FROM PARTIAL OBSERVATION TO COMPLETE UNDERSTANDING: UNIFIED 3D SEMANTIC AND AFFORDANCE REASONING FROM A SINGLE IMAGE

Anonymous authors
Paper under double-blind review

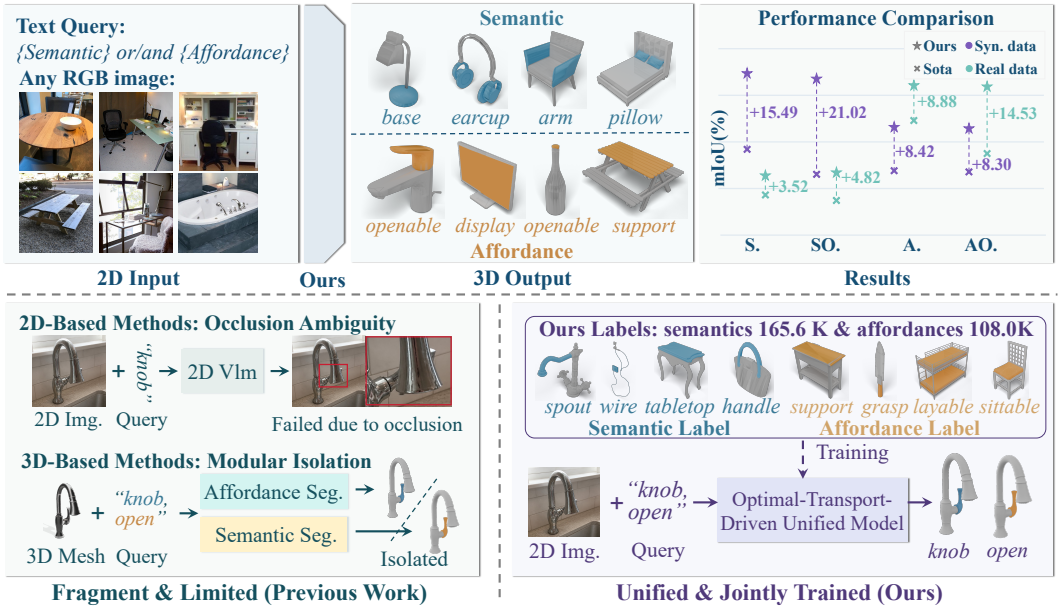


Figure 1: We propose a unified framework for 3D semantic and affordance segmentation from a single RGB image and an arbitrary text query. It predicts queried regions over the complete reconstructed object, including occluded surfaces, unifies previously isolated semantic and affordance supervision, and achieves state-of-the-art performance on both simulated and real-world benchmarks.

ABSTRACT

Real-world robots often receive **only partial observations** but **require complete 3D understanding of object semantics and affordances**, including in occluded regions, to guide actions. Existing methods **address only parts of this problem**: 2D approaches are limited to visible regions, while 3D approaches typically model semantics and affordances separately, making cascaded solutions prone to error propagation. We formulate a unified task that jointly grounds semantic and affordance queries over the full 3D object from a single image. This task involves two sources of heterogeneity. First, **features derived from 2D images and 3D shapes follow different representation mechanisms**. We align and fuse them via optimal transport, enforcing spatial and cross-instance consistency. Second, **semantic and affordance terms**, such as *arm* and *move*, **can be distant in language space despite referring to the same region**. We inject spatial–functional cues into text embeddings to capture such part–function correspondence while preserving part distinctions. We further construct a multi-level, dual-attribute dataset with semantic and affordance annotations at the region, instance, and category levels. Experiments in simulation and the real world demonstrate consistent improvements over 2D- and 3D-based baselines under occlusion.

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.

1 INTRODUCTION

In practical scenarios, robots often perceive objects from partial observations, such as a single RGB image of visible surfaces(Savva et al. (2019)). Yet interaction requires understanding the complete 3D object from two complementary perspectives: *what its regions are* and *what they can be used for*. A robot may need to localize an occluded mug handle as a semantic part, while also identifying the region that affords grasping or opening. Such semantic and functional understanding must extend beyond the observed view to the complete object.

Existing methods address only pieces of this problem, as illustrated in Fig. 2. **Image-based methods**, which predict semantic or functional regions from a single RGB image, are confined to the observed view and cannot naturally cover occluded surfaces(He et al. (2022); Li et al. (2024a)). **3D approaches**, which operate on explicit 3D representations, can reason over complete geometry, but typically treat semantic part and affordance segmentation as separate tasks with different annotations, models, and objectives(Mo et al. (2019); Deng et al. (2021)). As a result, neither paradigm readily supports unified semantic and functional reasoning over the complete object from a single image. We therefore formulate **unified 3D semantic and affordance segmentation from a single image**: given an RGB observation and an arbitrary query, predict the corresponding region on the complete reconstructed object, including invisible or occluded surfaces. The query may specify either a semantic part or an affordance, allowing both to be grounded within a shared framework.

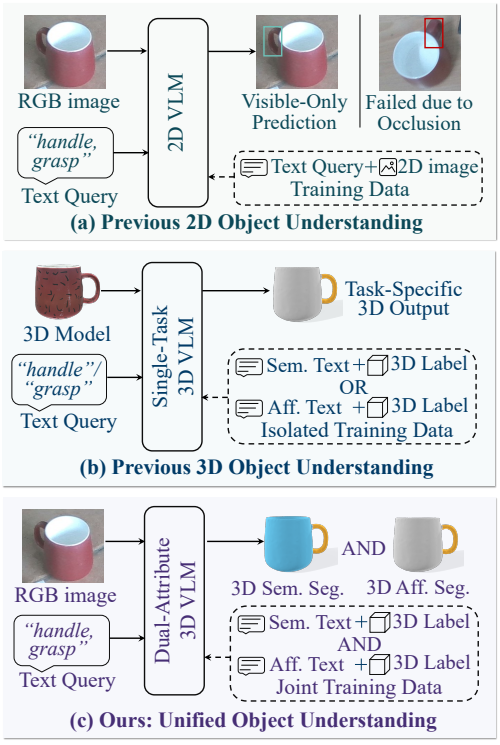


Figure 2: Comparison of object-understanding settings. 2D methods are limited to visible evidence, while 3D methods require explicit 3D input and separate semantic from affordance segmentation. From a single RGB image and text query, our method grounds either region type on the complete reconstructed object.

