

Intermediate Connectors and Geometric Priors for Language-Guided Affordance Segmentation on Unseen Object Categories

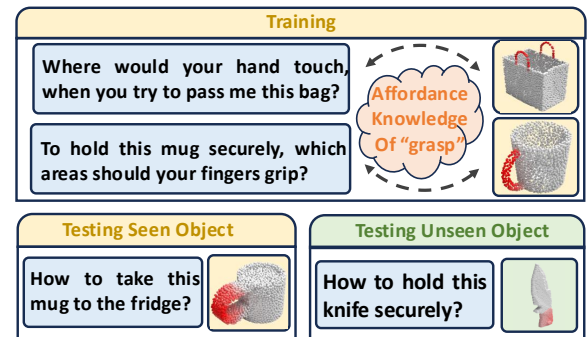
Yicong Li¹ Yiyang Chen¹ Zhenyuan Ma^{1,2} Junbin Xiao¹ Xiang Wang² Angela Yao²
¹National University of Singapore ²University of Science and Technology of China

Abstract

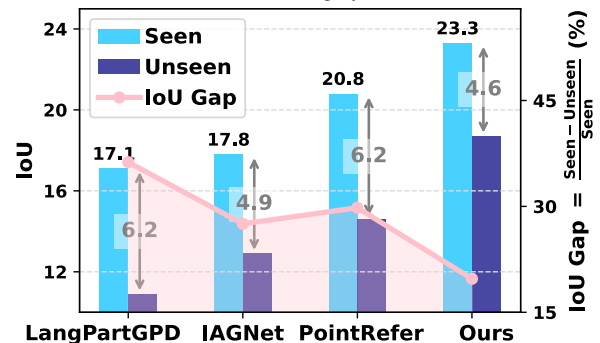
Language-guided Affordance Segmentation (LASO) aims to identify actionable object regions based on text instructions. At the core of its practicality is learning generalizable affordance knowledge that captures functional regions across diverse objects. However, current LASO solutions struggle to extend learned affordances to object categories that are not encountered during training. Scrutinizing these designs, we identify limited generalizability on unseen categories, stemming from (1) underutilized generalizable patterns in the intermediate layers of both 3D and text backbones, which impedes the formation of robust affordance knowledge, and (2) the inability to handle substantial variability in affordance regions across object categories due to a lack of structural knowledge of the target region. Towards this, we introduce a *Generalized framework on Unseen Categories (GLANCE)*, incorporating two key components: a cross-modal connector that links intermediate stages of the text and 3D backbones to enrich pointwise embeddings with affordance concepts, and a VLM-guided query generator that provides affordance priors by extracting a few 3D key points based on the intra-view reliability and cross-view consistency of their multi-view segmentation masks. Extensive experiments on two benchmark datasets demonstrate that GLANCE outperforms state-of-the-art methods (SoTAs), with notable improvements in generalization to unseen categories. Our code is available at <https://github.com/Monoxide-Chen/Affordance>.

1. Introduction

Embodied agents are now capable of breaking down high-level task commands into a series of executable instructions by using large language models (LLMs) as task planners [15, 27, 33]. Within this framework, Language-guided Affordance Segmentation (LASO) [16] is proposed as a crucial role that identifies the actionable parts of objects based on language instructions, which has significant implications for human-level task completion. [7]



(a) Seen vs Unseen category in LASO task.



(b) Performance gap between Seen and Unseen categories.

Figure 1. (a) Illustration of the unseen setting in LASO, where learned affordance knowledge is tested on novel object categories that is unseen during training; (b) Performance comparison of existing SoTA methods on LASO’s Seen and Unseen sets, showing that our method, GLANCE, not only improves on the seen setting but also achieves minimal IoU gap across the two settings.

Despite recent progress in Language-guided Affordance Segmentation (LASO) [16, 30], current approaches often memorize **category-specific cues** rather than learning generalizable affordance knowledge. Consequently, while performing well when training and testing cover the same set of objects, they struggle when testing with unseen object categories — even if the coupled affordance was encountered during training. Specifically, in the standard LASO “unseen setting” [16], models are explicitly evaluated on object categories withheld during training to assess generalization capability. For example, in Fig. 1a, a model trained to grasp

bags and mugs is expected to learn the handle-like concept for “grasping” and transfer it to knives, an unseen category during training. Yet, as shown in Fig. 1b, existing methods [16, 30] often fail in this scenario, leading to significant performance drops on unseen object categories.

Learning affordances for unseen object categories is inherently challenging. We argue that the performance degradation observed in existing LASO methods [16, 20, 30] stems primarily from two limitations. **First**, current point-text encoders process geometric and semantic cues separately, aligning them only at the final layers. Although intermediate layers capture generalizable geometric (*e.g.* edges, surfaces) and linguistic (*e.g.* syntax, semantics) patterns [11, 49], this late-stage fusion isolates such generalizable knowledge, causing the model to overfit category-specific patterns. **Second**, existing mask decoders rely predominantly on semantic cues, neglecting geometric knowledge of the target region. As affordance regions significantly vary across categories (*e.g.* in Fig. 1a round handle on mug versus elongated handle on knife), such structural variability further hinders generalization to unseen objects.

To this end, we propose a **Generalized Framework on Unseen Categories (GLANCE)** for the LASO task. We introduce two novel components on top of the existing LASO pipeline: (1) a **cross-modal connector (CMC)** that bridges the unimodal pre-trained backbones at *intermediate* stages, enhancing point-wise 3D embeddings with affordance concepts; and (2) a **geometric-aware query generator (GAQG)** that generates 3D key points from multi-view 2D segmentation results, providing a sparse but reliable structural prior for target region. Specifically, to preserve the generalizable knowledge of the backbones at the intermediate stage, we introduce a lightweight cross-modal connector that integrates across each stage of the frozen backbones. This connector promotes cross-modal alignment via a shared projection layer, creating a unified feature space where gradients from both modalities are jointly optimized. Notably, designing the cross-modal connector for 3D requires particular care. Compared to 2D images, 3D point clouds yield more tokens, while available 3D affordance annotations are limited, which makes directly adopting attention-based fusion designs commonly used in 2D [40, 44, 48] unsuitable—attention is computationally prohibitive with dense 3D tokens and prone to overfit under low-data regimes. Thus, we design the cross-modal connector (CMC) as a **lightweight MLP-based fusion module** tailored for high-resolution, data-scarce 3D inputs, which preserves intermediate geometric patterns while efficiently injecting language concepts. To mitigate the impact of affordance variability, we propose a VLM-guided query generator that provides affordance priors by extracting 3D key points from 2D multi-view segmentation results, based on their intra- and cross-view reliability. The features of these

key points represent the geometric patterns in the target region and, combined with language cues, guide the mask decoder in identifying affordance regions.

Our contributions are summarized as follows:

- We analyze the challenges in affordance understanding for unseen objects, with limited generalizability due to underutilized intermediate patterns in backbones and substantial variability in affordance across object categories.
- We propose GLANCE, which features a cross-modal connector that aligns language concepts with physical structures by connecting the inner layers of parallel backbones. Additionally, it introduces a geometric-aware query generator that provides 2D affordance priors based on multi-view VLM segmentation results.
- Extensive experiments show that GLANCE achieves SoTA performance in two benchmarks, significantly improving in generalizing to unseen object categories.

2. Related Work

3D Affordance Learning. 3D affordance learning aims to understand the functional properties of objects [8] and predict potential interactions. It benefits from richer geometric and spatial information available in point clouds, to achieve a more precise understanding and prediction of object interactions [4, 5, 12] for visual affordance understanding. Adding semantic or visual information can help understand the target affordance, making it extend beyond point cloud modality [2, 16, 20, 22, 30, 43, 45]. GEAL [20] employs Gaussian splatting to construct a 2D-3D Consistency Alignment Module. LangSHAPE [30] presents a large-scale dataset to learn 3D part-level affordances and enhance robotic grasp detection. LASO [16] uses carefully designed questions for affordance segmentation tasks. While prior works primarily focus on affordance understanding within training categories, our work is the first to emphasize cross-category affordance generalization. Unlike open-vocabulary 3D learning, which requires recognizing unseen affordance types, LASO’s unseen setting evaluates the ability to transfer known affordances to novel object categories, ensuring a controlled assessment of affordance generalization across object geometries.

Efficient Tuning with Adapters. The traditional approach to tuning deep learning models follows the “pretraining-finetuning” paradigm [1]. However, such approaches generally require substantial labeled data for each target task to attain strong performance in multimodal tasks [3, 24, 31, 36?–38]. For more data-efficient fine-tuning, network-based adaptation techniques have been proposed. Houlsby et al. [9] introduces adapter modules in NLP, which inserts learnable linear layers at each Transformer layer. In VLMs communities, Clip-Adapter [6] and Tip-Adapter [47] apply similar strategies by adding adapter layers after the im-

age encoder, focusing on uni-modal enhancements. These works extend adapter modules to multimodal settings via attention-based designs. While effective in data-rich domains like NLP and vision, such architectures tend to overfit in data-scarce tasks like LASO. In contrast, our work introduces an MLP-based adapter.

VLM-assisted 3D Segmentation. 3D segmentation faces significant challenges, primarily due to the lack of large-scale, high-precision datasets. By leveraging the rich contextual information provided by 2D visual features, VLM-assisted 3D segmentation enhances the capability of 3D models, enabling more accurate and robust segmentation outcomes. PartDistill [32] proposes a teacher-student framework to perform bidirectional distillation of multi-view 2D information. PartSLIP [18] fuses multi-view generated 2D bounding boxes with point clouds. We render multi-view images from point clouds and utilize the VLM-based segmentation model [14, 28] to obtain binary masks as the LASO task requires models with high-level reasoning capabilities. These point-wise features serve as target geometric patterns to identify affordance regions via a mask decoder. Unlike existing methods that project 2D features into point space [18, 35, 39, 42] or distill 2D features into 3D models [32], we bypass the direct use of 2D features, and instead take masks from different views as 3D geometry-aware voting weights, which avoids the computational overhead introduced by the 2D backbone. Additionally, VLMs demonstrate robust generalization capabilities, making them ideal for out-of-distribution cases, *e.g.*, images are rendered from 3D objects.

3. Method

Task Definition. As shown in Fig. 1a, given a question Q_{raw} and an object point cloud $P_{\text{raw}} \in \mathbb{R}^{N_P \times 3}$ with N_P points, the goal is to predict a binary mask of $\mathbf{M} \in \mathbb{R}^{N_P}$ that segments the aff part specified by the question.

Overview. As shown in Fig. 2, GLANCE first encodes the object points and question with pre-trained points and text backbones. To retain generalizable knowledge at the intermediate stage, we design a cross-modal connector (CMC) that bridges the Multi-Head Attention (MHA) and Feed Forward Network (FFN) layers across backbones side the transformer block using shared MLPs, facilitating the generation of enriched pointwise 3D features and question embeddings at the output of the backbones. During decoding, we first render multiview images of the object and apply a 2D VLM to segment question-referred regions, producing 2D binary masks. Then, we design a Geo-Aware Query Generator (GAQG) that projects a set of sparsely sampled object points onto each mask. Based on the intra-view reliability and cross-view consistency of these 2D projections, a subset of 3D key points is sampled, and their correspond-

ing pointwise features are used to form geometric queries. Finally, these geometric queries, together with the question embedding, are fed into a mask decoder to inform the 3D affordance region.

Feature Representation. Given a question Q_{raw} , we encode it as a sequence of N_Q tokens using a tokenizer from Roberta [19], producing the text embedding $\mathbf{Q} \in \mathbb{R}^{N_Q \times C_Q}$. For the object points $P_{\text{raw}} \in \mathbb{R}^{N_P \times 3}$, we adopt a commonly used tokenizer from [25, 46, 50], where Farthest Point Sampling (FPS) selects N_F center points from P_{raw} , each expanded into a local patch of k -nearest neighbors. The resulting N_F patches are encoded with PointNet [26], yielding the point embedding $\mathbf{P} \in \mathbb{R}^{N_F \times C_P}$. Here, C_Q and C_P denote the feature dimensions of the question and point embeddings, respectively.

3.1. Encoding with Cross-Modal Connector

To enrich the pointwise features while preserving generalizable knowledge at intermediate stages, we introduce a Cross-Modal Connector (CMC). The CMC modulates transformer-style backbones, integrating complementary affordance cues from text and 3D point representation.

Specifically, the tokenized points $\mathbf{P} \in \mathbb{R}^{N_F \times C_P}$ and questions embedding $\mathbf{Q} \in \mathbb{R}^{N_Q \times C_Q}$ are first processed through frozen text and point backbones, respectively. Inside the transformer block, CMC bridges the Multi-Head Attention (MHA) and Feed Forward Network (FFN) layers across the two backbones using a stack of linear layers. Taking the MHA in a transformer block as an example, since \mathbf{P} and \mathbf{Q} are processed with different feature dimensions, we first apply separate linear projections to map each embedding into a shared feature space. Next, a shared MLP layer integrates complementary signals from both modalities, followed by separate projections that restore the output dimensions and add back to each backbone. Formally, for the i -th transformer block with as input \mathbf{Q}^i and \mathbf{P}^i , the output \mathbf{Q}^{i+1} and \mathbf{P}^{i+1} are derived as:

$$\mathbf{Q}^{i+1} = \text{Up}_Q^i(\text{MLP}^i(\text{Down}_Q^i(\mathbf{Q}^i))) + \mathbf{Q}^i, \quad (1)$$

$$\mathbf{P}^{i+1} = \text{Up}_P^i(\text{MLP}^i(\text{Down}_P^i(\mathbf{P}^i))) + \mathbf{P}^i, \quad (2)$$

where $\text{Down}_P^i \in \mathbb{R}^{C_P \times C}$ and $\text{Up}_P^i \in \mathbb{R}^{C \times C_P}$ denote the linear projections in i -th MHA block of point backbone. While $\text{Down}_Q^i \in \mathbb{R}^{C_Q \times C}$ and $\text{Up}_Q^i \in \mathbb{R}^{C \times C_Q}$ represent linear projections for i -th MHA block in text backbone, with C as the shared feature dimension. The shared projection MLP^i , with a hidden dimension D , acts as a bridge, allowing gradients to propagate between the two modalities and facilitating the exchange of generalizable geometric cues with text representations. Similar to MHA, we also adapt this shared design to the FFN across two backbones.

Notably, since the text and point backbones may have different numbers of transformer blocks, the CMC mod-

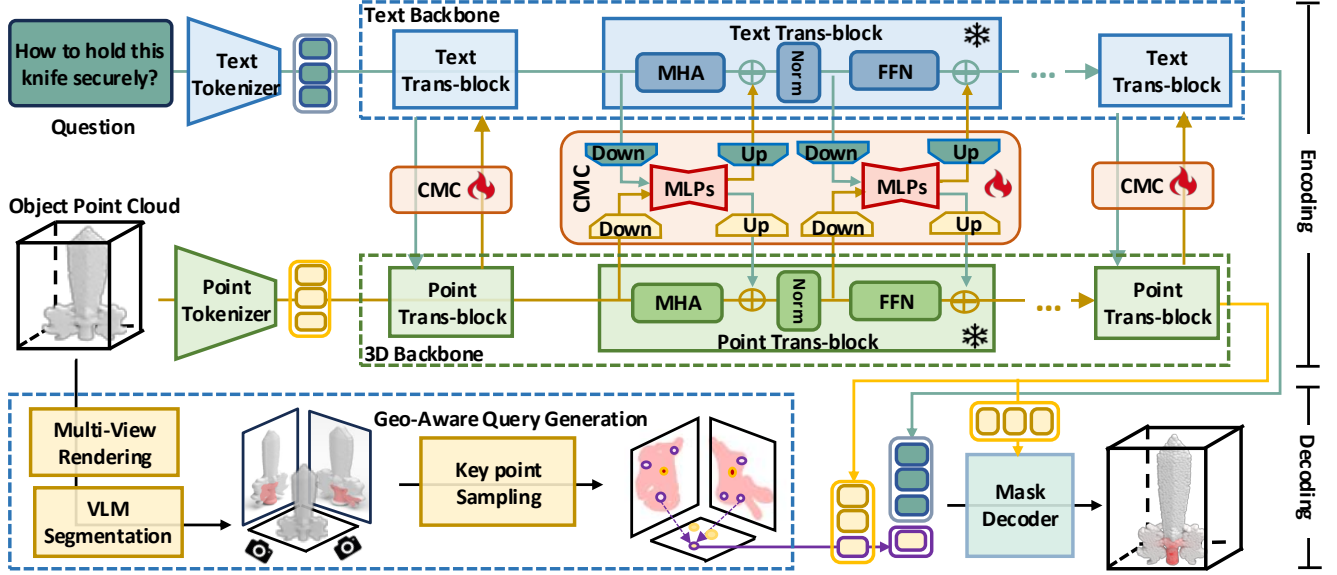


Figure 2. Overview of GLANCE. During encoding, GLANCE processes object points and the question through pre-trained 3D and text backbones, with CMC linking MHA and FFN layers via shared MLPs to enrich pointwise 3D features and question embeddings. In decoding, multi-view images are rendered, and a 2D VLM segments question-referred regions into 2D binary masks. GAQG projects sparse object points onto each mask, selecting key 3D points based on projection reliability and cross-view consistency. These points form geometric queries, which, along with the question embedding, guide the mask decoder in identifying the 3D affordance region.

ule is only applied to the last few layers of each backbone to ensure alignment. To acquire the final outputs, a feature propagation module [26] is employed to upsample the processed point tokens into the pointwise embedding. These pointwise embeddings, along with the text embeddings are mapped to C dimension via two linear layers, yielding pointwise feature $\mathbf{F}_P \in \mathbb{R}^{N_P \times C}$ and question feature $\mathbf{F}_Q \in \mathbb{R}^{N_Q \times C}$.

3.2. Decoding with Geo-Aware Query Generator

To equip the model with prior knowledge of unseen objects, we introduce the Geometry-Aware Query Generator. This query generator extracts informative key points from 2D segmentation masks. The feature of these key points acts as geometric queries to guide a DETR-style 3D affordance decoder, facilitating the identification of affordance regions on unseen objects.

Geo-Aware Query Generator (GAQG). As illustrated in Fig. 3, to inform the 3D affordance with 2D priors, we first render a set of L multi-view images $\{\mathbf{I}^l \mid \mathbf{I}^l \in \mathbb{R}^{H \times W}\}_{l=1}^L$ from the object point cloud. We then employ an off-the-shelf VLM-based image segmentation model [14, 28], which processes each multi-view image with the question to produce corresponding binary segmentation masks. This results in a set of 2D segmentation masks $\{\mathbf{S}^l \mid \mathbf{S}^l \in \{0, 1\}^{H \times W}\}_{l=1}^L$, with 1 indicating the question-referred region. Next, we project N_F FPS-sampled points (obtained from the tokenizer in Sec. 3) onto each 2D mask. Then,

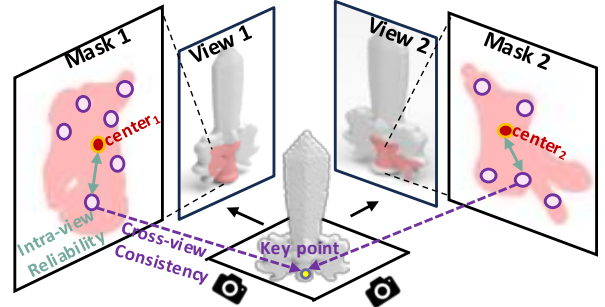


Figure 3. Illustration of key point sampling with Intra-View Reliability and Cross-View Consistency.

taking the projection result and multi-view masks, we sample K key points out of N_F candidates according to the following criteria:

- **Intra-View Reliability:** For points whose 2D projections lie within a masked region, those closer to the mask center are more likely to be sampled, as the center typically indicates higher confidence. Formally, for a point $\mathbf{p}_j \in \{\mathbf{p}_j\}_{j=1}^{N_F}$, we evaluate its intra-view score in the l -th view as:

$$S_{\text{intra}}^l(\mathbf{p}_j) = \delta_l(\mathbf{p}_j) \cdot \exp\left(-\frac{d(\mathbf{p}_j, \mathbf{c}_l)}{\sigma}\right), \quad (3)$$

where $\delta_l(\mathbf{p}_j)$ is an indicator function that equals 1 if the 2D projection of \mathbf{p}_j lies within the masked region and 0 otherwise, $d(\mathbf{p}_j, \mathbf{c}_l)$ is the Euclidean distance between the 2D projection of \mathbf{p}_j and the center \mathbf{c}_l of the masked region

in the l -th view, and σ is a scaling parameter that controls distance sensitivity.

- **Cross-View Consistency:** If a point is identified across masked regions of multiple views, its likelihood of being sampled accumulates, as such consistency across views suggests higher reliability. The cross-view score $S_{\text{cross}}(\mathbf{p}_j)$ is calculated by summing its intra-view score across all L views:

$$S_{\text{cross}}(\mathbf{p}_j) = \sum_{l=1}^L S_{\text{intra}}^l(\mathbf{p}_j). \quad (4)$$

Finally, we softmax-normalize the cross-view scores to obtain the probability $P(\mathbf{p}_j)$ for each of N_F candidates:

$$P(\mathbf{p}_j) = \frac{\exp(S_{\text{cross}}(\mathbf{p}_j))}{\sum_{k=1}^{N_F} \exp(S_{\text{cross}}(\mathbf{p}_k))}. \quad (5)$$

We then sample K points from N_F candidates based on their probabilities, and construct the geometric query $\mathbf{G} \in \mathbb{R}^{K \times C}$ by collecting their point-wise features from the encoded representation \mathbf{F}_P . These queries encapsulate geometric prior provided by the 2D VLM, guiding the identification of affordance regions. During training, this sampling is stochastic, while for inference, the top- K points with the highest probabilities are selected.

Affordance Mask Decoder. Inspired by the success of query-based segmentation methods [16, 34], which form target-specific cues as queries, our mask decoder uses affordance cues derived from VLM-elicited key points and question semantics to generate the segmentation mask for the object. Specifically, we concatenate the \mathbf{G} with text embedding \mathbf{F}_Q as a query to the decoder, along with the encoded point-wise feature \mathbf{F}_P used as key and value:

$$\mathbf{A} = \text{Transformer-Decoder}([\mathbf{G}; \mathbf{F}_Q], \mathbf{F}_P). \quad (6)$$

Following [16], the generated embedding $\mathbf{A} \in \mathbb{R}^{(K+N_Q) \times C}$ is passed through an MLP to produce $(K+N_Q)$ dynamic kernels $\Omega = \{\omega_i\}_{i=1}^{(K+N_Q)}$. Each kernel ω_i is used to convolute the point-wise feature \mathbf{F}_P , resulting $(K+N_Q)$ point-wise masks $\{\mathbf{M}_i\}_{i=1}^{(K+N_Q)}$:

$$\mathbf{M}_i = \{\mathbf{F}_P * \omega_i\}_{i=1}^{(K+N_Q)}. \quad (7)$$

Then, we apply mean pooling over all masks $\{\mathbf{M}_i\}$ followed by a sigmoid activation to obtain the final segmentation mask $\mathbf{M} \in \mathbb{R}^{N_P}$:

$$\mathbf{M} = \sigma(\text{Mean-Pool}(\{\mathbf{M}_i\}_{i=1}^{(K+N_Q)})), \quad (8)$$

where $\sigma(\cdot)$ denotes the sigmoid function.

In essence, the geometric query focuses on regions with similar geometry, while the text embedding ensures that this focus aligns with the affordance described in the question.

Objective. Following existing work [16], we use the Dice loss and Binary Cross-Entropy (BCE) loss to supervise the affordance prediction:

$$\mathcal{L} = \mathcal{L}_{BCE} + \mathcal{L}_{Dice} \quad (9)$$

4. Experiment

Implementation Details. We leverage pre-trained models for our backbone, specifically Uni3D [50] as the point cloud encoder and RoBERTa [19] as the text encoder. Our training procedure follows the configuration settings outlined in LASO [16]. The dimension of the point features C is set to 512. We utilize the Adam optimizer [13] with an initial learning rate of 1×10^{-4} . We train the model for 40 epochs using a mini-batch size of 16. We apply a StepLR scheduler with a step size of 10 epochs and a decay factor of 0.5. We set the number of sampled points per view to 5 for optimal performance in the main result. To render multi-view images, we employ RepKPU [29] to upsample the original point clouds by a factor of 64, producing a denser and more continuous representation of each object. We then use Mitsuba3 [10] to render the upsampled point clouds into images from four viewpoints. Image segmentation masks are generated for each view using LISA[14] and GLaMM [28] in response to specific questions.

Dataset. We evaluate our method on two 3d language-guided affordance segmentation datasets. **LASO** [16] comprises 19,751 paired samples across 23 categories, featuring 8,434 distinct point object shapes. There are 17 affordance types and 870 questions matching them. The difference between Seen and Unseen settings is that some seen objects are removed from the training dataset while keeping the testing dataset the same. **LangSHAPE** [30] divided into part-wise and object-wise grasp detection grain modes. The part-wise mode contains 42,109 objects and 547,417 point clouds from 35 categories. The language configuration has 1.38 million sentences, divided into 7 subsets. We use the part-wise mode with full sentences subset, sampling 3,735 point clouds and matching questions for training.

Seen vs. Unseen The standard LASO [16] dataset provides two settings: (1) **Seen**, where training and testing include the same set of affordance-object combinations, including 17 affordance types and 23 object categories, composing 58 unique affordance-object combinations; and (2) **Unseen**, for 11 out of 17 affordance types, objects of certain classes are omitted from training but evaluated at test time. Detailed statistics can be found in Appendix A.

Evaluation Metrics. Following the existing work [16], we assess our model by Mean Intersection Over Union (mIoU), Area Under the Curve (AUC), Similarity (SIM), and Mean Absolute Error (MAE), with mIoU being the key indicator.

4.1. Main Result

We present the results on the LASO [16] dataset in Tab. 1, demonstrating GLANCE’s strong generalization ability on

Table 1. Results on the LASO [16] dataset. The **best** and **second-best** performance are highlighted.

	Method	mIoU \uparrow	AUC \uparrow	SIM \uparrow	MAE \downarrow
Seen	ReferTrans [23]	13.7	79.8	0.497	0.124
	ReLA [17]	15.2	78.9	0.532	0.118
	3D-SPS [21]	11.4	76.2	0.433	0.138
	IAGNet [43]	17.8	82.3	0.561	0.109
	PointRefer [16]	20.8	87.3	0.629	0.093
	GEAL [20]	22.0	86.7	0.634	0.092
	GLANCE	23.3	88.1	0.632	0.089
Unseen	ReferTrans [23]	10.2	69.1	0.432	0.145
	ReLA [17]	10.7	69.7	0.429	0.144
	3D-SPS [21]	7.9	68.8	0.402	0.158
	IAGNet [43]	12.9	77.8	0.443	0.129
	PointRefer [16]	14.6	80.2	0.507	0.119
	GEAL [20]	16.7	80.9	0.567	0.106
	GLANCE	18.7	81.7	0.532	0.112

Table 2. Results on the LangSHAPE [30] dataset.

Method	mIoU \uparrow	AUC \uparrow	SIM \uparrow	MAE \downarrow
IAGNet [43]	59.2	90.1	0.712	0.158
PointRefer [16]	60.5	91.2	0.728	0.152
GLANCE	63.8	92.6	0.751	0.132

both seen and unseen objects. Key observations include: (1) GLANCE outperforms existing SoTAs, with a significant improvement in mIoU, confirming the effectiveness of our design. (2) The performance gain in unseen settings is more pronounced, as prior methods heavily relied on class-specific patterns, limiting their transferability. We also present the results on LangSHAPE in Tab. 2. All models were trained on the sampled training dataset and evaluated on the full test set. GLANCE outperforms all other methods across all four evaluation metrics, achieving a notably high mIoU of 63.8%. Compared to LASO, performance here is significantly higher, primarily because the affordance questions in LangSHAPE are more straightforward. Each question is given as a short phrase specifying both affordance type and object class, allowing the model to directly infer the target region without complex reasoning. (Detail per-class result can be found in Appendix B.)

4.2. In-Depth Study

Ablation Study. Tab. 3 presents our ablation study results. First, the baseline model, which excludes both the CMC and GAQG modules, shows the lowest accuracy, highlighting the effectiveness of our overall design. Removing the CMC module decreases performance, as it restricts the model’s ability to leverage generalizable patterns within the intermediate layers for cross-modal alignment, thus reducing its capacity to link affordance regions with textual cues. Similarly, removing the GAQG module results in a performance drop, as it eliminates the crucial 2D priors that enhance affordance visibility. We further examine the intra-view reliability and cross-view consistency within the GAQG module.

Table 3. Ablation results on the impact of different configurations for the CMC and GAQG modules. Tow marks (\checkmark , \times) indicate the presence or absence of each module. Feature or Random sampler replace the GAQG module with sampling strategies.

CMC	GAQG	mIoU \uparrow	AUC \uparrow	SIM \uparrow	MAE \downarrow
\times	\times	21.3	85.8	0.607	0.095
\times	\checkmark	22.3	87.1	0.627	0.095
\checkmark	\times	21.7	87.1	0.630	0.094
\checkmark	w/o Intra-View	22.4	87.6	0.630	0.091
\checkmark	w/o Cross-View	21.9	86.2	0.627	0.091
\checkmark	random sampler	21.5	87.0	0.631	0.093
\checkmark	feature sampler	21.9	86.7	0.620	0.097
\checkmark	\checkmark	23.3	88.1	0.632	0.089

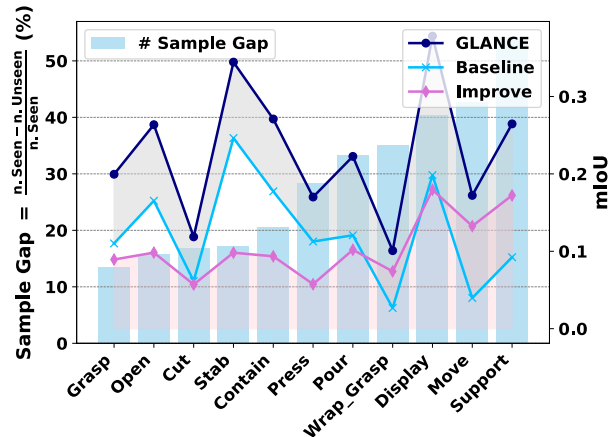


Figure 4. Illustration of how sample size differences between Seen and Unseen categories affect GLANCE’s performance gain over the Baseline. Bars represent sample size differences for each affordance type, while lines show test mIoU for models trained on Seen and Unseen. Improve (mIoU): GLANCE - Baseline.

For “w/o Intra-View,” we replace distance-based probability sampling with random sampling within each view, leading to a substantial performance decrease. For “w/o Cross-View,” we omit the cross-view consistency check, causing another severe performance decline. Additionally, we test a “Random Sampler” variant, which selects key points entirely at random, resulting in even worse performance. These outcomes validate the effectiveness of our key point sampling strategy. As a comparison, we also evaluate using point cloud features or random sampling in place of the GAQG’s 2D prior queries. These alternative methods perform poorly, as 2D priors derived from image masks provide high-confidence candidates, whereas raw point cloud features lack a direct correlation with affordance regions.

Impact of Sample Size Differences between Seen and Unseen Categories on Performance.

In LASO, the unseen data is created by omitting certain objects from the seen set. For instance, while the seen training set includes “grasp-mug” and “grasp-bag”, the unseen set is generated by removing “grasp-mug.” This setting expects the model to learn generalizable affordance knowledge, such as “grasp”, and apply it to an unseen object during testing. Note that

Table 4. In-depth study on CMC and GAQG. The best performance is in **bold**.

(a) CMC with last few blocks.					(b) Study on CMC hidden dimension D .					(c) Study on number of views L .								
Layer	mIoU \uparrow	AUC \uparrow	SIM \uparrow	MAE \downarrow	Dim	mIoU \uparrow	AUC \uparrow	SIM \uparrow	MAE \downarrow	L	Seen				Unseen			
											mIoU \uparrow	AUC \uparrow	SIM \uparrow	MAE \downarrow	mIoU \uparrow	AUC \uparrow	SIM \uparrow	MAE \downarrow
11 \rightarrow 12	22.2	87.1	0.626	0.098	32	22.1	87.0	0.622	0.096	1	22.4	86.7	0.623	0.093	18.0	80.9	0.520	0.125
9\rightarrow12	23.3	88.1	0.632	0.089	64	23.3	88.1	0.632	0.089	2	21.9	86.5	0.624	0.093	17.6	80.7	0.516	0.130
7 \rightarrow 12	22.5	87.1	0.622	0.095	128	21.7	86.8	0.623	0.093	3	22.7	86.7	0.629	0.094	18.3	81.0	0.521	0.123
5 \rightarrow 12	22.1	87.3	0.619	0.094	256	22.6	86.5	0.623	0.095	4	23.3	88.1	0.632	0.089	18.7	81.7	0.532	0.112

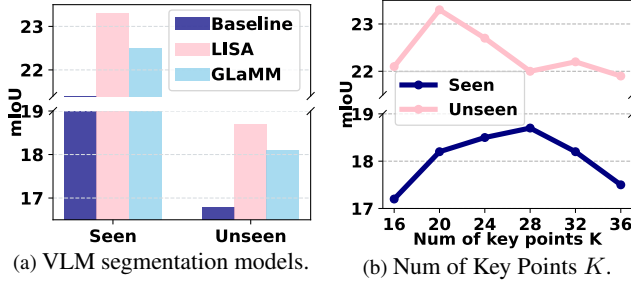


Figure 5. (a) Results of different VLM segmentation models, and (b) key point numbers K on the Geo-Aware Query Generator.

both the seen and unseen settings share the same validation and test sets. Fig. 4 shows the performance of the baseline (without CMC and GAQG) for each unseen affordance type. Each bar represents the difference in sample size for that affordance type, and the line indicates mIoU performance. We observe that GLANCE achieves greater improvements over the baseline for affordance types with larger sample gaps, validating our design’s effectiveness in handling unseen categories, even with limited samples of similar affordances. Additionally, the results highlight substantial improvements in affordance types like “Display” and “Stab”, where distinct visual patterns enable the model to effectively leverage prior cues. In contrast, affordances such as “Wrap Grasp” and “Move” show modest gains, as these require more context-specific interactions to accurately identify relevant regions.

Impact of CMC in Different blocks of Backbone. Inspired by [41], we equip our CMC to the last few blocks of the transformer-style backbone. Since our text backbone has 12 blocks, we study which layers to embed our CMC modules and show the results in Tab. 4a. For consistency, the hidden dimension of the shared projection layers is set to 64. Embedding CMC into 4 middle layers achieves the best performance, with an mIoU of 23.3%. In contrast, embedding CMC modules too early or too late yields relatively worse results. We hypothesize that this is because the low-level layers in the backbone capture modality-specific features, making CMC less effective. On the other hand, the high-level layers are more focused on semantic information than geometric details and local structural features, making it challenging for the CMC modules to effectively bridge modality differences. Embedding CMC in the middle layers allows for capturing generalizable geometry features and bridging text representations, which improves performance.

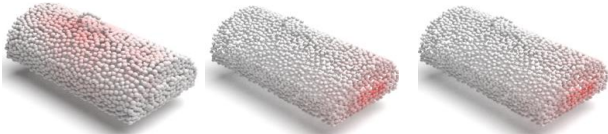
Impact of Hidden Dimension D in CMC. The CMC module incorporates a shared three-layer MLP with a hidden dimension of D . We investigate the impact of D with the embedded CMC fixed at the last 4 layers of the backbone for all runs. As depicted in Tab. 4b, the best performance is achieved at 64. We attribute this to a balance between expressiveness and overfitting risk. A larger hidden dimension increases parameter count, enhancing the network’s ability to model complex features but raising the risk of overfitting. Conversely, a smaller dimension may lead to insufficient representational capacity, limiting the network’s ability to capture and transfer essential cross-modal features.

Impact of Number of Views L . As shown in Tab. 4c, we analyze how the number of views L affects Cross-View Consistency. For consistency, total number of the key points K is fixed to 20. Increasing L enables the model to capture more diverse geometric affordance information while minimizing noise and occlusions, effectively utilizing 2D priors to find truth candidates. The best result is $L = 4$ on both two categories. Interestingly, using a single view ($L = 1$) achieves the second-highest mIoU of 22.4% on the Seen, as it probably captures a more focused mask.

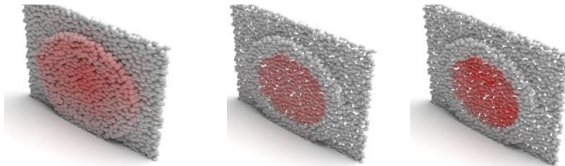
Different 2D segmentation masks. To evaluate GLANCE’s reliance on 2D segmentation quality, we replace the default LISA-generated [14] 2D masks with those produced by an alternative VLM-based segmentation model, GLaMM [28]. A manual inspection of their 2D segmentation results reveals that LISA provides more accurate and reasonable masks compared to GLaMM. Consequently, as shown in Fig. 5a, the LISA-based GLANCE model achieves superior performance in both seen and unseen settings. This indicates that GLANCE is affected by 2D segmentation results, and higher-quality 2D masks lead to improved affordance segmentation. (See more analysis on 2D segmentation in Appendix C.)

Impact of Number of key Points K on GAQG. We study the impact of sampling K key points out of N_F candidates on Fig. 5b. The number of views L is set to 4 for this experiment. The overall trend shows that mIoU initially increases with K but declines after reaching a peak. Sampling a moderate number of key points generates crucial prior queries, leading to better accuracy. However, as K increases, it introduces noisy points outside the target regions, resulting in a decline in the performance. Seen and Unseen settings achieve the highest mIoU at 20 and 28, respectively. This

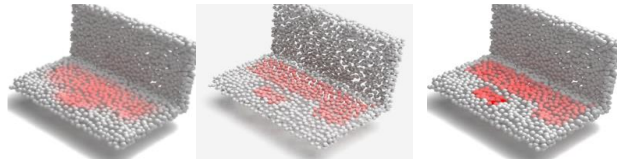
Considering the structure of the bag, what area would be most stable for opening?



If you want to look at the time, which points on this dock would you look at?



To type on a computer keyboard, which points on each key should your fingers apply pressure to?

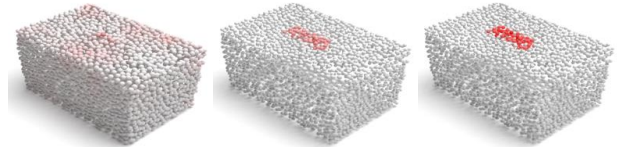


If you want to open the microwave, from which points on the door would you touch?



SoTA Ours Ground Truth

When you try to grab the bag, at which points will your palm position be?



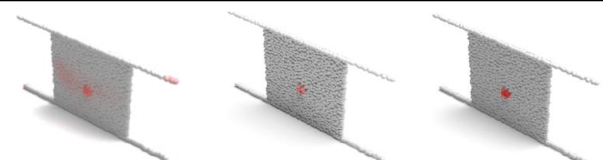
For the best control and safety when holding a knife, which part should your palm and fingers wrap around?



If you want to move this table, at which points on this table will you exert your strength?



To open a door with ease and control, where on the door should you apply force or grasp the handle?



SoTA Ours Ground Truth

Figure 6. Case Study of GLANCE: Each example includes a question and three shape predictions from PointRefer [16], GLANCE, and Ground Truth, with the segmented affordance part highlighted in red.

difference suggests that unseen objects benefit from a larger number of sampled points to gather target geometry.

4.3. Qualitative Results

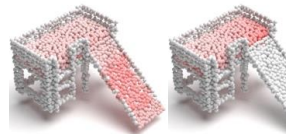
Case Study. We show the affordance labels for different problems in Fig. 6. Like the first case, PointRefer [16] fails to identify the actual opening area of the bag. In contrast, GLANCE uses multi-view image segmentation prior to accurately capturing the side region. As demonstrated by the computer keyboard example, GLANCE shows strong recognition capabilities for fine-grained affordance regions.

Failure Analysis. We present two failure cases in Fig. 7. The first one involves frequent over-segmentation issues. The model is not sensitive to the tilt of surfaces, so it assumes that the inclined plane is also suitable for sitting. The second one involves incorrect region predictions. However, in this case, GLANCE’s prediction gets a low score, though we still believe it is a reasonable prediction.

5. Conclusion

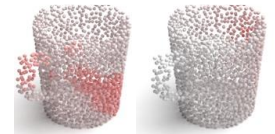
In this work, we analyzed the limited generalizability of LASO models on unseen categories, identifying chal-

What are the ideal locations for sitting to ensure both comfort and stability?



Ours Ground Truth

Which part of the mug allows for the most efficient wrap-grasping method?



Ours Ground Truth

Figure 7. Failure cases due to over-segmentation (left) and incorrect region predictions(right).

lenges with underutilized patterns in backbone layers and variability of affordance regions. To address these issues, we proposed GLANCE, which integrates a cross-modal connector for aligning text and 3D backbones and a geo-aware query generator using VLM-segmented multi-view masks. GLANCE achieves significant improvements over existing methods, especially on unseen categories, and we hope this work will inspire further research into enhancing the generalizability of LASO models.

References

- [1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS 2020*, 2020. 2
- [2] Meng Chu, Xuan Zhang, Zhedong Zheng, and Tat-Seng Chua. 3d-tafs: A training-free framework for 3d affordance segmentation, 2024. 2
- [3] Meng Chu, Yicong Li, and Tat-Seng Chua. Understanding long videos via llm-powered entity relation graphs. *arXiv preprint arXiv:2501.15953*, 2025. 2
- [4] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes. In *CVPR*, pages 14531–14542, 2024. 2
- [5] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *CVPR*, 2021. 2
- [6] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *Int. J. Comput. Vis.*, 132:581–595, 2024. 2
- [7] Yiran Geng, Boshi An, Haoran Geng, Yuanpei Chen, Yaodong Yang, and Hao Dong. Rlafford: End-to-end affordance learning for robotic manipulation. In *ICRA*, pages 5880–5886. IEEE, 2023. 1
- [8] J Gibson James. The ecological approach to visual perception, 1979. 2
- [9] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 2790–2799, 2019. 2
- [10] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. <https://mitsuba-renderer.org>. 5
- [11] Ganesh Jawahar, Benoît Sagot, and Djamel Seddah. What does BERT learn about the structure of language? In *ACL*, pages 3651–3657, 2019. 2
- [12] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *ICCV*, pages 14713–14724, 2023. 2
- [13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 5
- [14] 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. 3, 4, 5, 7, 1
- [15] Boyi Li, Philipp Wu, Pieter Abbeel, and Jitendra Malik. Interactive task planning with language models. *CoRR*, 2023. 1
- [16] Yicong Li, Na Zhao, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-Seng Chua. LASO: language-guided affordance segmentation on 3d object. In *CVPR*, pages 14251–14260. IEEE, 2024. 1, 2, 5, 6, 8
- [17] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *CVPR*, 2023. 6
- [18] Minghua Liu, Yin hao Zhu, Hong Cai, Shizhong Han, Zhan Ling, Fatih Porikli, and Hao Su. Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In *CVPR*, pages 21736–21746, 2023. 3
- [19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 3, 5
- [20] Dongyue Lu, Lingdong Kong, Tianxin Huang, and Gim Hee Lee. GEAL: generalizable 3d affordance learning with cross-modal consistency. In *CVPR*, 2025. 2, 6
- [21] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. *arXiv preprint arXiv:2204.06272*, 2022. 6
- [22] Teli Ma, Zifan Wang, Jiaming Zhou, Mengmeng Wang, and Junwei Liang. Glover: Generalizable open-vocabulary affordance reasoning for task-oriented grasping, 2024. 2
- [23] Li Muchen and Sigal Leonid. Referring transformer: A one-step approach to multi-task visual grounding. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 6
- [24] Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See-Kiong Ng, and Luu Anh Tuan. Video-language understanding: A survey from model architecture, model training, and data perspectives. *arXiv preprint arXiv:2406.05615*, 2024. 2
- [25] Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *ECCV*, pages 604–621, 2022. 3
- [26] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, pages 5099–5108, 2017. 3, 4
- [27] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. In *CoRL*, 2023. 1
- [28] 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. *CVPR*, 2024. 3, 4, 5, 7

- [29] Yi Rong, Haoran Zhou, Kang Xia, Cheng Mei, Jiahao Wang, and Tong Lu. Repkpu: Point cloud upsampling with kernel point representation and deformation. In *CVPR*, pages 21050–21060, 2024. 5
- [30] Yaoxian Song, Penglei Sun, Yi Ren, Yu Zheng, and Yue Zhang. Learning 6-dof fine-grained grasp detection zbased on part affordance grounding. *CoRR*, 2023. 1, 2, 5, 6
- [31] Pengzhan Sun, Junbin Xiao, Tze Ho Elden Tse, Yicong Li, Arjun Akula, and Angela Yao. Visual intention grounding for egocentric assistants. *arXiv preprint arXiv:2504.13621*, 2025. 2
- [32] Ardian Umam, Cheng-Kun Yang, Min-Hung Chen, Jen-Hui Chuang, and Yen-Yu Lin. Partdistill: 3d shape part segmentation by vision-language model distillation. In *CVPR*, pages 3470–3479, 2024. 3
- [33] Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *CoRR*, 2023. 1
- [34] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, pages 4964–4974. IEEE, 2022. 5
- [35] Xiaoyang Wu, Xin Wen, Xihui Liu, and Hengshuang Zhao. Masked scene contrast: A scalable framework for unsupervised 3d representation learning. In *CVPR*, 2023. 3
- [36] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2804–2812, 2022. 2
- [37] Jingqiao Xiu, Mengze Li, Zongxin Yang, Wei Ji, Yifang Yin, and Roger Zimmermann. Few-shot incremental learning via foreground aggregation and knowledge transfer for audio-visual semantic segmentation. In *AAAI Conference on Artificial Intelligence*, pages 8788–8796, 2025.
- [38] Jingqiao Xiu, Yicong Li, Na Zhao, Han Fang, Xiang Wang, and Angela Yao. Geometric alignment and prior modulation for view-guided point cloud completion on unseen categories. In *IEEE/CVF International Conference on Computer Vision*, 2025. 2
- [39] Xiuwei Xu, Huangxing Chen, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Embodiedsam: Online segment any 3d thing in real time. *arXiv preprint arXiv:2408.11811*, 2024. 3
- [40] Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *ICCV*, pages 17457–17466. IEEE, 2023. 2
- [41] Lingxiao Yang, Ru-Yuan Zhang, Yan Chen Wang, and Xiaohua Xie. MMA: multi-modal adapter for vision-language models. In *CVPR*, pages 23826–23837, 2024. 7
- [42] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. SAM3D: segment anything in 3d scenes. *CoRR*, 2023. 3
- [43] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3d object affordance from 2d interactions in images. In *ICCV*, pages 10905–10915, 2023. 2, 6
- [44] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. LAVT: language-aware vision transformer for referring image segmentation. In *CVPR*, pages 18134–18144. IEEE, 2022. 2
- [45] Chunlin Yu, Hanqing Wang, Ye Shi, Haoyang Luo, Sibe Yang, Jingyi Yu, and Jingya Wang. Seqafford: Sequential 3d affordance reasoning via multimodal large language model. In *CVPR*, 2025. 2
- [46] 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 *CVPR*. IEEE, 2022. 3
- [47] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *CoRR*, abs/2111.03930, 2021. 2
- [48] Yuxuan Zhang, Tianheng Cheng, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. EVF-SAM: early vision-language fusion for text-prompted segment anything model. *CoRR*, abs/2406.20076, 2024. 2
- [49] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *ICCV*, pages 16239–16248. IEEE, 2021. 2
- [50] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *ICLR*, 2024. 3, 5