.4458**	0.3749	0.3568
grasp	0.3515	0.2786	0.2348	**0.1106**	0.3344	0.5334	0.5163	**0.7618**	0.6068	0.8109	0.8897	**0.9422**	0.1016	0.3512	0.3390	**0.6756**
lift	0.1011	0.0926	0.1336	**0.0396**	0.1087	0.4079	0.3106	**0.5703**	0.6183	0.8257	**0.9383**	0.7938	0.0075	0.3611	0.2564	**0.5052**
contain	0.2084	0.2793	0.2073	**0.1059**	0.1268	0.2325	0.4035	**0.6087**	0.5115	0.4083	0.8829	**0.9278**	0.0198	0.1586	0.2789	**0.4848**
open	0.1103	0.1066	0.1210	**0.0728**	0.0604	0.2942	0.2242	**0.3884**	0.4301	0.8508	0.7891	**0.8717**	0.0137	0.2044	0.1259	**0.2736**
wrap_grasp	0.3380	0.3415	0.2642	**0.1796**	0.3089	0.3601	0.5869	**0.6799**	0.5390	0.7056	0.8977	**0.9157**	0.0605	0.1367	0.4142	**0.5475**
pour	0.2163	0.1549	0.1558	**0.1236**	0.1050	0.4497	0.4502	**0.5186**	0.4265	0.9105	0.9068	**0.9214**	0.0192	0.3691	0.3603	**0.4242**
move	0.3758	0.2235	0.2132	**0.1459**	0.2208	0.6300	0.6101	**0.6902**	0.4848	0.9034	0.8995	**0.9123**	0.0450	0.5170	0.4947	**0.5851**
display	0.2147	0.1742	0.2623	**0.1735**	0.1599	0.3329	0.2799	**0.3555**	0.5660	0.7497	0.7774	**0.7915**	0.0544	0.1622	0.1543	**0.2388**
push	0.2032	0.0661	**0.0620**	0.0824	0.0357	0.2233	0.2948	**0.3024**	0.4001	0.7836	0.6772	**0.8963**	0.0146	0.0045	0.1765	**0.1992**
pull	-	0.0109	**0.0085**	0.0499	-	0.0285	**0.4782**	0.3824	-	0.6882	**0.9604**	0.9529	-	0.0006	**0.3519**	0.2529
listen	0.3968	**0.3041**	0.4586	0.3160	0.1261	**0.5734**	0.3564	0.5446	0.4932	**0.8121**	0.4670	0.7860	0.0357	**0.2926**	0.1336	0.2725
wear	0.7050	0.3131	0.5000	**0.1033**	0.4788	0.7598	0.6752	**0.8807**	0.6206	0.8890	0.4678	**0.9644**	0.0487	0.5809	0.3845	**0.8130**
press	0.2686	0.2582	0.2960	**0.1900**	0.1885	0.4434	0.3144	**0.5416**	0.5107	0.8177	0.6620	**0.8844**	0.0163	0.3400	0.1771	**0.3830**
cut	0.4487	0.2194	0.2207	**0.0696**	0.3431	0.6987	0.7169	**0.8846**	0.5044	**0.9967**	0.9961	0.9966	0.0358	0.5027	0.5296	**0.8060**
stab	0.2691	0.1212	0.1225	**0.0506**	0.0543	0.5527	0.5591	**0.7167**	0.1790	0.9208	0.9812	**0.9862**	0.0184	0.5041	0.4392	**0.6259**

Table 6: Per-object-category comparison on the seen-category test set.

Object	MAE ↓				SIM ↑				AUC ↑				IoU ↑			
	IAGNet	LASO	3DAffordSplat	ST-Afford	IAGNet	LASO	3DAffordSplat	ST-Afford	IAGNet	LASO	3DAffordSplat	ST-Afford	IAGNet	LASO	3DAffordSplat	ST-Afford
bag	0.1493	0.1432	0.1686	**0.0630**	0.1190	0.3405	0.3130	**0.5581**	0.5342	0.6756	**0.8968**	0.8697	0.0455	0.2461	0.2154	**0.4626**
bed	0.2280	0.3514	0.2256	**0.2023**	0.1409	0.2860	**0.3379**	0.3173	0.5262	0.7114	**0.7724**	0.7198	0.0232	**0.2195**	0.1754	0.1886
bottle	0.1542	0.1370	0.1266	**0.0775**	0.0997	0.3236	0.3009	**0.5254**	0.4989	0.7674	0.8524	**0.9375**	0.0163	0.2236	0.1587	**0.3993**
bowl	0.2985	0.4053	0.2947	**0.2108**	0.2026	0.3096	0.4506	**0.4864**	0.4728	0.5639	0.8137	**0.8402**	0.0446	0.1618	0.3377	**0.3860**
chair	0.2908	0.1775	0.1864	**0.1498**	0.1614	**0.6443**	0.6120	0.6347	0.4203	**0.9420**	0.9191	0.9279	0.0274	**0.5408**	0.4746	0.5147
clock	0.1823	0.1891	0.2870	**0.1563**	0.0862	0.1549	**0.2018**	0.1884	0.4905	0.6514	**0.7395**	0.6768	0.0100	0.0214	0.0976	**0.1138**
dishwasher	0.1473	0.0320	0.0892	**0.0279**	0.1357	**0.6409**	0.4493	0.6301	0.4407	**0.9745**	0.9470	0.9660	0.0070	0.5330	0.3455	**0.5337**
display	0.3254	**0.2661**	0.3292	0.3464	0.1698	0.2846	**0.3269**	**0.3269**	0.4901	0.6509	0.6633	**0.7060**	0.0647	0.0868	0.1329	**0.1637**
door	0.1606	0.0295	**0.0262**	0.0602	0.0193	0.0860	**0.4206**	0.3607	0.3786	0.7043	0.8676	**0.9337**	0.0079	0.0019	**0.3021**	0.2411
earphone	0.4564	0.2931	0.4258	**0.2603**	0.3153	0.6515	0.4587	**0.6628**	0.5581	**0.8529**	0.6308	0.8387	0.1064	0.4387	0.2401	**0.4715**
faucet	0.3603	0.3028	0.2432	**0.2039**	0.2556	0.4658	0.5430	**0.5629**	0.5221	0.7471	**0.8158**	0.8140	0.0521	0.2020	0.3596	**0.4703**
hat	0.4596	0.2805	0.3640	**0.0950**	0.3818	0.5962	0.5733	**0.8091**	0.6301	0.8554	0.6927	**0.9723**	0.1004	0.4383	0.3410	**0.7138**
keyboard	0.4697	0.4041	0.3628	**0.3111**	0.4083	0.5389	0.5782	**0.6078**	0.5987	0.7087	0.7785	**0.8164**	0.0312	0.3902	0.3770	**0.4336**
knife	0.3464	0.2128	0.1976	**0.0753**	0.2590	0.5731	0.6043	**0.7858**	0.4684	0.9010	0.9330	**0.9887**	0.0284	0.4432	0.4219	**0.7062**
laptop	0.1467	0.1182	0.2130	**0.0947**	0.1527	0.4823	0.2434	**0.5254**	0.6007	0.9149	0.7762	**0.9453**	0.0517	0.3482	0.1548	**0.3881**
microwave	0.1299	0.1406	0.1458	**0.0918**	0.1871	0.3882	0.3204	**0.5116**	0.6249	0.9037	0.8127	**0.9188**	0.0091	0.2057	0.2125	**0.3579**
mug	0.2115	0.3009	0.1751	**0.0628**	0.1446	0.2519	0.4892	**0.7852**	0.4481	0.4679	0.9393	**0.9857**	0.0207	0.2121	0.3763	**0.6832**
storagefurniture	0.3229	0.2814	0.2779	**0.2119**	0.1158	0.4163	0.4158	**0.4936**	0.3658	0.7435	0.7693	**0.7840**	0.0226	0.2554	0.2084	**0.3551**
table	0.3727	0.2780	0.2656	**0.2145**	0.1470	0.5225	0.5328	**0.5536**	0.4240	0.8566	**0.8604**	0.8340	0.0368	**0.4496**	0.3958	0.4095
trashcan	0.1600	0.2021	0.1740	**0.1080**	0.0549	0.2097	0.2718	**0.3923**	0.4248	0.6986	0.8482	**0.8917**	0.0108	0.1443	0.1967	**0.2887**
vase	0.3476	0.2506	0.1926	**0.1168**	0.2790	0.3726	0.5550	**0.6915**	0.5775	0.6442	0.9572	**0.9655**	0.0732	0.2721	0.4499	**0.5744**

D PER-AFFORDANCE AND PER-CATEGORY RESULTS

Tables 5 and 6 provide a more detailed breakdown of the seen-category results. At the affordance level, ST-Afford shows particularly strong performance on interactions such as *grasp*, *wear*, *cut*, and *stab*, where successful prediction depends on identifying localized functional structures rather than only the overall object extent. Performance is less dominant on several categories such as *lay*, *sit*, *support*, and *pull*, indicating that broad support regions or sparsely represented interaction types remain more challenging.

At the object-category level, ST-Afford performs strongly across objects with substantially different scales and shapes, including compact objects such as mugs, knives, bottles, and hats, as well as larger articulated or container-like objects. The model is particularly effective on categories with localized and geometrically distinctive functional components. At the same time, categories such as doors, beds, chairs, and displays retain metric-specific gaps to the strongest comparison method, suggesting that large or spatially diffuse functional regions remain an important direction for further improvement.