# CAN 3D GEOMETRY-AWARE FEATURES HELP WITH AFFORDANCE PREDICTION?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Visual affordance prediction enables embodied agents to identify actionable objects and localize interaction regions from limited visual observations. However, existing affordance prediction methods typically rely on RGB appearance cues or explicit 3D representations. Since embodied agents often need to act before a complete 3D scene model is available, extracting useful geometric information directly from RGB observations is particularly valuable. To this end, we integrate reconstruction-pretrained geometry-aware features with visual representations without constructing an explicit 3D scene model. We evaluate this integration across multiple datasets, predictors, and prediction settings. Our experiments indicate that geometry-aware features can improve visual affordance prediction. Furthermore, the gains depend on data characteristics and view sampling and do not increase monotonically with the number of views. Notably, fine-grained geometric features provide greater benefits than global geometric features.

## 1 INTRODUCTION

Affordance refers to the potential actions that an agent can perform in interaction with an object or its surrounding environment (Gibson, 1977), and has become an important concept in physical AI (Nagarajan & Grauman, 2020; Wang et al., 2024c). Visual affordance prediction aims to infer these action possibilities from visual observations. It encompasses a range of tasks, including affordance classification, grounding, and segmentation, which respectively determine what actions an object supports, where the action-relevant regions are, and which pixels correspond to those regions (Apicella et al., 2025). Accurate prediction therefore requires both recognizing an object's function and identifying the spatial regions that support the intended interaction. In this work, we investigate how incorporating geometry-aware features improves visual affordance prediction from limited RGB observations.

Existing visual affordance prediction methods primarily rely on appearance, object semantics, part structure, and human–object interaction cues in RGB images Nagarajan et al. (2019); Goyal et al. (2022); Luo et al. (2022) to predict objects that support specific actions and their local interaction regions. Fully supervised methods typically depend on dense object or affordance annotations, whereas weakly supervised methods learn from category labels, interaction cues, or knowledge transferred between exocentric interaction images and target object images (Luo et al., 2022; 2024; Xu & MU, 2025; Li et al., 2023). These methods can effectively predict object regions and interaction locations, but their spatial reasoning mainly relies on appearance features and task supervision. In these cross-view transfer settings, the images need not depict the same scene or object instance. Such knowledge transfer differs from aggregating geometrically related observations of the same scene. Some RGB-based affordance prediction methods further introduce estimated single-view depth to provide spatial information (Li et al., 2025b; Qian et al., 2024; Zhang et al., 2026a). This offers a way to incorporate geometric cues from a single image, while multiple observations can additionally reveal changes in visibility, viewpoint, and object state. Understanding how geometry-aware features use these different sources of information is important for determining when and how they improve affordance prediction. Another line of work directly uses point clouds, reconstructed scenes, or other explicit 3D representations for affordance prediction Deng et al. (2021); Nguyen et al. (2023); Li et al. (2024b); Delitzas et al. (2024); Yang et al. (2024); Lu et al. (2025); Shao et al. (2025); Zhu et al. (2025); Zhang & Lee (2025); Li et al. (2025a). These representations provide explicit spatial structure, but they typically require geometric data or a reconstructed repre-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sentation as input. For embodied agents operating with limited observations, a complete 3D scene model may not be readily available. This motivates augmenting image-based affordance predictors with geometry-aware features extracted directly from RGB observations, without constructing an explicit 3D scene.

Geometry-aware features can support prediction under both single-image and multiple-observation settings (Hartley & Zisserman, 2004; Snavely et al., 2006; Agarwal et al., 2009; Yao et al., 2018; Ummenhofer et al., 2017; Wang et al., 2024a; Mildenhall et al., 2020; Wang et al., 2024b; Duisterhof et al., 2025; Wang & Agapito, 2025; Yang et al., 2025; Ye et al., 2024; Wang et al., 2025; 2026b). With a single image, they encode learned spatial priors, with auxiliary views or video frames, they also incorporate information from associated observations. Their usefulness depends on both the information available and how it is integrated into the predictor. We therefore investigate three questions: How do geometry-aware features improve visual affordance prediction? How do auxiliary observations affect different prediction outputs? How should geometry-aware features from different depths be selected and integrated?

As illustrated in Fig. 1, we augment existing affordance predictors (Luo et al., 2022; Xu & MU, 2025; Qian & Fouhey, 2023; Wu et al., 2025) with reconstruction-pretrained geometry-aware features extracted by a frozen VGGT-Ω encoder (Wang et al., 2026b) to address these questions. The encoder processes the target image together with any available auxiliary views of the same scene or frames from the same interaction sequence. We retain the target image's patch features and align their spatial dimensions with the predictor's feature grid. Here, target-view-aligned refers to preserving the target patch layout and matching feature-grid dimensions, rather than camera-based geometric registration. The resulting geometry-aware features are fused with visual features to produce predictions for the target image, without constructing an explicit 3D scene. We evaluate this geometry-aware augmentation on 3DOI (Qian & Fouhey, 2023), HANDAL (Guo et al., 2023; Wu et al., 2025), and AGD20K (Luo et al., 2022), covering scene-level images and videos, object-centric multi-view segmentation, and single-view weakly supervised grounding. We retain each predictor's task definition and training objectives, while specifying the training and evaluation protocol for each benchmark. Within each benchmark, compared configurations use the same data split. Comparisons among visual features, geometry-aware features, visual multi-view aggregation, and their fusion examine the benefits of geometry-aware augmentation and auxiliary observations. We further compare single-layer late fusion, multi-layer late fusion, and progressive fusion to study feature selection and fusion placement.

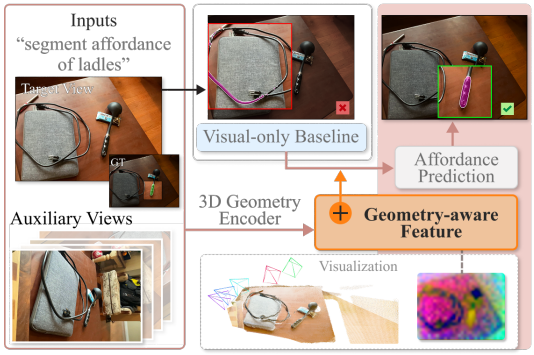


Figure 1: **Enhancing visual affordance prediction with geometry-aware features.** A visual foundation model misidentifies a cable as the ladle's interaction region (top). Fusing visual features with geometry-aware features extracted from the target and auxiliary views correctly localizes the ladle handle (bottom). The reconstruction is shown only for visualization, while affordance prediction uses intermediate features without constructing an explicit 3D scene.

Our main contributions are summarized as follows: (1) We show that geometry-aware features improve visual affordance prediction across multiple datasets and predictors, even with a single image and without explicit 3D scene reconstruction. (2) We show that the benefits of auxiliary observations depend on observation content, sampling strategy, and prediction output, rather than increasing monotonically with the number of views. (3) We find that shallow-layer features perform better in the single-layer comparison on HANDAL, while multi-layer late fusion effectively integrates features from different depths.

# 2 RELATED WORKS

## 2.1 VISUAL AFFORDANCE UNDERSTANDING

Visual affordance prediction combines recognizing object functions with localizing interaction regions. Fully supervised methods formulate the task as instance-level affordance segmentation, but require dense annotations and predefined categories (Do et al., 2018; Wu et al., 2025). Later approaches learn functional cues from human-object interaction videos (Nagarajan et al., 2019; Goyal et al., 2022; Wang et al., 2026a) or transfer interaction knowledge from exocentric images to target-object images (Luo et al., 2022; Li et al., 2023; Luo et al., 2024; Xu & MU, 2025; Li et al., 2025b). Vision-language models further introduce pretrained semantic priors for affordance reasoning (Chen et al., 2024; Qian et al., 2024; Li et al., 2024a; Zhang et al., 2026a). Although effective, these methods primarily exploit appearance, interaction, or semantic cues and do not explicitly model geometric correspondences among observations of the same scene.

In contrast, 3D affordance methods typically ground predictions in explicit 3D representations. Research has progressed from human-pose prediction in indoor environments to point-wise affordance prediction for objects and scenes, covering complete and partial shapes as well as rotation-invariant, open-vocabulary, and language-conditioned settings (Li et al., 2019; Deng et al., 2021; Nguyen et al., 2023; Li et al., 2024b; Delitzas et al., 2024). Egocentric interaction videos further extended affordance modeling to 3D human–environment relations, while language, interaction images, and vision foundation models improve semantic conditioning, intent reasoning, and robustness (Yang et al., 2024; Liu et al., 2024; Lu et al., 2025; Shao et al., 2025; Zhu et al., 2025). Complementary approaches introduce geometric cues without requiring complete 3D representations (Qian & Fouhey, 2023; Qian et al., 2024; Zhang et al., 2026b). Nevertheless, existing methods either rely on explicit 3D inputs or do not disentangle the contribution of geometry-aware representations from that of additional observations, and they generally do not explicitly model cross-view correspondences within the same scene. We therefore investigate whether frozen reconstruction-pretrained features can transfer useful geometric information to RGB affordance predictors without constructing an explicit 3D scene at inference time.

## 2.2 MULTI-VIEW GEOMETRIC REPRESENTATIONS

Multi-view geometric representation methods seek a scene-consistent explanation from multiple observations. Classical approaches rely on structure-from-motion and multi-view stereo, while learning-based methods predict geometry directly or make the reconstruction process differentiable (Hartley & Zisserman, 2004; Snavely et al., 2006; Agarwal et al., 2009; Schönberger & Frahm, 2016; Yao et al., 2018; Ummenhofer et al., 2017; Teed & Deng, 2020; 2021; Wang et al., 2024a). Neural implicit methods instead represent scenes with continuous radiance or signed-distance fields (Mildenhall et al., 2020; Wang et al., 2021). More recent feed-forward models infer geometric structure directly from one or more images and expose intermediate features for downstream perception (Wang et al., 2024b; Duisterhof et al., 2025; Wang & Agapito, 2025; Yang et al., 2025). Among them, VGGT jointly predicts multiple geometric quantities from one or more images (Wang et al., 2025), while VGGT-$\Omega$ further aggregates cross-view information (Wang et al., 2026b). We use its frozen intermediate features to test whether reconstruction knowledge can improve affordance localization by providing single-image geometric priors and multi-view context, without constructing an explicit 3D scene at inference time.

