# VideoAfford: Grounding 3D Affordance from HOI Videos via Multimodal Large Language Model

**Hanqing Wang**[1] **Mingyu Liu**[2] **Xiaoyu Chen**[1] **Chengwei Ma**[1]
Yiming Zhong[3] Wenti Yin[4] Yuhao Liu[5] Zhiqing Cui[1]
Jiahao Yuan[1] Xing Mu[1] Lu Dai[1,6] Zhiyuan Ma[4,*] Hui Xiong[1,6,*]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Zhejiang University
[3]ShanghaiTech University
[4]Huazhong University of Science and Technology
[5]Shandong University
[6]The Hong Kong University of Science and Technology

hwang201@connect.hkust-gz.edu.cn

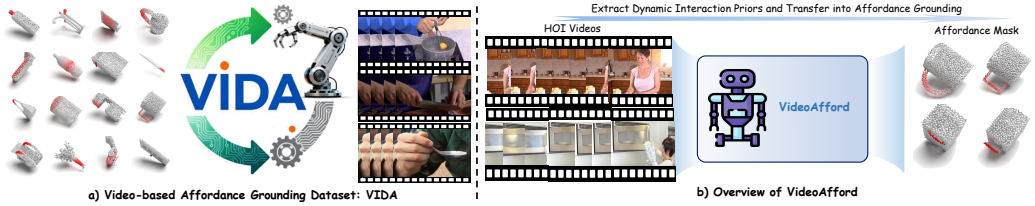



Figure 1: We introduce the task of grounding 3D affordance from human-object-interaction (HOI) videos, and collect a large-scale dataset VIDA (left), which contains 38K HOI videos spanning 16 affordance categories and 22K 3D point clouds. Based on this, we proposed the VideoAfford (right), which can transfer the HOI interaction priors into 3D affordance grounding.

## Abstract

3D affordance grounding aims to highlight the actionable regions on 3D objects, which is crucial for embodied AI. While previous research primarily leverages static cues from language or single images, which often lack the rich interaction context necessary to accurately locate functional zones. To alleviate this predicament, we collect a comprehensive video-based 3D affordance dataset, *VIDA*, which contains 38K human-object-interaction (HOI) videos covering 16 affordance types, 38 object categories, and 22K point clouds. Based on *VIDA*, we propose a strong baseline: **VideoAfford**, a unified framework that extends Multimodal Large Language Models (MLLMs) with fine-grained affordance segmentation capabilities. VideoAfford incorporates a latent action encoder to distill dynamic interaction knowledge from demonstration videos and a spatial-aware loss to encourage geometrically consistent affordance predictions. Extensive experiments on VIDA show that VideoAfford significantly outperforms strong baselines in both seen and unseen settings, demonstrating its effectiveness in video-driven 3D affordance reasoning and open-world generalization. The dataset and code are available at `https://github.com/hq-King/VideoAfford`.

*Corresponding authors: Zhiyuan Ma and Hui Xiong.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

# 1 Introduction

Recent years have witnessed the remarkable success of affordance grounding [1, 2], which serves as a vital bridge between visual perception and robotic manipulation for embodied intelligence. The primary goal is to identify and localize operable object regions in response to human commands or demonstrations. While early research focused on 2D affordance segmentation from images [3–7] and videos [8–10], subsequent efforts have leveraged foundation models like CLIP [11] and Multimodal Large Language Models (MLLMs) [12] to enhance open-vocabulary reasoning. Furthermore, the learning-from-demonstration paradigm has prompted research [13–18] to exploit human-object-interaction (HOI) images as interaction priors. Despite the utility of 2D cues, 3D affordance grounding provides more precise and intuitive guidance for physical task execution, establishing a robust foundation for downstream applications such as robotic grasping [19, 20] and manipulation [21–25].

However, existing 3D affordance learning methods rely heavily on static modalities, such as HOI images [26, 27] or linguistic instructions [28, 29]. These static sources often fail to capture the nuanced "how-to" knowledge—the temporal and dynamic interaction context—inherent in human demonstrations. While several recent works have explored HOI videos [8, 9, 30] to mitigate this, they face significant hurdles: methods like OPPA [31] and 2HandedAfforder [32] remain confined to 2D planes. Although EGO-SAG [30] pioneered 3D affordance learning from unlabeled egocentric videos, its focus on coarse-grained scene masks falls short of the fine-grained, object-centric precision required for dexterous manipulation.

To bridge these gaps, we formulate a novel task: *Grounding 3D object-centric affordance from HOI demonstration videos*. This task seeks to distill object-centric interaction priors from large-scale video corpora for 3D affordance reasoning. Accordingly, we introduce VIDA, a comprehensive video-point cloud affordance dataset. By aggregating diverse HOI videos from Internet, HOIGEN-1M [33], and TASTE-RoB [34], and employing an automated pipeline supported by GPT-4o and rigorous manual verification, we ensure both high-scale and high-quality annotations.

Building on this, we propose VideoAfford, a strong baseline that transfers the latent interaction knowledge within video-based MLLMs to the 3D domain. VideoAfford incorporates a latent action encoder to extract embeddings from HOI videos and a spatial-aware loss mechanism to refine geometric reasoning. This synergy enables the model to achieve a deep understanding of 3D spatial structures, resulting in superior performance on both in-distribution and out-of-distribution data.

In summary, our contributions are:

- We formulate the task of video-based 3D object affordance grounding, bridging the gap between dynamic human demonstrations and static 3D geometric representations.
- We present **VIDA**, the first large-scale benchmark for this task, featuring 38K HOI videos and 22K annotated affordance point clouds.
- We propose **VideoAfford**, a framework that integrates a latent action encoder and spatial-aware loss to exploit dynamic interaction priors for 3D localization.
- Extensive evaluations demonstrate that our method significantly outperforms existing baselines and exhibits robust open-vocabulary generalization.

# 2 Related Works

**Affordance Learning.** Affordance is a pivotal bridge linking perception, reasoning, and physical manipulation. Early research primarily focuses on detecting affordance regions from 2D sources, such as images [3–6, 35, 36] and videos [8–10, 37], with some works incorporating natural language guidance [38]. However, affordance knowledge derived from 2D planes is difficult to project onto precise 3D interaction coordinates. To bridge this gap, 3D AffordanceNet [39] introduced a large-scale, fine-grained 3D dataset, significantly catalyzing advancements in 3D affordance learning. Subsequently, researchers have explored learning universal affordance from HOI images [26, 27, 40, 41] and language instructions [42, 28, 29, 43–50], while others have leveraged the rich world priors within foundation models like VLMs [29, 43] and diffusion models [51, 52]. Despite these efforts, existing works often overlook the dynamic interaction flow, which is essential for robust affordance reasoning. While EGO-SAG [30] attempts to learn from egocentric views, it mainly targets scene-level affordance and lacks sufficient object-centric priors for manipulation. To fill this

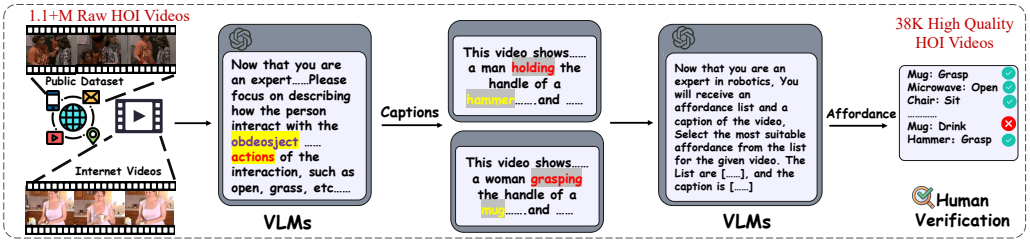



Figure 2: **Data Collection Pipeline.** We show the whole data collection and verification pipeline here. We utilize VLMs to caption each video and extract keywords about action and objects. We then utilize the VLMs to pair the video to an affordance type. Finally, we manually check the results to ensure correctness.

void, we present VIDA, the first video-based object-centric 3D affordance dataset, aiming to advance 3D affordance research and further empower embodied intelligence. VIDA will be publicly released once the paper is accepted.

**Multimodal Large Language Model.** MLLMs have achieved remarkable success in both 2D [53–55] and 3D understanding [56–59]. By leveraging large-scale web data, these models exhibit human-like reasoning and perception across diverse modalities. However, their potential for embodied perception—specifically in the context of affordance reasoning—remains largely under-explored. While pioneering efforts such as AffordanceLLM [12] and 3DAffordanceLLM [43] have begun to utilize MLLMs for affordance grounding, they predominantly rely on static images or textual descriptions, thereby overlooking the dynamic interaction cues that are indispensable for real-world manipulation. To bridge this gap, our work aims to empower MLLMs with affordance-aware reasoning by learning from video-based human-object interactions. This approach enables the models to effectively interpret and interact with 3D objects in complex, context-sensitive scenarios.

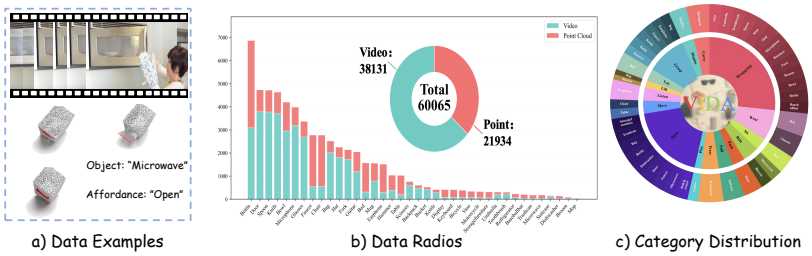
## 3 Dataset

Existing 3D affordance datasets mainly rely on static inputs, such as point clouds, images, or language descriptions. Although these annotations have greatly advanced object-level affordance perception, they provide limited evidence about *how* humans interact with objects. In contrast, HOI videos naturally capture the pre-contact, contact, and post-contact stages of manipulation, revealing action intention, contact evolution, and functional object regions. Motivated by this observation, we construct ***VIDA***, a large-scale benchmark designed to exploit HOI videos for 3D object-centric affordance grounding. As shown in Table 1, VIDA is distinguished from prior datasets by jointly providing object-level 3D affordance annotations and dynamic interaction information.

### 3.1 Collection Details

**Object Category Selection.** We select 38 object categories according to two principles. First, the objects should frequently appear in daily scenarios and robotic manipulation tasks. Second, each object should contain clear and distinguishable functional regions that can support affordance annotation. The selected categories, such as mug, microwave, door, and knife, are aligned with representative 3D affordance benchmarks, including 3D AffordanceNet [39], PIAD1 [27], and PIAD2 [26]. This design ensures both comparability with existing benchmarks and sufficient diversity for video-driven affordance learning.

**Video Collection and Filtering.** VIDA contains approximately **38K** HOI videos collected from public sources, including HOIGEN-1M [33], TASTE-Rob [34], and Internet videos. HOIGEN-1M provides a large-scale, diverse set of human-object interaction videos with rich action semantics, while TASTE-Rob offers hand-centric interaction sequences that are particularly valuable for understanding functional use in real-world tasks. Because these two public datasets provide the captions themselves, we need to caption the collected internet videos only. To ensure that each clip provides reliable




a) Data Examples b) Data Radios c) Category Distribution

Figure 3: **VIDA Dataset.** We provide an overview of VIDA. (a) Examples of HOI videos and their corresponding affordance point clouds. (b) The ratio of videos and point clouds across affordance categories. (c) The category distribution of videos and point clouds in VIDA.

Table 1: **Comparison of Existing 3D Affordance Datasets with Ours.** #Point Cloud, #Aff, and #Classes denote the number of point clouds, affordance types, and object categories, respectively. #Dynamic Information indicates whether the dataset provides dynamic interaction cues for affordance grounding. ✓/× denotes whether the corresponding attribute is available.

| Dataset | #Publication Year | #Dynamic Information | #Granularity | #Input Source | #Classes | #Aff | #Point Cloud |
|---|---|---|---|---|---|---|---|
| 3D AffordanceNet [39] | CVPR 2021 | × | Object | Point Cloud | 23 | 18 | 23k |
| O2O-Afford [60] | CoRL 2021 | × | Object | Point Cloud | 18 | – | 1.7k |
| PartAfford [61] | ECCVW 2022 | × | Object | Point Cloud | 23 | 24 | 25k |
| IAGNet [27] | ICCV 2023 | × | Object | Image, Point Cloud | 23 | 17 | 7k |
| LASO [28] | CVPR 2025 | × | Object | Language, Point Cloud | 23 | 17 | 8.4k |
| SceneFun3D [44] | CVPR 2024 | × | Scene | Language, Point Cloud | – | 9 | 710 |
| SeqAfford [29] | CVPR 2025 | × | Object | Language, Point Cloud | 23 | 28 | 180k |
| 3DAffordSplat [45] | ACM MM 2025 | × | Object | Language, 3D Gaussian | 21 | 18 | 24k |
| SeqAffordSplat [46] | Arxiv 2025 | × | Object | Language, 3D Gaussian | 21 | 18 | 14k |
| GREAT [26] | CVPR 2025 | × | Object | Language,Image, Point Cloud | 43 | 24 | 39k |
| AGPIL [62] | CVPR 2025 | × | Object | Language,Image, Point Cloud | 23 | 17 | 31k |
| Affogato [63] | Arxiv 2025 | × | Object | Language,Image, Point Cloud | >450 | >350 | 150k |
| VSAD [30] | Arxiv 2024 | ✓ | Scene | Videos, Point Cloud | 16 | 17 | 2k |
| **VIDA** (Ours) | Neurips 2026 | ✓ | Object | Language, Videos, Point Cloud | 38 | 16 | 22k |

interaction evidence, we apply a three-step filtering pipeline. We first remove low-resolution, blurry, or excessively short clips and use MLLM to give captions for the internet videos. Next, we extract object and action keywords from video captions and use GPT-4o [53] to infer the corresponding affordance labels. Finally, we recruit five human experts to perform the final confirmation of the filtered video clips and labels, ensuring that the resulting dataset preserves high-quality interaction evidence for affordance grounding. More details can be seen in Sec. H.

**Point Clouds and Training Pair Construction.** We collect 3D object point clouds from PIADv1 [27] and PIADv2 [26], retaining only instances that belong to the 38 predefined categories. Since videos and point clouds are collected from different sources, strict instance-level alignment is generally unavailable. Therefore, we construct video-point cloud pairs at the **object-category level**. This strategy is consistent with the training protocol of PIADv1 [27], where an HOI image is treated as the anchor and point clouds from the same object category are randomly sampled for training. Similarly, during training in VIDA, one HOI video can be paired with multiple point clouds from the same category, which increases instance diversity and encourages the model to transfer interaction priors across object shapes. During testing, we use **fixed one-to-one pairs** to ensure rigorous and reproducible evaluation.

### 3.2 Statistics and Analysis

As illustrated in Fig. 3, ***VIDA*** contains over 38.1K videos and 21.9K object point clouds, covering 38 object categories and 16 affordance types. Since videos and point clouds are sampled from different instances, VIDA does not require fixed one-to-one correspondence during training. As shown in Fig. 3(a), a single HOI video can correspond to multiple point clouds from the same category. For evaluation, however, we strictly use fixed one-to-one video-point cloud pairs to ensure rigor and reproducibility. Fig. 3(b) reports the ratio of videos and point clouds for each affordance category, while Fig. 3(c) summarizes the distribution of affordance labels across both modalities. Following PIADv1 [27] and PIADv2 [26], we divide VIDA into **seen** and **unseen** settings. In the seen setting, object categories and affordance labels in the training and test sets are consistent. In the unseen

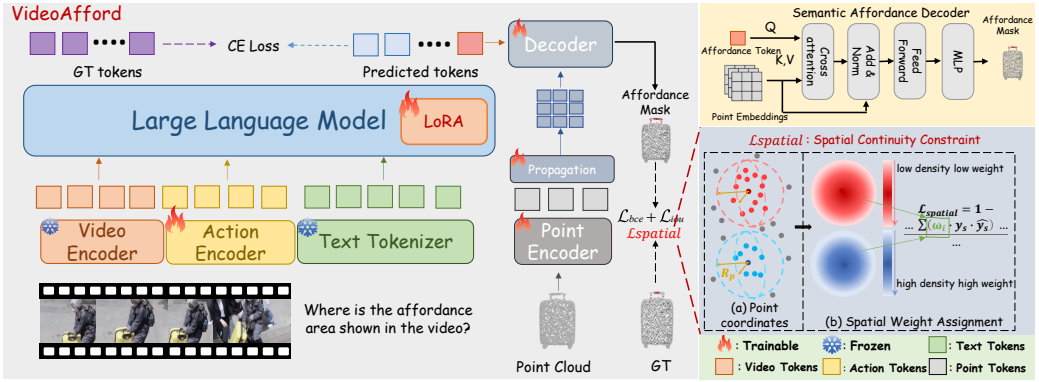



Figure 4: **Overview of VideoAfford.** VideoAfford extracts video embeddings and latent action embeddings, which are fed into an LLM with the language instruction to produce textual responses and an affordance token. In parallel, a pre-trained 3D encoder extracts semantic point embeddings, which are propagated to dense point features through a geometry-guided upsampling module. Finally, the affordance token and dense point features are fused by an affordance decoder to predict the 3D affordance mask. More details of the propagation process are provided in the Appendix.

setting, the test set contains novel objects or affordances that do not appear during training, enabling evaluation of open-world generalization. More details can be seen in the Sec. B.

# 4 Method

## 4.1 Architecture Overview

VideoAfford is designed to transfer dynamic interaction priors from HOI videos to static 3D object geometry. Different from previous methods that mainly rely on language, images, or other static cues, our model uses video demonstrations to infer *where* an action should be performed on the target object. Formally, given an HOI video $\mathcal{V}$, a textual instruction $\mathcal{T}$, and a target object point cloud $\mathcal{P}_c \in \mathbb{R}^{N\times 3}$, the model $g$ predicts a point-wise affordance mask $\mathcal{A}_{\text{mask}} \in [0,1]^N$:

$$\mathcal{A}_{\text{mask}} = g(\mathcal{T}, \mathcal{V}, \mathcal{P}_c). \tag{1}$$

Each value in $\mathcal{A}_{\text{mask}}$ represents the probability that the corresponding 3D point belongs to the actionable region implied by the video and instruction.

As illustrated in Fig. 4, VideoAfford consists of four key components: (1) a **semantic-aware point encoder** that extracts 3D semantic features from the target object; (2) a **latent action encoder** that distills motion and interaction priors from HOI videos; (3) an **action-conditioned multimodal LLM** that integrates video, action, and language tokens for affordance reasoning; and (4) a **cross-modal affordance decoder** that projects the resulting affordance query onto dense 3D point features to obtain fine-grained masks.

## 4.2 Network Architecture

**Semantic-Aware Point Encoder.** To obtain a strong 3D semantic representation, we adopt a pre-trained point cloud encoder trained on large-scale text-image-point paired data. Given the target point cloud $\mathcal{P}_c \in \mathbb{R}^{N\times 3}$, the encoder $f$ extracts sparse semantic point embeddings $\mathcal{P}_s$:

$$\mathcal{P}_s = f(\mathcal{P}_c). \tag{2}$$

Since affordance grounding requires dense point-wise prediction, we further propagate the sparse semantic features back to the original point resolution with a geometry-guided upsampling module:

$$\mathcal{P}_d = \text{Up}(\mathcal{P}_s, o(\mathcal{P}_c)), \quad \mathcal{P}_d \in \mathbb{R}^{N\times D}, \tag{3}$$

where $o(\cdot)$ denotes geometric propagation based on the 3D structure of the input point cloud, and $D$ is the feature dimension of each point. This step preserves the semantic knowledge of the pre-trained 3D

encoder while recovering dense geometric details required for mask prediction. And the geometric propagation details are shown in the Sec. G.

**Latent Action Encoder.** While point features describe the geometry and semantics of the target object, affordance reasoning also depends on the interaction dynamics in the demonstration video. For example, the actionable region of a handle is more clearly revealed by the hand trajectory and contact pattern than by object appearance alone. To capture such cues, we build a latent action encoder $m$ that factorizes a video into compact interaction states and state-transition action tokens. Inspired by [36], the encoder uses DINOv2 [64] as an image backbone to extract frame-level visual tokens, followed by a lightweight transformer compressor that maps each frame into a small set of compact state tokens. We first train this on robot manipulation videos [65] and Ego-centric videos [66] with self-supervised state reconstruction. During this pretraining stage, a diffusion decoder is attached to reconstruct each input frame from its compact state tokens, forcing the tokens to retain the interaction state while remaining low-dimensional. After pretraining, the decoder is discarded, and only the trained state encoder is used to produce action tokens for VideoAfford. This design exposes the encoder to hand-object and robot-object interactions that are close to our video-driven 3D affordance setting while avoiding the cost of image decoding during affordance grounding. More architectural and training details are provided in Appendix D.

Given an HOI video $\mathcal{V} = \{I_1, I_2, \ldots, I_T\}$, the encoder first maps sampled frames into compact interaction states:

$$\mathbf{s}_t = \phi_\theta(I_t) \in \mathbb{R}^{D_s}, \tag{4}$$

where $\phi_\theta$ denotes our trained state encoder and $D_s$ is the state dimension. We then represent action as the transition between consecutive compact states and project these transitions into action tokens:

$$\mathcal{A}_c = [\psi(\mathbf{s}_2 - \mathbf{s}_1); \ldots; \psi(\mathbf{s}_{K+1} - \mathbf{s}_K)] \in \mathbb{R}^{K \times D_a}, \tag{5}$$

where $\psi$ is a learnable projection layer, $K$ denotes the number of temporal transitions, and $D_a$ denotes the action feature dimension. In this way, the action tokens focus on how the interaction evolves, rather than only encoding static appearance. These tokens capture motion patterns, contact evolution, and functional interaction semantics, providing dynamic priors that complement the sparse video tokens extracted by the video encoder.

**Action-Conditioned Multimodal LLM.** We use LanguageBind [67] as the video encoder and Llama [68] as the language model backbone. Given the video sequence, LanguageBind extracts video tokens $e(\mathcal{V})$, while the latent action encoder produces action tokens $\mathcal{A}_c$. We concatenate the video tokens, action tokens, and textual instruction tokens into a unified multimodal sequence:

$$\mathcal{Z} = [e(\mathcal{V}); \mathcal{A}_c; \mathcal{T}]. \tag{6}$$

The sequence is then processed by the LLM to reason jointly over visual dynamics, latent action priors, and language semantics.

To obtain an affordance-specific representation, we introduce a learnable special token $\langle\text{AFF}\rangle$. The hidden state corresponding to this token, denoted as $\mathcal{H}_{\text{aff}}$, summarizes the model's contextual understanding of the interaction and its associated 3D functional region. We project it into an affordance query embedding:

$$\mathcal{A}_m = \text{Proj}(\mathcal{H}_{\text{aff}}) \in \mathbb{R}^D. \tag{7}$$

This query serves as the cross-modal bridge between the video-language reasoning space and the 3D point feature space.

**Cross-Modal Affordance Decoder.** The affordance decoder maps the affordance query to dense 3D predictions. Specifically, we use a lightweight transformer decoder in which the affordance query $\mathcal{A}_m$ acts as the query feature, and the dense point features $\mathcal{P}_d$ serve as keys and values:

$$\mathcal{F}_{\text{aff}} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}}\right)\mathbf{V}, \tag{8}$$

where $\mathbf{Q}$ is projected from $\mathcal{A}_m$, and $\mathbf{K}$ and $\mathbf{V}$ are projected from $\mathcal{P}_d$. The fused representation is then passed through an MLP to predict the point-wise affordance probability:

$$\mathcal{A}_{\text{mask}} = \text{MLP}(\mathcal{F}_{\text{aff}}). \tag{9}$$

Through this design, the decoder grounds the affordance semantics inferred from the HOI video onto the geometric structure of the target 3D object.

Table 2: **Main Results.** AUC and mIoU are reported as percentages. The best results are highlighted in **bold** with a light pink font color, and the second-best results are underlined with a light blue color.

| Setting | Metric | XMF | PFusion | IAGNet | LASO | GREAT | Seqafford | LMAfford3D | **Ours** |
|---|---|---|---|---|---|---|---|---|---|
| Seen | *mIoU*↑ | 14.41$^{\uparrow 15.09}$ | 16.33$^{\uparrow 13.17}$ | 20.39$^{\uparrow 9.11}$ | 18.65$^{\uparrow 10.85}$ | 23.62$^{\uparrow 5.88}$ | 23.03$^{\uparrow 6.47}$ | 22.74$^{\uparrow 6.76}$ | **29.51** |
| | *AUC*↑ | 71.47$^{\uparrow 13.30}$ | 78.43$^{\uparrow 6.34}$ | 80.22$^{\uparrow 4.55}$ | 78.44$^{\uparrow 6.33}$ | 81.41$^{\uparrow 3.36}$ | 81.17$^{\uparrow 3.60}$ | 80.74$^{\uparrow 4.03}$ | **84.77** |
| | *SIM*↑ | 41.10$^{\uparrow 18.20}$ | 46.28$^{\uparrow 12.02}$ | 50.11$^{\uparrow 9.19}$ | 49.46$^{\uparrow 9.84}$ | 51.25$^{\uparrow 8.05}$ | 47.71$^{\uparrow 11.59}$ | 47.28$^{\uparrow 12.02}$ | **59.34** |
| | *MAE*↓ | 0.281$^{\downarrow 0.129}$ | 0.264$^{\downarrow 0.112}$ | 0.188$^{\downarrow 0.036}$ | 0.257$^{\downarrow 0.105}$ | 0.173$^{\downarrow 0.021}$ | 0.227$^{\downarrow 0.075}$ | 0.234$^{\downarrow 0.082}$ | **0.152** |
| Unseen | *mIoU*↑ | 6.010$^{\uparrow 7.20}$ | 7.270$^{\uparrow 5.94}$ | 7.970$^{\uparrow 5.24}$ | 7.410$^{\uparrow 5.80}$ | 8.220$^{\uparrow 4.99}$ | 8.070$^{\uparrow 5.14}$ | 8.110$^{\uparrow 5.10}$ | **13.21** |
| | *AUC*↑ | 53.41$^{\uparrow 21.21}$ | 56.69$^{\uparrow 17.93}$ | 68.97$^{\uparrow 5.65}$ | 69.21$^{\uparrow 5.41}$ | 70.19$^{\uparrow 4.43}$ | 65.53$^{\uparrow 9.09}$ | 66.42$^{\uparrow 8.20}$ | **74.62** |
| | *SIM*↑ | 31.53$^{\uparrow 11.64}$ | 34.05$^{\uparrow 9.12}$ | 34.85$^{\uparrow 8.32}$ | 33.77$^{\uparrow 9.40}$ | 35.08$^{\uparrow 8.09}$ | 32.40$^{\uparrow 10.77}$ | 33.61$^{\uparrow 9.56}$ | **43.17** |
| | *MAE*↓ | 0.388$^{\downarrow 0.141}$ | 0.371$^{\downarrow 0.124}$ | 0.277$^{\downarrow 0.030}$ | 0.288$^{\downarrow 0.041}$ | 0.269$^{\downarrow 0.022}$ | 0.286$^{\downarrow 0.039}$ | 0.278$^{\downarrow 0.031}$ | **0.247** |

**Spatial Continuity Constraint.** A core challenge in 3D affordance grounding is fragmented prediction, commonly observed as isolated noisy points around the target region. Standard point-wise objectives such as binary cross-entropy treat each point independently and do not explicitly encourage spatially coherent masks. To address this issue, we introduce a spatial continuity constraint that incorporates local geometric neighborhoods into the segmentation loss. For each point $\mathbf{x}_i \in \mathbb{R}^3$, we define its closed neighborhood as

$$\widetilde{\mathcal{N}}_i = \{j \mid \|\mathbf{x}_i - \mathbf{x}_j\|_2 \leq \mathcal{R}_p\}, \tag{10}$$

where $\mathcal{R}_p$ is a predefined radius and the point itself is included in $\widetilde{\mathcal{N}}_i$. We then estimate the local kernel density and convert it into a geometry-adaptive weight:

$$\rho_i = \sum_{j \in \widetilde{\mathcal{N}}_i} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right), \quad \omega_i = \frac{\rho_i}{\frac{1}{N}\sum_{k=1}^{N} \rho_k + \epsilon}, \tag{11}$$

where $\sigma$ controls the Gaussian decay and is set to $0.1\mathcal{R}_p$ by default. This formulation preserves the neighborhood cardinality through the summation, rather than averaging it away, so denser local regions receive larger weights. Since $\mathbf{x}_i \in \widetilde{\mathcal{N}}_i$, every point has at least one kernel contribution, and the additional $\epsilon$ keeps the normalization numerically stable.

We integrate these weights into a Dice-style objective and obtain the spatial-aware loss:

$$\mathcal{L}_{\text{spatial}} = 1 - \frac{2\sum_{i=1}^{N} \omega_i y_i \hat{y}_i}{\sum_{i=1}^{N} \omega_i y_i^2 + \sum_{i=1}^{N} \omega_i \hat{y}_i^2 + \epsilon}, \tag{12}$$

where $y_i$ and $\hat{y}_i$ denote the ground-truth label and predicted probability of the $i$-th point, respectively, and $\epsilon$ is a small constant for numerical stability. Compared with standard segmentation losses, this loss encourages the model to predict compact and spatially continuous affordance regions that better match the physical morphology of objects. Refer to Table 6 for more details.

### 4.3 Training Objectives

VideoAfford is trained to align multimodal affordance reasoning with accurate 3D grounding. The overall objective combines language modeling supervision and point-wise segmentation supervision:

$$\mathcal{L} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{iou}}\mathcal{L}_{\text{iou}} + \lambda_{\text{spatial}}\mathcal{L}_{\text{spatial}}. \tag{13}$$

$\mathcal{L}_{\text{ce}}$ supervises the textual output of the LLM, $\mathcal{L}_{\text{bce}}$ and $\mathcal{L}_{\text{iou}}$ guide binary affordance segmentation, and $\mathcal{L}_{\text{spatial}}$ improves geometric continuity of the predicted 3D masks. See Table 6 for more details.

## 5 Experiments

### 5.1 Benchmark Setting

**Baselines.** Since video-driven object-centric 3D affordance grounding is a newly introduced task, there is no directly comparable prior method designed for the same setting. The most related work is EGO-SAG [30], but its code has not been released. To provide a comprehensive comparison, we

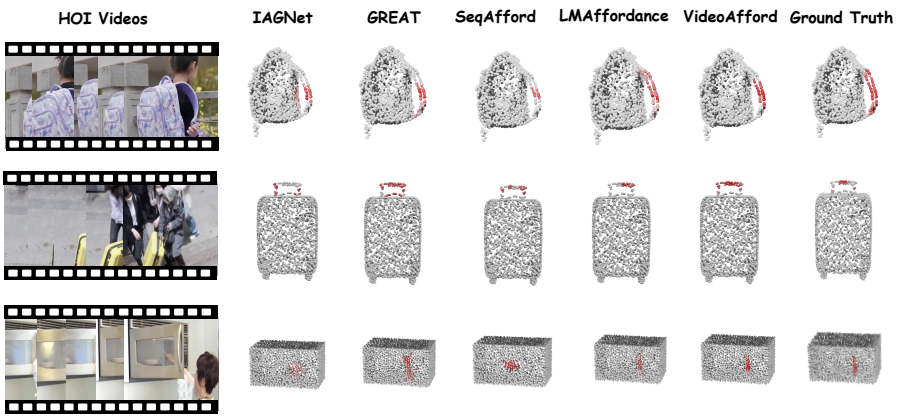



Figure 5: **Visualization Results.** Darker red indicates higher affordance probability. More visualizations are provided in the supplementary materials.

adapt several representative 3D affordance grounding methods, including IAGNet [27], GREAT [26], SeqAfford [29], and LMAfford3D [62], as modular baselines. Following previous research, we also chose XMF [69] and PFusion [70] as baselines. For each method, we apply the same frame sampling strategy to obtain $N$ video frames, encode each frame with the corresponding image encoder, and fuse the resulting embeddings for 3D affordance prediction.

**Evaluation Metrics.** Following previous 3D affordance grounding works, we use four evaluation metrics: Area Under the Curve (AUC) [71], mean Intersection over Union (mIoU) [72], SIMilarity (SIM) [73], and Mean Absolute Error (MAE) [74].

### 5.2 Implementation Details.

Following VideoLLaVA [75], we use LanguageBind [67] as the video encoder and Llama [68] as the large language model backbone and utilize the default Pytorchvideo to sample the videos. For latent action modeling, we follow the setting of Stamo [36] to model the actions. We adopt Uni3D [76] as the 3D vision encoder to enhance point cloud understanding. All encoders except the Action Encoder are frozen during training. We use LoRA [77] for parameter-efficient fine-tuning and set the LoRA rank to 128 by default. The model is optimized with AdamW [78], where the learning rate and weight decay are set to 0.0002 and 0, respectively. We use a cosine learning rate scheduler with a warm-up ratio of 0.03. During training, all attention modules are implemented with FlashAttention [79]. The main experiments are trained for 10 epochs on four H200 GPUs, taking approximately 11 hours in total. Additional implementation details are provided in the Sec. C and E.

### 5.3 Comparison Results

Table 2 reports the quantitative results on VIDA. VideoAfford consistently outperforms all baselines across both seen and unseen settings. The weaker performance of frame-based methods shows that sampled frames alone are insufficient for capturing interaction dynamics, while our latent action modeling provides stronger cues for 3D affordance grounding.

The unseen setting is more challenging because it contains novel objects or affordance types not observed during training. Existing baselines struggle to transfer static image or language cues to these new cases. By leveraging dynamic HOI evidence, VideoAfford achieves more robust generalization.

Qualitative results in Fig. 5 further show that VideoAfford produces more complete and compact masks, while baselines often yield scattered or incomplete predictions. This indicates that video demonstrations provide useful interaction priors for localizing actionable 3D regions.

Table 3: Ablation study on the latent action encoder and the proposed spatial constraint loss.

| | Action Encoder | Spatial Loss | *mIoU*↑ | *AUC*↑ | *SIM*↑ | *MAE*↓ |
|---|---|---|---|---|---|---|
| Seen | × | × | 21.22 | 80.11 | 45.21 | 0.226 |
| | × | ✓ | 22.87 | 81.35 | 52.34 | 0.172 |
| | ✓ | × | 24.02 | 82.71 | 53.65 | 0.162 |
| | ✓ | ✓ | **29.51** | **84.77** | **59.34** | **0.152** |
| Unseen | × | × | 7.710 | 67.52 | 34.24 | 0.298 |
| | × | ✓ | 8.260 | 69.73 | 35.19 | 0.291 |
| | ✓ | × | 10.11 | 71.92 | 38.79 | 0.271 |
| | ✓ | ✓ | **13.21** | **74.62** | **43.17** | **0.247** |

### 5.4 Ablation Study

**Effectiveness of Action Encoder and Spatial Loss.** Table 3 evaluates the impact of the latent action encoder and spatial loss. Introducing either component improves the baseline, and combining them yields the best performance across all metrics. This confirms that latent action modeling helps the model capture dynamic interaction priors from HOI videos, while the spatial loss encourages geometrically coherent predictions on point clouds. Without the action encoder, the model has difficulty identifying the interaction pattern behind the video and may activate irrelevant object regions. Without spatial loss, the model mainly optimizes point-wise correctness but lacks explicit constraints on spatial continuity, leading to fragmented masks and lower IoU scores.

**Effectiveness of Interaction Logic.** To verify that temporal interaction logic benefits affordance grounding, we randomly shuffle the order of video frames during inference. As shown in Table 4, disrupting the natural frame sequence leads to a clear performance drop across all metrics. This degradation suggests that the model relies not only on static object appearance but also on coherent dynamic interaction priors, such as hand motion, contact evolution, and causal relations between humans and objects. Preserving the original frame order enables the model to capture meaningful interaction evolution, confirming that temporal logic provides essential context for robust affordance grounding.

Table 4: Ablation study on frame order.

| Metric | Original | Shuffled |
|---|---|---|
| *mIoU*↑ | **29.51** | 18.12 |
| *AUC* ↑ | **84.77** | 75.86 |
| *SIM* ↑ | **59.34** | 43.33 |
| *MAE* ↓ | **0.152** | 0.185 |

**Choice of Sampled Frames.** Due to the existing video MLLMs that process videos by extracting frames into images, we further investigate the impact of the number of sampled frames. As shown in Table 5, sampling only 2 or 4 frames provides insufficient temporal context, making it difficult for the model to capture complete interaction dynamics. In contrast, using too many frames (e.g., 16) introduces redundant and potentially noisy information, while also increasing computational cost. Sampling 8 frames achieves the best balance between temporal coverage and information efficiency, and we therefore use it as the default setting.

Table 5: Ablation study on the number of sampled video frames.

| | Metric | 2 frames | 4 frames | 8 frames | 16 frames |
|---|---|---|---|---|---|
| Seen | *mIoU*↑ | 22.47 | 24.72 | **29.51** | 19.67 |
| | *AUC*↑ | 81.03 | 82.97 | **84.77** | 79.21 |
| | *SIM*↑ | 52.41 | 53.46 | **59.34** | 43.02 |
| | *MAE*↓ | 0.181 | 0.182 | **0.152** | 0.194 |
| Unseen | *mIoU*↑ | 8.780 | 10.12 | **13.21** | 7.170 |
| | *AUC*↑ | 70.11 | 71.93 | **74.62** | 71.92 |
| | *SIM*↑ | 36.72 | 40.11 | **43.17** | 39.34 |
| | *MAE*↓ | 0.257 | 0.255 | **0.247** | 0.268 |

## 6 Conclusion, Limitation and Future Work

**Summary.** In this paper, we introduce the task of *grounding 3D object affordance from human-object interaction demonstration videos*, which aims to harness large-scale demonstration video corpora to advance object-centric 3D affordance reasoning from static interaction knowledge to complex, dynamic interaction priors. We collect and construct the first large-scale video-based 3D object affordance dataset **VIDA**, and based on this strong dataset, we introduce **VideoAfford**, the first MLLM to reason fine-grained 3D affordance for this new paradigm. Bolstered by novel Spatial

Constraint Loss and the powerful world knowledge within MLLM, our method establishes a strong baseline for this new but meaningful task.

**Limitations and Future Directions.** Despite promising results on VIDA, VideoAfford relies on heavy video-3D paired annotations. Future work will explore weak/self-supervised learning from unlabeled HOI videos to cut annotation cost, scale up datasets and interaction types, and better exploit unlabeled videos for embodied perception and real-world embodied AI.

## 7 Acknowledgements

This work was supported in part by the National Natural Science Foundation of ChinaGrant No.92370204, in part by the National Natural Science Foundation of China (No. 62406161), in part by the National Key RD Program of ChinaGrant No.2023YFF0725001in part by Guangdong Provincial Key Laboratory of Frontier Basic Science for All-domain Intelligence, in part by the guangdong Basic and Applied Basic Research FoundationGrant No.2023B1515120057,in part by the Key-Area Special Project of Guangdong Provincial Ordinary Universities(2024ZDZX1007).

Competing interests: The authors declare that they have no competing financial or non-financial interests, including financial relationships outside the submitted work during the past 36 months, that could be perceived as influencing this work.

## References

[1] James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977.

[2] Gen Li, Hanqing Wang, Jingliang Li, Yifan Han, Jindou Jia, Tao Lin, Yuanzhe Liu, Yutong Wang, Bo Zhao, Fangqiang Ding, et al. From passive perception to active interaction: A survey of affordance learning for embodied ai. 2026.

[3] Spyridon Thermos, Petros Daras, and Gerasimos Potamianos. A deep learning approach to object affordance segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2358–2362. IEEE, 2020.

[4] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5882–5889. IEEE, 2018.

[5] Hongchen Luo, Wei Zhai, Jiao Wang, Yang Cao, and Zheng-Jun Zha. Visual-geometric collaborative guidance for affordance learning. *arXiv preprint arXiv:2410.11363*, 2024.

[6] Xue Zhao, Yang Cao, and Yu Kang. Object affordance detection with relationship-aware network. *Neural Computing and Applications*, 32(18):14321–14333, 2020.

[7] Litao Liu, Yifan Han, Pengfei Yi, Wenbo Yu, Hanqing Wang, Haoran Du, Enze Yuan, Zilin Yuan, Ruiding Feng, Michael Liu, et al. Affordance2action: Task-conditioned scene-level affordance grounding for real-time manipulation. *arXiv preprint arXiv:2606.04172*, 2026.

[8] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.

[9] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning visual affordance grounding from demonstration videos. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[10] Joya Chen, Difei Gao, Kevin Qinghong Lin, and Mike Zheng Shou. Affordance grounding from demonstration video to target image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6799–6808, 2023.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

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

[13] Jiajin Tang, Zhengxuan Wei, Ge Zheng, and Sibei Yang. Closed-loop transfer for weakly-supervised affordance grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9530–9539, 2025.

[14] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022.

[15] Peiran Xu and MU Yadong. Weakly-supervised affordance grounding guided by part-level semantic priors. In *The Thirteenth International Conference on Learning Representations*.

[16] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10922–10931, 2023.

[17] Yuxuan Wang, Aming Wu, Muli Yang, Yukuan Min, Yihang Zhu, and Cheng Deng. Reasoning mamba: Hypergraph-guided region relation calculating for weakly supervised affordance grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27618–27627, 2025.

[18] WonJun Moon, Hyun Seok Seong, and Jae-Pil Heo. Selective contrastive learning for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5210–5220, 2025.

[19] Yi-Lin Wei, Mu Lin, Yuhao Lin, Jian-Jian Jiang, Xiao-Ming Wu, Ling-An Zeng, and Wei-Shi Zheng. Afforddexgrasp: Open-set language-guided dexterous grasp with generalizable-instructive affordance. *arXiv preprint arXiv:2503.07360*, 2025.

[20] Haoyu Zhao, Linghao Zhuang, Xingyue Zhao, Cheng Zeng, Haoran Xu, Yuming Jiang, Jun Cen, Kexiang Wang, Jiayan Guo, Siteng Huang, et al. Towards affordance-aware robotic dexterous grasping with human-like priors. *arXiv preprint arXiv:2508.08896*, 2025.

[21] Yueh-Hua Wu, Jiashun Wang, and Xiaolong Wang. Learning generalizable dexterous manipulation from human grasp affordance. In *Conference on Robot Learning*, pages 618–629. PMLR, 2023.

[22] Ran Xu, Yan Shen, Xiaoqi Li, Ruihai Wu, and Hao Dong. Naturalvlm: Leveraging fine-grained natural language for affordance-guided visual manipulation. *IEEE Robotics and Automation Letters*, 2024.

[23] Kaijun Wang, Liqin Lu, Mingyu Liu, Jianuo Jiang, Zeju Li, Bolin Zhang, Wancai Zheng, Xinyi Yu, Hao Chen, and Chunhua Shen. Odyssey: Open-world quadrupeds exploration and manipulation for long-horizon tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 18602–18610, 2026.

[24] Mingyu Liu, Zheng Huang, Xiaoyi Lin, Muzhi Zhu, Canyu Zhao, Yating Wang, Haoyi Zhu, Hao Chen, and Chunhua Shen. Gae: Unleashing physical potential of vlm with generalizable action expert. In *Forty-third International Conference on Machine Learning*, 2025.

[25] Mingyu Liu, Zeju Li, Jiuhe Shu, Hanqing Wang, Yuhao Chao, Hao Chen, and Chunhua Shen. Perfect demo makes poor teacher: Learning robust alignment from critical motion segments. *arXiv preprint arXiv:2606.15587*, 2026.

[26] Yawen Shao, Wei Zhai, Yuhang Yang, Hongcheng Luo, Yang Cao, and Zhengjun Zha. 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)*, pages 17326–17336, 2025.

[27] 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*, pages 10905–10915, 2023.

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

[29] Chunlin Yu, Hanqing Wang, Ye Shi, Haoyang Luo, Sibei Yang, Jingyi Yu, and Jingya Wang. Seqafford: Sequential 3d affordance reasoning via multimodal large language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1691–1701, 2025.

[30] Cuiyu Liu, Wei Zhai, Yuhang Yang, Hongchen Luo, Sen Liang, Yang Cao, and Zheng-Jun Zha. Grounding 3d scene affordance from egocentric interactions. *arXiv preprint arXiv:2409.19650*, 2024.

[31] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2139–2147, 2018.

[32] Marvin Heidinger, Snehal Jauhri, Vignesh Prasad, and Georgia Chalvatzaki. 2handedafforder: Learning precise actionable bimanual affordances from human videos. *arXiv preprint arXiv:2503.09320*, 2025.

[33] Kun Liu, Qi Liu, Xinchen Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, and Wu Liu. Hoigen-1m: A large-scale dataset for human-object interaction video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24001–24010, 2025.

[34] Hongxiang Zhao, Xingchen Liu, Mutian Xu, Yiming Hao, Weikai Chen, and Xiaoguang Han. Taste-rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27683–27693, 2025.

[35] Hanqing Wang, Shaoyang Wang, Yiming Zhong, Zemin Yang, Jiamin Wang, Zhiqing Cui, Jiahao Yuan, Yifan Han, Mingyu Liu, and Yuexin Ma. Affordance-r1: Reinforcement learning for generalizable affordance reasoning in multimodal large language model. *arXiv preprint arXiv:2508.06206*, 2025.

[36] Mingyu Liu, Jiuhe Shu, Hui Chen, Zeju Li, Canyu Zhao, Jiange Yang, Shenyuan Gao, Hao Chen, and Chunhua Shen. Stamo: Unsupervised learning of generalizable robot motion from compact state representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 35014–35024, 2026.

[37] Yifan Han, Litao Liu, Yuqi Gu, Ye Lu, Hanqing Wang, Sidney Wai, Ishaan Myrie, Qi Zhang, Jingjin Yu, and Gen Li. A4a: Cross-embodiment transfer of action-oriented 4d affordances from human demonstrations. *arXiv preprint arXiv:2609.05892*, 2026.

[38] Changmao Chen, Yuren Cong, and Zhen Kan. Worldafford: Affordance grounding based on natural language instructions. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 822–828. IEEE, 2024.

[39] 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*, pages 1778–1787, 2021.

[40] Xianqiang Gao, Pingrui Zhang, Delin Qu, Dong Wang, Zhigang Wang, Yan Ding, Bin Zhao, and Xuelong Li. Learning 2d invariant affordance knowledge for 3d affordance grounding. *arXiv preprint arXiv:2408.13024*, 2024.

[41] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16284–16295, 2024.

[42] Tongxuan Tian, Xuhui Kang, and Yen-Ling Kuo. $O^3$ afford: One-shot 3d object-to-object affordance grounding for generalizable robotic manipulation. *arXiv preprint arXiv:2509.06233*, 2025.

[43] Hengshuo Chu, Xiang Deng, Xiaoyang Chen, Yinchuan Li, Jianye Hao, and Liqiang Nie. 3d-affordancellm: Harnessing large language models for open-vocabulary affordance detection in 3d worlds. *arXiv preprint arXiv:2502.20041*, 2025.

[44] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. Scenefun3d: fine-grained functionality and affordance understanding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14531–14542, 2024.

[45] Zeming wei, Junyi Lin, Yang Liu, Weixing Chen, Jingzhou Luo, Guanbin Li, and Liang Lin. 3daffordsplat: Efficient affordance reasoning with 3d gaussians, 2025. URL `https://arxiv.org/abs/2504.11218`.

[46] Di Li, Jie Feng, Jiahao Chen, Weisheng Dong, Guanbin Li, Yuhui Zheng, Mingtao Feng, and Guangming Shi. Seqaffordsplat: Scene-level sequential affordance reasoning on 3d gaussian splatting. *arXiv preprint arXiv:2507.23772*, 2025.

[47] Xinyi Wang, Xun Yang, Yanlong Xu, Yuchen Wu, Zhen Li, and Na Zhao. Affordbot: 3d fine-grained embodied reasoning via multimodal large language models. *arXiv preprint arXiv:2511.10017*, 2025.

[48] Shang-Ching Liu, Wenkai Chen, Wei-Lun Cheng, Yen-Lin Huang, I-Bin Liao, Yung-Hui Li, Jianwei Zhang, et al. Pavlm: Advancing point cloud based affordance understanding via vision-language model. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4299–4306. IEEE, 2025.

[49] Mingyu Liu, Chonghao Sima, Tianjian Feng, Hanqing Wang, Cong Chen, Hao Chen, and Chunhua Shen. Perturbot: Breaking shortcut priors in vision-language-action models with perturbative training. *arXiv preprint arXiv:2610.04616*, 2026.

[50] Zhenhao Zhang, Ye Shi, Lingxiao Yang, Suting Ni, Qi Ye, and Jingya Wang. OpenHOI: Open-world hand-object interaction synthesis with multimodal large language model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=0biUwyjKkm`.

[51] Hanqing Wang, Zhenhao Zhang, Kaiyang Ji, Mingyu Liu, Wenti Yin, Yuchao Chen, Zhirui Liu, Xiangyu Zeng, Tianxiang Gui, and Hangxing Zhang. Dag: Unleash the potential of diffusion model for open-vocabulary 3d affordance grounding. *arXiv preprint arXiv:2508.01651*, 2025.

[52] Yuanchen Ju, Kaizhe Hu, Guowei Zhang, Gu Zhang, Mingrun Jiang, and Huazhe Xu. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In *European Conference on Computer Vision*, pages 222–239. Springer, 2024.

[53] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[54] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[55] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[56] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer, 2024.

[57] Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *European Conference on Computer Vision*, pages 214–238. Springer, 2024.

[58] 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. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

[59] Zheng Huang, Mingyu Liu, Xiaoyi Lin, Muzhi Zhu, Canyu Zhao, Zongze Du, Ye Lin, Xiaoman Li, Yiduo Jia, Hao Zhong, et al. Notvla: Semantics-preserving robot adaptation via narrative action interfaces. *arXiv preprint arXiv:2510.03895*, 2025.

[60] Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2o-afford: Annotation-free large-scale object-object affordance learning. In *Conference on robot learning*, pages 1666–1677. PMLR, 2022.

[61] 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.

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

[63] Junha Lee, Eunha Park, Chunghyun Park, Dahyun Kang, and Minsu Cho. Affogato: Learning open-vocabulary affordance grounding with automated data generation at scale. *arXiv preprint arXiv:2506.12009*, 2025.

[64] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[65] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[66] Xiaofeng Wang, Kang Zhao, Feng Liu, Jiayu Wang, Guosheng Zhao, Xiaoyi Bao, Zheng Zhu, Yingya Zhang, and Xingang Wang. Egovid-5m: A large-scale video-action dataset for egocentric video generation. *arXiv preprint arXiv:2411.08380*, 2024.

[67] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023.

[68] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[69] Emanuele Aiello, Diego Valsesia, and Enrico Magli. Cross-modal learning for image-guided point cloud shape completion. *Advances in Neural Information Processing Systems*, 35:37349–37362, 2022.

[70] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 244–253, 2018.

[71] Jorge M Lobo, Alberto Jiménez-Valverde, and Raimundo Real. Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2):145–151, 2008.

[72] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*, pages 234–244. Springer, 2016.

[73] Michael J Swain and Dana H Ballard. Color indexing. *International journal of computer vision*, 7(1): 11–32, 1991.

[74] Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005.

[75] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[76] Bo Zhang, Jiakang Yuan, Botian Shi, Tao Chen, Yikang Li, and Yu Qiao. Uni3d: A unified baseline for multi-dataset 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9253–9262, 2023.

[77] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[78] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[79] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35: 16344–16359, 2022.

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

[81] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19313–19322, 2022.

# A Overview of Appendix

This appendix provides additional details that complement the main paper. We include dataset construction and evaluation protocols, implementation details, the complete hyperparameter configuration, prompts used for annotation assistance, the point feature propagation algorithm, additional experiment results, and more qualitative visualizations.

The remainder of this supplementary material is organized as follows. Section B describes the VIDA construction pipeline and evaluation settings in more detail. Section C provides the computational resources and training cost. Section E lists the hyperparameters used in our experiments. Section F presents the detailed prompts for caption generation and affordance type analysis. Section G gives more details about point feature propagation. Section I reports additional results. Section J presents more visualizations.

# B Dataset Construction and Evaluation Protocol

## B.1 VIDA Construction Pipeline

VIDA is designed to connect dynamic human-object interaction (HOI) evidence with object-centric 3D affordance. The raw videos are collected from public HOI sources and Internet videos, while the point clouds are collected from existing 3D affordance benchmarks. Because the two modalities are obtained from different sources, the dataset is constructed at the object-category level rather than requiring strict instance-level video-to-point-cloud alignment. This design reflects a realistic transfer setting: a demonstration video shows how an object category is used, and the model must infer the corresponding functional region on a potentially different 3D instance from the same category.

The video filtering and annotation pipeline contains three stages:

1. **Interaction filtering.** We keep clips that contain visually clear human-object interaction evidence, including pre-contact motion, contact formation, and post-contact manipulation when available.
2. **Caption and label proposal.** We use GPT-4o to assist in generating interaction-centric captions and affordance label proposals from the filtered clips.
3. **Manual verification.** Human experts verify the object category, action semantics, and affordance label. Clips with inconsistent or uncertain labels are removed from the benchmark.

For point clouds, we retain instances belonging to the selected 38 object categories and use the provided point-wise affordance masks. Each point cloud is normalized before being fed into the model. During training, a video can be paired with multiple point clouds from the same object category, encouraging VideoAfford to learn category-level interaction priors rather than memorizing a particular object shape. During evaluation, each sample uses a fixed video-point-cloud pair to ensure deterministic and reproducible measurement. The overall video duration distribution information is shown in Figure 6.

## B.2 Seen and Unseen Settings

We follow the seen/unseen evaluation protocol used in prior 3D affordance grounding benchmarks. In the **seen** setting, object categories and affordance types appearing in the test set are also present during training. This setting evaluates whether the model can accurately ground affordance regions when both the object semantics and affordance labels are familiar.

In the **unseen** setting, test samples contain novel objects or affordances that are not observed during training. This setting is more challenging because the model must transfer dynamic interaction knowledge from demonstrations to previously unseen semantic or geometric configurations. We use the same evaluation metrics and inference procedure for both settings.

## B.3 Evaluation Details

For each evaluated pair, the model predicts a point-wise affordance probability map over the target point cloud. We report four metrics: mIoU, AUC, SIM, and MAE. mIoU evaluates the overlap

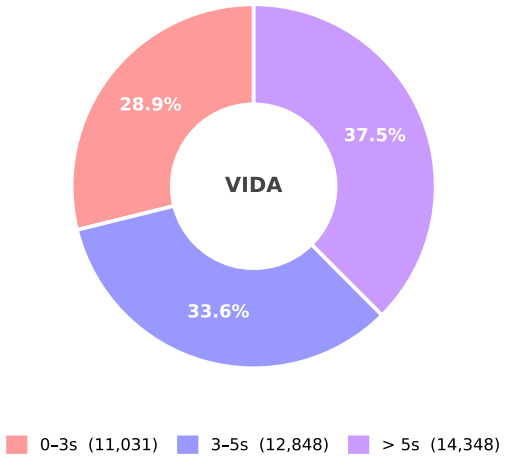



Figure 6: video duration distribution information. Average Duration seconds is 5.28s, Minimum Duration: 2.33s, Maximum Duration: 9.98s.

between predicted and ground-truth affordance masks after thresholding, AUC measures ranking quality over point-wise predictions, SIM measures distributional similarity between predicted and ground-truth affordance maps, and MAE measures the average absolute point-wise prediction error. For all metrics, higher values are better except MAE, where lower values are better.

All methods are evaluated using the same point clouds, fixed video-point-cloud pairs, and sampled video frames. For image-based or language-image-based baselines adapted to our task, we sample frames from the HOI video using the same temporal sampling strategy and fuse the resulting frame-level features for 3D affordance prediction. If the model is language-only, we follow the LMAffordance to utilize the clip image encoder to encode the frames and fuse the embeddings with point clouds. This ensures that performance differences mainly reflect the ability to exploit dynamic interaction information rather than differences in input access.

## C Computational Resources

To ensure reproducibility, we provide detailed information on the computational resources used in our experiments. For all experiments, including training and inference, we used 4 NVIDIA H200 GPUs.

The main training runs are conducted for 10 epochs and take approximately 11 hours on 4 NVIDIA H200 GPUs. The reported memory consumption corresponds to the full VideoAfford training configuration with video tokens, latent action tokens, language tokens, and point-cloud features enabled. During training, attention modules use FlashAttention to reduce memory usage and improve throughput. Unless otherwise specified, we use 8 uniformly sampled frames for each HOI video, which provides a balance between temporal coverage and computational cost as shown in our frame-number ablations.

**Trainable parameters.** We freeze the video encoder, the 3D point encoder, and the LLM backbone, except for the LoRA adapters and newly introduced projection and decoding modules. The latent action encoder is optimized in our framework, enabling the model to adapt its action representation to the video-based 3D affordance grounding task while keeping the majority of pretrained model parameters fixed.

## D Latent Action Encoder Details

**Compact state tokenizer.** The latent action encoder is designed to describe an interaction through two levels of representation: a compact state for each sampled frame and a transition token between

**Input:** an HOI video clip.
**Instruction:** Please describe the human-object interaction in this video. Focus on the manipulated object, the action being performed, the body part or tool involved in the interaction, the temporal evolution of contact, and the object region that appears functionally relevant. Avoid describing irrelevant background details. If the interaction is ambiguous or the target object cannot be identified, respond with "uncertain".
**Output format:** {object category; action; interaction description; likely functional region; confidence}.

Figure 7: Prompt for generating interaction-centric captions.

neighboring states. For a sampled frame $I_t$, we first extract dense visual features with a frozen DINOv2 image backbone:

$$\mathbf{F}_t = h(I_t) \in \mathbb{R}^{L \times C}, \tag{14}$$

where $L$ is the number of visual tokens and $C$ is the feature dimension. A lightweight transformer-based compressor then maps these dense features into a small set of compact state tokens:

$$\mathbf{S}_t = q_\theta(\mathbf{F}_t) \in \mathbb{R}^{K_s \times D_s}. \tag{15}$$

In our implementation, $K_s$ is much smaller than $L$, so the state representation removes most appearance redundancy while preserving information useful for reconstructing object pose, hand contact, and manipulation-relevant spatial layout. Compared with directly using dense frame features, this compact bottleneck encourages the encoder to retain interaction state information that can support state transition reasoning.

# E Hyperparameters

In Table 6, we present the hyperparameters used in our experiments.

Table 6: Details of Hyperparameters

| Hyperparameter | Value |
|---|---|
| batch size | 8 |
| precision | fp32 |
| optimizer | AdamW |
| learning rate | 0.0002 |
| weight decay | 0 |
| scheduler | cosine |
| warm-up ratio | 0.03 |
| sampled video frames | 8 |
| model max length | 2048 |
| tokenizer model max length | 3072 |
| lora r | 128 |
| lora alpha | 256 |
| epochs | 10 |
| lora dropout | 0.05 |
| beta1 | 0.9 |
| beta2 | 0.95 |
| out dim | 256 |
| mm vision select layer | -2 |
| lora target module | qproj, pproj |
| $\lambda_{ce}$ | 1 |
| $\lambda_{bce}$ | 1 |
| $\lambda_{spatial}$ | 2 |
| $\lambda_{iou}$ | 1 |
| $\mathcal{R}_p$ | 0.5 |

**Input:** an interaction-centric caption and the candidate affordance label set.
**Instruction:** Given the caption, infer the most appropriate affordance type. Base the decision on the action intention and the functional object region implied by the interaction, rather than only on object category names. Select one label from the candidate affordance set. If multiple labels are plausible, choose the one most directly supported by the observed contact and manipulation. If no label is supported, respond with "uncertain".
**Output format:** {affordance label; supporting evidence; confidence}.

Figure 8: Prompt for affordance type analysis.

Table 7: Seen and unseen object categories used in our experiments.

| Seen object categories | Unseen object categories |
|---|---|
| spoon, door, knife, microphone, bottle, glasses, bowl, bag, fork, hat, guitar, mug, scissors, backpack, bucket, hammer, kettle, toothbrush, umbrella, earphone, suitcase, refrigerator, vase, keyboard, broom, microwave, bed, mop, baseballbat, chair, bicycle, TrashCan, StorageFurniture, dishwasher, faucet, display, motorcycle, table | knife, microphone, bowl, bag, fork, mug, hammer, kettle, earphone, suitcase, vase, keyboard, broom, microwave, bed, baseballbat, chair, bicycle, TrashCan, StorageFurniture, faucet, display, motorcycle, table |

# F Prompts

We carefully design prompts to help GPT-4o generate captions and analyze affordance types during dataset construction. The prompts are used only for annotation assistance; the proposed VideoAfford model is trained and evaluated on the verified annotations. All automatically generated outputs are manually checked before being included in VIDA.

## F.1 Prompts for Caption

**Prompt for generating interaction-centric captions.** The caption prompt encourages the model to focus on functional evidence in the video, including the manipulated object, the action, the contact pattern, and the likely object part involved in the interaction. This reduces the influence of irrelevant background content and makes the caption more suitable for affordance label assignment. The detailed prompt template is shown in Figure 7.

## F.2 Prompts for analyzing affordance types

**Prompt for affordance type assignment.** Given an interaction-centric caption, the affordance analysis prompt asks the model to choose the most appropriate affordance label according to the observed action intention and functional contact region. The prompt explicitly discourages decisions based only on object category names, since the same object can support multiple affordances depending on the demonstrated interaction. The detailed prompt template is shown in Figure 8.

**Manual verification protocol.** After automatic captioning and affordance type proposal, annotators verify whether the caption matches the video and whether the selected affordance label is supported by visible interaction evidence. Samples are discarded if the target object is unclear, the interaction is incomplete, the model output conflicts with the video, or no candidate affordance label is appropriate. This verification step improves the reliability of video-affordance correspondence in the benchmark.

# G Point feature propagation algorithm

We provide the detailed point feature algorithm in the following Algorithm 1

# H Source HOI Video Datasets

VIDA collects HOI videos from several public sources. Among them, HOIGen-1M [33] and TASTE-Rob [34] are two important sources because they provide large-scale interaction videos with rich

Table 8: Seen and unseen affordance categories used in our experiments.

| Seen affordance categories | Unseen affordance categories |
|---|---|
| wrapgrasp, grasp, wear, open, push, play, lift, carry, pull, listen, press, ride, move, lay, sit, display | wrapgrasp, open, lift, pull, listen, press, ride, move, lay, sit, display |

**Algorithm 1** Upsample and Propagation

**Require:** Input point cloud pts $\in \mathbb{R}^{B\times C\times N}$ (batch size $B$, channels $C$, points $N$),
Class labels cls_label $\in \mathbb{R}^{B\times 16}$
**Ensure:** Output logits x $\in \mathbb{R}^{B\times N\times K}$ (classes $K$),
High-level features f_level_3 $\in \mathbb{R}^{B\times D\times G}$ (dimensions $D$, groups $G$)
Extract dimensions: $B, C, N \leftarrow$ size(pts)
Reshape point cloud: pts $\leftarrow$ permute(pts, $[0, 2, 1]$) Transform to $B \times N \times 3$
$(\text{neighborhood}, \text{center}) \leftarrow$ GroupDivider(pts) Partition into local groups and centers
group_tokens $\leftarrow$ Encoder(neighborhood) Encode local groups to $B \times G \times N$
group_tokens $\leftarrow$ ReduceDim(group_tokens) Reduce feature dimensions
cls_tokens $\leftarrow$ Expand(cls_token, $[B, 1, D]$) Class token initialization
cls_pos $\leftarrow$ Expand(cls_pos, $[B, 1, D]$) Class position embedding
center_pos $\leftarrow$ PosEmbed(center) Position embedding for group centers
x $\leftarrow$ Concat($[\text{cls_tokens}, \text{group_tokens}], \text{dim} = 1$) Combine class token and group features
pos $\leftarrow$ Concat($[\text{cls_pos}, \text{center_pos}], \text{dim} = 1$) Combine position embeddings
feature_list $\leftarrow$ TransformerBlocks(x, pos) Multi-stage feature extraction
feature_list $\leftarrow$ [Norm($f$)[:, 1 :].$permute$($[0, 2, 1]$) | $f \in$ feature_list] Adjust feature dimensions
cls_onehot $\leftarrow$ Repeat(cls_label, $[1, 1, N]$) Expand labels to $B \times 16 \times N$
center_0 $\leftarrow$ permute(pts, $[0, 2, 1]$); $\text{f}_0 \leftarrow$ Concat($[\text{cls_onehot}, \text{center_0}], \text{dim} = 1$)
center_1 $\leftarrow$ permute(FPS(pts, 512), $[0, 2, 1]$); $\text{f}_1 \leftarrow$ center_1 Farthest Point Sampling
center_2 $\leftarrow$ permute(FPS(pts, 256), $[0, 2, 1]$); $\text{f}_2 \leftarrow$ center_2
center_3 $\leftarrow$ permute(center, $[0, 2, 1]$)
$\text{f}_3 \leftarrow$ feature_list[2]
$\text{f}_2 \leftarrow$ Propagate(center_2, center_3, $\text{f}_2$, feature_list[1]) Level 2 from Level 3
$\text{f}_1 \leftarrow$ Propagate(center_1, center_3, $\text{f}_1$, feature_list[0]) Level 1 from Level 3
$\text{f}_2 \leftarrow$ DGCNN_Fuse(center_3, $\text{f}_3$, center_2, $\text{f}_2$)
$\text{f}_1 \leftarrow$ DGCNN_Fuse(center_2, $\text{f}_2$, center_1, $\text{f}_1$)
$\text{f}_0 \leftarrow$ Propagate(center_0, center_1, $\text{f}_0$, $\text{f}_1$) Level 0 from Level 1
feat $\leftarrow$ ReLU(BN(Conv1($\text{f}_0$)))
feat $\leftarrow$ Dropout(feat)
x $\leftarrow$ Conv2(feat)
x $\leftarrow$ LogSoftmax(x, dim $= 1$)
x $\leftarrow$ permute(x, $[0, 2, 1]$) Final shape $B \times N \times K$
**return** x, $\text{f}_3$

human-object manipulation cues. In addition, we collect Internet videos to cover object categories, affordance types, and interaction styles that are under-represented in curated HOI datasets. We briefly introduce these sources below and discuss how they complement our video-based 3D affordance grounding benchmark.

### H.1 HOIGen-1M

HOIGen-1M [33] is a large-scale dataset designed for human-object interaction video generation. It contains over one million high-quality HOI videos collected from diverse sources. The motivation behind HOIGen-1M is that existing text-to-video models often struggle to generate physically plausible and semantically accurate HOI videos, largely because available training data lacks accurate interaction-centric captions and sufficient coverage of fine-grained human-object behaviors.

To improve data quality, HOIGen-1M uses an automatic curation pipeline based on multimodal large language models, followed by human cleaning. The dataset further introduces a Mixture-of-Multimodal-Experts captioning strategy to produce expressive HOI descriptions while reducing hallucinations from individual MLLMs. In addition to the dataset itself, HOIGen-1M proposes

evaluation metrics for generated HOI videos, aiming to assess video quality from coarse interaction correctness to finer interaction details.

### H.2 TASTE-Rob

TASTE-Rob [34] is a task-oriented hand-object interaction video dataset for robotic manipulation and video generation. It contains 100,856 egocentric hand-object interaction videos, each aligned with a language instruction. Unlike general Internet-scale HOI collections, TASTE-Rob emphasizes a consistent camera viewpoint and clear task-oriented interactions, which are important for imitation learning and robotic manipulation.

### H.3 Internet Videos

Besides existing HOI video datasets, we also collect videos from the Internet to improve the diversity and coverage of VIDA. This additional source is important because public HOI datasets are usually shaped by their original goals, such as video generation, action recognition, or robotic demonstration. As a result, some object categories, affordance types, camera viewpoints, and manipulation styles remain sparse. Internet videos provide a complementary long-tail source of natural interactions, including daily tool use, household manipulation, food preparation, appliance operation, and instructional demonstrations.

**Query construction.** We construct search queries according to the object categories and affordance labels in VIDA. For each target object category, we combine the object name with action verbs and task phrases that indicate functional use. For example, for objects with graspable regions, we search for phrases such as "hold mug", "pick up knife", or "grasp handle". For objects involving opening or closing, we use queries such as "open microwave door", "open cabinet", or "close drawer". For tool-use objects, we include task-level descriptions such as "cut with knife", "write with pen", "pour from bottle", or "use scissors". This query design helps retrieve videos where the interaction itself, rather than only the object appearance, is visible.

**Initial retrieval.** We retrieve candidate clips from public web video platforms and search results using category-action queries. We prefer videos that contain a single dominant manipulated object, visible human hands or tools, and a complete interaction process. Long videos are temporally trimmed into shorter clips when the relevant interaction only occupies a small segment. We discard clips where the target object appears only as background or where the action can be inferred only from narration rather than visible manipulation.

**Temporal trimming.** For each retained video, we identify the temporal segment that contains the key interaction. The selected segment typically includes three stages: approach before contact, contact or manipulation, and a short post-contact state. Keeping these stages is useful for affordance grounding because the model can observe not only the final object state but also how the human approaches and uses the functional region. Clips that only show a static object after the action has already happened are removed unless the contact region remains visually unambiguous.

**Quality filtering.** We apply several filtering criteria to ensure that Internet videos provide reliable supervision. First, the object must be visually identifiable and match one of the VIDA object categories. Second, the interaction must be physically plausible and directly related to one of our affordance labels. Third, the functional region should be visible for at least part of the clip. Fourth, videos with heavy motion blur, severe occlusion, extreme camera shake, very low resolution, or strong editing effects are removed. Fifth, if multiple objects are manipulated simultaneously and the target object cannot be determined, the clip is discarded.

**Affordance label assignment.** After filtering, we assign an affordance type to each Internet video. We first use an MLLM-assisted pipeline to generate an interaction-centric caption and propose a candidate affordance label. The caption emphasizes the manipulated object, human action, contact pattern, and likely functional region. Human annotators then verify whether the proposed label is supported by visible evidence. If the object category and action are correct but the affordance type is ambiguous, the sample is either reassigned after discussion or removed from the dataset.

Table 9: **Main Results with reproduced EGO-3D.**

| Setting | Metric | PFusion | IAGNet | LASO | GREAT | Seqafford | LMAfford3D | EGO-3D | **Ours** |
|---|---|---|---|---|---|---|---|---|---|
| Seen | *mIoU*↑ | 16.33 | 20.39 | 18.65 | 23.62 | 23.03 | 22.74 | 24.89 | **29.51** |
| | *AUC*↑ | 78.43 | 80.22 | 78.44 | 81.41 | 81.17 | 80.74 | 82.01 | **84.77** |
| | *SIM*↑ | 46.28 | 50.11 | 49.46 | 51.25 | 47.71 | 47.28 | 50.13 | **59.34** |
| | *MAE*↓ | 0.264 | 0.188 | 0.257 | 0.173 | 0.227 | 0.234 | 0.168 | **0.152** |
| Unseen | *mIoU*↑ | 7.270 | 7.970 | 7.410 | 8.220 | 8.070 | 8.110 | 8.180 | **13.21** |
| | *AUC*↑ | 56.69 | 68.97 | 69.21 | 70.19 | 65.53 | 66.42 | 68.91 | **74.62** |
| | *SIM*↑ | 34.05 | 34.85 | 33.77 | 35.08 | 32.40 | 33.61 | 33.22 | **43.17** |
| | *MAE*↓ | 0.371 | 0.277 | 0.288 | 0.269 | 0.286 | 0.278 | 0.274 | **0.247** |

**Why Internet videos matter.** Internet videos increase the diversity of VIDA in several ways. They provide natural camera viewpoints beyond the egocentric or curated settings of existing datasets, include more varied human strategies for completing the same task, and cover long-tail object-action combinations that may appear rarely in HOI generation datasets. They also help reduce overfitting to the visual style of any single source dataset. This is especially important for video-based 3D affordance grounding, where the model should learn general interaction principles such as contact evolution and functional region use, rather than source-specific biases.

# I More Experiments Results

As mentioned in the paper, the most related work to ours is EGO-3D, but it has not released the code, so we reproduced its code here and reported the results in the table 9. As shown in the table, our VideoAfford shows the best performance.

**Ablation on 3D backbone.** We conducted ablation experiments to study the influence of the 3D Vision Encoder backbone, and we investigated the performance of VideoAfford with some alternative backbones. As shown in the table. 10, Uni3D [76] performs better in this task due to its strong representation ability. Thus, we set it as our default 3D vision encoder backbone.

Table 10: Ablation study on the choices of 3D vision backbone.

| 3D Vision Backbone | *mIoU*↑ | *AUC*↑ | *SIM*↑ | *MAE*↓ |
|---|---|---|---|---|
| PointNet++ [80] | 22.11 | 79.21 | 49.23 | 0.192 |
| PointBert [81] | 23.19 | 81.34 | 51.13 | 0.184 |
| Uni3D [76] | **29.51** | **84.77** | **59.34** | **0.152** |

**Ablation about $\mathcal{R}_p$.** Here we conduct the experiments to explore the influence of the choices of $\mathcal{R}_p$.

Table 11: Ablation study on the choices of $\mathcal{R}_p$ .

| $\mathcal{R}_p$ | *mIoU*↑ | *AUC*↑ | *SIM*↑ | *MAE*↓ |
|---|---|---|---|---|
| 0.1 | 27.85 | 83.14 | 57.22 | 0.159 |
| 0.5 | **29.51** | **84.77** | **59.34** | **0.152** |
| 1 | 23.23 | 78.92 | 52.09 | 0.211 |

# J More Visualization Results

We provide more visualization results of VideoAfford in Figure 10, Figure 9, and Figure 11. These examples cover both seen and unseen settings and include different object categories and affordance types. The qualitative results show that VideoAfford can transfer interaction priors from HOI videos to 3D point clouds and localize compact functional regions. The visualizations also illustrate challenging cases where the model must infer the same affordance on geometrically different instances from the same category.

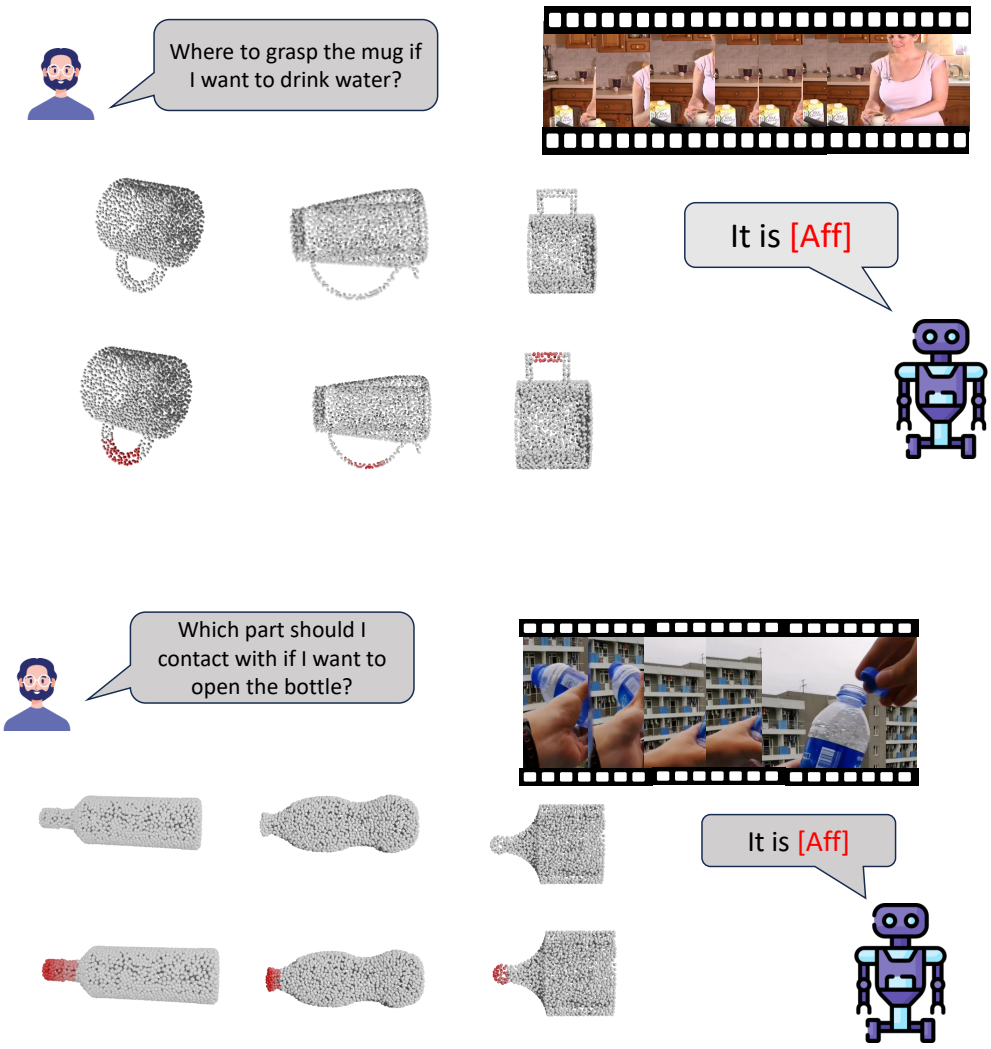



Figure 9: Videoafford can infer the corresponding affordance on the different 3D instances from the same category based on a demonstration video

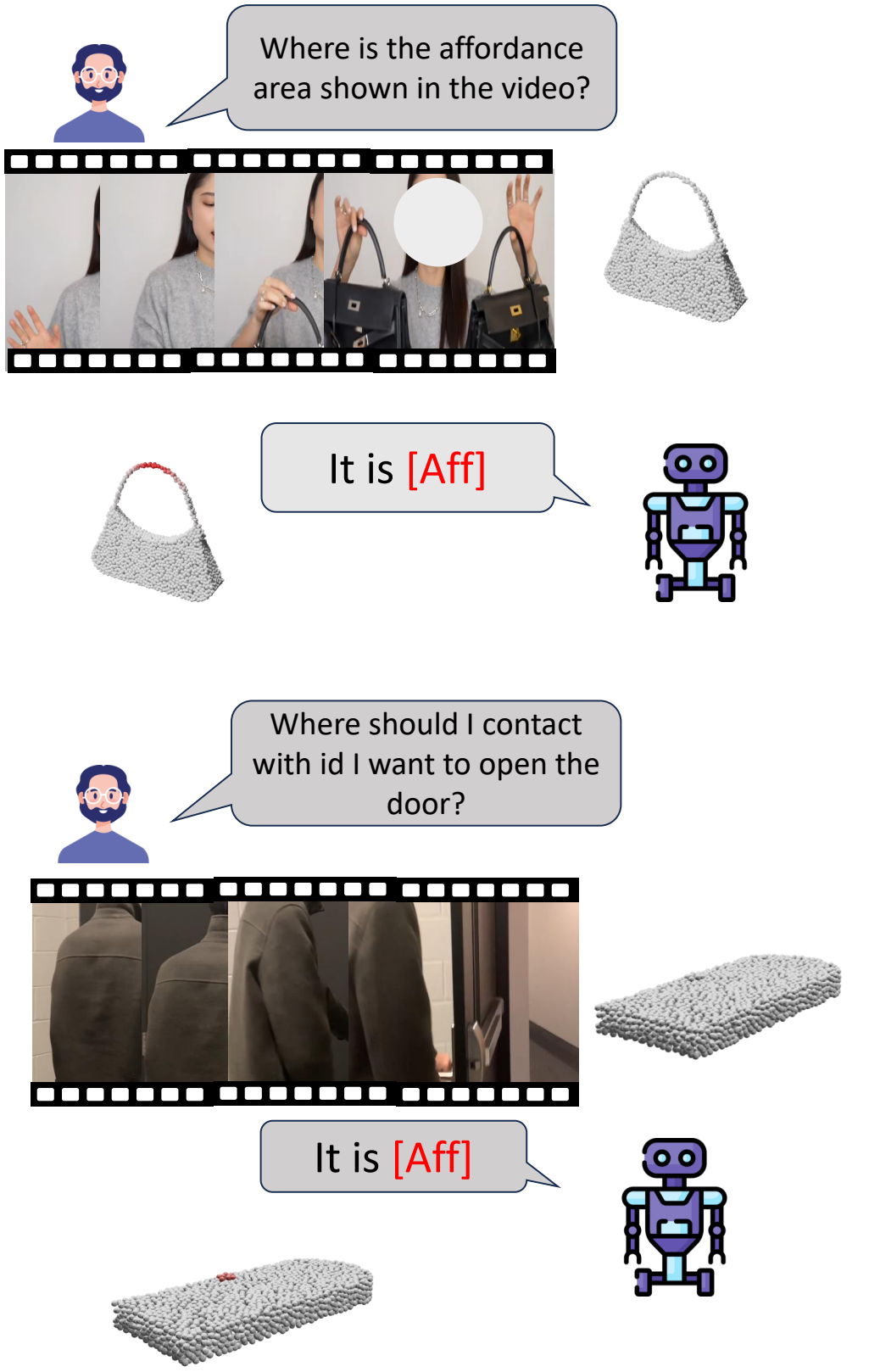





Figure 10: Visualization Results.

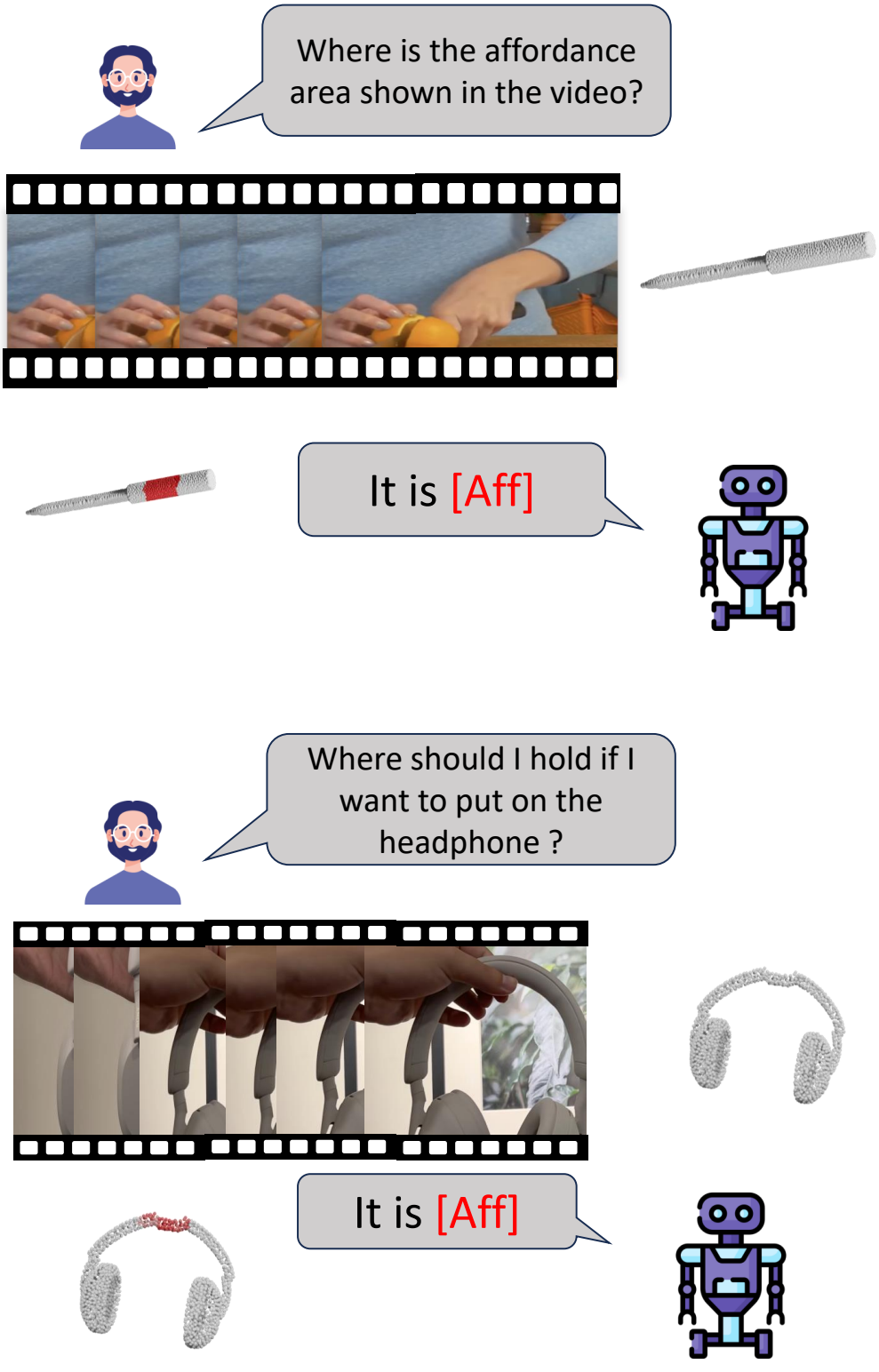



Figure 11: Visualization Results

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's main contributions: introducing the task of grounding 3D affordance from HOI videos, constructing the VIDA dataset, and proposing the VideoAfford model. These claims accurately reflect the experimental results and scope of the work.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated "Limitations and Future Directions" section that discusses several important limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper focuses on empirical machine learning research and does not include theoretical results, theorems, or formal proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed information about the experimental setup, including dataset construction, baseline methods, evaluation metrics, hyperparameters, and computational resources in the main paper and appendix. The VIDA dataset construction process is described in detail, and all experimental settings are fully specified.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The anonymous submission does not link a public bundle that contains the full training corpus, distilled trajectories, and fine-tuned checkpoints needed for bit-for-bit replication; we will release all training data and trained models together with code and README instructions in the open repository at camera-ready time after acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides comprehensive details about the experimental setup, including data splits (seen/unseen settings), hyperparameters (batch size, learning rate, optimizer), model architecture details, and training procedures in both the main paper and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The paper reports quantitative results with clear performance metrics but does not include error bars or statistical significance tests, as the focus is on demonstrating the effectiveness of the proposed method rather than rigorous statistical analysis across multiple runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix provides detailed information about computational resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research involves standard machine learning practices on publicly available datasets, with proper attribution and no involvement of human subjects or sensitive data collection.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses positive impacts through improved video affordance understanding for robotics and AI applications, while addressing negative impacts related to dataset biases and potential misuse in the limitations section.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The contribution is standard open-source research code and configuration, not a new high-risk standalone generative model or scraped dataset release in the sense targeted by this question, so no safeguards are required.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Datasets and models are cited from their sources; the software release notes third-party terms where relevant.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces the VIDA dataset and will be released or provide documentation alongside the assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: We are recruiting five human experts for label confirmation.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not applicable: no human-subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are part of the studied system; any LLM use for writing or polishing is ancillary and not a core methodological component requiring separate declaration.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.