

UniPart: Towards Zero-shot Language-Grounded 3D Part Segmentation for Embodied Interaction

Xinqiang Yu^{1,4,6*}, Zekun Qi^{3,4*}, Jiawei He⁶, Wenyao Zhang⁵, Xuchuan Chen^{3,4}, Guocai Yao⁶, Li Yi^{3†}, Zhaoxiang Zhang^{1†}, and He Wang^{2,4†}

¹ NLPR, Institute of Automation, Chinese Academy of Sciences, China

² Peking University, China

³ Tsinghua University, China

⁴ Galbot Inc., China

⁵ Shanghai Jiao Tong University, China

⁶ Beijing Academy of Artificial Intelligence, China

* Equal contribution. † Corresponding authors.

{yuxinqiang, qizekun, chenxuchuan, yili, wanghe}@galbot.com

{jwhe2024, wyzhang2022, }@gmail.com

gcyao1@baai.ac.cn

zhaoxiang.zhang@ia.ac.cn

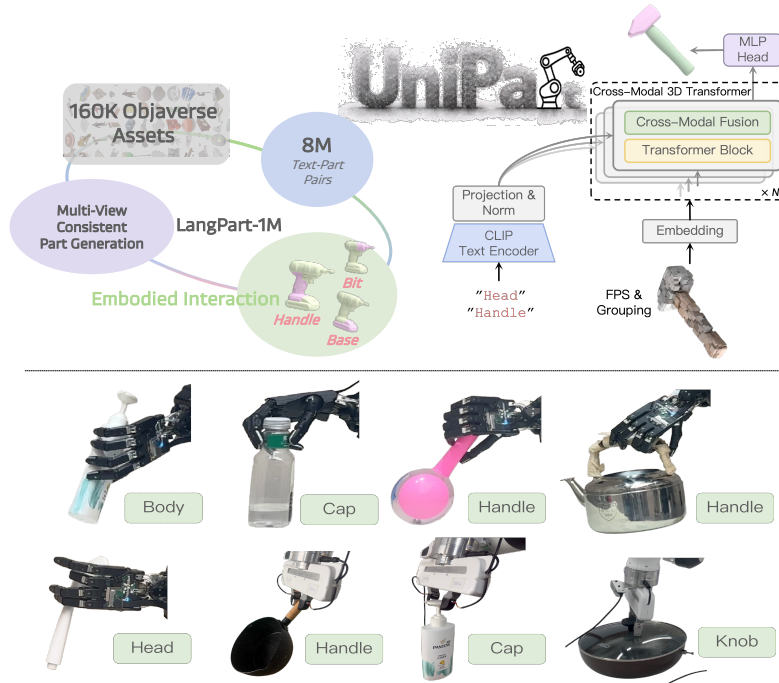


Fig. 1: Overview. We construct LangPart-1M from 160K+ Objaverse assets and generate 8M text to part pairs using a scalable multi-view consistent part generation pipeline. UniPart is a feed-forward text-conditioned 3D part segmentation model trained on this dataset. The predicted part masks provide fine-grained geometric cues that support embodied interaction.

Abstract. Fine-grained robotic manipulation depends on understanding parts, not only whole objects. Existing 3D foundation models tend to be either generalized but object-aware, or part-aware but limited to closed-set taxonomies, which weakens zero-shot transfer. We study text-conditioned 3D part segmentation, where a free-form phrase selects a functional part on point cloud. We introduce UniPart, a feed-forward cross-modal 3D Transformer that conditions CLIP text embedding. To scale supervision, we build LangPart-1M with 160K+ Objaverse assets and 8M text to part pairs using multi-view consistent part generation. We further manually label a high-quality subset, LangPart-4K, for fine-tuning and evaluation. UniPart achieves strong zero-shot results on open-vocabulary part benchmarks and transfers to language-conditioned part grasping in real world.

Keywords: Zero-Shot · 3D Part Segmentation · Embodied Interaction

1 Introduction

The quest for truly general-purpose embodied intelligence hinges on a robot’s ability to not merely perceive objects, but to comprehend their function and composition at a granular level [18, 44, 51, 54]. In open-world settings, manipulation depends on generalizing to unseen objects and to the parts that matter for an action. A robot must identify the semantically relevant object part on a previously unseen instance, conditioned on a human instruction.

Foundation models such as SAM [29] and CLIP [57] have made 2D perception broadly generalizable, and several lines of work extend these priors to 3D. Yet fine-grained part understanding remains difficult [4, 51, 54]. Existing approaches often fall into two patterns that leave a gap for dexterous manipulation.

- **Promptable but coarse** Models that inherit SAM-style prompting in 3D, such as Point-SAM [86], are strong at separating object instances but do not resolve internal semantically relevant parts within one object.
- **Part-aware but closed-set** Part and affordance models trained on datasets such as ShapeNetPart [72] provide fine granularity, yet they rely on fixed taxonomies and struggle with open-vocabulary generalization.