# 3 METHOD

## 3.1 TASK FORMULATION

We consider visual affordance prediction from a target RGB image $I_0$, an optional set of auxiliary observations

$$\mathcal{A}_K = \{I_i\}_{i=1}^{K},$$

and a task condition $q$. An overview of the proposed framework is shown in Fig. 2. The auxiliary observations are views of the same static scene or frames from the same interaction sequence. The

predictor produces outputs for the target image:

$$\widehat{\mathcal{Y}}_0 = f\left(I_0, \mathcal{A}_K, q; \theta\right), \tag{1}$$

where $\theta$ denotes the trainable parameters and $q$ specifies a task-specific condition, such as an instance query point (Qian & Fouhey, 2023), affordance labels (Luo et al., 2022; Li et al., 2024a) or a textual instruction (Wu et al., 2025). The condition is omitted when unnecessary. The output $\widehat{\mathcal{Y}}_0$ may include affordance predictions and other task-specific outputs, such as object regions or interaction attributes. When $K = 0$, the predictor uses only the target image. When $K > 0$, the auxiliary observations provide additional context without requiring extra task annotations. In all cases, prediction uses RGB observations and does not construct an explicit 3D scene.

### 3.2 Geometry-Aware Feature Extraction

We extract visual and geometry-aware features using a task-specific visual backbone $E_{\mathrm{v}}$ and a frozen, reconstruction-pretrained VGGT-$\Omega$ encoder $E_{\mathrm{g}}$ (Wang et al., 2026b). The visual backbone processes only the target image:

$$F_{\mathrm{v},0}^{(\ell)} = E_{\mathrm{v}}^{(\ell)}(I_0), \qquad \ell \in \mathcal{S}_{\mathrm{v}}, \tag{2}$$

where $\mathcal{S}_{\mathrm{v}}$ denotes the selected visual layers. For Transformer backbones, patch tokens are reshaped into two-dimensional spatial feature maps.

The geometry-aware encoder processes the target image and its auxiliary observations, $(I_0, \ldots, I_K)$, with $I_0$ placed first. At each selected block $m$, it produces token states $H_{\mathrm{frame}}^{(m)}$ and $H_{\mathrm{inter}}^{(m)}$ after frame attention and subsequent inter-frame attention, respectively. We concatenate these states and extract the target-image features as

$$F_{\mathrm{g},0}^{(m)} = \mathcal{R}_0\left(\left[H_{\mathrm{frame}}^{(m)} \| H_{\mathrm{inter}}^{(m)}\right]\right), \qquad m \in \mathcal{S}_{\mathrm{g}}, \tag{3}$$

where $\mathcal{S}_{\mathrm{g}}$ denotes the selected geometry blocks and $\|$ denotes channel-wise concatenation. The operator $\mathcal{R}_0$ retains the target-image patch tokens, removes camera and register tokens, and restores their two-dimensional spatial layout. When $K = 0$, the resulting features encode geometric priors learned from reconstruction pretraining. When $K > 0$, they additionally aggregate information from the auxiliary observations while preserving the target-image patch layout. Thus, target-view alignment refers to layout correspondence rather than camera-based geometric registration.

### 3.3 Feature Adaptation and Fusion

We compare single-layer and multi-layer late fusion to study feature selection, and late and progressive fusion to study the placement of geometry-aware features.

**Feature adaptation.** We use branch- and layer-specific residual adapters to map visual and geometry-aware features to a common target-image feature grid. For multi-layer fusion, adapted geometry-aware features are aggregated before fusion. The adaptation operates on target-aligned feature maps and requires neither geometric warping nor explicit 3D projection. Detailed adapter architectures are provided in Appendix B.

**Late fusion.** Single-layer late fusion uses one geometry feature level, whereas multi-layer late fusion aggregates multiple levels at the same prediction interface. Let $V$ and $G$ denote the visual and adapted geometry-aware representations at this interface. We define the fused representation as

$$Z = \begin{cases} \Phi_{\mathrm{fuse}}([V\|G]), & \text{3DOI}, \\ V + P_{\mathrm{H}}(G), & \text{HANDAL}, \\ V + \Delta_{\mathrm{A}}(V, G), & \text{AGD20K}, \end{cases} \tag{4}$$

where $\|$ denotes channel-wise concatenation. The predictor-specific fusion interfaces are described in Appendix B. In brief, 3DOI combines the two feature branches through a fusion block, HANDAL projects the geometry features to the SAM embedding dimension, and AGD20K uses a residual fusion module. The fusion outputs are zero-initialized where applicable so that the original predictors are preserved at initialization. we treat four-layer late fusion as the primary configuration and progressive fusion as a comparison or ablation setting.

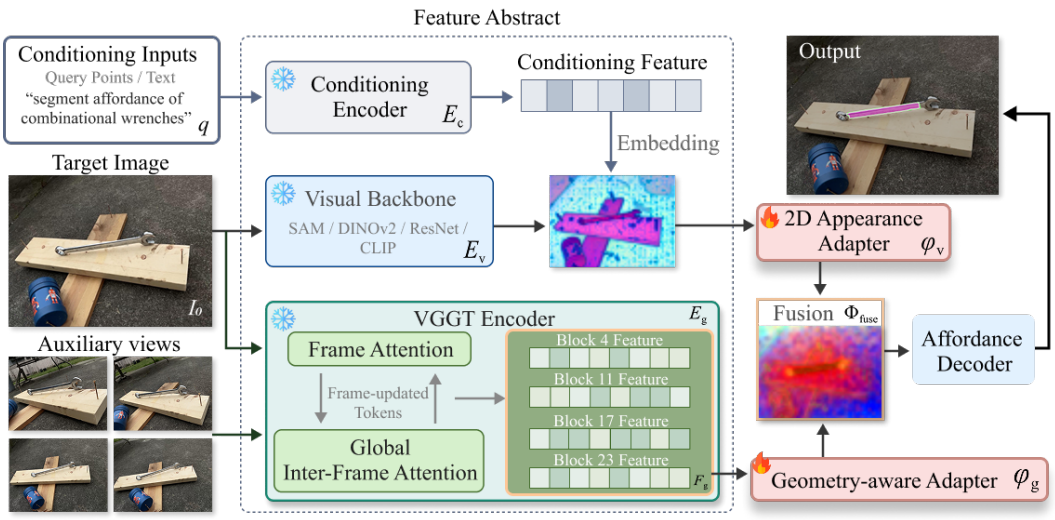


Figure 2: Overview of our geometry-aware affordance prediction framework. Given a target image $I_0$, optional auxiliary observations $\mathcal{A}_K$, and a task condition $q$, the frozen visual backbone $E_v$ encodes $I_0$, while the frozen VGGT-Ω encoder $E_g$ extracts target-aligned geometry-aware features from $(I_0, \mathcal{A}_K)$. The adapted features are fused to produce task-specific predictions for $I_0$ without constructing an explicit 3D scene.

**Progressive fusion.** Progressive fusion injects geometry-aware features into intermediate visual backbone stages:

$$X_{j+1} = B_{j+1}(X_j + \Delta_j(X_j, G_j)), \qquad j = 0, \ldots, J-1, \tag{5}$$

where $X_j$ is the visual feature before the $j$-th injection, $G_j$ is the adapted geometry-aware feature at the corresponding spatial grid, and $\Delta_j$ is a residual fusion module. The subsequent backbone segment $B_{j+1}$ processes the fused representation, allowing geometry-aware features to influence later visual computation. Detailed projection, normalization, and initialization settings are given in Appendix B.

### 3.4 Training and Inference

**Training.** We freeze the pretrained visual (Qian & Fouhey, 2023; Wu et al., 2025; Luo et al., 2022; Xu & MU, 2025; Li et al., 2024a) and geometry-aware encoders (Wang et al., 2026b; Oquab et al., 2023) and optimize the added adapters and fusion modules. The original downstream modules are also trained for 3DOI (Qian & Fouhey, 2023), whereas the HANDAL (Wu et al., 2025) and AGD20K (Luo et al., 2022; Li et al., 2024a; Xu & MU, 2025) predictors remain frozen.

**Inference.** At inference time, the predictor receives $I_0$, the required task condition $q$, and the auxiliary observations specified by each evaluation protocol. Predictions use intermediate geometry-aware features directly, without executing the geometric prediction heads of VGGT-Ω or constructing an explicit 3D scene.

## 4 Experiments

We evaluate geometry-aware feature augmentation on three affordance benchmarks with distinct task settings. 3DOI covers scene-level images and manipulation videos (Qian & Fouhey, 2023). HANDAL focuses on object-centric affordance segmentation (Guo et al., 2023; Wu et al., 2025). And AGD20K addresses single-view, weakly supervised affordance grounding Luo et al. (2022). Across these benchmarks, we systematically examine the effects of auxiliary observations, feature depth, and fusion strategy while preserving the original task definitions and evaluation protocols. Implementation details for 3DOI, HANDAL, and AGD20K are provided in Appendix B, while details of dataset processing are provided in Appendix A.