This unified formulation presents two key challenges. On the representation side, image-derived and shape-derived features follow different representation mechanisms, requiring an effective fusion strategy to align and integrate their complementary information. On the attribute side, semantic parts and affordances are not naturally aligned in language space and often exhibit many-to-many relationships. For example, *handle* and *grasp* may be linguistically distant despite referring to the same physical region, while different semantic parts, such as the *back* and *arm* of a chair, may both support the affordance *support*. The model therefore needs to capture these cross-attribute correspondences while preserving the distinction between queries that refer to different physical regions. We refer to these two challenges as **representation heterogeneity** and **attribute heterogeneity**, respectively.

To tackle these challenges, we build a unified framework that jointly aligns heterogeneous visual representations and bridges semantic and functional concepts for object-centric grounding. Specifically, **Optimal Transport Fusion** aligns image- and shape-derived features in a shared prototype space and integrates them into a unified 3D representation. **Spatio-Functional Adaptation** addresses attribute heterogeneity by reshaping the language representation according to the current object, allowing semantically distinct part and affordance concepts to establish object-dependent functional correspondence. Together, these two modules address representation and attribute heterogeneity from complementary perspectives, followed by a shared matching head that predicts either semantic or affordance regions on the reconstructed 3D object.

Training a unified model also calls for datasets where semantic-part and affordance annotations coexist for each object, a setting rarely covered by existing datasets. Therefore, we construct a **multi-level dual-attribute dataset** by cross-annotating objects from PartNet and 3DAffordanceNet, covering 16 categories with about 165.6K semantic labels and 108.0K affordance labels. Experiments on simulated single-view observations and real images show consistent gains over representative baselines on both tasks, with further generalization to real observations and unseen categories. Experimental results demonstrate that our framework outperforms state-of-the-art methods by 15.49 mIoU on semantic queries and by 8.42 mIoU on affordance queries.

2 Related Work

2D semantic and affordance understanding. 2D methods ground semantic descriptions and functional queries in image regions. Semantic methods transfer vision–language knowledge through multi-granularity supervision (VLPart (Sun et al., 2023)) or explicit object–region modeling (PartGLEE (Li et al., 2024b)), while affordance methods transfer interaction cues (LOCATE (Li et al., 2023)) or align visual features with affordance text (OOAL (Li et al., 2024a)). General-purpose multimodal models, including Qwen2.5-VL, GPT-4, and Gemini 2.5 (Bai et al., 2025; OpenAI, 2023; Gemini Team, Google, 2025), provide visual understanding and language reasoning for both query types. LISA and PixelLM (Lai et al., 2024; Ren et al., 2024) further connect such reasoning to pixel-level masks, enabling grounding from explicit descriptions and implicit intentions; AffordanceVLM in RAGNet (Wu et al., 2025) extends this paradigm to functional regions by combining generic segmentation data with affordance supervision. However, standard image grounding labels only visible pixels, leaving occluded surfaces unlocalized even when their semantics or affordances are understood. Amodal methods such as pix2gestalt (Ozguroglu et al., 2024) reconstruct whole objects under occlusion, but do not jointly ground both query types on completed 3D surfaces, as required for whole-object robotic interaction.

3D semantic and affordance understanding. 3D methods take an object and text prompt and predict the referred region. Semantic segmentation follows two pipelines: inference-time 2D-to-3D lifting, which renders the shape, applies a pretrained vision–language model, and lifts predictions to 3D (Liu et al., 2023; Abdelreheem et al., 2023; Kim & Sung, 2024); and direct 3D prediction. The latter includes distilling 2D predictions into a 3D student (PartDistill (Umam et al., 2024)) and learning reusable 3D features through point–text alignment, canonical spatial modeling, or language-aligned patch embeddings (Find3D, CoSMo3D, and PatchAlign3D (Ma et al., 2025; Jin et al., 2026; Hadgi et al., 2026)). Affordance grounding follows three directions: aligning point features with affordance text, optionally extending to language-conditioned 6-DoF poses (Nguyen et al., 2023b;a); transferring 2D knowledge through rendered-view consistency (GEAL (Lu et al., 2025)) or interaction cues and geometry (GREAT (Shao et al., 2025)); and using language-model reasoning, as 3D-AffordanceLLM (Chu et al., 2025) does to predict instruction-conditioned masks after pretraining on referring semantic segmentation. This pretraining connects the attributes, but the final task remains affordance grounding. Most systems still target either semantics or affordances. Joint grounding must align linguistically distant queries for the same region (e.g., *handle*/*grasp*) while distinguishing regions sharing an affordance (e.g., a chair's *back* and *arm* both afford *support*). This complementarity motivates our shared supervision and object-conditioned grounding over complete objects reconstructed from single RGB images.

3 Method

Overview. Given a single RGB image I and a text query q, our goal is to predict a mask $\hat{\mathbf{y}} \in [0,1]^N$ over the complete reconstructed object, including surfaces occluded in I. The query may describe either a semantic part or an affordance. These two attributes are complementary but not interchangeable: one semantic region may support multiple actions, while one affordance may span multiple semantic regions. We therefore use a shared representation while predicting a separate binary mask for each query, allowing semantic and affordance regions to overlap.