We frame the missing capability as a semantic interface for fine-grained 3D geometry. Instead of predicting from a fixed label set, a robot should take a natural language instruction and return the corresponding 3D part mask. We study *language-grounded zero-shot 3D part segmentation* and tackle it with a paired design of supervision and model.

On the supervision side, we build LangPart-1M. The key challenge is not only scale, but also multi-view consistency of part masks when lifting from 2D to 3D. Our pipeline combines multi-view rendering, SAM region proposals, Set-of-Mark prompting, and a text-guided merging step followed by lifted to 3D. This produces point-level part masks aligned to language at scale. We then curate LangPart-4K with manual verification so that training can benefit from both

breadth and high-fidelity corrections. On the model side, we introduce UniPart, which is a feed-forward cross-modal 3D Transformer that injects a frozen CLIP text embedding into every layer through simple addition. This choice keeps the architecture lightweight while maintaining strong language conditioning. We also include a controlled study that separates the effect of the dataset from the effect of the model, which is critical for positioning against closely related work.

In summary, our key contributions are:

- **LangPart-1M and LangPart-1K** We introduce a large-scale dataset with 160K objects and 8M text-to-part pairs, and a benchmark split with 1K objects and more than 4K annotations for open-vocabulary evaluation.
- **Scalable data engine and high-quality subset** We present a multi-view pipeline that aligns language to 3D part masks using SAM proposals, Set-of-Mark prompting, and text-guided merging for view consistency. We additionally build LangPart-4K with ten annotators and assign text labels to part masks, providing high-fidelity language-grounded 3D part supervision.
- **UniPart model and controlled positioning** We propose a feed-forward cross-modal 3D Transformer with layer-wise additive text injection, and we report a matched supervision study that disentangles the gains from data and from architecture.
- **Embodied validation** We show that UniPart can produce part masks on real sensor point clouds that are useful for language-conditioned grasping.

2 Related Works

2.1 3D Part Segmentation

Classic 3D part segmentation relies on closed-set supervision, such as ShapeNet-Part [72] and PartNet [42], and achieves fine granularity within fixed taxonomies. More recent work leverages 2D vision language priors to reduce labeling cost, yet open-vocabulary generalization is still limited [34]. In parallel, promptable and open-vocabulary 3D segmentation extends SAM [28] and CLIP [56] into 3D through multi-view lifting or 3D scene pipelines, and it is mainly evaluated on instance or scene semantics rather than object internal parts [17, 55, 87]. Recent efforts also explore open-vocabulary part segmentation in narrower domains or synthetic settings [65].

Our work is closest to concurrent text-prompted part segmentation such as FIND3D [38]. We share the goal of querying arbitrary parts with language, while differing in two aspects that we evaluate under matched supervision. First, we construct LangPart-1M using a multi-view pipeline that enforces 3D consistency through text guided mask merging. Second, we train a feed-forward point-based model with layer-wise additive text injection, which provides dense part masks without multi-view querying at inference. We include a controlled study that isolates the effects of supervision and architecture.

Related but distinct, language-guided 3D affordance segmentation focuses on action regions rather than semantic parts and often depends on task-specific

definitions [37, 66]. UniPart targets category-agnostic part masks as a perception primitive that can be composed into downstream manipulation.

2.2 Language-Grounded Robot Manipulation

Language-grounded robot Manipulation adopts the human language as a general instruction interface. Existing works can be categorized into two groups: (i) *End-to-end* models like RT-series [2, 3, 89] built upon unified cross-modal Transformers with tokenized actions [33, 58, 82], large vision-language-action models built from VLMs [27, 81], or 3D representations [75, 83]. Training on robot data such as Open X-Embodiment [5] and DROID [26], a remarkable process has been made. However, the data *scale* is still limited compared to in-the-wild data for training VLMs. (ii) *Decoupled* high-level reasoning and low-level actions in large VLMs and small off-the-shelf policy models, primitives [16, 20, 22, 24, 25, 32, 40, 71, 76, 85], or articulated priors [23, 31]. Our model lies in this group, where an open-world generalization property emerges from VLMs and our proposed UniPart is empowered by fine-grained understanding.

2.3 3D Representation Learning

Research on 3D Representation Learning encompasses various methods, including point-based [48, 49], voxel-based [41], and multiview-based approaches [19, 60]. Point-based methods [14, 53] have gained prominence in object classification [62, 68] due to their sparsity yet geometry-informative representation. On the other hand, voxel-based methods [8, 52, 73] offer dense representation and translation invariance, leading to a remarkable performance in object detection [6] and segmentation [1, 72]. The evolution of attention mechanisms [63] has also contributed to the development of effective representations for downstream tasks, as exemplified by the emergence of 3D Transformers [14, 36, 39]. Notably, 3D self-supervised representation learning has garnered significant attention in recent studies. PointContrast [69] utilizes contrastive learning across different views to acquire discriminative 3D scene representations. Innovations such as Point-BERT [74] and Point-MAE [45] introduce masked modeling [9, 21] pretraining into the 3D domain. ACT [12] pioneers cross-modal geometry understanding through 2D or language foundation models such as CLIP [56] or BERT [9]. Following ACT, RECON [50] further proposes a learning paradigm that unifies generative and contrastive learning. PPT [79] highlights the significance of positional encoding in 3D representation learning. Additionally, leveraging foundation vision-language models like CLIP [12, 56] has spurred the exploration of a new direction in open-world 3D representation learning. This line of work seeks to extend the applicability and adaptability of 3D representations in diverse and open-world/vocabulary scenarios [10, 11, 15, 46, 78, 80].

Table 1: Properties and inference time of UniPart and baselines. We report whether each method is open-world, supports cross-category transfer, predicts part-level masks, and runs as a feed-forward 3D model without multi-view rendering and backprojection at inference. We also report a representative runtime and the number of training objects when applicable.

Method	Time	Open-world	Cross-category	Part-level	Feed-forward	Num. Object
PointCLIPV2	5.4s	✓	✓	✓	✗	-
PartSLIP++	174.3s	✓	✗	✓	✗	-
OpenMask3D	296.5s	✓	✓	✗	✗	-
PartDistill	0.7s	✗	✗	✓	✓	-
PointNext	1.4s	✗	✓	✓	✓	-
FIND3D	0.9s	✓	✓	✓	✓	36k
UniPart(Ours)	0.4s	✓	✓	✓	✓	160k+

3 UniPart: Language-Grounded 3D Part Segmentation for Embodied Interaction

3.1 Datasets

Scaling text-part alignment dataset to hundreds of thousands of 3D assets requires balancing annotation scalability and quality. Fully manual labeling at this scale is impractical, whereas fully automated annotation inevitably introduces noise. To address this trade-off, we construct our dataset in two complementary stages. The first subset **LangPart-1M** is generated through a fully automated pipeline leveraging large-scale foundation models, providing broad coverage and diversity across object categories. While scalable, this subset may contain minor inconsistencies due to model-induced noise. To provide high-fidelity semantic supervision, we additionally curate a manually annotated subset **LangPart-4K** with carefully verified part labels, which is split into **LangPart-3K** for training and **LangPart-1K** as benchmark.

During training, we adopt a **three-stage curriculum**. We first align point-cloud representations with RGB features to inject rich cross-modal semantic cues into the 3D backbone, establishing an initialization that is both geometrically grounded and semantically aware. We then pretrain the model on large-scale data to acquire broad structural and semantic priors at scale. Finally, we fine-tune on the manually annotated subset to correct the noise of automatic supervision and sharpen part-level localization accuracy. This staged design progressively transfers knowledge from cross-modal semantic alignment to scalable representation learning, and ultimately to high-fidelity part understanding.

Dataset Statistics To facilitate rigorous evaluation and training, LangPart-1M is curated with extensive coverage and fine-grained annotations. Specifically, the dataset contains over 160K+ unique 3D objects spanning more than 250 distinct object categories (e.g., furniture, tools, appliances, vehicles), as shown in Tab. 1. Within these objects, we annotate over 1,000 unique part types (e.g.,

handle, button, spout, screen), yielding a total of approximately 8 million text-part pairs. We train on the remaining data after holding out the benchmark split, and report results on our proposed benchmark.

Data Source To enable large-scale training, we construct the LangPart-1M dataset based on Objaverse [7], which contains approximately 800K Internet-sourced 3D models across diverse categories. However, Internet-sourced models often exhibit noisy annotations and inconsistent visual quality. To obtain clean and consistent supervision, we re-render the assets in Blender with controlled lighting, generating over 8M high-fidelity images.

Data Filtering To clean the data that can be better used for generating fine-grained part annotations, we first clean the data by using a dedicated filtering strategy that filters data to preserve samples that satisfy the following six requirements. ❶ Clean objects without the ground for auxiliary visualization. ❷ Reasonable objects that have sufficient spatial reasoning potentials. ❸ High-quality objects. Low-quality objects such as blurry and wrong samples are filtered. ❹ Distinguishable objects. Abstract objects such as meaningless solids are filtered. ❺ Non-scene objects. Samples that describe a 3D scene are filtered for object-centric understanding purposes. ❻ No test leakage. We proactively remove all asset IDs from evaluation benchmarks, including Objaverse-General, from the training pool.

Filtering at this scale is impractical to do fully by hand. Inspired by recent work on using VLMs as image-based judges [47, 67, 84], we use GPT-5.2-Pro [59] with a fixed prompt and deterministic decoding to decide whether an asset passes the criteria. For reproducibility, we will release the full filtering prompt, the list of retained asset IDs, and the per-asset decisions so that the training set can be reconstructed without re-running the VLM.

LangPart-1M Dataset Generation To generate the fine-grained annotations for our dataset, we devise a novel pipeline that systematically leverages large-scale foundation models. As shown in Fig. 2, this process follows four stages:

- (i) **Part-List Querying:** For each cleaned 3D object, we first employ GPT-5.2-Pro [59] to query and generate its object name and a comprehensive list of its potential functional or meaningful parts.
- (ii) **2D Unsupervised Segmentation** For each object, we render twelve randomly sampled single-view images in Blender. For each view, we apply SAM [28] to obtain all possible segmentations, producing multiple fine-grained part candidates per image. To enable large language models to reason about these spatial regions, we employ the Set-of-Mark [70] prompting technique, assigning a unique visual identifier to each generated segment.
- (iii) **Multi-View Consistent Part Generation** Lifting 2D masks into 3D often produces view-dependent inconsistencies. We address this with a text guided

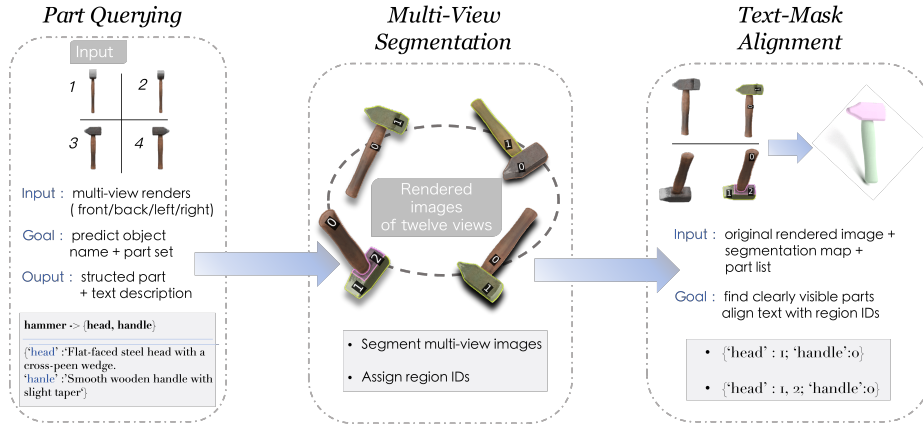


Fig. 2: Pipeline of **LangPart-1M** dataset generation, including Part Querying, Multi-View Segmentation and Text-Mask Alignment.

merging step followed by lifted to 3D. For each view, we provide the SoM labeled image to GPT-5.2-Pro and ask it to select the visible parts from the queried part list and to merge the region IDs that belong to the same part. This step improves cross-view coherence, merges over-segmented regions, and suppresses isolated outliers.

(iv) **Text to Point Cloud Alignment** Finally, using the camera intrinsics, extrinsics, and depth maps from our renderer, we aggregate the back-projected points and obtain a unified point cloud with per-point part labels. To support reproducibility, we will release the intermediate artifacts needed to replay the pipeline, including render metadata, depth maps, SAM masks, SoM overlays, and the merged region sets produced during alignment.

Manual Annotation for Quality Improvement After data filtering in Sec. 3.1, we obtain over 160K 3D assets from Objaverse. However, not all objects are equally relevant for functional robotic manipulation. We therefore manually select 4K objects and manually construct and label a high-quality subset for post-training, named **LangPart-4K** by two stages.

(i) **Part-Aware Object Filtering** In practice, robots primarily interact with objects that exhibit clear functional or structurally meaningful parts. Therefore, from the 160K objects, we further select 4K objects that contain *explicit functional components or well-defined part structures*. The subset is manually annotated to provide high-fidelity supervision for the third training stage.

(ii) **Refinement of Auto Annotation** For each object in LangPart-4K, we start from the automatically generated multi-view masks from Sec. 3.1. Annotators inspect the correspondence between each part name and its regions in every rendered view and correct mismatches when needed. This yields a verified set of part masks that is used for fine-tuning and for constructing our benchmark.

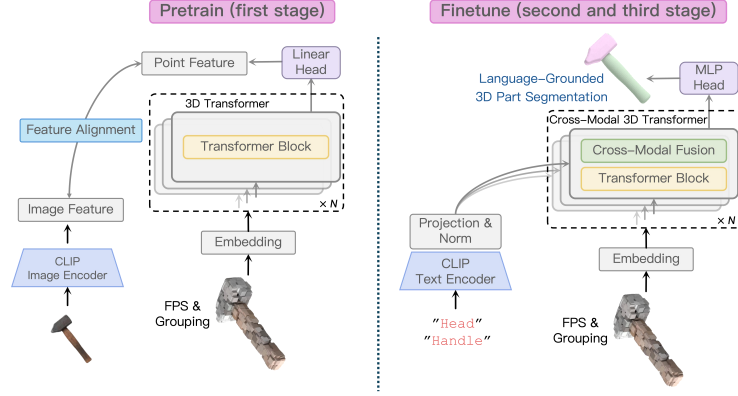


Fig. 3: UniPart architecture for part segmentation. During the pretraining stage, UniPart is trained to align image feature and point feature. And for the finetuning stage, UniPart takes 3D & text embeddings as inputs of a N -layer plain Transformer model, and uses a token-level sum as cross-modal fusion in all layers. A PointNet decoder predicts the logits per point.

Text-Alignment Part Segmentation Benchmark Existing 3D part segmentation benchmarks are limited in diversity and flexibility. They cover only a small set of object categories, rely on closed-set part taxonomies, and often contain canonicalized CAD models. ShapeNetPart has 16 categories and 41 part labels [72]. PartNet-E expands to 45 categories but still uses a fixed vocabulary shared across categories [42]. After manual verification in Sec. 3.1, we split LangPart-4K into LangPart-3K for training and LangPart-1K with 1K objects for evaluation. LangPart-1K is designed to evaluate both seen and unseen categories and part types, and we will release the split specification together with the asset IDs and text prompts used for evaluation.

3.2 A Cross-Modal 3D Transformer for Language-Grounded Part Segmentation

We introduce UniPart, a simple Transformer-based architecture [63] with cross-modal point cloud–language fusion for open-vocabulary text-conditioned part segmentation. As illustrated in Fig. 3, UniPart takes an object’s point cloud and a language description as input and predicts the corresponding text-conditioned segmentation mask.

Language and Point Cloud Embeddings Given an object point cloud $X = \{\mathbf{x}_i \in \mathbb{R}^3 | i = 1, 2, \dots, N\}$ in Cartesian coordinates and an arbitrary language description ℓ , we first tokenize both modalities into discrete embeddings. For the 3D input, following [12, 50, 74], we sample N_s seed points via Farthest Point Sampling (FPS) and form local neighborhoods using k -nearest Neighbors (kNN).

Each neighborhood is encoded into a local geometric token using a lightweight PointNet-style feature extractor [48, 49], yielding a sequence of 3D tokens. For language, we adopt OpenAI CLIP [56] and use its global text embedding as the conditioning signal for cross-modal fusion with the 3D tokens. The fused representations are then processed by a Transformer encoder. Finally, a MLP-style decoder head upsamples [CLS] token features back to per-point predictions and outputs text-conditioned binary mask over the input points, producing an open-vocabulary part segmentation mask corresponding to ℓ . Similar to Vision Transformer [13], we adopt a Large configuration with more Transformer layers, larger hidden dimensions, and a stronger CLIP text backbone.

Cross-Modal Fusion We adopt a dense, layer-wise fusion strategy that injects the global text embedding into each layer of the 3D Transformer. In principle, cross-modal fusion can be implemented in many ways, e.g., cross-attention, adapters, or feature concatenation along the channel dimension. Empirically, we find that a simple additive injection—adding the text feature to every point token at each layer—yields the best trade-off between simplicity and effectiveness. We hypothesize this is because the language input is relatively short, and per-token addition strengthens language conditioning throughout the network.

3.3 Optimization

Let $\mathcal{F}_{\text{UP}}$ represent the UniPart model parameterized by θ_{UP} (the CLIP is kept frozen and thus its parameters are not included). Given every object point cloud $X_i \in \mathcal{D}_{\text{LangPart-1M}}$ in the LangPart-1M dataset, where each object is labeled with a part text set $L_i = \{\ell_j^i, j = 1, 2, \dots, P_i\}$ where P_i is the part number of point cloud X_i . And the corresponding ground truth 3D part segmentation mask $S_i = \{s_j^i, j = 1, 2, \dots, P_i\}$ where S_i is an one-hot label. For the second and third **finetune stage**, the optimization is to minimize the Binary Cross-Entropy (BCE) loss between predicted segmentation mask and ground truth label, where $\mathcal{L}_{\text{BCE}}(\hat{S}_j^i, S_j^i) = -[S_j^i \log(\hat{S}_j^i) + (1 - S_j^i) \log(1 - \hat{S}_j^i)]$. And the final optimization function is

$$\min_{\theta_{\text{UP}}} \sum_{X_i \in \mathcal{D}_{\text{LangPart-1M}}} \sum_{S_j^i \in S_i} \mathcal{L}_{\text{BCE}}(\hat{S}_j^i, S_j^i) \quad (1)$$

During the first **pretraining stage**, for each single view point cloud $X_i^{\text{sin}} \in \mathcal{D}_{\text{LangPart-1M}}$, and a corresponding RGB image I_i rendered from the same view, and a mapping $\pi(i) \in \{1, \dots, M\}$ that assigns each point to an image patch index, where M is the image patch number. We extract patch-level CLIP features from the image $\{c_j\}_{j=1}^M$ and align them with per-point features produced by the point encoder, encouraging the 3D representation to inherit CLIP’s semantic structure. This initialization significantly facilitates subsequent text-guided part understanding during fine-tuning. And the loss function is

$$\mathcal{L}_{\text{align}} = \max\left(0, 1 - \frac{1}{N} \sum_{i=1}^N \langle \hat{\mathbf{z}}_i, \hat{\mathbf{c}}_{\pi(i)} \rangle\right) \quad (2)$$

where $\hat{\mathbf{z}}_i$ is the normalized point feature, $\hat{\mathbf{c}}_{\pi(i)}$ is its corresponding normalized CLIP patch feature, N is the point number, and $\langle \cdot, \cdot \rangle$ denotes the inner product.

3.4 Comparison with Related Works

Recent work such as FIND3D [38] also studies text-prompted 3D part segmentation. We focus on two differentiators that we evaluate under matched supervision. LangPart-1M provides large-scale point-level supervision constructed with a multi-view consistency mechanism based on text-guided merging. We report a controlled study that isolates the contribution of the supervision engine and the contribution of the architecture, and we test the resulting part masks in a real robotic grasping pipeline.

4 Experiments

4.1 Experimental Settings

Benchmarks We evaluate on three settings. We report results on **Objaverse-General** [38] to compare with prior open-world baselines. We additionally evaluate on **ShapeNetPart** [72] and PartNet-E [42] to test transfer to classic part benchmarks. Finally, we evaluate on **LangPart-1K**, which is constructed from LangPart-4K with manual verification.

Metric We report mean Intersection over Union. For each input point cloud, we compute the mean IoU across its annotated parts, then average over the point cloud.

4.2 Part Segmentation Baselines

Open-world Baselines To ensure a consistent and fair comparison, we follow the baseline selection and evaluation protocol in FIND3D [38] and detail the adopted methods below. We use the same two prompt formats, a compositional query of the form `{part} of a {object}` and a short query `{part}`.

- **PointCLIPV2** [88] is an open-world 2D-to-3D framework that repeatedly queries CLIP [56] across rendered views. It relies on a prompt-selection procedure that retains the top- k textual prompts for each object, with $k = 1400 \times n_{\text{parts}}$, where the prompt set is chosen using the ShapeNetPart test split.
- **PartSLIP++** [35] follows a detection-driven design, leveraging GLIP [30] together with a dedicated superpoint generation module to transfer 2D detections back to 3D. The method additionally trains category-specific models on PartNet-E [43]; for consistency and fairness, we report results using its released zero-shot checkpoint.
- **OpenMask3D** [61] is an open-vocabulary 2D-to-3D approach originally trained on 3D scenes. Both PointCLIPV2 and OpenMask3D produce dense predictions by assigning a label to every point. When the benchmark includes unlabeled regions, we add an “other” text query to represent the no-label option.

Table 2: Performance comparison with open-world methods on Objaverse-General. UniPart performs best on Objaverse-General benchmark. We show results evaluated with 2 common query prompts: “{part} of a {object}” and “{part}” for all methods following FIND3D [38].

Methods	Objaverse-General			
	Seen Categories		Unseen Categories	
	{part} of a {object}	{part}	{part} of a {object}	{part}
PointCLIPV2	9.81	11.27	10.27	11.09
PartSLIP++	2.69	15.03	0.57	10.43
OpenMask3D	11.81	11.93	7.01	10.31
FIND3D	33.78	34.10	26.21	27.41
UniPart(ours)	46.43	49.27	41.46	45.23

Table 3: Performance comparison with open-world methods on ShapeNetPart. UniPart also performs best on this benchmark. And we also evaluate with prompts “{part} of a {object}” and “{part}”.

Methods	ShapeNetPart					
	Canonical		Rotated		ShapeNetPart-V2	
	{part} of a {object}	{part}	{part} of a {object}	{part}	{part} of a {object}	{part}
PointCLIPV2	16.91	20.22	16.88	18.19	15.14	17.11
PartSLIP++	1.43	6.46	0.94	6.03	1.54	11.62
OpenMask3D	8.94	10.37	6.75	14.56	15.87	13.77
FIND3D	28.39	24.09	29.64	23.71	42.15	30.02
UniPart(ours)	40.07	41.75	42.43	39.56	55.63	51.29

Table 4: Comparison on LangPart-1K benchmark.

Methods	LangPart-1K benchmark			
	Single-View point cloud		Whole Object point cloud	
	{part} of a {object}	{part}	{part} of a {object}	{part}
FIND3D	10.39	8.97	14.36	17.28
UniPart(ours)	31.23	33.56	28.69	27.17

4.3 Experimental Results

Results on Objaverse-General, ShapeNetPart, and PartNet-E Table 2 and Table 3 report mIoU on Objaverse-General and ShapeNetPart. As shown in this table, our method UniPart can perform better than others. As shown in Table 5, we compare open-world part segmentation methods on PartNet-E. Among these approaches, Our method has best performance.

Qualitative results Figure 4 shows qualitative examples from our model predictions. Our benchmark and dataset has more objects and more parts than FIND3D. And our datasets include single-view and whole point cloud.

Results on LangPart-1K Our dataset provides both single-view point clouds and complete 3D point clouds. Accordingly, we train two separate models—one

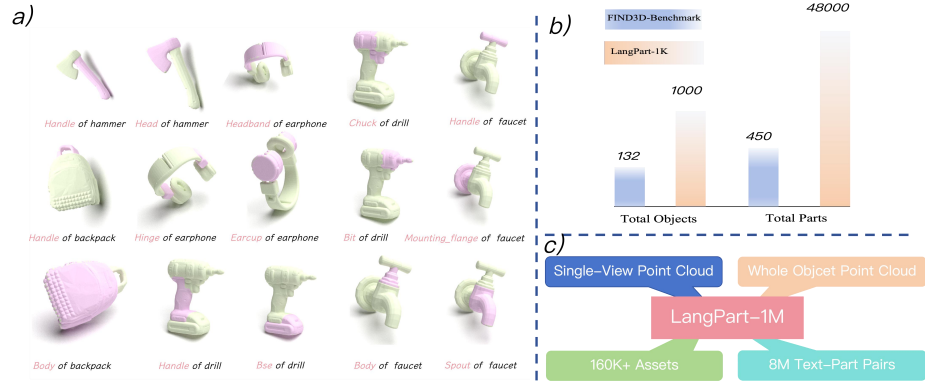


Fig. 4: Qualitative results and our dataset. a) Examples of our dataset. b) Our benchmark has more objects and parts than FIND3D. c) Our datasets has single-view or whole point cloud, more objects and more pairs.

Table 5: Comparison of open-world methods on PartNet-E.

Methods	Canonical Orientation		Rotated	
	{part} of a {object}	{part}	{part} of a {object}	{part}
Find3D	16.86	16.38	17.62	17.16
PartSLIP++	5.12	32.71	3.87	23.03
PointCLIPV2	11.28	9.70	10.32	10.22
OpenMask3D	12.54	11.24	11.93	11.67
UniPart(ours)	23.15	22.90	24.28	21.72

tailored to partial observations and the other operating on full geometry. For a fair and comprehensive evaluation, we further report results under both test protocols, as shown in Table 4. And Our method has best performance.

4.4 Fine-Grained Robotic Manipulation

Real World Robot Setups We perform object manipulation tasks using the Franka Panda equipped with a gripper and the UR robot arm with a Shadow Hand. All the robot arms mount a RealSense D415 camera at their end for image capturing. In Fig. 5, we present the workspace and robotics for real-world object manipulation. We utilize only a single D415 camera. This setup significantly reduces the additional overhead associated with environmental setup and multi-camera calibration, and it is more readily reproducible.

Language-Conditioned Grasping in Real-World Scenarios We evaluate robotic grasping tasks to test whether UniPart produces usable part masks on real sensor point clouds. We integrate UniPart with off-the-shelf grasp generators, GraspNet [64] and DexGraspNet2.0 [77]. Given a language query, UniPart



Fig. 5: Object Manipulation robot setup.

Table 6: Controlled study evaluated on Objaverse-General (zero-shot mIoU). Both our dataset and architecture contribute significantly to the final performance.

Model Architecture	Training Data	mIoU
FIND3D	FIND3D Data	30.75
UniPart(Ours)	FIND3D Data	35.27
FIND3D	LangPart-1M (Ours)	38.42
UniPart(Ours)	LangPart-1M (Ours)	47.25

predicts a target part point cloud, which is then used as the input to the grasp generator to produce and execute a grasp.

Quantitative Results We evaluate 20 real objects with language prompts that specify a target part. We report part segmentation accuracy by inspection of the predicted part, and we report grasp success as the fraction of successful grasps that land on the intended part. UniPart achieves 90.5% part segmentation accuracy and the downstream pipeline achieves 85.0% grasp success.

Qualitative Results UniPart delineates parts such as handles and grips under sensor noise and occlusion, which supports its use as a perception module for language-conditioned grasping.

4.5 Ablation Study

Controlled Study To strictly isolate the performance gains brought by our proposed architecture and our large-scale dataset, we conduct a controlled cross-evaluation against the FIND3D [38] baseline on the Objaverse-General zero-shot benchmark. We evaluate four settings: (1) FIND3D model trained on FIND3D data, (2) UniPart trained on FIND3D data, (3) FIND3D model trained on LangPart-1M, and (4) UniPart trained on LangPart-1M.

As shown in Tab. 6, when restricted to the same FIND3D training data, UniPart still outperforms the FIND3D model by 14.69% mIoU, demonstrating the superior cross-modal alignment capability of our additive text-injection Transformer. Conversely, when the FIND3D model is trained on our LangPart-1M

Table 7: Ablation study of pretraining for UniPart. **Table 8:** Ablation study of multi-modal fusion for UniPart.

	mIoU	Fusion Method	mIoU
with pretrain + partial PC	33.56	Cross-attn	29.22
with pretrain + complete PC	27.17	Multiplication	28.13
without pretrain + partial PC	26.34	Concat	30.31
without pretrain + complete PC	23.39	Addition	33.56

dataset, its performance improves by 24.94% mIoU, validating the high quality and scalability of our automated data engine. Ultimately, UniPart trained on LangPart-1M achieves the best performance, proving that our architecture is uniquely suited to absorb the rich supervision from large-scale text-part data.

Ablation study on Pretraining To isolate the contribution of the proposed cross-modal pretraining, we conduct ablations under both partial and complete point-cloud settings. Although the pretraining stage is performed only on single-view partial point clouds (as it relies on point-to-image patch correspondences), we evaluate its effect on models trained and tested with partial inputs as well as on models operating on complete geometry. As shown in Tab. 7, pretraining on single-view partial point clouds still improves fine-tuning on complete point clouds by transferring CLIP semantics into the 3D encoder. Gains are typically larger for partial inputs but remain consistent on complete geometry.

Cross-Modal Fusion Choices We further conduct an ablation study on the multi-modal fusion methods in UniPart, testing commonly used feature fusion techniques such as cross-attention, multiplication, addition, and concatenation, as shown in Tab. 8. The results indicate that simple addition achieves the best performance. This may be attributed to the fact that instructions in the semantic domain are typically composed of short phrases or sentences, and the text CLS token already encodes sufficiently high-level semantic information.

5 Conclusion and Future Work

We introduced **UniPart** for zero-shot, language-grounded 3D part segmentation. Our goal is to provide a simple semantic interface that maps a natural language phrase to a fine-grained 3D part mask, which is a useful primitive for embodied interaction. A key enabler is **LangPart-1M**, a large-scale resource with 8 million text to part alignments on 160K+ objects, constructed with a multi-view pipeline that improves 3D consistency through text guided merging. We further curate **LangPart-4K** with manual verification and establish **LangPart-1K** for evaluation. UniPart, a feed-forward cross-modal 3D Transformer with layer-wise additive text conditioning, achieves strong generalization

across open-vocabulary benchmarks and produces usable part masks on real sensor point clouds for language-conditioned grasping. Future work will extend part understanding beyond isolated objects and strengthen closed-loop integration with robotic control.

References

1. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I.K., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 1534–1543. IEEE Computer Society (2016)
2. Belkhale, S., Ding, T., Xiao, T., Sermanet, P., Vuong, Q., Tompson, J., Chebotar, Y., Dwibedi, D., Sadigh, D.: RT-H: action hierarchies using language. CoRR **abs/2403.01823** (2024)
3. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N.J., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, K., Levine, S., Lu, Y., Malla, U., Manjunath, D., Mordatch, I., Nachum, O., Parada, C., Peralta, J., Perez, E., Pertsch, K., Quiambao, J., Rao, K., Ryoo, M.S., Salazar, G., Sanketi, P.R., Sayed, K., Singh, J., Sontakke, S., Stone, A., Tan, C., Tran, H.T., Vanhoucke, V., Vega, S., Vuong, Q., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., Zitkovich, B.: RT-1: robotics transformer for real-world control at scale. In: Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023 (2023)
4. Chu, H., Deng, X., Lv, Q., Chen, X., Li, Y., Hao, J., Nie, L.: 3d-affordancellm: Harnessing large language models for open-vocabulary affordance detection in 3d worlds. arXiv preprint arXiv:2502.20041 (2025)
5. Collaboration, O.X., Padalkar, A., Pooley, A., Jain, A., Bewley, A., Herzog, A., Irpan, A., Khazatsky, A., Raj, A., Singh, A., Brohan, A., Raffin, A., Wahid, A., Burgess-Limerick, B., Kim, B., Schölkopf, B., Ichter, B., Lu, C., Xu, C., Finn, C., Xu, C., Chi, C., Huang, C., Chan, C., Pan, C., Fu, C., Devin, C., Driess, D., Pathak, D., Shah, D., Büchler, D., Kalashnikov, D., Sadigh, D., Johns, E., Ceola, F., Xia, F., Stulp, F., Zhou, G., Sukhatme, G.S., Salhotra, G., Yan, G., Schiavi, G., Kahn, G., Su, H., Fang, H., Shi, H., Amor, H.B., Christensen, H.I., Furuta, H., Walke, H., Fang, H., Mordatch, I., Radosavovic, I., et al.: Open x-embodiment: Robotic learning datasets and RT-X models. CoRR **abs/2310.08864** (2023)
6. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T.A., Nießner, M.: ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 2432–2443. IEEE Computer Society (2017)
7. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 13142–13153. IEEE (2023)
8. Deng, J., Shi, S., Li, P., Zhou, W., Zhang, Y., Li, H.: Voxel R-CNN: towards high performance voxel-based 3d object detection. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative

- Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. pp. 1201–1209. AAAI Press (2021)
9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019)
 10. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: PLA: language-driven open-vocabulary 3d scene understanding. In: CVPR (2023)
 11. Ding, R., Yang, J., Xue, C., Zhang, W., Bai, S., Qi, X.: Lowis3d: Language-driven open-world instance-level 3d scene understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**(12), 8517–8533 (2024)
 12. Dong, R., Qi, Z., Zhang, L., Zhang, J., Sun, J., Ge, Z., Yi, L., Ma, K.: Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? In: ICLR (2023)
 13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
 14. