Geo-DPO: Aligning Semantic Intent with Geometry for 3D Affordance Segmentation

Ziqian Yang[1,2], Xiaolei Wang[1,2], Xianglin Qiu[1,2], Weiguang Zhao[1,2], Quan Zhang[1†], and Jimin Xiao[1†]

[1] Xi'an Jiaotong-Liverpool University, China
[2] University of Liverpool, UK

Abstract. 3D sequential affordance reasoning requires a model to locate a logical series of functional parts based on language instructions. Current Multimodal Large Language Models (MLLMs) tackle this by generating a special `<SEG>` token to guide the final mask prediction. However, compressing complex reasoning into a single token loses crucial low-level structural details. This leads to severe geometric ambiguity, causing the predicted masks to have inaccurate boundaries. To solve this, we propose the Geo-DPO framework, which addresses the ambiguity from two directions. First, we introduce the Hierarchical Geometry Adapter (HGA) to structurally inject multi-scale 3D physical features back into the `<SEG>` token. Second, we propose Contrastive Affordance Preference Optimization (CAPO), a reward-based training objective that forces the segmentation to align with real geometric boundaries. Extensive experiments on the standard SeqAfford dataset demonstrate that our method achieves state-of-the-art performance, generating highly precise masks and effectively resolving geometric ambiguity. Code is available here.

Keywords: Affordance Segmentation · Point Cloud · Multimodal Large Language Model

1 Introduction

3D affordance segmentation aims to identify the actionable and functional regions of objects within a 3D scene. To enable more flexible and intuitive interaction, language-based 3D affordance segmentation [14,33] has recently been introduced. Given a 3D point cloud and a natural language instruction or query, the goal of this task is to predict a point-wise binary mask that precisely highlights the specific area where the text-described action can be performed. This fine-grained, text-driven localization is crucial for robotic manipulation [13,29,45,46], as it directly bridges high-level human commands with the physical object parts required for task execution.

Recently, with their strong reasoning abilities and understanding of 3D objects, 3D Multimodal Large Language Models (3D MLLMs) are being applied to

†: Corresponding authors

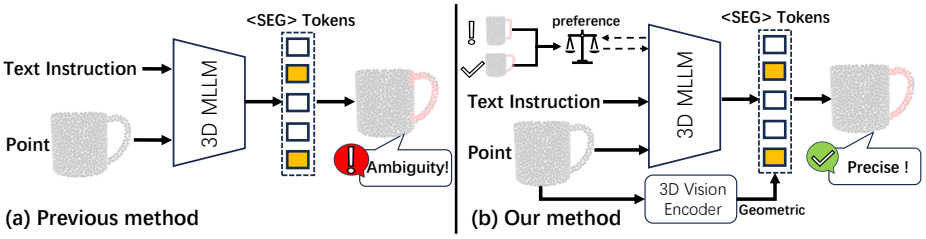

Fig. 1: Our motivation. (a) Previous method: Relying solely on compressed semantic <SEG> tokens inherently leads to severe geometric ambiguity (e.g., mask overflow). **(b) Our method**: By injecting fine-grained geometric features from a 3D vision encoder and employing preference optimization, our Geo-DPO successfully resolves this ambiguity to produce precise affordance masks.

3D affordance tasks. Ingrained with internalized common-sense knowledge encoded from vast text data corpora, these models can naturally perform sequential affordance reasoning. Unlike single-step tasks, sequential affordance requires the model [6, 22] to deduce a logical series of functional parts to complete a complex instruction. For example, to "pour water from the watering can", the model should first locate the "handle" to lift it, and then the "spout" to pour. To achieve this, current 3D MLLM frameworks usually generate a special <SEG> token at the end of the text response. This token aggregates the semantic and target part information to prompt a downstream decoder for mask prediction. However, this operation creates a critical problem. While the <SEG> token contains high-level semantics, this heavy compression loses the essential low-level structural details and precise spatial coordinates of the 3D scene.

Consequently, the loss of these essential low-level details leads to severely inaccurate segmentation. As illustrated in Figure 1 (a), relying solely on the heavily compressed <SEG> token causes current models to struggle with precise physical boundaries. Rather than cleanly isolating the target part, the predicted masks frequently bleed into spatially adjacent but functionally distinct regions. For instance, when attempting to locate the "handle" of a mug, the baseline prediction often spills over onto the main cup body. We formally refer to this phenomenon of boundary confusion and spatial imprecision as *geometric ambiguity*. This geometric ambiguity fundamentally limits the reliability of current MLLM-based frameworks, as they fail to capture the sharp, fine-grained geometries required for real-world 3D affordance tasks.

To address this limitation, we propose the Geometric Direct Preference Optimization (Geo-DPO) framework, designed to explicitly align the semantic intent of the <SEG> token with the actual 3D geometry. Specifically, we resolve the geometric ambiguity from two complementary directions: structural design and training objective. First, from a structural perspective, we introduce the Hierarchical Geometry Adapter (HGA). HGA extracts multi-scale dense features from the intermediate layers of the 3D encoder and injects them back into the <SEG> token. This operation restores the missing low-level structural details, equipping

the semantic query with accurate spatial awareness. Second, from a training perspective, we propose Contrastive Affordance Preference Optimization (CAPO). Unlike standard point-wise losses that evaluate points in isolation, CAPO formulates 3D affordance segmentation as a preference ranking problem. By utilizing a hybrid reward, it strictly penalizes predictions that overflow into geometrically distinct regions. This forces the model to respect fine-grained physical boundaries. Together, these two modules effectively eliminate geometric ambiguity. As contrasted in Figure 1 (b), this enables Geo-DPO to generate highly precise 3D affordance masks that respect the actual geometric boundaries, fixing the bleeding issue.

In summary, the main contributions of our work are threefold:

- We identify the *geometric ambiguity* issue caused by token compression in current MLLM-based pipelines. To solve this, we propose the **Geo-DPO** framework, which formally aligns semantic intent with 3D geometry for precise 3D affordance segmentation.
- We design two complementary modules to resolve this ambiguity: the **Hierarchical Geometry Adapter (HGA)** structurally restores missing low-level physical details, while **Contrastive Affordance Preference Optimization (CAPO)** utilizes a hybrid reward to penalize mask overflow and enforce boundary alignment.
- Extensive experiments on the standard SeqAfford dataset demonstrate that our Geo-DPO framework achieves state-of-the-art performance. It effectively eliminates boundary confusion, yielding highly accurate segmentation masks that respect real physical geometries.

2 Related Work

3D Affordance Segmentation. Early 3D affordance tasks [4,6,26] primarily relied on predefined categories, predicting fixed functional labels for 3D point clouds. To support open-vocabulary and flexible instructions, language-based 3D affordance segmentation was recently introduced. Current methods [19,30,37] encode both the 3D scene and the text query to predict a dense probability mask. Notably, SeqAfford [56] pioneered sequential affordance reasoning, enabling models to localize a logical series of functional parts for complex tasks. However, these existing approaches focus heavily on aligning high-level semantics while neglecting the underlying 3D geometry. Trained with isolated point-wise losses (e.g., BCE), they struggle to capture sharp part contours, often suffering from geometric ambiguity. In contrast, our Geo-DPO forces the predicted masks to align with real geometric boundaries.

3D Multimodal Large Language Models. Driven by the success of 2D vision-language models [1, 21, 27, 58], 3D MLLMs (e.g., 3D-LLM [9], PointLLM [51]) have been developed to bridge 3D point clouds with LLMs. These models [7, 42] demonstrate strong capabilities in 3D scene understanding [15], reasoning, and

question answering. To adapt these text-generation models for dense segmentation tasks, current frameworks [3, 56] typically generate a special <SEG> token at the end of the text response. This token aggregates the reasoning results to query a downstream decoder. While highly effective for semantic reasoning, this heavy token compression loses crucial dense spatial coordinates and structural details. To solve this, our Hierarchical Geometry Adapter (HGA) structurally injects multi-scale features from the 3D encoder back into the <SEG> token, restoring its spatial awareness.

Direct Preference Optimization. Direct Preference Optimization (DPO) [34] was originally proposed in natural language processing to align LLMs [28] with human preferences without complex reinforcement learning. Recently, preference learning has been extended to visual domains [47, 48, 59], primarily to improve image generation quality [39] or 2D visual-text alignment [17, 53, 54]. However, exploring preference optimization for dense 3D segmentation remains largely untouched. In this work, we introduce CAPO to adapt the DPO paradigm for 3D affordance segmentation. Instead of aligning with human subjective preferences, we use the actual 3D geometry as the preference target. By utilizing a contrastive shape reward, we penalize predictions that overflow into incorrect regions, effectively resolving geometric ambiguity from a training objective perspective.

3 Methodology

3.1 Preliminaries

Given a 3D point cloud $\mathcal{X} \in \mathbb{R}^{N\times 3}$ and a natural language instruction $\mathcal{T}$, our primary objective is to predict a point-wise binary segmentation mask $\mathbf{P} \in [0,1]^N$ that localizes the target affordance region.

Following the paradigm of SeqAfford [56], we employ a Multimodal Large Language Model (MLLM) [40, 55] as the core reasoning engine. Specifically, the raw point cloud is first encoded into a sequence of 3D visual tokens and projected into the identical embedding space as the text tokens of $\mathcal{T}$. The MLLM processes this interleaved multimodal input to perform complex reasoning regarding the 3D scene and the sequential action requirements [36, 49, 50].

Upon comprehending the semantic intent, the MLLM auto-regressively generates a text response that culminates in a special <SEG> token. The final hidden state of this token, denoted as $\mathbf{h}_{\text{raw}} \in \mathbb{R}^{1\times D}$ (where D represents the hidden dimension of the MLLM), encapsulates the macroscopic concept of the target object part (e.g., "the handle of the mug"). We extract $\mathbf{h}_{\text{raw}}$ to serve as the initial semantic query for the subsequent fine-grained segmentation module.

3.2 Overview

As illustrated in Figure 2, we propose a Geometric Direct Preference Optimization (Geo-DPO) framework designed to bridge the granularity gap between high-level semantic planning and precise physical grounding. Building upon a 3D

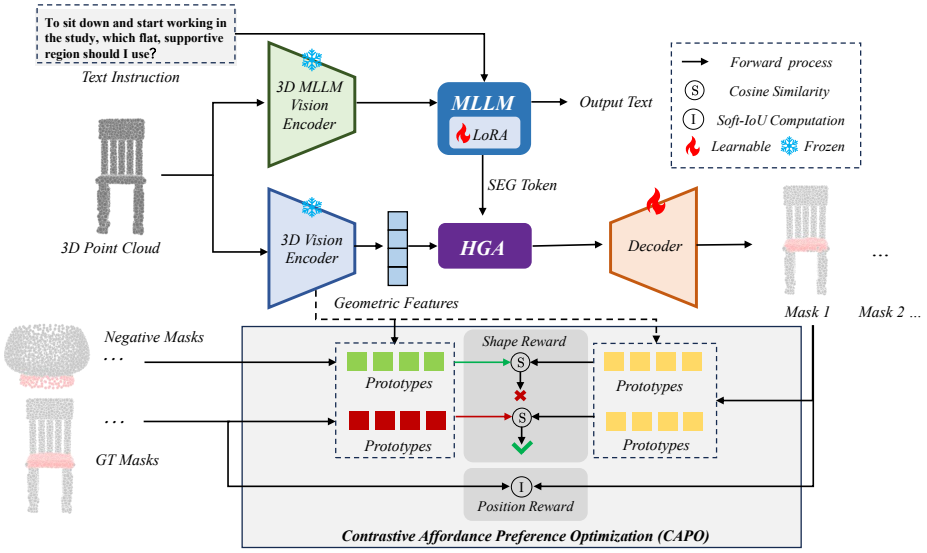

Fig. 2: Overview of the Geo-DPO framework. Given a 3D point cloud and a text instruction, the MLLM processes the multi-modal inputs to auto-regressively generate an output text culminating in a semantic `<SEG>` token. To alleviate geometric ambiguity, the **Hierarchical Geometry Adapter (HGA)** integrates multi-scale geometric features from the frozen 3D vision encoder into the token. During training, the **Contrastive Affordance Preference Optimization (CAPO)** module regularizes the decoder's predictions. CAPO computes a Position Reward (via Soft-IoU) and a Shape Reward (via cosine similarity between geometric prototypes) by contrastively supervising the predicted masks between Ground Truth (GT) and Negative Masks, thereby penalizing geometric overflow and ensuring precise physical alignment.

MLLM backbone, we first introduce the **Hierarchical Geometry Adapter (HGA)**, which acts as a cross-modal bridge to recover fine-grained structural details lost in the semantic space. To further enforce physical consistency, we propose **Contrastive Affordance Preference Optimization (CAPO)**, a novel training paradigm that functions as a physical critic to align predicted affordances with geometric reality through preference ranking.

3.3 Hierarchical Geometry Adapter (HGA)

While the MLLM exhibits strong reasoning capabilities to comprehend sequential instructions and isolate distinct action steps, the generated `<SEG>` token primarily operates in a highly compressed semantic space. In the standard SeqAfford [56] pipeline, this `<SEG>` token neglects low-level visual structures and precise positional coordinates. This semantic over-compression usually leads to geometric ambiguity, resulting in coarse and imprecise segmentation masks that fail to delineate sharp object boundaries.

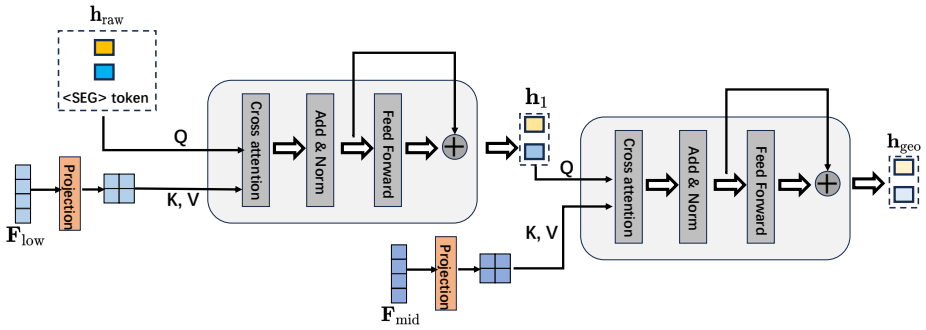

Fig. 3: Architecture of the Hierarchical Geometry Adapter (HGA). The HGA employs a cascaded cross-attention mechanism to progressively inject multi-scale geometric priors into the highly compressed semantic space. Specifically, the raw semantic `<SEG>` token ($\mathbf{h}_{\text{raw}}$) first serves as the query to attend to the projected shallow features ($\mathbf{F}_{\text{low}}$), capturing fine-grained local structures to yield the intermediate token $\mathbf{h}_1$. Subsequently, $\mathbf{h}_1$ queries the projected intermediate features ($\mathbf{F}_{\text{mid}}$) to incorporate broader part-level geometries. This hierarchical injection culminates in the final geometry-aware token ($\mathbf{h}_{\text{geo}}$), successfully bridging the semantic-geometric gap.

To alleviate this limitation, we first design the Hierarchical Geometry Adapter (HGA), as shown in Figure 3. Previous studies in 3D representation learning [18] have demonstrated that the intermediate layers of 3D encoders inherently excel at capturing local geometric patterns, such as contours, edges, and curvatures. Leveraging this property, HGA is formulated as a cross-modal framework [31,41] designed to efficiently inject these multi-scale geometric priors back into the semantic `<SEG>` token, equipping it with accurate structural awareness.

Specifically, let $\mathbf{F}_{\text{low}} \in \mathbb{R}^{N_1 \times C_1}$ and $\mathbf{F}_{\text{mid}} \in \mathbb{R}^{N_2 \times C_2}$ denote the shallow and intermediate geometric features extracted from the frozen 3D visual encoder, respectively. To align the feature spaces, we first apply linear projections to map these geometric features to the MLLM's hidden dimension D. The refinement process is formulated as a cascaded cross-attention mechanism, where the `<SEG>` token iteratively queries the physical features:

$$\mathbf{h}_1 = \mathbf{h}_{\text{raw}} + \text{Attention}(\mathbf{Q} = \mathbf{h}_{\text{raw}}, \mathbf{K} = \mathbf{F}_{\text{low}}, \mathbf{V} = \mathbf{F}_{\text{low}}), \tag{1}$$

$$\mathbf{h}_{\text{geo}} = \mathbf{h}_1 + \text{Attention}(\mathbf{Q} = \mathbf{h}_1, \mathbf{K} = \mathbf{F}_{\text{mid}}, \mathbf{V} = \mathbf{F}_{\text{mid}}), \tag{2}$$

In this formulation, the raw semantic query $\mathbf{h}_{\text{raw}}$ first aggregates fine-grained contour and edge information from the dense shallow layer $\mathbf{F}_{\text{low}}$. Subsequently, the updated query $\mathbf{h}_1$ attends to the intermediate layer $\mathbf{F}_{\text{mid}}$ to incorporate broader local part structures, culminating in the refined token $\mathbf{h}_{\text{geo}} \in \mathbb{R}^{1 \times D}$.

Through this hierarchical injection, the resulting $\mathbf{h}_{\text{geo}}$ transcends a purely abstract semantic concept; it becomes grounded in the geometric reality of the 3D scene. Finally, this geometry-aware token acts as the conditional query in the decoder. It interacts with the high-resolution dense point features $\mathbf{F}_{\text{point}} \in \mathbb{R}^{N \times D}$ via a dot-product operation to generate the precise affordance prediction mask

$\mathbf{P} \in [0,1]^N$:

$$\mathbf{P} = \sigma(\mathbf{F}_{\text{point}} \cdot \mathbf{h}_{\text{geo}}^{\top}). \tag{3}$$

where $\sigma(\cdot)$ denotes the Sigmoid activation function.

3.4 Contrastive Affordance Preference Optimization (CAPO)

While the structural design of HGA successfully equips the `<SEG>` token with multi-scale physical priors, fully resolving geometric ambiguity requires a mechanism that teaches the model to respect fine-grained physical boundaries. To complement the structural enhancement from a preference learning perspective, we introduce Contrastive Affordance Preference Optimization (CAPO). By establishing a paradigm of preference-based geometric supervision, CAPO formulates 3D affordance segmentation as a reward ranking problem, directly forcing the semantic intent to align with physical reality and resolving the inherent geometric ambiguity.

Hybrid Preference Reward The core of CAPO is a hybrid reward function $R(\mathbf{P}, \mathbf{G})$ that comprehensively evaluates the prediction mask $\mathbf{P}$ against the binary ground-truth $\mathbf{G} \in \{0,1\}^N$ to guide the preference alignment.
Position Reward (R_{pos}). To ensure macroscopic spatial coverage, we employ the Soft-IoU score:

$$R_{\text{pos}}(\mathbf{P}, \mathbf{G}) = \frac{\sum_{i=1}^{N} \mathbf{P}_i \mathbf{G}_i}{\sum_{i=1}^{N} \mathbf{P}_i + \sum_{i=1}^{N} \mathbf{G}_i - \sum_{i=1}^{N} \mathbf{P}_i \mathbf{G}_i + \epsilon}, \tag{4}$$

where ϵ is a small smoothing factor.
Shape Reward (R_{shape}). Although Soft-IoU encourages spatial overlap, it cannot penalize the remaining geometric ambiguity (e.g., predictions bleeding into physically distinct parts). To address this, we propose a feature-based shape reward. We leverage the intermediate features $\mathbf{F}_{\text{mid}}$ from the 3D encoder, which inherently encode rich physical attributes (e.g., curvatures, edges). We extract the physical prototypes by performing mask-weighted pooling over these dense features:

$$\mathbf{v}_{\mathbf{P}} = \frac{\sum_{i=1}^{N} \mathbf{P}_i \cdot \mathbf{F}_{\text{mid},i}}{\sum_{i=1}^{N} \mathbf{P}_i}, \quad \mathbf{v}_{\mathbf{G}} = \frac{\sum_{i=1}^{N} \mathbf{G}_i \cdot \mathbf{F}_{\text{mid},i}}{\sum_{i=1}^{N} \mathbf{G}_i}, \tag{5}$$

where $\mathbf{v}_{\mathbf{P}}, \mathbf{v}_{\mathbf{G}} \in \mathbb{R}^{C_2}$. Here, $\mathbf{v}_{\mathbf{P}}$ acts as the predicted physical prototype of the selected region, while $\mathbf{v}_{\mathbf{G}}$ represents the ground-truth physical prototype. The shape reward is then defined as the cosine similarity between them:

$$R_{\text{shape}}(\mathbf{P}, \mathbf{G}) = \frac{\mathbf{v}_{\mathbf{P}}^{\top} \mathbf{v}_{\mathbf{G}}}{\|\mathbf{v}_{\mathbf{P}}\|_2 \|\mathbf{v}_{\mathbf{G}}\|_2}, \tag{6}$$

This metric regularizes geometric consistency. If the mask $\mathbf{P}$ overflows into a geometrically distinct region, the pooled prototype $\mathbf{v}_{\mathbf{P}}$ will deviate significantly from $\mathbf{v}_{\mathbf{G}}$, resulting in a lower reward. The total hybrid reward is formulated as $R = R_{\text{pos}} + \alpha R_{\text{shape}}$, where α is a fixed weight for the shape reward.

Contrastive Preference Objective Building upon the hybrid reward, we optimize the network using a direct preference objective. To implement this *contrastive reward supervision*, we establish an In-Batch Negative Sampling strategy.

Consider a training batch of size B. For the i-th instruction-scene pair, the matched ground truth $\mathbf{G}_i$ serves as the positive target, while the ground truths of other instances in the batch, denoted as $\mathbf{G}_j$ $(j \neq i)$, act as the negative targets. The preference loss is formulated to maximize the reward margin:

$$\mathcal{L}_{\text{CPO}} = -\frac{1}{B(B-1)} \sum_{i=1}^{B} \sum_{j \neq i} \log \sigma\Big(\beta\big[R(\mathbf{P}_i, \mathbf{G}_i) - R(\mathbf{P}_i, \mathbf{G}_j)\big]\Big). \tag{7}$$

where $\sigma(\cdot)$ is the Sigmoid function and β is a temperature hyperparameter. By maximizing the reward margin between the correct affordance mask and semantically distinct negative masks, this objective provides strict geometric supervision. The contrastive push-and-pull mechanism eliminates geometric ambiguity, ensuring the `<SEG>` token yields precise 3D affordance segmentation.

3.5 Total Loss

The entire Geo-DPO framework is trained in an end-to-end manner. The final objective function combines three distinct loss components. First, following SeqAfford [56], we utilize a standard auto-regressive text generation loss ($\mathcal{L}_{\text{txt}}$) to train the MLLM for sequential reasoning and `<SEG>` token generation. Second, a basic segmentation loss ($\mathcal{L}_{\text{seg}}$), comprising Binary Cross Entropy (BCE) and Dice loss, provides foundational point-wise spatial supervision while mitigating class imbalance between the affordance target and the background. Finally, our proposed preference loss ($\mathcal{L}_{\text{CPO}}$) applies contrastive reward supervision to penalize geometric ambiguity, ensuring the predicted mask aligns tightly with the actual physical boundaries. The total loss is formulated as a weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{txt}} + \mathcal{L}_{\text{seg}} + \lambda \mathcal{L}_{\text{CPO}}. \tag{8}$$

where λ is a hyperparameter balancing the contributions of the segmentation and preference optimization tasks.

4 Experiments

4.1 Dataset

To evaluate our Geo-DPO framework, we conduct extensive experiments on the large-scale SeqAfford dataset [56]. This benchmark is highly comprehensive, as it naturally encompasses both single-step affordance grounding and complex sequential affordance reasoning. Therefore, it serves as a unified and sufficient testbed to fully validate our method. The dataset contains approximately 180K instruction-point cloud pairs built upon 18K diverse 3D point clouds. Following

its official evaluation protocol, the dataset is divided into standard training and testing sets. To rigorously assess the model's segmentation accuracy and reasoning generalization, the test set is carefully categorized into three distinct settings: *Seen* (evaluating on familiar object categories and instructions), *Unseen* (testing generalization on novel object categories), and *Sequential* (evaluating long-horizon, multi-step instructions).

4.2 Implementation Details

Our Geo-DPO framework is built upon the ShapeLLM-7B [32] architecture, which serves as the foundational 3D MLLM. During the training phase, the original 3D encoder of ShapeLLM is kept frozen. To facilitate precise 3D dense predictions, we integrate Uni3D [57] as the primary vision encoder, connecting the modalities via a standard multi-layer perceptron (MLP) projection layer. Within this visual processing pipeline, our proposed Hierarchical Geometry Adapter (HGA) is embedded to extract multi-scale representations from the intermediate blocks of the 3D encoder for structural feature injection.

For parameter-efficient training, we apply LoRA [12] with a default rank of 8. The network is optimized using the AdamW [24] optimizer, configured with a learning rate of 2×10^{-4} and 0 weight decay. We implement a cosine annealing learning rate schedule featuring a 0.03 warm-up iteration ratio. To accelerate the training process, the standard attention mechanisms within ShapeLLM are substituted with flash-attention [5]. Finally, the scaling weight λ, which controls the contrastive preference optimization objective, is empirically set to 0.2.

4.3 Evaluation Metrics and Baselines

To provide a comprehensive and effective evaluation, we follow previous works [56] to choose four evaluation metrics: Area Under the Curve (AUC) [23], Mean Intersection Over Union (mIoU) [35], Similarity (SIM) [38], and Mean Absolute Error (MAE) [43]. Among these, mIoU and MAE are particularly crucial for assessing the model's precise boundary alignment and its ability to resolve geometric ambiguity. We conduct comparisons based on its setup in the Single Affordance segmentation task, and comparisons with other foundational baselines were also implemented following the approach mentioned in it. While in the sequential affordance segmentation task, to ensure a fair evaluation, we offer decomposed sequential information to these foundational models, enabling them to perform sequential reasoning.

4.4 Results on Language-Guided Single Affordance Segmentation Task

As detailed in Table 1, our Geo-DPO framework demonstrates superior performance across all evaluation metrics compared to the baseline methods, particularly outperforming LASO [19] and the previous state-of-the-art, SeqAfford [56].

Table 1: Main Results on the SeqAfford Dataset. The overall results of all comparative methods are reported, with the best results highlighted in **bold**. Seen and Unseen are two partitions of the Single Affordance segmentation dataset. AUC and mIoU are shown in percentage (%). * indicates that baseline methods use ground-truth sequential order, as existing 2D/3D methods cannot naturally predict sequential affordances.

Task	Method	mIoU (↑)	AUC (↑)	SIM (↑)	MAE (↓)
Seen	ReferTrans [16]	11.4	77.2	0.449	0.135
	ReLA [20]	12.1	76.3	0.480	0.130
	3D-SPS [25]	10.1	75.2	0.413	0.141
	IAGNet [52]	14.2	81.7	0.510	0.117
	LASO [19]	16.3	84.3	0.568	0.108
	SeqAfford [56]	19.5	86.9	0.594	0.098
	Geo-DPO (Ours)	**20.9**	**88.1**	**0.610**	**0.091**
Unseen	ReferTrans [16]	9.1	67.4	0.427	0.151
	ReLA [20]	9.3	68.2	0.423	0.147
	3D-SPS [25]	7.1	66.9	0.397	0.162
	IAGNet [52]	11.7	73.6	0.438	0.143
	LASO [19]	12.4	76.1	0.502	0.132
	SeqAfford [56]	13.8	82.4	0.518	0.128
	Geo-DPO (Ours)	**14.8**	**85.6**	**0.528**	**0.126**
Sequential	ReferTrans* [16]	10.8	74.5	0.425	0.142
	ReLA* [20]	11.4	74.8	0.463	0.136
	3D-SPS* [25]	9.9	73.1	0.407	0.148
	IAGNet* [52]	13.5	78.2	0.496	0.131
	LASO* [19]	14.3	80.7	0.521	0.124
	SeqAfford [56]	14.6	84.2	0.573	0.118
	Geo-DPO (Ours)	**15.1**	**86.5**	**0.601**	**0.115**

Unlike conventional segmentation tasks, language-guided single affordance segmentation demands not just semantic identification but precise spatial localization. While existing MLLM-based methods like SeqAfford successfully integrate perception and cognition, their reliance on a heavily compressed `<SEG>` token inevitably leads to geometric ambiguity and boundary overflow. In contrast, our model explicitly addresses this limitation. By utilizing the Hierarchical Geometry Adapter (HGA) to structurally inject dense spatial features, and employing Contrastive Affordance Preference Optimization (CAPO) to penalize mask overflow, Geo-DPO achieves significantly higher mIoU and SIM scores. This indicates a much stronger capability in capturing fine-grained physical boundaries compared to existing paradigms.

4.5 Results on Language-Guided Sequential Affordance Reasoning Task

The sequential affordance reasoning task implies the ability to infer multiple affordances from a single text, which is significantly more challenging as geometric errors can easily accumulate over multiple reasoning steps. As shown in the sequential setting of Table 1, our model consistently outperforms all comparative methods. Previous models like LASO rely on purely linguistic encoders and lack deep integration with 3D geometry. On the other hand, SeqAfford, despite introducing a 3D MLLM, still struggles with precise boundary alignment during long-horizon tasks due to the lack of geometric supervision. Our Geo-DPO framework substantially mitigates these issues. Because our CAPO objective forces the predicted masks to respect real geometric boundaries via preference ranking, our model maintains highly accurate segmentation even when decomposing complex, multi-step user intentions. The substantial improvements confirm that our method not only understands the sequential logic but also effectively grounds each step into precise 3D physical regions without spatial confusion.

4.6 Visualization

To intuitively demonstrate the effectiveness of our Geo-DPO in bridging the semantic-geometric gap, we present a qualitative comparison with SeqAfford [56] in Figure 4. While SeqAfford successfully comprehends high-level semantic intents, it consistently suffers from severe geometric ambiguity due to the over-compression of the `<SEG>` token. For instance, in multi-step tasks like operating scissors, the baseline's predicted mask for the handle inevitably bleeds into the crossing blades. Similarly, when tasked with carrying a bag by hand, it struggles to cleanly delineate the handle from the main body, resulting in semantic overflow. In contrast, our Geo-DPO consistently yields highly precise and boundary-aware affordance masks. As shown in Figure 4, Geo-DPO isolates the bag's handle without touching the main body. Furthermore, for heavily intersecting structures like the scissors' hinge, our method decisively disentangles the adjacent components into distinct interaction regions.

4.7 Ablation Study

To thoroughly investigate the effectiveness of our proposed modules and hyperparameter settings, we conduct extensive ablation studies on the SeqAfford dataset. All ablation experiments are evaluated on the most challenging Sequential setting to rigorously test both reasoning and geometric boundary alignment.

Effectiveness of Core Components. We first ablate the two key components of the Geo-DPO framework: the Hierarchical Geometry Adapter (HGA) and Contrastive Affordance Preference Optimization (CAPO). As demonstrated in Table 2, the baseline model without any geometric enhancements struggles with severe mask overflow, yielding an mIoU of 14.5%. Integrating HGA alone

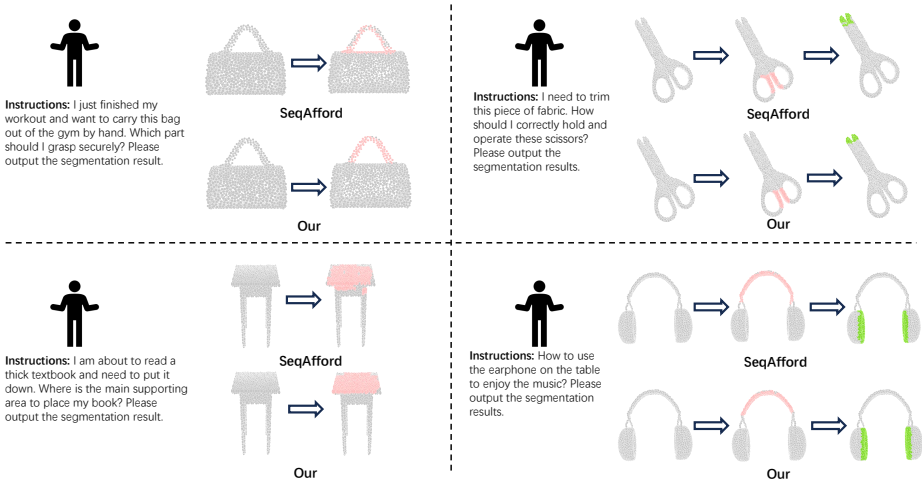

Fig. 4: Qualitative comparison of 3D affordance segmentation.

brings a slight improvement to 14.7% mIoU, proving that structurally injecting multi-scale 3D features effectively restores low-level spatial awareness. Conversely, applying only the CAPO loss without HGA also improves the performance to 14.8% mIoU, confirming that penalizing boundary overflow via preference optimization forces the model to respect geometric limits. Ultimately, combining both HGA and CAPO achieves the best performance (15.1% mIoU). This confirms that structural feature restoration and preference-based geometric supervision are highly complementary in resolving geometric ambiguity.

Impact of the Preference Optimization Weight (λ). We further analyze the impact of λ, the critical balancing weight that controls the CAPO loss weight. As shown in Table 3, relying on point-wise supervision ($\lambda = 0.0$) fails to constrain the physical boundaries. Gradually increasing λ to 0.2 yields the optimal performance across all metrics, as it perfectly balances the foundational segmentation accuracy with the contrastive boundary reward. However, increasing λ to 0.4 degrades the performance (dropping to 14.8% mIoU), as an excessively high weight destabilizes the baseline semantic alignment. Therefore, $\lambda = 0.2$ is adopted as our default setting.

Feature Fusion Strategy within HGA. A core design of our Hierarchical Geometry Adapter is the extraction and fusion of multi-scale dense features from the 3D encoder. To validate our design choice, we ablate different feature aggregation strategies within the HGA module, as shown in Table 4. A simple element-wise addition (Add) or channel-wise concatenation (Concat) fails to fully exploit the spatial correlation across different receptive fields, resulting in suboptimal boundary precision (14.4% and 14.5% mIoU, respectively). In contrast, our adopted Cross-Attention mechanism enables dynamic and context-aware feature aggregation, allowing the model to adaptively focus on the most relevant

Table 2: Ablation study on the core components of Geo-DPO (evaluated on the Sequential task). HGA: Hierarchical Geometry Adapter. CAPO: Contrastive Affordance Preference Optimization.

HGA	CAPO	mIoU (↑)	AUC (↑)	SIM (↑)	MAE (↓)
		14.5	84.0	0.572	0.117
✓		14.7	84.9	0.585	0.116
	✓	14.8	85.4	0.591	0.117
✓	✓	**15.1**	**86.5**	**0.601**	**0.115**

Table 3: Ablation study on the balancing weight λ for the CAPO objective.

Setting	mIoU (↑)	AUC (↑)	SIM (↑)	MAE (↓)
$\lambda = 0.0$ (w/o CAPO)	14.7	84.9	0.585	0.116
$\lambda = 0.05$	14.7	85.1	0.583	0.116
$\lambda = 0.2$ (Default)	**15.1**	**86.5**	**0.601**	**0.115**
$\lambda = 0.4$	14.8	86.1	0.601	0.116

geometric cues. This strategy achieves the peak performance of **15.1**% mIoU, demonstrating its superiority in handling complex 3D structures.

Robustness to Instruction Complexity. To further demonstrate the robustness of our framework in resolving geometric ambiguity under varying levels of intent complexity, we analyze the model's performance across different instruction lengths (measured by the number of sequential affordance steps). As illustrated in Table 5, while the performance of the baseline SeqAfford [56] drops precipitously as the reasoning steps increase from 2 to 3 (a drop of 0.3% mIoU), our Geo-DPO maintains a significantly more stable performance (only dropping by 0.1% mIoU). This confirms that our CAPO objective effectively prevents error accumulation and spatial confusion, enabling robust geometric alignment even in highly complex, long-horizon tasks.

Table 4: Ablation on different multi-scale feature fusion strategies within the HGA module.

Fusion Strategy	mIoU (↑)	AUC (↑)	SIM (↑)	MAE (↓)
Addition	14.4	84.1	0.572	0.118
Concatenation	14.5	84.5	0.579	0.118
Cross-Attention (Ours)	**15.1**	**86.5**	**0.601**	**0.115**

Table 5: Performance decay across different instruction complexities (number of sequential reasoning steps).

Reasoning Steps	Baseline mIoU	Geo-DPO mIoU	Δ Improvement
1 Step (Single)	19.5	**20.9**	+1.4
2 Steps	14.8	**15.2**	+0.4
$\geq$ 3 Steps	14.5	**15.1**	+0.6

4.8 Discussion and Limitations

While Geo-DPO bridges the semantic-geometric gap in 3D affordance segmentation [2, 8, 44], there are still several avenues for future exploration. First, our framework relies on the auto-regressive decoding of a large multi-modal model (e.g., a 7B-parameter MLLM). Although this architecture provides powerful implicit reasoning capabilities for complex instructions, it introduces non-negligible inference latency. In real-world embodied AI scenarios, where a robot must dynamically and rapidly interact with its environment, this computational overhead could become a bottleneck. Future work will explore knowledge distillation techniques to compress the MLLM's heavy reasoning capabilities into a lightweight, real-time geometry-aware segmentation network. Second, our current Contrastive Affordance Preference Optimization (CAPO) primarily focuses on static and rigid objects. Extending this preference learning mechanism to highly deformable objects (e.g., cloth manipulation) or complex articulated structures (e.g., multi-joint robotic arms [10, 11]) remains an open challenge. Beyond affordance segmentation, the proposed preference-based boundary regularization may also be applicable to generic 3D semantic or part segmentation, where adjacent regions often require precise geometric separation. We leave this broader extension to future work.

5 Conclusion

In this paper, we address the persistent challenge of geometric ambiguity and mask overflow in language-driven 3D sequential affordance segmentation. To tackle this, we propose the Geo-DPO framework, which seamlessly aligns the high-level semantic intents generated by 3D Multimodal Large Language Models with fine-grained physical boundaries. Specifically, we introduce a Hierarchical Geometry Adapter (HGA) to restore multi-scale structural awareness from the visual encoder, and design a Contrastive Affordance Preference Optimization (CAPO) objective to penalize spatial confusion. Extensive experiments on the large-scale SeqAfford benchmark demonstrate that our method establishes a new state-of-the-art, significantly outperforming existing paradigms in both single-step and complex sequential reasoning tasks. Our future work will explore extending this preference-based geometric alignment to more dynamic, real-time interactive environments, further advancing the capabilities of embodied 3D scene understanding.

Acknowledgement This work was supported by the National Natural Science Foundation of China (No.62331003, 62471405, 62301451), Basic Research Program of Jiangsu (BK20241814), Suzhou Basic Research Program (SYG202316) and XJTLU REF-22-01-010.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Bińkowski, M.a., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems (2022)
2. Chen, D., Kong, D., Li, J., Yin, B.: Maskprompt: open-vocabulary affordance segmentation with object shape mask prompts. In: Proceedings of the AAAI conference on artificial intelligence (2025)
3. Chen, Y., Yang, S., Huang, H., Wang, T., Xu, R., Lyu, R., Lin, D., Pang, J.: Grounded 3d-llm with referent tokens. arXiv preprint arXiv:2405.10370 (2024)
4. Chu, H., Deng, X., Lv, Q., Chen, X., Li, Y., Hao, J., Nie, L.: 3d-affordancellm: Harnessing large language models for open-vocabulary affordance detection in 3d worlds. arXiv preprint arXiv:2502.20041 (2025)
5. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in neural information processing systems (2022)
6. Deng, S., Xu, X., Wu, C., Chen, K., Jia, K.: 3d affordancenet: A benchmark for visual object affordance understanding. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2021)
7. Fu, R., Liu, J., Chen, X., Nie, Y., Xiong, W.: Scene-llm: Extending language model for 3d visual understanding and reasoning. arXiv preprint arXiv:2403.11401 (2024)
8. He, L., Liu, M., Ye, Q., Zhou, Y., Deng, X., Ding, G.: Task-aware 3d affordance segmentation via 2d guidance and geometric refinement. arXiv preprint arXiv:2511.11702 (2025)
9. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems (2023)
10. Hou, Y., Ma, J., Sun, H., Wu, F.: Effective offline robot learning with structured task graph. IEEE Robotics and Automation Letters (2024)
11. Hou, Y., Sun, H., Ma, J., Wu, F.: Improving offline reinforcement learning with inaccurate simulators. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE (2024)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr p. 3 (2022)
13. Ju, Y., Hu, K., Zhang, G., Zhang, G., Jiang, M., Xu, H.: Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In: European Conference on Computer Vision (2024)
14. Li, G., Sun, D., Sevilla-Lara, L., Jampani, V.: One-shot open affordance learning with foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)

15. Li, H., Xie, H., Xu, J., Wen, B., Hong, F., Liu, Z.: MonoArt: progressive structural reasoning for monocular articulated 3d reconstruction. arXiv preprint arXiv 2603.19231 (2026)
16. Li, M., Sigal, L.: Referring transformer: A one-step approach to multi-task visual grounding. Advances in neural information processing systems (2021)
17. Li, S., He, Y., Huang, H., Bu, X., Liu, J., Guo, H., Wang, W., Gu, J., Su, W., Zheng, B.: 2d-dpo: Scaling direct preference optimization with 2-dimensional supervision. In: Findings of the Association for Computational Linguistics: NAACL 2025 (2025)
18. Li, Y., Chen, Y., Ma, Z., Xiao, J., Wang, X., Yao, A.: Intermediate connectors and geometric priors for language-guided affordance segmentation on unseen object categories. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
19. Li, Y., Zhao, N., Xiao, J., Feng, C., Wang, X., Chua, T.s.: Laso: Language-guided affordance segmentation on 3d object. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
20. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2023)
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. pp. 34892–34916 (2023)
22. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
23. Lobo, J.M., Jiménez-Valverde, A., Real, R.: Auc: a misleading measure of the performance of predictive distribution models. Global ecology and Biogeography (2008)
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
25. Luo, J., Fu, J., Kong, X., Gao, C., Ren, H., Shen, H., Xia, H., Liu, S.: 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 (2022)
26. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2019)
27. Nasiriany, S., Kirmani, S., Ding, T., Smith, L., Zhu, Y., Driess, D., Sadigh, D., Xiao, T.: Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). IEEE (2025)
28. Nguyen, T.T., Wilson, C., Dalins, J.: Aligning large vision-language models by deep reinforcement learning and direct preference optimization. arXiv preprint arXiv:2509.06759 (2025)
29. Ning, C., Wu, R., Lu, H., Mo, K., Dong, H.: Where2explore: Few-shot affordance learning for unseen novel categories of articulated objects. Advances in Neural Information Processing Systems (2023)
30. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2023)

31. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems (2017)
32. Qi, Z., Dong, R., Zhang, S., Geng, H., Han, C., Ge, Z., Yi, L., Ma, K.: Shapellm: Universal 3d object understanding for embodied interaction. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024 (2025)
33. Qian, S., Chen, W., Bai, M., Zhou, X., Tu, Z., Li, L.E.: Affordancellm: Grounding affordance from vision language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
34. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems (2023)
35. Rahman, M.A., Wang, Y.: Optimizing intersection-over-union in deep neural networks for image segmentation. In: International symposium on visual computing (2016)
36. Roy, A., Todorovic, S.: A multi-scale cnn for affordance segmentation in rgb images. In: European conference on computer vision (2016)
37. Shao, Y., Zhai, W., Yang, Y., Luo, H., Cao, Y., Zha, Z.J.: Great: Geometry-intention collaborative inference for open-vocabulary 3d object affordance grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
38. Swain, M.J., Ballard, D.H.: Color indexing. International journal of computer vision (1991)
39. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
40. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. Advances in Neural Information Processing Systems (2023)
41. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics (tog) (2019)
42. Wang, Z., Huang, H., Zhao, Y., Zhang, Z., Zhao, Z.: Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. arXiv preprint arXiv:2308.08769 (2023)
43. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. Climate research (2005)
44. Wu, D., Fu, Y., Huang, S., Liu, Y., Jia, F., Liu, N., Dai, F., Wang, T., Anwer, R.M., Khan, F.S., et al.: Ragnet: Large-scale reasoning-based affordance segmentation benchmark towards general grasping. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
45. Wu, R., Cheng, K., Zhao, Y., Ning, C., Zhan, G., Dong, H.: Learning environment-aware affordance for 3d articulated object manipulation under occlusions. Advances in Neural Information Processing Systems (2023)
46. Wu, S., Zhu, Y., Huang, Y., Zhu, K., Gu, J., Yu, J., Shi, Y., Wang, J.: Afforddp: Generalizable diffusion policy with transferable affordance. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)

47. Xie, Y., Li, G., Xu, X., Kan, M.Y.: V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In: Findings of the Association for Computational Linguistics: EMNLP 2024 (2024)
48. Xing, S., Li, P., Wang, Y., Bai, R., Wang, Y., Hu, C.W., Qian, C., Yao, H., Tu, Z.: Re-align: Aligning vision language models via retrieval-augmented direct preference optimization. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (2025)
49. Xu, C., Chen, Y., Wang, H., Zhu, S.C., Zhu, Y., Huang, S.: Partafford: Part-level affordance discovery from 3d objects. arXiv preprint arXiv:2202.13519 (2022)
50. Xu, P., Mu, Y.: Weakly-supervised affordance grounding guided by part-level semantic priors. arXiv preprint arXiv:2505.24103 (2025)
51. Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D.: Pointllm: Empowering large language models to understand point clouds. In: European Conference on Computer Vision (2024)
52. Yang, Y., Zhai, W., Luo, H., Cao, Y., Luo, J., Zha, Z.J.: Grounding 3d object affordance from 2d interactions in images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
53. Yang, Z., Qiu, X., Zhao, X., Wang, X., Zhang, Q., Xiao, J.: Frequency-aware affinity for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20468–20477 (2026)
54. Yang, Z., Zhao, X., Wang, X., Zhang, Q., Xiao, J.: Leveraging class distributions in clip for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 34714–34723 (2026)
55. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. arXiv preprint arXiv:2310.07704 (2023)
56. Yu, C., Wang, H., Shi, Y., Luo, H., Yang, S., Yu, J., Wang, J.: Seqafford: Sequential 3d affordance reasoning via multimodal large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
57. Zhou, J., Wang, J., Ma, B., Liu, Y.S., Huang, T., Wang, X.: Uni3d: Exploring unified 3d representation at scale. In: International Conference on Learning Representations (2024)
58. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
59. Zhu, L., Chen, T., Xu, Q., Liu, X., Ji, D., Wu, H., Soh, D.W., Liu, J.: Popen: Preference-based optimization and ensemble for lvlm-based reasoning segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)