As shown in Fig. 3, we employ SAM3D (Chen et al., 2026) to reconstruct the complete object and extract image-derived features F^S, and Utonia (Zhang et al., 2026) to encode the reconstructed geometry into shape-derived features F^D. Although complementary, the two features lie in different

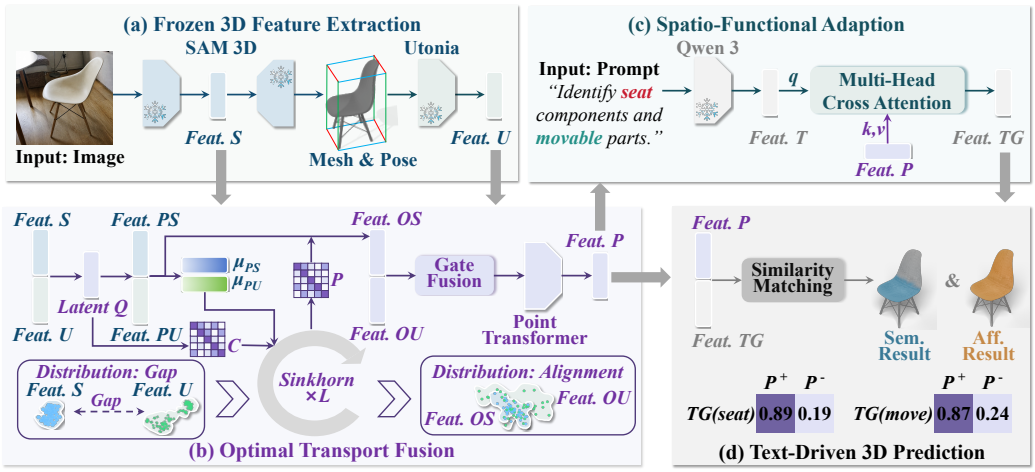


Figure 3: Method overview. Given a single RGB image and an arbitrary semantic or affordance query, our framework first reconstructs the complete object and extracts two features using frozen pretrained models (a). It then aligns and fuses the heterogeneous feature distributions through optimal transport (b). Next, the text query is adapted to the spatial and functional context of the current object via multi-head cross-attention (c). Finally, point-text similarity matching predicts the queried semantic or affordance regions on both visible and occluded surfaces (d).

representation spaces and cannot be reliably fused directly. We propose the Optimal Transport Feature Fusion module (Sec. 3.1) to align and fuse them into a unified point representation F_P. A second challenge arises on the language side: semantic–affordance correspondence is object-dependent and often many-to-many, while semantic and affordance queries for the same region may still lie far apart in the text embedding space. We propose Spatio-Functional Adaptation (Sec. 3.2) to address this mismatch and adapt the text representation to the current object, yielding F_{TG}. Finally, we compute point–text similarity between F_P and F_{TG} to localize the queried region, using the same formulation for both semantic and affordance queries. To support joint training, we further construct a larger-scale text–mask dataset, detailed in Sec. 3.3.

3.1 Optimal Transport Feature Fusion

Let $F^S, F^D \in \mathbb{R}^{N\times d}$ denote the projected SAM3D and Utonia features on the same N sampled surface points. F^S provides image-derived reconstruction cues, while F^D captures geometric structure. Although complementary, their mismatched distributions make direct fusion unreliable. Thus, we align the two representations through a shared latent before fusing them.

Mapping into latent. To align the two feature spaces efficiently, we avoid direct point-to-point transport in the original feature space. Such dense transport would require an $N \times N$ cost matrix and incur high computational and memory costs. We therefore introduce a learnable shared latent $Q = \{q_k\}_{k=1}^{K} \in \mathbb{R}^{K\times d}$ with $K \ll N$. After ℓ_2 normalization, the point-wise soft assignments are

$$F^{PS} = \mathrm{softmax}\left(\bar{F}^S \bar{Q}^\top\right), \qquad F^{PU} = \mathrm{softmax}\left(\bar{F}^U \bar{Q}^\top\right). \tag{1}$$

Here, the bar denotes ℓ_2 normalization. Softmax is applied over the latent dimension, yielding $F^{PS}, F^{PU} \in \mathbb{R}^{N\times K}$, where each row gives the soft assignment of one point to the K latent vectors. This reduces dense point matching to compact assignment in the shared latent space.

Sinkhorn optimal transport. Despite reducing computational and memory costs, the shared latent does not by itself resolve the distribution gap between the two feature sets, as they may still distribute assignment weights differently across the same latent vectors. We therefore use entropy-regularized optimal transport with Sinkhorn iterations (Cuturi, 2013) to align the two latent occupancy distributions by redistributing assignment mass across latent vectors. This procedure takes the two occupancy distributions and a transport kernel as input, and after L Sinkhorn iterations, yields a transport plan P describing the redistribution, as detailed below.

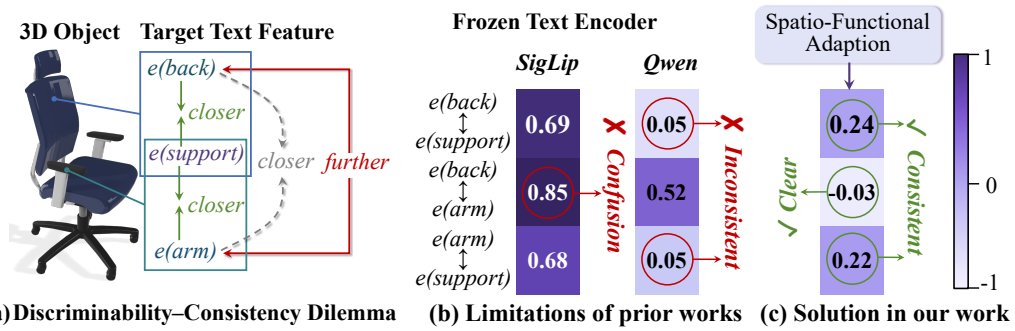


Figure 4: Figure 4. Motivation for Spatio-Functional Adaptation. Target text features of related semantic parts and affordances should be aligned while those of distinct regions remain discriminative (a). Frozen text encoders fail to balance both properties (b), whereas our adaptation achieves this balance using object-specific spatial context (c).

Given the marginals $\mu_{PS}, \mu_{PU} \in \mathbb{R}^K$ obtained by averaging F^{PS} and F^{PU} over points, we define the ground cost C and the corresponding entropic kernel G from the latent vectors $\{q_i\}_{i=1}^K$ in Q as

$$C_{ij} = \|\bar{q}_i - \bar{q}_j\|_2, \qquad G_{ij} = \exp(-C_{ij}/\epsilon), \tag{2}$$

where ϵ controls the entropic regularization. Starting from $u = v = \mathbf{1}_K$, Sinkhorn alternates

$$v \leftarrow \mu_{PS} \oslash (G^\top u), \qquad u \leftarrow \mu_{PU} \oslash (Gv), \quad \text{for } L \text{ iterations}, \tag{3}$$

where $\oslash$ denotes element-wise division. After L iterations, we obtain the transport plan $P = \mathrm{Diag}(u)G\,\mathrm{Diag}(v)$, where $\mathrm{Diag}(\cdot)$ forms a diagonal matrix and u and v rescale G to match the prescribed marginals. We use P to transform the SAM3D latent assignments F^{PS} toward the Utonia latent distribution represented by F^{PU}, and decode the aligned representations as

$$F^{OS} = (F^{PS}P^\top)\bar{Q}, \qquad F^{OU} = F^{PU}\bar{Q}. \tag{4}$$

This remapping is intentionally asymmetric: F^{PU} serves as the geometric reference, so only F^{PS} is transported by P, while F^{PU} is decoded directly from the shared latent. Thus, F^{OS} and F^{OU} are brought into a common feature space through the shared latent.

Optimal transport regularization. We further regularize latent alignment with

$$\mathcal{L}_{\mathrm{OT}} = \lambda_{\mathrm{feat}}\mathcal{L}_{\mathrm{feat}} + \lambda_{\mathrm{cost}}\mathcal{L}_{\mathrm{cost}} + \lambda_{\mathrm{ent}}\mathcal{L}_{\mathrm{ent}} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}. \tag{5}$$

Here, $\mathcal{L}_{\mathrm{feat}}$ encourages consistency between F^{OS} and F^{OU}, $\mathcal{L}_{\mathrm{cost}}$ penalizes transport between distant latent vectors, $\mathcal{L}_{\mathrm{ent}}$ regularizes the transport plan, and $\mathcal{L}_{\mathrm{KL}}$ aligns the two latent occupancy distributions. $\mathcal{L}_{\mathrm{OT}}$ is combined with the mask loss in the overall training objective (Sec. 3.3); detailed definitions and weights are provided in the supplementary material.

Fusion and Extraction. Although latent alignment makes F^{OS} and F^{OU} more compatible, their relative importance may vary across points. We employ a point-wise gate to adaptively adjust the weighting between the two feature sources. Point Transformer V3 (Wu et al., 2024) then aggregates spatial context and produces the unified point representation $F^P \in \mathbb{R}^{N\times d}$ for subsequent query adaptation and mask prediction.

3.2 Spatio-Functional Adaptation and 3D Prediction

We observe that frozen text representations often fail to reflect the object-specific correspondence between semantic parts and affordances. As illustrated in Fig. 4(a), the chair semantic parts *back* and *arm* are both associated with the affordance *support*. Thus, their point embeddings should be close to *support* while remaining distinct as they refer to different physical regions. However, directly enforcing such cross-attribute consistency in language space can inadvertently pull $e(\textit{back})$ and $e(\textit{arm})$ closer together. More generally, semantic–affordance relations are often object-dependent and many-to-many, creating a tension between cross-attribute consistency and inter-region discriminability.

To illustrate this limitation, Fig. 4(b) compares two representative pretrained text encoders, SigLIP(Zhai et al., 2023) and Qwen(Team, 2025), using the *back–arm–support* example. They exhibit different failure modes: SigLIP places *back* and *arm* too close, confusing distinct semantic regions, whereas Qwen better separates the two but leaves their related affordance *support* weakly aligned. These observations suggest that generic text representations alone do not reliably capture object-specific semantic–affordance correspondence. We therefore introduce Spatio-Functional Adaptation to adapt the text representation using the reconstructed object before the final mask generation via point–text matching.

Our Spatio-Functional Adaptation module is designed as follows. A frozen Qwen3-8B encoder first maps a semantic or affordance query q to a text feature $F^T \in \mathbb{R}^{1\times d}$. Instead of matching F^T directly with the unified point representation $F^P \in \mathbb{R}^{N\times d}$, we adapt the query to the current object through multi-head attention:

$$F^{TG} = F^T + \mathrm{MHA}\left(F^T,\, F^P,\, F^P\right). \tag{6}$$

Here, $\mathrm{MHA}(\cdot)$ uses F^T as the query and F^P as key and value. The residual preserves the original linguistic cue, while the attended point context injects object-specific spatial evidence. As shown in Fig. 4(c), this adaptation improves the consistency between related semantic and affordance queries while preserving the separation of queries referring to different physical regions. This allows one affordance to associate with multiple relevant semantic parts without collapsing their representations.

For final mask prediction, we match the adapted text feature F^{TG} with the unified point feature F^P to localize the queried region. The point–text similarity is computed by inner product, yielding

$$\hat{\mathbf{y}} = \sigma\big(F^P (F^{TG})^\top\big) \in [0,1]^N, \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid function. As F^{TG} is adapted to the current object, a shared matching head can predict either a semantic-part or affordance mask without attribute-specific decoders.

3.3 Training Data and Objective

To support unified semantic–affordance training, we construct a larger-scale text-mask dataset from existing 3D object datasets with PartNet (Mo et al., 2019) hierarchical semantics and 3DAffordanceNet (Deng et al., 2021) affordance labels. By cross-annotating the two sources, our constructed dataset covers 16 object categories and 16 affordance types, with broader supervision than the native PhysXNet (Cao et al., 2025) annotations. In total, we obtain about 165.6K semantic and 108.0K affordance text–mask pairs for joint training. More details are provided in the supplementary material.

Each query is paired with a binary mask $\mathbf{y} \in \{0,1\}^N$. We supervise the predicted mask $\hat{\mathbf{y}}$ with binary cross-entropy and Dice losses:

$$\mathcal{L}_{\mathrm{mask}} = \lambda_{\mathrm{BCE}}\mathcal{L}_{\mathrm{BCE}} + \lambda_{\mathrm{Dice}}\mathcal{L}_{\mathrm{Dice}}. \tag{8}$$

Together with the OT regularizer $\mathcal{L}_{\mathrm{OT}}$ in Sec. 3.1, the overall objective is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{mask}} + \lambda_{\mathrm{OT}}\mathcal{L}_{\mathrm{OT}}. \tag{9}$$

λ_{OT} balances alignment regularization. Detailed loss weights are in the supplementary material.

4 Experiments

We evaluate our jointly trained model on semantic and affordance grounding, real-image generalization, and the contributions of the proposed designs.

4.1 Experimental Setup

Datasets and baselines. We evaluate semantic segmentation and affordance segmentation on the simulated datasets PartNet (Mo et al., 2019) and 3DAffordanceNet (Deng et al., 2021), respectively, and test generalization on the real-world captured dataset DREDS (Dai et al., 2022). We compare our method with the 2D-based methods Qwen3-VL-8B-Instruct (Qwen3; Team, 2025) and Gemini (Gemini Team, Google, 2025), and with 3D-based methods. For a fair comparison, the 2D-based methods take multiple uniformly rendered views as input and back-project their predictions onto the 3D object, while the 3D-based methods take the SAM3D-reconstructed object as input.

Table 1: Semantic segmentation results (mIoU, %) on PartNet under two prompt settings. **Input** indicates whether a method is 2D-based or 3D-based. **S.**: {Semantic}; **SO.**: {Semantic} of {Object}. We report 7 representative categories for space; "· · ·" indicates omitted categories. **Avg. (all)** is the mean mIoU over *all* categories (not only those shown).

		Lamp		*Faucet*		*Chair*		*Bottle*		*Display*		*Table*		*Knife*		· · ·	**Avg. (all)**	
Input	Method	**S.**	**SO.**	**S.**	**SO.**	**S.**	**SO.**	**S.**	**SO.**	**S.**	**SO.**	**S.**	**SO.**	**S.**	**SO.**		**S.**	**SO.**
3D	PCV2.	20.41	18.61	14.05	9.96	9.19	8.89	17.16	43.30	25.06	31.14	21.22	21.29	19.24	18.59		25.59	27.71
3D	Find3D	32.13	31.74	15.20	17.32	22.82	23.08	41.85	56.19	41.70	37.56	37.89	38.29	20.95	31.05		35.11	37.21
3D	PA3D.	25.93	35.42	14.89	29.60	16.61	23.30	41.14	47.52	49.90	53.69	25.99	35.88	26.52	29.68		35.10	43.16
3D	CoS.	29.05	30.18	21.45	21.50	24.59	18.57	50.35	54.24	27.67	22.08	36.56	36.43	29.13	23.48		38.07	38.06
3D	CoS.*	46.17	36.33	35.56	26.50	51.35	41.83	53.22	29.85	55.76	52.55	48.65	37.80	49.15	35.65		49.70	40.27
2D	Qwen3	24.77	25.81	17.49	17.74	12.17	11.83	48.27	47.70	44.57	44.47	24.21	24.73	52.15	54.11		37.81	37.90
2D	Gemini	30.88	31.87	22.22	21.02	10.83	11.44	53.50	52.42	48.50	49.60	28.09	28.58	45.46	45.89		39.09	38.65
2D	**Ours**	**64.37**	**62.43**	**53.63**	**52.69**	**63.95**	**61.40**	**72.94**	**71.36**	**76.53**	**76.58**	**64.35**	**64.05**	**66.95**	**64.28**		**65.19**	**64.18**

Table 2: Affordance segmentation results (mIoU, %) on 3DAffordanceNet under two prompt settings. **A.**: {Affordance}; **AO.**: {Affordance} of {Object}. We report 8 representative categories for space; "· · ·" indicates omitted categories. **Avg. (all)** is the mean mIoU over *all* categories (not only those shown).

		lift		*sittable*		*stab*		*cut*		*listen*		*openable*		*press*		*wrap grasp*	· · ·	**Avg. (all)**	
Input	Method	**A.**	**AO.**	**A.**	**AO.**	**A.**	**AO.**	**A.**	**AO.**	**A.**	**AO.**	**A.**	**AO.**	**A.**	**AO.**	**A.**	**AO.**	**A.**	**AO.**
3D	Seq.	6.43	3.01	30.27	19.37	26.31	1.28	2.05	2.28	49.20	48.91	8.65	5.44	13.28	0.46	**81.01**	**80.72**	32.03	23.02
3D	CoS.*	6.74	6.74	31.64	31.64	48.07	48.07	48.90	48.90	49.20	**49.20**	14.29	14.32	13.68	13.69	80.77	80.43	44.48	44.46
2D	Qwen3	5.22	2.00	8.02	7.83	46.22	46.34	44.10	42.35	18.59	17.45	9.42	8.77	7.07	7.03	15.53	15.80	17.42	16.53
2D	Gemini	3.72	10.26	7.83	5.18	12.50	15.33	54.95	53.41	30.56	23.00	7.03	7.43	5.44	6.04	30.23	30.57	17.90	17.30
2D	**Ours**	**32.34**	**43.34**	**62.12**	**62.12**	**80.64**	**81.10**	**78.98**	**79.21**	**59.27**	40.35	**27.93**	**28.30**	**23.28**	**23.53**	75.13	77.16	**52.90**	**52.76**

For semantic segmentation, we compare with Find3D (Ma et al., 2025), PatchAlign3D (Hadgi et al., 2026), CoSMo3D (Jin et al., 2026), and PointCLIP V2 (Zhu et al., 2023). For affordance segmentation, we compare with SeqAfford (Yu et al., 2025). We also train CoSMo3D on our constructed dataset (Sec. 3.3), denoted CoSMo3D*.

Metrics and query settings. Following Find3D (Ma et al., 2025) and CoSMo3D (Jin et al., 2026), we measure prediction accuracy by the mIoU (%) between predicted and ground-truth masks. Affordance methods and 2D-based methods output the probability that each point belongs to the queried label. We assign a point to the label at a threshold of 0.5, the same threshold used by our method. Queries in the form {`Semantic`} or {`Affordance`}, denoted S. and A., evaluate in-distribution estimation. Queries in the form {`Semantic`} `of` {`Object`} or {`Affordance`} `of` {`Object`}, denoted SO. and AO., evaluate accuracy when the text input is out of distribution.

4.2 Comparative Results

Results on simulated datasets. Tables 1 and 2 report the results on the simulated datasets. Compared with previous 3D methods specialized separately for semantic or affordance segmentation, our unified model improves average mIoU by 27.12 and 21.02 on semantic queries without and with object context, over CoSMo3D and PatchAlign3D, and by 20.87 and 30.18 on affordance queries over SeqAfford. These gains show the effectiveness of the unified model. We further retrain CoSMo3D on our dual-attribute dataset, denoted CoSMo3D*. Our method still improves over this retrained model by 15.49 and 23.91 on semantic queries and by 8.42 and 8.30 on affordance queries, without and with object context. The remaining margin shows that the result comes from the method design as well as the dual-attribute dataset. Against the 2D models Qwen3 and Gemini, the corresponding gains are 26.10 and 35.00 on semantic and affordance queries. These models have broad empirical generalization, and part-level 3D perception still leaves them a large margin for improvement. Interestingly, these 2D models produce accurate labels when a 3D part is annotated directly, as in the dataset annotation pipeline. The gap shows that their difficulty lies in delineating the 3D part boundary from 2D views.

Table 3: Semantic and affordance segmentation on DREDS (mIoU, %).

(a) Semantic Segmentation												(b) Affordance Segmentation										
		Aeroplane		*Bottle*		*Car*		*Mug*		**Avg.**			*contain*		*grasp*		*lift*		*openable*		**Avg.**	
Input	Method	**S.**	**SO.**	**S.**	**SO.**	**S.**	**SO.**	**S.**	**SO.**	**S.**	**SO.**	Method	**A.**	**AO.**	**A.**	**AO.**	**A.**	**AO.**	**A.**	**AO.**	**A.**	**AO.**
3D	CoS.	**22.36**	**16.37**	37.27	46.77	30.84	28.76	**89.93**	**89.14**	39.19	38.93	Seq.	**83.97**	**83.95**	1.02	21.95	53.72	0.17	7.45	0.36	53.39	47.69
2D	Qwen3	1.48	1.51	38.11	37.85	5.17	4.10	4.28	4.65	12.48	12.37	Qwen3	0.91	1.00	0.33	0.78	0.37	0.97	0.59	0.10	1.01	1.11
2D	Gemini	4.60	3.36	37.56	37.61	9.35	10.27	10.05	9.28	15.38	15.24	Gemini	2.34	2.25	2.27	2.16	2.70	2.33	0.14	0.15	2.58	2.58
2D	**Ours**	10.68	11.07	**70.34**	**70.22**	**32.13**	**32.83**	86.00	86.01	**42.71**	**43.75**	**Ours**	79.66	73.82	**72.85**	**72.81**	**59.69**	**66.40**	**30.98**	**30.98**	**62.27**	**62.22**

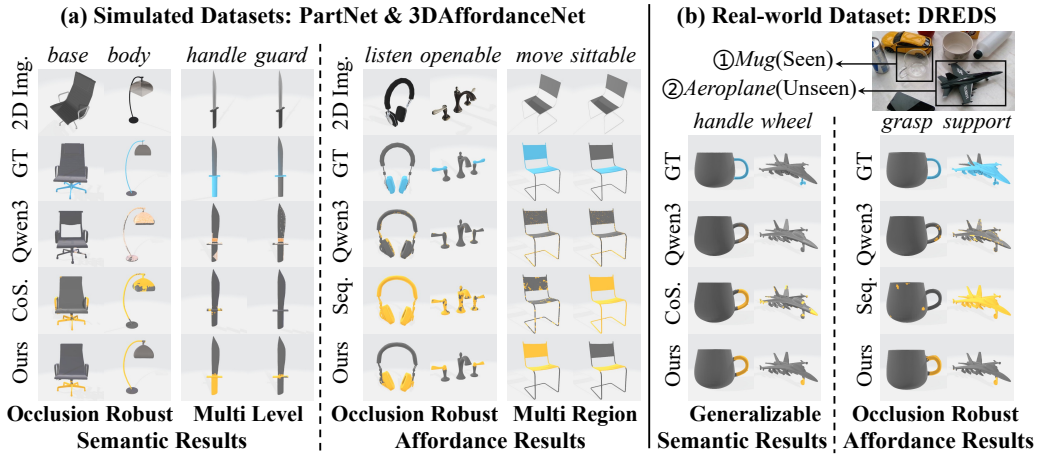


Figure 5: Qualitative results on simulated and real-world datasets.

Further analysis is shown in Fig. 5. Under partial observation, our method grounds both semantic and affordance queries on the complete reconstructed object, including surfaces absent from the input view. The prediction also follows the query: semantic masks separate a knife's *handle* from its finer *guard*, and affordance masks can cross semantic boundaries, with *move* covering the chair seat and back and *support* concentrating on the seat.

Results on real-world images. Table 3 reports results on captured DREDS images. Compared with the strongest 3D baselines, our method improves average mIoU by 3.52 on semantic queries over CoSMo3D and by 8.88 on affordance queries over SeqAfford. Against the 2D models Qwen3 and Gemini, the gains are 27.33 and 59.69 on semantic and affordance queries. Real captured objects have more severe viewpoint variation and blur, so segmenting and understanding the object directly in the 2D image is more difficult. The model also generalizes to unseen categories such as *Can* and *Car*. Fig. 5 shows that our predictions more completely localize the grasp region of a mug and the structure of an aeroplane for both semantic and affordance queries. Competing methods often miss occluded regions or produce fragmented masks.

4.3 Ablation and Analysis

Ablation on feature fusion. To verify the effectiveness of optimal-transport fusion, we ablate the fusion strategy in the third and fourth rows of Table 4. Fusing the two features with optimal transport improves semantic mIoU by 8.81 over direct addition. Affordance differs from semantics in its stronger one-to-many correspondence. The *move* region, for example, overlaps both the chair seat and the back in the point cloud. Feature fusion does not resolve this correspondence; it is handled by Spatio-Functional Adaptation.

Ablation on Spatio-Functional Adaptation. To verify Spatio-Functional Adaptation, we compare the sixth and seventh rows of Table 4. Adding this module improves semantic mIoU by 10.85 and affordance mIoU by 7.73. The module uses the current object to associate an affordance such as *move* with multiple overlapping semantic regions, including the chair seat and back.

Ablation on the dataset. We ablate the training data in the fifth and sixth rows of Table 4. Adding our dual-attribute dataset improves semantic mIoU by 1.34 and affordance mIoU by 2.43.

Table 4: Ablation study. Feat. U./S.: Utonia/SAM3D features; Fusion A./OT.: additive/optimal-transport fusion. Qwen3: replacing SigLIP with Qwen3-8B. New Data: our dataset with online text-feature augmentation; Spatio-func.: Spatio-Functional Adaptation.

Feat. U.	Feat. S.	Fusion A.	Fusion OT.	Qwen3	New Data	Spatio-func.	Sem. (**S.**)	Sem. (**SO.**)	Afford. (**A.**)	Afford. (**AO.**)
✓	×	×	×	×	×	×	46.04	42.56	39.92	38.13
×	✓	×	×	×	×	×	46.32	43.20	38.43	34.50
✓	✓	✓	×	×	×	×	44.54	38.29	36.73	33.65
✓	✓	×	✓	×	×	×	53.35	49.89	35.02	31.81
✓	✓	×	✓	✓	×	×	53.00	39.86	42.74	37.64
✓	✓	×	✓	✓	✓	×	54.34	51.74	45.17	43.05
✓	✓	×	✓	✓	✓	✓	**65.19**	**64.18**	**52.90**	**52.76**

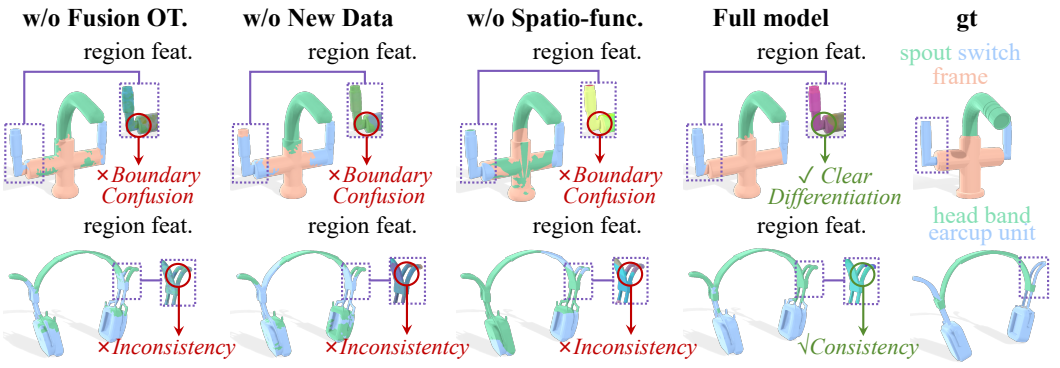


Figure 6: Qualitative ablation. From left to right: ground truth, without Fusion OT., our constructed data, Spatio-functional adaption models and the full model.

As shown in Table 5, joint training improves semantic mIoU by 1.31 and affordance mIoU by 13.97 over task-specific training. The two forms of supervision improve joint training, and the affordance gain is larger.

Table 5: Comparison of task-specific training and joint semantic-affordance training (mIoU, %). "–" denotes a setting not applicable under the corresponding training protocol.

Training data	Sem.(**S.**)	Sem.(**SO.**)	Afford.(**A.**)	Afford.(**AO.**)
Semantic-only	63.88	62.01	–	–
Affordance-only	–	–	38.93	38.34
Joint training (Ours)	**65.19**	**64.18**	**52.90**	**52.76**

Qualitative analysis. Fig. 6 shows the effects of feature fusion, the constructed dataset, and Spatio-Functional Adaptation. Removing any component weakens feature consistency within a region and feature separability across region boundaries. The full model improves both.

5 Conclusion

We presented a unified framework for semantic and affordance grounding over complete 3D objects from a single RGB image. Our method aligns and fuses 2D- and 3D-derived features through optimal transport, while adapting text representations with object-specific spatial and functional cues. We further constructed a dual-attribute dataset that provides enough supervision for both tasks. Experiments on synthetic and real-world benchmarks demonstrate consistent improvements in semantic and affordance segmentation, highlighting the benefits of jointly learning semantic and functional knowledge for complete 3D object understanding.

AI use statement

In this work, we used generative AI tools for limited early-stage assistance in dataset construction: Gemini-3.5-Flash helped with bidirectional cross-annotation between semantic parts and affordance labels from multi-view renderings. All such outputs were verified, filtered, and corrected by the authors before inclusion in the final corpus.

We have not used generative AI tools to formulate mathematical claims, assist in writing proofs, propose hypotheses, design experiments, interpret results, or produce scientific claims; other required disclosure tasks are not applicable.

Additionally, we used generative AI tools for literature search and minor readability edits. All citations, factual checks, and final text were handled by the authors.

We have reviewed all AI-assisted work and take responsibility for the final content of this work.

Ethics statement

This work studies open-vocabulary semantic and affordance segmentation on 3D object shapes and does not involve human subjects, personal data, or user studies. Our experiments use existing public 3D object datasets together with annotations constructed for research purposes; we do not collect private or sensitive information. All authors acknowledge and adhere to the ICLR Code of Ethics.

Reproducibility statement

To facilitate reproducibility, we summarize where the relevant materials can be found. The main paper presents the key method and experimental findings. Detailed implementation and training settings are provided in the supplementary material. A complete description of the dataset construction and processing pipeline is also included in the supplementary material. We will release the code, pretrained weights, and the constructed dataset upon acceptance. Together, these materials provide the information needed to reproduce our results.

References

Ahmed Abdelreheem, Ivan Skorokhodov, Maks Ovsjanikov, and Peter Wonka. SATR: Zero-shot semantic segmentation of 3d shapes. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Ziang Cao, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-3d: Physical-grounded 3d asset generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Xingyu Chen, FU-JEN CHU, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jia-Wei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. Sam 3d: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7220–7232, June 2026.

Hengshuo Chu, Xiang Deng, Qi Lv, Xiaoyang Chen, Yinchuan Li, Jianye Hao, and Liqiang Nie. 3D-AffordanceLLM: Harnessing large language models for open-vocabulary affordance detection in 3d worlds. In *International Conference on Learning Representations*, 2025.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, pp. 2292–2300, 2013.

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, June 2021.

Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL https://arxiv.org/abs/2507.06261.

Souhail Hadgi, Bingchen Gong, Ramana Sundararaman, Emery Pierson, Lei Li, Peter Wonka, and Maks Ovsjanikov. PatchAlign3D: Local feature alignment for dense 3d shape understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, Qihang Yu, and Alan Yuille. PartImageNet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, 2022.

Li Jin, Weikai Chen, Yujie Wang, Yingda Yin, Zeyu Hu, Runze Zhang, Keyang Luo, Shengju Qian, Xin Wang, and Xueying Qin. CoSMo3D: Open-world promptable 3d semantic part segmentation through LLM-guided canonical spatial modeling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Hyunjin Kim and Minhyuk Sung. PartSTAD: 2d-to-3d part segmentation task adaptation. In *European Conference on Computer Vision*, 2024.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lai_LISA_Reasoning_Segmentation_via_Large_Language_Model_CVPR_2024_paper.html.

Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. LOCATE: Localize and transfer object parts for weakly supervised affordance grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Junyi Li, Junfeng Wu, Weizhi Zhao, Song Bai, and Xiang Bai. PartGLEE: A foundation model for recognizing and parsing any objects. In *European Conference on Computer Vision*, 2024b.

Minghua Liu, Yinhao 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Dongyue Lu, Lingdong Kong, Tianxin Huang, and Gim Hee Lee. GEAL: Generalizable 3d affordance learning with cross-modal consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Ziqi Ma, Yisong Yue, and Georgia Gkioxari. Find any part in 3d. In *IEEE/CVF International Conference on Computer Vision*, 2025.

Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 909–918, June 2019.

Toan Nguyen, Minh Nhat Vu, Baoru Huang, Tuan Van Vo, Vy Truong, Ngan Le, Thieu Vo, Bac Le, and Anh Nguyen. Language-conditioned affordance-pose detection in 3d point clouds. *arXiv preprint arXiv:2309.10911*, 2023a.

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 *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2023b.

OpenAI. GPT-4 technical report. Technical report, OpenAI, 2023. URL https://cdn.openai.com/papers/gpt-4.pdf.

Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2401.14398.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. PixelLM: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2312.02228.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *IEEE/CVF International Conference on Computer Vision*, 2019.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Peize Sun, Shoufa Chen, Chenchen Zhu, Fanyi Xiao, Ping Luo, Saining Xie, and Zhicheng Yan. Going denser with open-vocabulary part segmentation. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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*, 2025. URL https://arxiv.org/abs/2507.23734.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4840–4851, June 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

Yujia Zhang, Xiaoyang Wu, Yunhan Yang, Xianzhe Fan, Han Li, Yuechen Zhang, Zehao Huang, Naiyan Wang, and Hengshuang Zhao. Utonia: Toward one encoder for all point clouds. In *ICML*, 2026.