Table 1: Object localization and affordance prediction on 3DOI-Taskonomy. Input configurations denote one target image with zero or two auxiliary views. All predictions are evaluated on the target image. Arrows indicate the preferred direction of each metric. Darker = better within each metric.

| Method | Views | Movable | Localization | | Action | Affordance | | |
|---|---|---|---|---|---|---|---|---|
| | | Cat. ↑ | Box ↑ | Mask ↑ | Cat. ↑ | SIM ↑ | Distance ↓ | KLD ↓ |
| 3DOI-DINOv2 (ViT-B) Qian & Fouhey (2023); Oquab et al. (2023) | 1 | 0.8522 | 0.6018 | 0.6019 | 0.9220 | 0.1282 | 0.0502 | 3.5312 |
| + Geometry-aware Feature | 1 | 0.8423 | 0.6704 | 0.7203 | 0.9290 | 0.1312 | 0.0415 | 2.9833 |
| | $1+2$ | 0.8589 | 0.6820 | 0.7278 | 0.9322 | 0.1332 | 0.0418 | 3.0011 |

Table 2: Object localization and affordance prediction on 3DOI-Articulation, a third-person manipulation video subset. Input configurations denote one target frame with zero, two, or six auxiliary frames. All predictions are evaluated on the target frame. Arrows indicate the preferred direction of each metric. Darker = better within each metric.

| Method | Views | Movable | Localization | | Action | Affordance | | |
|---|---|---|---|---|---|---|---|---|
| | | Cat. ↑ | Box ↑ | Mask ↑ | Cat. ↑ | SIM ↑ | Distance ↓ | KLD ↓ |
| 3DOI-DINOv2 (ViT-B) Qian & Fouhey (2023); Oquab et al. (2023) | 1 | 0.8846 | 0.6102 | 0.6060 | 0.9105 | 0.0700 | 0.0730 | 4.3438 |
| | 1 | 0.9246 | 0.6994 | 0.7451 | 0.9312 | 0.0937 | 0.0630 | 3.6777 |
| + Geometry-aware Feature | $1+2$ | 0.9127 | 0.7054 | 0.7346 | 0.9280 | 0.1015 | 0.0646 | 3.5374 |
| | $1+6$ | 0.9253 | 0.7160 | 0.7415 | 0.9379 | 0.1172 | 0.0597 | 3.4299 |

### 4.1 Scene-Level Images and Videos

We evaluate geometry-aware features on 3DOI (Qian & Fouhey, 2023), which contains static indoor images, third-person manipulation videos, and first-person kitchen videos. The original 3DOI framework uses a SAM visual backbone (Kirillov et al., 2023). For the main experiments, we replace it with a frozen DINOv2 ViT-B backbone (Oquab et al., 2023) to test whether geometry-aware features transfer beyond a segmentation-specialized encoder. Results with the original SAM backbone are reported in Table 7. They show the same overall trend, with geometry-aware features improving the SAM-based baseline. Following the original task definitions and objectives, we freeze the visual and geometry-aware encoders and train the adapters, fusion modules, and downstream task modules. All predictions and supervision correspond to the target image $I_0$. Auxiliary observations are sampled from the same scene or video without additional task annotations, and $I_0$ is always provided as the first input to VGGT-$\Omega$. Qualitative comparisons of object masks and affordance heatmaps across the three subsets are provided in Appendix C.1. Additional 3DOI comparisons with a SAM backbone, multi-view DINOv2 feature stacking, and VGGT-$\Omega$ as a standalone backbone are reported in Tables 7, 8, and 9 in the appendix, respectively. Implementation details for the 3DOI experiments are provided in Appendix B.1.

**Experiments on 3DOI-Taskonomy.** On static indoor scenes, we compare target-only input (1) with two auxiliary views ($1+2$), following the sampling protocol in Appendix A.1. Table 1 shows that single-view geometry-aware fusion improves object localization and all affordance heatmap metrics over DINOv2 (Oquab et al., 2023). Adding two auxiliary views further improves Box IoU, Mask IoU, and SIM, but slightly worsens Peak Distance and KLD. Thus, substantial gains already arise from a single image, while auxiliary views yield smaller, metric-dependent benefits. Qualitative results are presented in Fig. 4.

**Experiments on 3DOI-Articulation.** On third-person manipulation videos, we compare target-only input (1) with two and six auxiliary frames ($1+2$ and $1+6$), following the sampling protocol in Appendix A.2. Table 2 shows that geometry-aware fusion improves object movable, localization, action and affordance heatmap prediction over DINOv2. Increasing the number of input frames from one to seven raises SIM by roughly one quarter and reduces Peak Distance and KLD, while the single-view setting yields the highest Mask IoU. These results indicate that geometry-aware features are effective without auxiliary frames, whereas additional temporal context primarily benefits affordance heatmap prediction. Qualitative results are presented in Fig. 5.

Table 3: Object localization and affordance prediction on 3DOI-EpicKitchen, a first-person kitchen manipulation video subset. Input configurations denote one target frame with zero, two, or six auxiliary frames. All predictions are evaluated on the target frame. Arrows indicate the preferred direction of each metric. Darker = better within each metric.

| Method | Views | Movable | Localization | | Action | Affordance | | |
|---|---|---|---|---|---|---|---|---|
| | | Cat. ↑ | Box ↑ | Mask ↑ | Cat. ↑ | SIM ↑ | Distance ↓ | KLD ↓ |
| 3DOI-DINOv2 (ViT-B) Qian & Fouhey (2023); Oquab et al. (2023) | 1 | 0.8703 | 0.6283 | 0.6140 | 0.9799 | 0.0410 | 0.0765 | 4.3815 |
| + Geometry-aware Feature | 1 | 0.8922 | 0.7580 | 0.7784 | 0.9839 | 0.0456 | 0.0644 | 4.0483 |
| | $1+2$ | 0.8693 | 0.6871 | 0.7664 | 0.9809 | 0.0470 | 0.0630 | 4.0325 |
| | $1+6$ | 0.8882 | 0.7241 | 0.7765 | 0.9809 | 0.0628 | 0.0632 | 3.7309 |

Table 4: Test-set performance (%). Higher is better. Bold and underlined values indicate the best and second-best results among the geometry-aware configurations, respectively.

| Method | Views | GIoU ↑ | cIoU ↑ |
|---|---|---|---|
| VLPart (Sun et al., 2023) + SAM2 (Ravi et al., 2024) | 1 | 40.9 | 28.9 |
| Grounding DINO (Liu et al., 2023b) + SAM2 (Ravi et al., 2024) | 1 | 34.7 | 26.8 |
| Florence 2 (Xiao et al., 2024)+ SAM2 (Ravi et al., 2024) | 1 | 39.7 | 22.4 |
| LISA (Lai et al., 2023) | 1 | 16.2 | 12.0 |
| GLaMM (Rasheed et al., 2024) | 1 | 24.9 | 17.2 |
| AffordanceNet (Wu et al., 2025) | 1 | 60.3 | 60.8 |
| AffordanceNet (our split) (Wu et al., 2025) | 1 | 31.0 | 13.3 |
| + Geometry-aware Feature | 1 | 67.5 | 69.7 |
| | $1+2$ | <u>68.1</u> | <u>71.0</u> |
| | $1+4$ | **68.7** | **71.7** |
| | $1+6$ | 67.3 | 69.9 |

**Experiments on 3DOI-EpicKitchen.** On first-person kitchen manipulation videos, we compare target-only input (1) with two and six auxiliary frames ($1+2$ and $1+6$), following the sampling protocol in Appendix A.2. The videos involve camera motion, hand occlusion, and changing object states, resulting in substantial variation across observations. As shown in Table 3, single-frame geometry-aware fusion improves all five spatial prediction metrics over 3DOI-DINOv2 (Qian & Fouhey, 2023; Oquab et al., 2023). However, increasing the number of views does not consistently improve the metrics on 3DOI-EpicKitchen dataset. Qualitative results are presented in Fig. 6.

### 4.2 Object-Centric Multi-View Images

RAGNet-HANDAL provides object-centric observations of the same object instance (Guo et al., 2023; Wu et al., 2025). We compare target-only input (1) with two, four, and six auxiliary views ($1+2$, $1+4$, and $1+6$), while keeping the target images fixed across configurations. The observation sampling and boundary-handling procedures are described in Appendix A.3. Implementation details for the RAGNet-HANDAL experiments are provided in Appendix B.2.

We use multi-layer late fusion to inject VGGT-$\Omega$ features from blocks $\{4, 11, 17, 23\}$ into the SAM ViT-H (Li et al., 2025c) image embeddings of AffordanceVLM (Wu et al., 2025). The original predictor and geometry-aware encoder remain frozen, and only the adapters and fusion modules are trained. As shown in Table 4, adding two auxiliary views further improves both metrics, with the best performance obtained using four auxiliary views. Increasing the number of auxiliary views to six reduces both scores. These results show that geometry-aware features substantially improve object-centric affordance segmentation, while a moderate number of auxiliary views provides additional but non-monotonic gains. Qualitative results are presented in Fig. 8.

### 4.3 Cross-View Weakly Supervised

We evaluate geometry-aware features on AGD20K (Luo et al., 2022) using Cross-View-AG (Luo et al., 2022) and WSAG-PLSP (Xu & MU, 2025). We retain their original training objectives and

Table 5: Affordance grounding on AGD20K with different geometry-aware fusion configurations. Arrows indicate the preferred direction of each metric. Best results within the Cross-View-AG, WSAG-PLSP, and OOAL groups are bold.

| Predictor | Fusion configuration | Seen | | | Unseen | | |
|---|---|---|---|---|---|---|---|
| | | mKLD ↓ | mSIM ↑ | mNSS ↑ | mKLD ↓ | mSIM ↑ | mNSS ↑ |
| Hotspots (Nagarajan et al., 2019) | None | 1.7730 | 0.2780 | 0.6150 | 1.9940 | 0.2370 | 0.5770 |
| Cross-View-AG+ (Luo et al., 2024) | None | 1.4890 | 0.3420 | 0.9810 | 1.7650 | 0.2790 | 0.8820 |
| LOCATE (Li et al., 2023) | None | 1.2260 | 0.4010 | 1.1770 | 1.4050 | 0.3720 | 1.1570 |
| OOAL (Li et al., 2024a) | None | 0.7978 | 0.5649 | 1.6773 | 1.1894 | **0.4454** | 1.3545 |
| | 4 layer stage-wise fusion | **0.7595** | **0.5824** | **1.7287** | **1.1480** | 0.4433 | **1.4040** |
| Cross-View-AG (Luo et al., 2022) | None | 1.5080 | 0.3440 | 0.9350 | 1.8490 | 0.2720 | 0.7170 |
| | Single-layer late fusion | 1.4960 | 0.3460 | 0.9480 | 1.8390 | 0.2760 | 0.7450 |
| | 4 layer stage-wise fusion | 1.5000 | 0.3410 | 0.9500 | 1.8020 | 0.2800 | 0.8150 |
| | 4 layer late fusion | **1.4510** | **0.3570** | **0.9950** | **1.7710** | **0.2890** | **0.8160** |
| WSAG-PLSP (Xu & MU, 2025) | None | 0.8900 | 0.5100 | 1.5470 | 1.1530 | 0.4370 | 1.4180 |
| | Single-layer late fusion | 0.8760 | 0.5100 | 1.5700 | **1.1320** | 0.4510 | **1.4410** |
| | 4 layer stage-wise fusion | 0.8560 | 0.5240 | 1.5770 | 1.1520 | 0.4460 | 1.4270 |
| | 4 layer late fusion | **0.8520** | **0.5280** | **1.5840** | 1.1460 | **0.4590** | 1.4250 |

Table 6: Effect of geometry-aware feature depth on HANDAL. Results are reported in percentages. Higher is better. Best results are bold.

| Configuration | Test gIoU ↑ | Test cIoU ↑ |
|---|---|---|
| AffordanceNet Baseline (Wu et al., 2025) | 31.0 | 13.3 |
| + VGGT-Ω feature 4 Wang et al. (2026b) | 67.4 | 69.6 |
| + VGGT-Ω feature 11 Wang et al. (2026b) | 66.7 | 69.0 |
| + VGGT-Ω feature 17 Wang et al. (2026b) | 63.3 | 65.0 |
| + VGGT-Ω feature 23 Wang et al. (2026b) | 61.1 | 62.1 |
| + VGGT-Ω features 4/11/17/23 Wang et al. (2026b) | **67.5** | **69.7** |

extract geometry-aware features only from the target ego image, while exo images remain part of training (Appendix A.4). We compare single-layer late fusion, multi-layer late fusion, and progressive fusion on the Seen and Unseen object-category splits. Implementation details for the AGD20K experiments are provided in Appendix B.3.

As shown in Table 5, multi-layer late fusion improves grounding performance across both splits and predictors and generally performs better than progressive fusion. On the Seen split, adding geometry-aware features improves the prediction performance of all three baseline methods. On the Unseen split, however, the gains tend to be weaker and may be less consistent. These results show that geometry-aware features enhance single-view weakly supervised affordance grounding and that integrating multiple feature levels at the final prediction interface provides an effective fusion strategy.

## 5 Discussion

**How do geometry-aware features improve affordance prediction?** Geometry-aware features improve affordance localization across multiple datasets and predictors. These gains already appear with a single image, showing that geometry-aware features enhance the spatial representation of the target image. Auxiliary observations provide additional scene or interaction information. Geometry-aware features thus support prediction by capturing spatial cues within a single image and incorporating information from associated observations. Figure 3 presents representative examples from 3DOI-Taskonomy. The reconstructions and predictions illustrate the relationship between object structure, local details, and interaction-region localization. Reconstructions are used for qualitative analysis. Affordance prediction directly uses geometry-aware features.

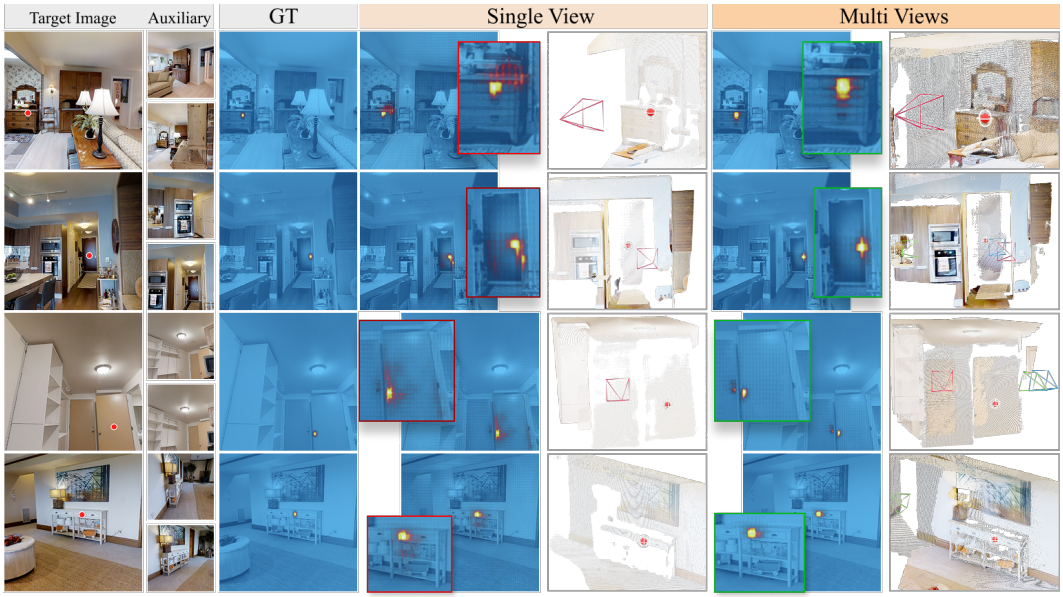


Figure 3: Quantitative comparison of single-view and multi-view geometry-aware affordance prediction on 3DOI-Taskonomy. In the single-view setting, AffordanceNet uses geometry-aware features from the target image only. In the multi-view setting, it uses features from the target image and auxiliary views. In both settings, prediction does not require explicit 3D reconstruction. Instead, point clouds are shown only to visualize the geometry-aware features.

**Do more views always help the performance of affordance prediction?** The benefits of auxiliary observations are task- and data-dependent rather than monotonic. This may be because static views primarily provide structural and occlusion cues, whereas first-person video frames also capture contact and action-state changes. However, aggregating too many observations can introduce noise, causing performance to peak at an intermediate number of views or decline. These findings suggest that view selection should match both the prediction target and the information provided by the observations. Qualitative comparisons in Appendix A.4 and Figs. 8 suggest that input views providing complementary information for occluded regions and covering near, intermediate, and distant views of the target object are more likely to yield better predictions.

**How do features from different geometry layers behave?** Table 6 compares features from VGGT-Ω blocks 4, 11, 17, and 23 on HANDAL, using zero-based indices. Each single-layer late fusion configuration combines features from one block with SAM features. Block 4 performs best among the single-layer configurations, and performance decreases for the deeper blocks tested. Multi-layer late fusion combines all four blocks and achieves the highest scores. These results support the complementary value of features from different depths. The advantage of shallow features suggests that local structure and boundary details are useful for affordance localization. Multi-layer aggregation provides an effective way to combine these details with deeper contextual information. Qualitative comparisons across feature depths are provided in Figure 7 in the appendix.

## 6 Conclusion

We studied whether reconstruction-pretrained geometry-aware features can improve visual affordance prediction without explicit 3D reconstruction. Augmenting existing predictors with target-aligned VGGT-Ω features improved affordance localization across scene-level, object-centric, and weakly supervised benchmarks, with gains often emerging from a single image. Our analyses further indicate that the effects of auxiliary observations depend on the task and data. Additional views can help but may introduce aggregation noise, and feature depth also matters, with shallow and multi-layer features providing complementary benefits. These results support geometry-aware feature augmentation as a practical bridge between RGB affordance prediction and 3D reasoning.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with the formulation and presentation of mathematical claims and their proofs, as well as to improve the clarity, grammar, and readability of the manuscript. We also used generative AI tools for literature retrieval and discovery of potentially relevant prior work. All AI-assisted mathematical content was independently checked and verified by the authors. In particular, the assumptions, derivations, proofs, and final statements were reviewed by the authors, who take full responsibility for their correctness. All references suggested or retrieved with AI assistance were verified against the original publications. All AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including all claims, proofs, citations, and other materials produced with the aid of generative AI.

### ETHICS STATEMENT

This work uses publicly available datasets and does not involve the collection of personally identifiable information or direct interaction with human subjects. We identify no immediate significant ethical concerns.

### REPRODUCIBILITY STATEMENT

The data processing procedures and experimental details are described in the main paper and supplementary material. The source code will be publicly released upon acceptance.

## REFERENCES

Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. Building rome in a day. In *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, pp. 72–79, 2009. doi: 10.1109/ICCV.2009.5459148.

Tommaso Apicella, Alessio Xompero, and Andrea Cavallaro. Visual affordance prediction: Survey and reproducibility. *arXiv preprint arXiv:2505.05074*, 2025. URL `https://arxiv.org/abs/2505.05074`.

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*, pp. 822–828. IEEE, 2024.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert W. 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 (CVPR)*, pp. 14531–14542, 2024.

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

Thanh-Toan Do, Anh Nguyen, and Ian Reid. AffordanceNet: An end-to-end deep learning approach for object affordance detection. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 5882–5889, 2018. doi: 10.1109/ICRA.2018.8460902.

Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jérôme Revaud. MASt3R-SfM: A fully-integrated solution for unconstrained structure-from-motion. In *2025 International Conference on 3D Vision (3DV)*, pp. 1–10, 2025. doi: 10.1109/3DV66043.2025.00008.

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10786–10796, 2021.

James J. Gibson. The theory of affordances. In Robert Shaw and John Bransford (eds.), *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, pp. 67–82. Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.

Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3293–3303, June 2022.

Andrew Guo, Bowen Wen, Jianhe Yuan, Jonathan Tremblay, Stephen Tyree, Jeffrey Smith, and Stan Birchfield. HANDAL: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions. In *IROS*, 2023.

Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2004. doi: 10.1017/CBO9780511811685.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

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*, 2025a.

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 (CVPR)*, pp. 10922–10931, June 2023.

Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3086–3096, June 2024a.

Gen Li, Nikolaos Tsagkas, Jifei Song, Ruaridh Mon-Williams, Sethu Vijayakumar, Kun Shao, and Laura Sevilla-Lara. Learning precise affordances from egocentric videos for robotic manipulation. In *ICCV*, pp. 10581–10591, 2025b.

Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3D indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12368–12376, 2019.

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*, pp. 14251–14260, 2024b.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10486–10496, 2025c.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023b.

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

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*, pp. 2252–2261, 2022.

Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Grounded affordance from exocentric view. *International Journal of Computer Vision*, 132(6):1945–1969, 2024. doi: 10.1007/s11263-023-01962-z.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Computer Vision – ECCV 2020*, pp. 405–421, 2020. doi: 10.1007/978-3-030-58452-8_24.

Tushar Nagarajan and Kristen Grauman. Learning affordance landscapes for interaction exploration in 3D environments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2005–2015, 2020.

Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Toan Nguyen, Minh Nhat Vu, An Vuong, Dzung Nguyen, Thieu Vo, Ngan Le, and Anh Nguyen. Open-vocabulary affordance detection in 3D point clouds. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5692–5698, 2023. doi: 10.1109/IROS55552.2023.10341553.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Shengyi Qian and David F. Fouhey. Understanding 3D object interaction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21753–21763, October 2023.

Shengyi Qian, Linyi Jin, Chris Rockwell, Siyi Chen, and David F. Fouhey. Understanding 3d object articulation in internet videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1599–1609, June 2022.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL `https://proceedings.mlr.press/v139/radford21a.html`.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4104–4113, 2016. doi: 10.1109/CVPR.2016.445.

Yawen Shao, Wei Zhai, Yuhang Yang, Hongchen Luo, Yang Cao, and Zheng-Jun 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)*, pp. 17326–17336, 2025.

Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3D. *ACM Transactions on Graphics*, 25(3):835–846, 2006. doi: 10.1145/1141911.1141964.

Peize Sun, Shoufa Chen, Chenchen Zhu, Fanyi Xiao, Ping Luo, Saining Xie, and Zhicheng Yan. Going denser with open-vocabulary part segmentation. *arXiv preprint arXiv:2305.11173*, 2023.

Zachary Teed and Jia Deng. DeepV2D: Video to depth with differentiable structure from motion. In *International Conference on Learning Representations (ICLR)*, 2020.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16558–16569, 2021.

Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *2017 IEEE conference on computer vision and pattern recognition*, pp. 5622–5631. IEEE, 2017.

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

Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89, 2025. doi: 10.1109/3DV66043.2025.00013.

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. VGGSfM: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21686–21697, 2024a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-$\Omega$. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21486–21499, 2026b.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27171–27183, 2021.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024b.

Xiaohan Wang, Yuehu Liu, Xinhang Song, Yuyi Liu, Sixian Zhang, and Shuqiang Jiang. An interactive navigation method with effect-oriented affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16446–16456, 2024c.

Dongming Wu, Yanping Fu, Saike Huang, Yingfei Liu, Fan Jia, Nian Liu, Feng Dai, Tiancai Wang, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Jianbing Shen. RAGNet: Large-scale reasoning-based affordance segmentation benchmark towards general grasping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11980–11990, 2025.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4818–4829, June 2024.

Peiran Xu and Yadong MU. Weakly-supervised affordance grounding guided by part-level semantic priors. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 6741–6769, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/13b28ea4105bf409ab6b8ac8184af812-Paper-Conference.pdf.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935, 2025.

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 (CVPR)*, pp. 16284–16295, 2024.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *Computer Vision – ECCV 2018*, pp. 785–801, 2018. doi: 10.1007/978-3-030-01237-3_47.

Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3D scenes. In *Computer Vision – ECCV 2024*, volume 15087 of *Lecture Notes in Computer Science*, pp. 162–179. Springer, 2024. doi: 10.1007/978-3-031-73397-0_10.

Can Zhang and Gim Hee Lee. IAAO: Interactive affordance learning for articulated objects in 3D environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12132–12142, 2025.

Qing Zhang, Xuesong Li, and Jing Zhang. Probing and bridging geometry-interaction cues for affordance reasoning in vision foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2526–2536, June 2026a.

Zixin Zhang, Chenfei Liao, Hongfei Zhang, Harold H Chen, Kanghao Chen, Zichen Wen, Litao Guo, Bin Ren, Xu Zheng, Yinchuan Li, et al. Panoramic affordance prediction. In *European Conference on Computer Vision*, pp. 288–305. Springer, 2026b.

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

# A OBSERVATION SAMPLING

We denote each input configuration by $1 + K$, consisting of one target image $I_0$ and $K$ auxiliary observations. The target image is always the first input to VGGT-$\Omega$, and predictions are evaluated on this image. Auxiliary observations provide context without additional task annotations.

## A.1 3DOI-TASKONOMY

The target and auxiliary RGB images are drawn from OmniData Taskonomy scenes (Eftekhar et al., 2021) and share the target image's scene and point identifiers. For each target image, we sort the available view IDs and select the entries immediately preceding and following the target view in circular order. We evaluate target-only input ($K = 0$) and two auxiliary views ($K = 2$).

## A.2 3DOI-ARTICULATION AND 3DOI-EPICKITCHEN

We sample auxiliary frames from the source videos in the 3DOI-Articulation (Qian et al., 2022) and 3DOI-EpicKitchen (Damen et al., 2018) subsets. Both subsets use the same temporal sampling protocol. Videos are sampled at 10 FPS, and we evaluate target-only, $1+2$, and $1+6$ configurations. Relative to the target frame, the $1+2$ configuration uses auxiliary frames at temporal offsets of $\pm 0.3$ seconds, while the $1 + 6$ configuration additionally includes frames at $\pm 0.7$ and $\pm 1.0$ seconds. All auxiliary frames come from the same source video as the target.

## A.3 HANDAL

HANDAL (Guo et al., 2023) provides object-centric videos in which consecutive frames depict the same object instance. We construct each multi-view input from one target frame and auxiliary frames from the same video. The target frame is always placed first in the input sequence and remains the only frame associated with the prediction target and task annotation.

Following the RAGNet-HANDAL test split (Wu et al., 2025), we divide the original test videos within each object category into validation and test subsets. Videos are sorted by their scene identifiers, with the first half used for validation and the remaining videos used for testing. This produces 2,816 target frames in the test subset. The target-frame set is fixed across all observation configurations.

For a target frame at index $t$, auxiliary frames are selected using index offsets $\pm 4$, $\pm 8$, and $\pm 12$ within the same video. We denote an input with one target frame and $K$ auxiliary frames by $1 + K$. The $1 + 2$ configuration uses the pair $(t - 4, t + 4)$, $1 + 4$ additionally uses $(t - 8, t + 8)$, and $1 + 6$ uses all three offset pairs:

$$(t - 4, t + 4), \quad (t - 8, t + 8), \quad (t - 12, t + 12).$$

Auxiliary frames are ordered by their temporal indices after sampling. When an requested index falls outside the video, it is replaced by the nearest valid sampled frame, allowing repeated auxiliary frames near video boundaries. This padding strategy keeps the target-frame set and the number of input frames identical across configurations. Auxiliary frames provide contextual information only and introduce no additional task annotations.

## A.4 AGD20K

The geometry-aware encoder receives only the target ego image ($K = 0$), with no auxiliary observations. Exocentric images remain part of each predictor's original weakly supervised training branches but are not provided to the geometry-aware encoder. Inference requires no exocentric images (Luo et al., 2022).

# B Implementation Details

## B.1 3DOI Experimental Setup

We conduct experiments using the official training, validation, and test splits of 3DOI (Qian & Fouhey, 2023). To investigate the effect of geometry-aware features across datasets from different sources, we train and evaluate separately on the three 3DOI subsets: 3DOI-AR, 3DOI-EK, and 3DOI-Taskonomy. Following the official protocol, each image contains at most 15 interaction-instance queries. Invalid queries are padded and ignored during training. We consider three backbone configurations: the original SAM ViT-B/16 prompt-and-mask decoder (Li et al., 2025c), and two variants in which the original SAM backbone is replaced with DINOv2 (ViT-B) (Oquab et al., 2023) or VGGT-$\Omega$ (Wang et al., 2026b), respectively, allowing us to evaluate geometry-aware features across different visual backbones. In multi-view experiments, the labelled target image is accompanied by up to six unlabelled neighboring frames, which are processed by a frozen VGGT-$\Omega$ Aggregator. Supervision is applied only to the target view. The VGGT-$\Omega$ features and the corresponding frozen SAM or DINOv2 ViT-B/14 features are processed by layer-wise $1 \times 1$ convolution–GroupNorm–GELU adapters, followed by spatial alignment, channel concatenation, and a $512 \rightarrow 256$ fusion projection. All pretrained visual backbones remain frozen, and only the feature adapters, fusion layers, and task decoders are optimized. The training objective consists of mask focal and Dice losses, affordance focal loss, attribute classification cross-entropy, motion-axis regression, and, where applicable, box L1/GIoU and MiDaS/VNL losses on valid depth samples. Depth losses are applied only to samples with valid depth annotations in 3DOI-Taskonomy. Models are trained with AdamW for 50 epochs using a learning rate of $10^{-4}$, weight decay of $10^{-4}$, StepLR scheduling, fp16 mixed precision, and gradient clipping with a maximum norm of 0.1. The batch size is set to 4, and all experiments are conducted on a single NVIDIA RTX 5090 GPU with 32 GB of memory.

## B.2 HANDAL Experimental Setup

We evaluate our method on the HANDAL dataset using a deterministic fixed-11 multi-view protocol (Guo et al., 2023). All official training videos are used for training. Since RAGNet-HANDAL does not provide a validation split (Wu et al., 2025), we divide the official test videos into validation and test sets in a 1:1 ratio according to their scene IDs. Each video is partitioned into non-overlapping consecutive segments of 11 frames, with the sixth frame serving as the query view. The remaining frames are sampled from the same video as auxiliary views using different strategies, including nearest-neighbor sampling and sampling with temporal strides of 2 or 4.

We initialize our model from the publicly released AffordanceVLM (Wu et al., 2025), which consists of LLaVA-7B (Liu et al., 2023a) and SAM ViT-H (Li et al., 2025c). The query image is first processed by SAM at a resolution of $1024 \times 1024$, while the query and auxiliary images are jointly fed into a frozen VGGT-$\Omega$ Aggregator in the order of query view first, using an input resolution of $384 \times 512$. We extract intermediate features from Aggregator blocks $\{4, 11, 17, 23\}$, concatenate the 1024-dimensional representations from the frame and inter-frame branches into 2048-dimensional tokens, remove the camera and register tokens from each frame, and retain only the patch tokens corresponding to the query view. The resulting $24 \times 32$ feature maps are projected to 256 channels using independent $1 \times 1$ convolutional adapters, followed by GroupNorm and GELU. After bilinear interpolation, the features are aligned with the valid spatial region of the SAM embedding and fused along the channel dimension. Finally, the fused features are injected into the SAM image embedding through a zero-initialized $1 \times 1$ residual projection, ensuring that the model is strictly equivalent to the original AffordanceVLM at initialization.

During training, the SAM image encoder, prompt encoder, and mask decoder, as well as the VGGT Aggregator and language model, are kept frozen. Only the feature adapters, fusion layers, and residual projection are optimized. The training objective is defined as

$$\mathcal{L} = 2.0\mathcal{L}_{\mathrm{BCE}} + 0.5\mathcal{L}_{\mathrm{Dice}},$$

and the checkpoint with the highest validation gIoU is selected for evaluation. At test time, the predicted masks are binarized using zero as the threshold. We use AdamW with a learning rate of $1 \times 10^{-4}$, a weight decay of $1 \times 10^{-4}$, a batch size of 4, and train for 2 epochs. All experiments are conducted on a single NVIDIA RTX 5090 GPU with 32 GB of memory.

### B.3 AGD20K Experimental Setup

We conduct experiments on the Seen and Unseen splits of AGD20K (Luo et al., 2022), which contain 36 and 25 affordance categories, respectively. Unless otherwise specified, input images are first resized such that their shorter side is 256 pixels, then randomly cropped to $224 \times 224$, randomly horizontally flipped, and normalized using the ImageNet mean and standard deviation. During inference, images are deterministically resized to $224 \times 224$. For the VGGT-$\Omega$-enhanced experiments, we use a frozen VGGT-$\Omega$ Aggregator (Wang et al., 2026b) to extract features from the egocentric images at layers 4, 11, 17, and 23 offline. Given a $512 \times 512$ input and a patch size of 16, the features from each selected layer are reshaped into two-dimensional native feature maps of size $2048 \times 32 \times 32$ and stored in FP16 format. During training, the cached features undergo the same cropping and horizontal-flipping transformations as the corresponding RGB images to preserve spatial alignment. Both the VGGT-$\Omega$ backbone and the original backbone of each baseline are kept frozen, while only lightweight channel-projection layers and zero-initialized residual-fusion modules are optimized, ensuring that each model strictly reduces to its corresponding baseline at initialization. For Cross-view-AG (Luo et al., 2022), we retain the original egocentric/exocentric cross-view supervision, classification, feature-distillation, and ACP losses, and inject VGGT-$\Omega$ features either after the final ResNet layer4 or after ResNet layers 1–4. The model is trained for 35 epochs with a batch size of 16 using SGD with momentum and an initial learning rate of $10^{-3}$. For OOAL (Li et al., 2024a), we adopt the one-shot training setting, freeze the DINOv2 ViT-B/14 (Oquab et al., 2023) and CLIP text encoder (Radford et al., 2021), and optimize the segmentation head, prompt embeddings, and VGGT-$\Omega$-fusion modules using only the mask BCE loss. The model is trained for 20,000 iterations with a batch size of 1 using SGD with a cosine learning-rate schedule. For WSAG-PLSP (Xu & MU, 2025), we use refined pseudo labels, object boxes, and verb/noun/part semantic features, and initialize the model with a pretrained CLIP ViT-B/16 encoder and semantic/SAM decoder. The original WSAG parameters remain frozen, and only the VGGT-$\Omega$ adapters and residual-fusion modules are optimized with AdamW for 40 epochs using a batch size of 20. All experiments are conducted on a single NVIDIA RTX 5090 GPU with 32 GB of memory.

## C Additional Experiments on 3DOI

**Geometry-aware fusion with a SAM backbone.** The main experiments use a frozen DINOv2 (Oquab et al., 2023) backbone to test whether reconstruction-pretrained geometry-aware features transfer beyond the segmentation-specialized prior of SAM (Li et al., 2025c). Here, we restore the original SAM backbone while keeping the task heads, objectives, data splits, and evaluation protocols unchanged. Table 7 shows that geometry-aware fusion improves object localization and classification across all three 3DOI subsets, extending its benefits to the SAM-based predictor. Affordance prediction also generally improves, although the lower SIM on Articulation (Qian & Fouhey, 2023; Qian et al., 2022) reveals a trade-off: more accurate interaction-peak localization and lower distributional divergence do not necessarily yield greater heatmap overlap. This contrasts with the consistent improvements on EpicKitchen (Qian & Fouhey, 2023; Damen et al., 2018) and Taskonomy (Qian & Fouhey, 2023; Eftekhar et al., 2021), indicating that the benefits vary across datasets and prediction metrics. Overall, these results support the applicability of geometry-aware fusion across visual backbones. Because the fused configurations also introduce auxiliary observations, the improvements reflect the combined use of reconstruction-pretrained features and additional context.

**Multi-view stacking versus geometry-aware fusion.** Table 8 compares geometry-aware fusion with a multi-view DINOv2 baseline (Oquab et al., 2023) that stacks features without an aggregation module (Oquab et al., 2023). Simple stacking yields mixed results relative to single-view prediction: improvements in object classification or localization can coincide with degraded affordance maps. Additional observations therefore do not automatically translate into better interaction-region prediction. With the same number of input views, geometry-aware fusion consistently outperforms stacking in object localization and affordance prediction across all three subsets, although movability classification does not uniformly improve. These results indicate that the gains extend beyond increasing the observation count and support reconstruction-pretrained features as an effective way to incorporate auxiliary context. The comparison evaluates the complete configurations rather than isolating the effects of pretraining and aggregation.

Table 7: Geometry-aware feature fusion with 3DOI-SAM(Qian & Fouhey, 2023; Li et al., 2025c) on three 3DOI subsets. Best results within each subset are shown in bold.

| Method | Views | Movable | Localization | | Action | Affordance | | |
|---|---|---|---|---|---|---|---|---|
| | | Cat. ↑ | Box ↑ | Mask ↑ | Cat. ↑ | SIM ↑ | Distance ↓ | KLD ↓ |
| **3DOI-Articulation** | | | | | | | | |
| 3DOI-SAM (Qian & Fouhey, 2023) | 1 | 0.9009 | 0.6771 | 0.7464 | 0.8136 | **0.1046** | 0.0721 | 3.8960 |
| 3DOI-SAM + VGGT-Ω (Qian & Fouhey, 2023; Wang et al., 2026b) | $1+6$ | **0.9157** | **0.6987** | **0.7793** | **0.9349** | 0.0829 | **0.0628** | **3.6127** |
| **3DOI-EpicKitchen** | | | | | | | | |
| 3DOI-SAM (ep. 49) (Qian & Fouhey, 2023) | 1 | 0.8752 | 0.7247 | 0.7852 | 0.9688 | 0.0414 | 0.0782 | 4.0679 |
| 3DOI-SAM + VGGT-Ω (Qian & Fouhey, 2023; Wang et al., 2026b) | $1+6$ | **0.8882** | **0.7653** | **0.8067** | **0.9789** | **0.0515** | **0.0677** | **3.8384** |
| **3DOI-Taskonomy** | | | | | | | | |
| 3DOI-SAM (Qian & Fouhey, 2023) | 1 | 0.7954 | 0.6696 | 0.7565 | 0.8838 | 0.1196 | 0.0443 | 3.2823 |
| 3DOI-SAM + VGGT-Ω (Qian & Fouhey, 2023; Wang et al., 2026b) | $1+2$ | **0.8532** | **0.6840** | **0.7676** | **0.9220** | **0.1468** | **0.0397** | **2.8156** |

Table 8: Comparison of single-view and multi-view DINOv2 (Oquab et al., 2023), while geometry-aware feature fusion on three 3DOI subsets (Qian & Fouhey, 2023). Best results within each subset are shown in bold.

| Method | Views | Movable | Localization | | Action | Affordance | | |
|---|---|---|---|---|---|---|---|---|
| | | Cat. ↑ | Box ↑ | Mask ↑ | Cat. ↑ | SIM ↑ | Distance ↓ | KLD ↓ |
| **3DOI-Articulation** (Qian & Fouhey, 2023) | | | | | | | | |
| DINOv2 (ViT-B) (Oquab et al., 2023) | 1 | 0.8846 | 0.6102 | 0.6060 | 0.9105 | 0.0700 | 0.0730 | 4.3438 |
| Multi-view DINOv2 (ViT-B) (Oquab et al., 2023) | $1+2$ | **0.9260** | 0.6146 | 0.6112 | 0.9182 | 0.0555 | 0.0762 | 4.5930 |
| DINOv2 (ViT-B) + VGGT-Ω (Oquab et al., 2023; Wang et al., 2026b) | $1+2$ | 0.9127 | **0.7054** | **0.7346** | **0.9280** | **0.1015** | **0.0646** | **3.5374** |
| **3DOI-EpicKitchen**(Qian & Fouhey, 2023) | | | | | | | | |
| DINOv2 (ViT-B) (Oquab et al., 2023) | 1 | **0.8703** | 0.6283 | 0.6140 | 0.9799 | 0.0410 | 0.0765 | 4.3815 |
| Multi-view DINOv2 (ViT-B) (Oquab et al., 2023) | $1+2$ | 0.8693 | 0.6261 | 0.6256 | 0.9759 | 0.0354 | 0.0740 | 4.4746 |
| DINOv2 (ViT-B) + VGGT-Ω (Oquab et al., 2023; Wang et al., 2026b) | $1+2$ | 0.8693 | **0.6871** | **0.7664** | **0.9809** | **0.0470** | **0.0630** | **4.0325** |
| **3DOI-Taskonomy**(Qian & Fouhey, 2023) | | | | | | | | |
| DINOv2 (ViT-B) (Oquab et al., 2023) | 1 | 0.8522 | 0.6018 | 0.6019 | 0.9220 | 0.1282 | 0.0502 | 3.5312 |
| Multi-view DINOv2 (ViT-B) (Oquab et al., 2023) | $1+2$ | 0.8540 | 0.5874 | 0.5869 | 0.9244 | 0.1199 | 0.0486 | 3.3891 |
| DINOv2 (ViT-B) + VGGT-Ω (Oquab et al., 2023; Wang et al., 2026b) | $1+2$ | **0.8589** | **0.6820** | **0.7278** | **0.9322** | **0.1332** | **0.0418** | **3.0011** |

**VGGT-Ω as a standalone backbone.** Table 9 shows that auxiliary observations affect VGGT-Ω-only (Wang et al., 2026b) prediction differently across subsets. Additional views generally improve spatial prediction on EpicKitchen (Qian & Fouhey, 2023; Damen et al., 2018) and Taskonomy (Qian & Fouhey, 2023; Eftekhar et al., 2021), whereas Articulation (Qian & Fouhey, 2023; Qian et al., 2022) exhibits mixed trends, with some outputs deteriorating as more frames are introduced. Thus, increasing observation count does not consistently improve the transfer of reconstruction-pretrained features to affordance prediction. Comparing these results with Table 8, fusion with DINOv2 (Oquab et al., 2023) under the same three-image configuration improves object localization and reduces both interaction-peak error and heatmap divergence across all three subsets. Heatmap overlap, however, does not uniformly improve. Together, these findings favor using reconstruction-pretrained features alongside visual representations rather than as a direct backbone replacement for the evaluated spatial prediction tasks.

### C.1 Qualitative Comparison on 3DOI

Figures 4, 5, and 6 compare object masks and affordance heatmaps across the three 3DOI (Qian & Fouhey, 2023). On Taskonomy Eftekhar et al. (2021), the illustrated examples show that geometry-aware augmentation produces more coherent object masks and suppresses some spurious heatmap responses relative to the visual-only baseline. These improvements are already visible with target-only input. Adding auxiliary views further refines some predictions, although the changes are less pronounced and vary across examples.

On Articulation (Qian et al., 2022) and EpicKitchen (Damen et al., 2018), the examples include clutter, human occlusion, and changes in object state. Geometry-aware augmentation can concentrate affordance responses closer to the annotated interaction regions, as illustrated by the utensil and kitchen-object examples in Figure 6. Additional frames refine some interaction-region predic-

Table 9: Effect of auxiliary observations on VGGT-Ω-only (Wang et al., 2026b) prediction across three 3DOI subsets. Best results within each subset are shown in bold.

| Method | Views | Movable | Localization | | Action | Affordance | | |
|---|---|---|---|---|---|---|---|---|
| | | Cat. ↑ | Box ↑ | Mask ↑ | Cat. ↑ | SIM ↑ | Distance ↓ | KLD ↓ |
| **3DOI-Articulation** (Qian & Fouhey, 2023) | | | | | | | | |
| VGGT-Ω (Wang et al., 2026b) | 1 | 0.9149 | **0.6310** | 0.6124 | **0.9308** | **0.0724** | **0.0720** | 4.5826 |
| | $1+2$ | 0.9172 | 0.6244 | **0.6189** | 0.9293 | 0.0696 | 0.0769 | 4.7178 |
| | $1+6$ | **0.9201** | 0.6196 | 0.6149 | 0.9105 | 0.0708 | 0.0792 | **4.5744** |
| **3DOI-EpicKitchen** (Qian & Fouhey, 2023) | | | | | | | | |
| VGGT-Ω (Wang et al., 2026b) | 1 | 0.8732 | 0.6192 | 0.6020 | **0.9830** | 0.0342 | 0.0749 | 4.4320 |
| | $1+2$ | 0.8802 | 0.6324 | 0.6333 | 0.9819 | 0.0368 | 0.0740 | **4.3455** |
| | $1+6$ | **0.8902** | **0.6432** | **0.6404** | 0.9799 | **0.0375** | **0.0736** | 4.3528 |
| **3DOI-Taskonomy** (Qian & Fouhey, 2023) | | | | | | | | |
| VGGT-Ω (Wang et al., 2026b) | 1 | **0.8589** | 0.5864 | 0.5947 | 0.9236 | 0.1265 | **0.0463** | 3.3462 |
| | $1+2$ | 0.8484 | **0.6110** | **0.6063** | **0.9316** | **0.1340** | **0.0463** | **3.3286** |

tions, but do not consistently improve object masks. In particular, increasing the number of auxiliary frames can change mask boundaries without a corresponding improvement in their agreement with the ground truth.

These examples illustrate both the benefits and limitations of auxiliary observations: object segmentation and affordance localization can respond differently to the same additional context. The reconstructions are included only for visual context. Affordance prediction uses intermediate geometry-aware features without executing the geometric prediction heads or constructing an explicit 3D scene.

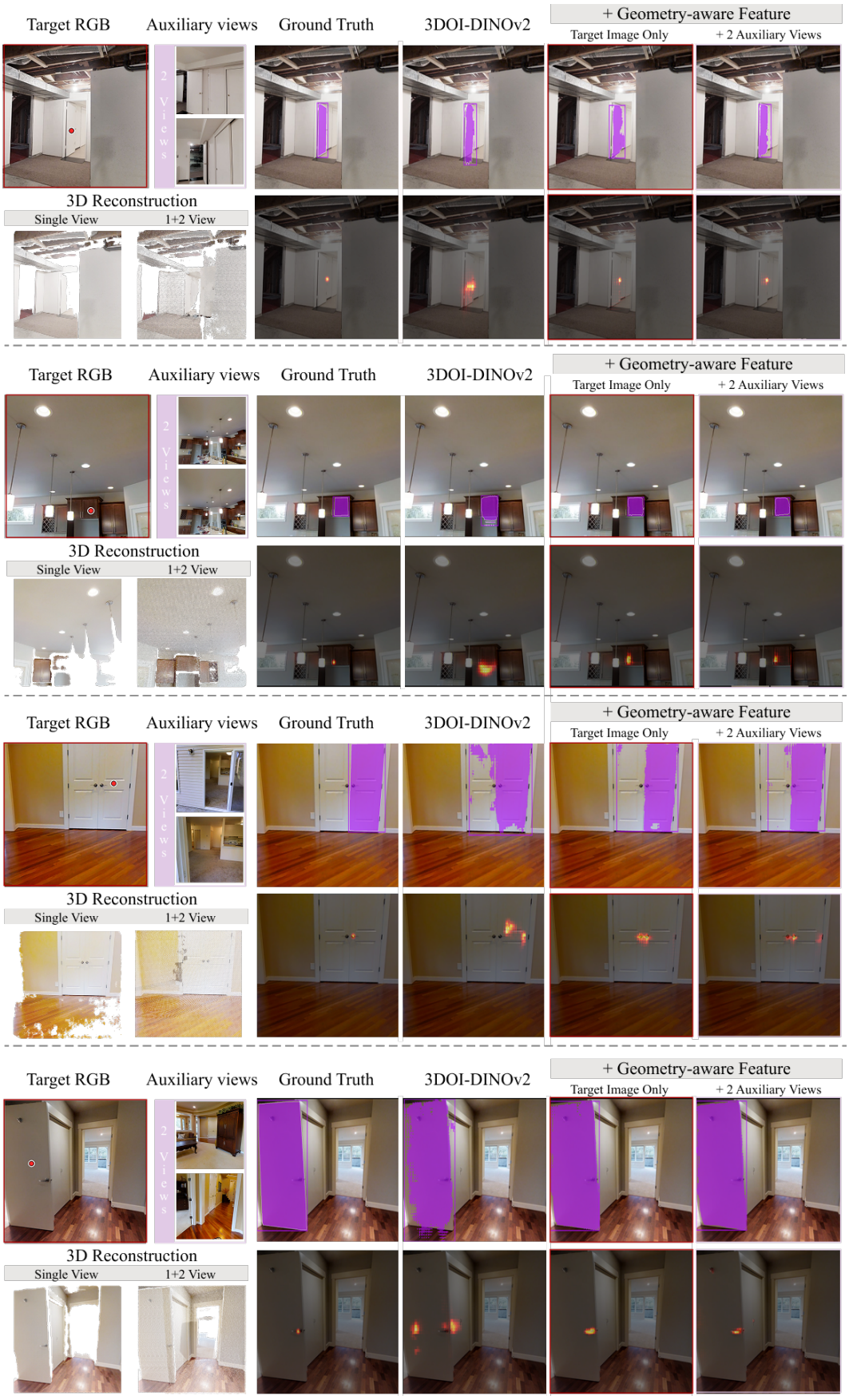


Figure 4: **Qualitative comparison on 3DOI-Taskonomy.** We compare the visual-only baseline with geometry-aware augmentation using 1 target view and 1 target view + 2 auxiliary views for object segmentation and affordance localization. Reconstructions are shown for visualization only.

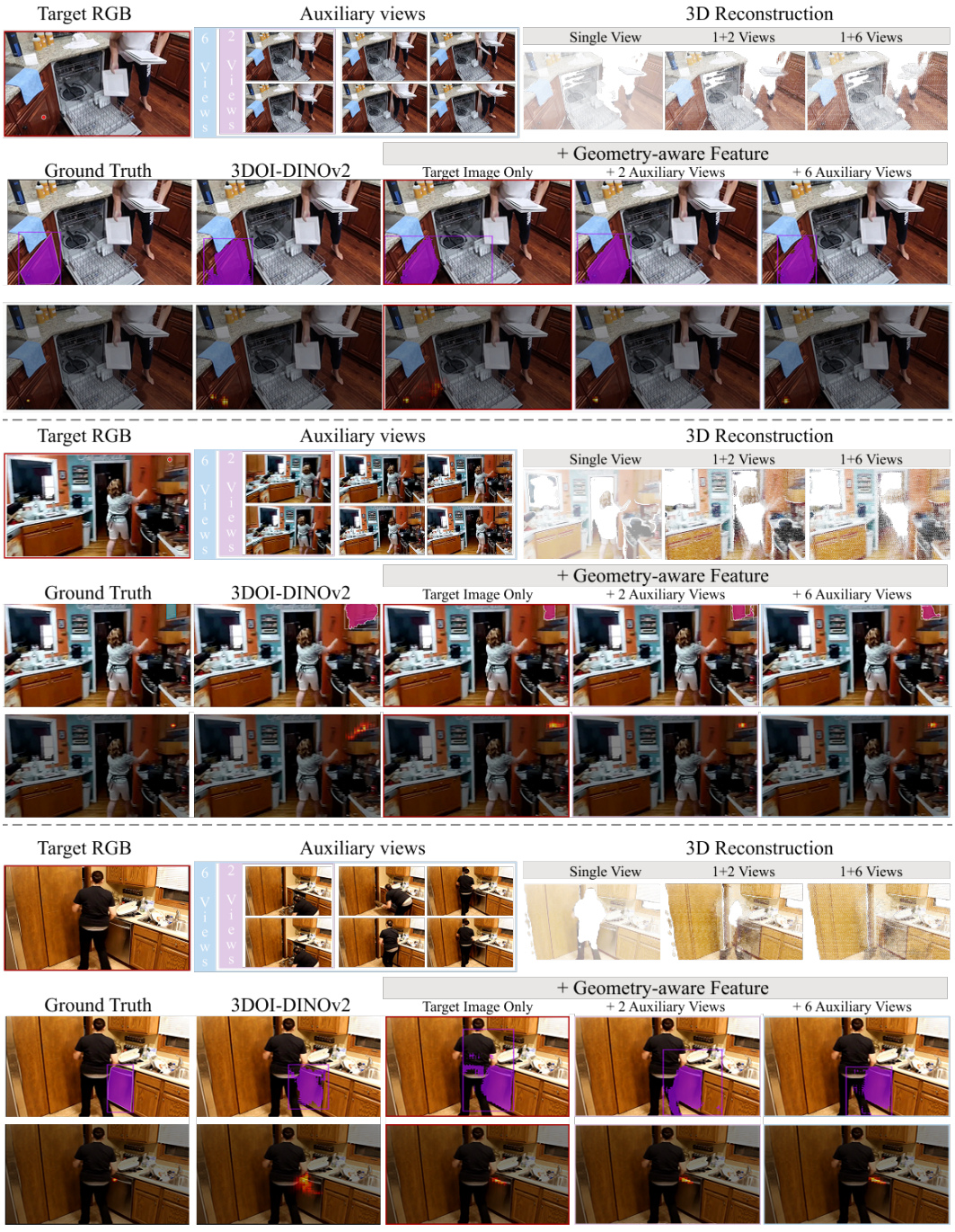


Figure 5: **Qualitative comparison on 3DOI-Articulation.** We compare the visual-only baseline with geometry-aware augmentation using 1 target view, 1 target view + 2 auxiliary views, and 1 target view + 6 auxiliary views for object segmentation and affordance localization. Reconstructions are shown for visualization only.

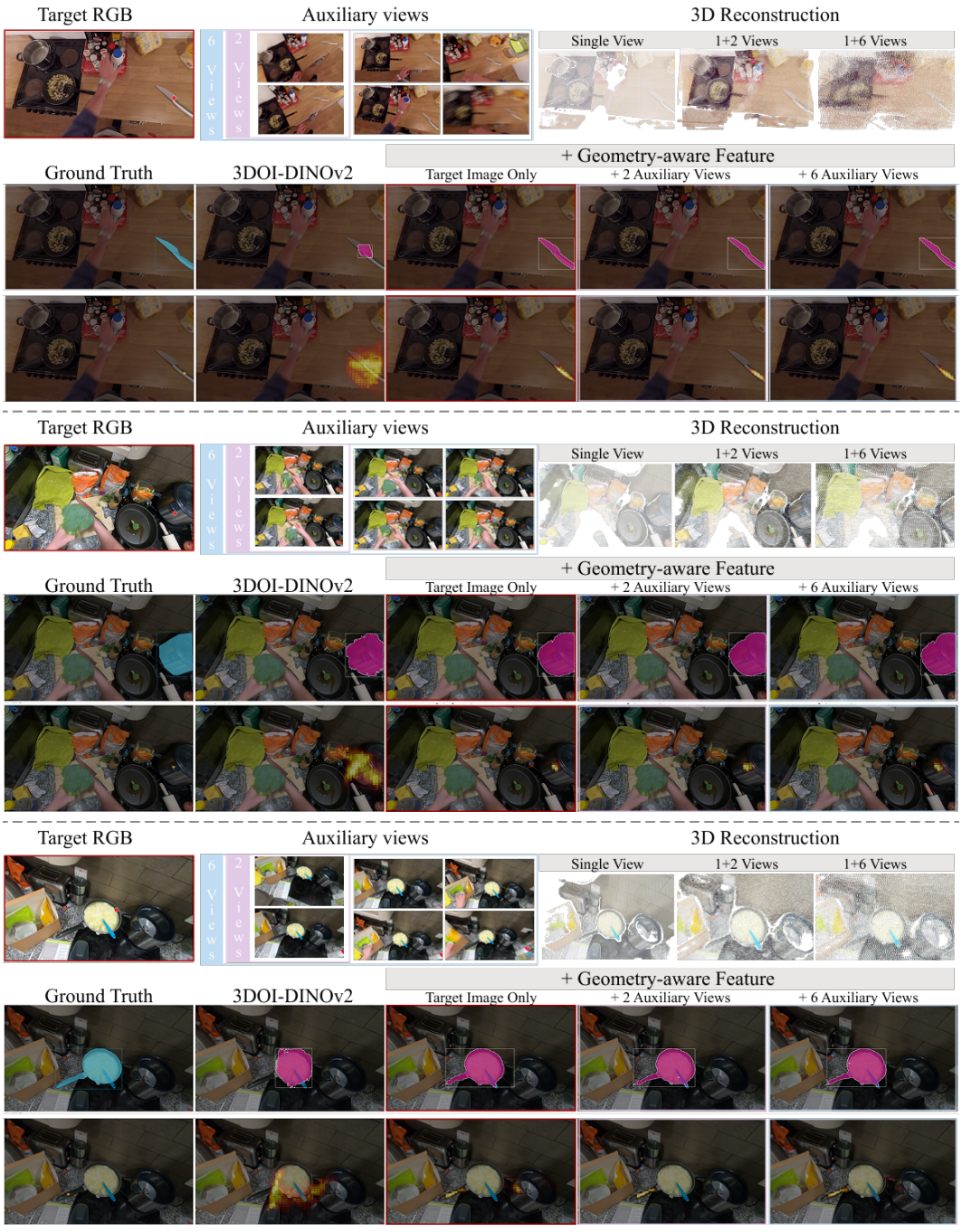


Figure 6: **Qualitative comparison on 3DOI-EpicKitchen.** We compare the visual-only baseline with geometry-aware augmentation using 1 target view, 1 target view + 2 auxiliary views, and 1 target view + 6 auxiliary views for object segmentation and affordance localization. Reconstructions are shown for visualization only.

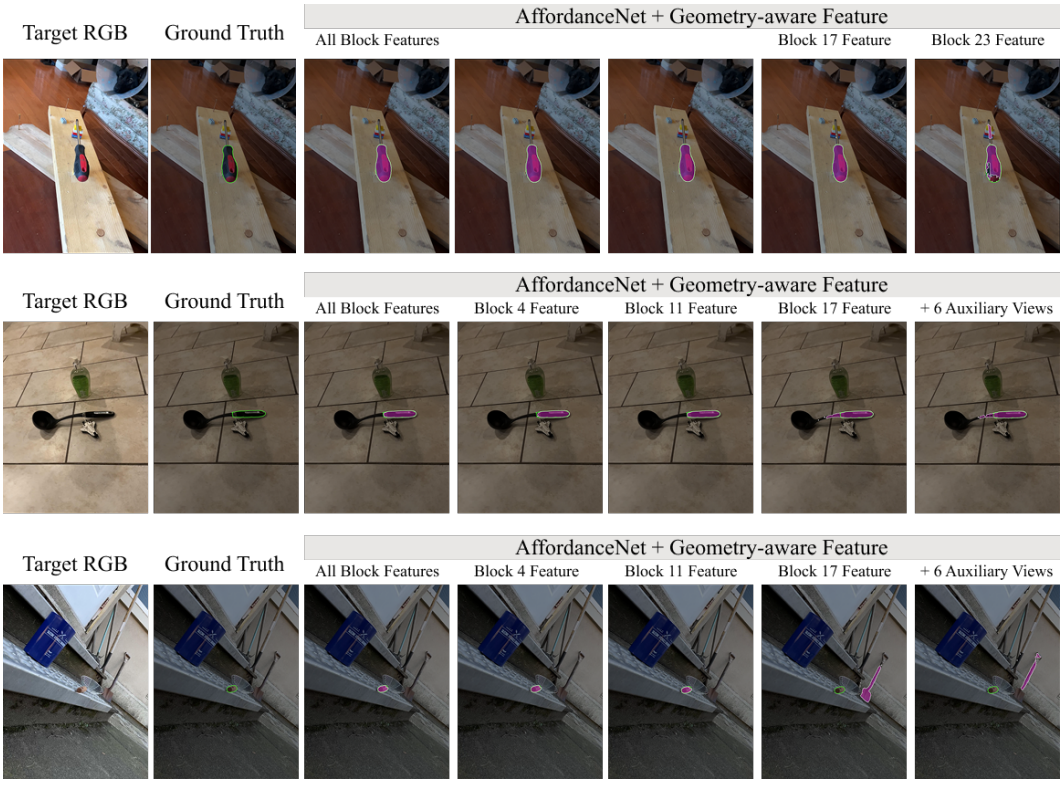


Figure 7: **Effect of geometry-aware feature depth on HANDAL.** We compare affordance segmentation using individual VGGT-Ω layers and multi-layer aggregation, illustrating how feature selection affects interaction-region localization.

## D Qualitative Comparison on HANDAL

**Effect of auxiliary views.** Figure 8 compares affordance segmentation with different numbers of auxiliary views. In the second and third examples, geometry-aware augmentation suppresses background responses and predictions on unrelated objects, producing more localized masks even with target-only input. The first example illustrates the non-monotonic effect of additional observations: four auxiliary views help identify the annotated interaction region, whereas target-only input and configurations with two or six auxiliary views select an incorrect region. Thus, additional observations can help resolve ambiguity, but their benefits vary across examples and do not increase consistently with view count. Reconstructions are shown for visual context only and are not used as inputs to affordance prediction.

**Effect of feature depth.** Figure 7 compares individual VGGT-Ω blocks with multi-layer aggregation (Wang et al., 2026b). The predictions differ in mask extent and responses outside the annotated interaction regions. In the illustrated examples, shallow-layer features and multi-layer aggregation produce compact masks, whereas some deeper-layer predictions include extraneous regions. Multi-layer aggregation yields results similar to the stronger single-layer configurations in several examples, with relatively small visual differences. These observations complement the quantitative feature-depth analysis without implying that aggregation improves every prediction.




Figure 8: **Qualitative comparison on HANDAL.** We compare the visual-only baseline with geometry-aware augmentation for affordance segmentation, using 1 target view with 0, 2, 4, or 6 auxiliary views. Reconstructions are for visualization only.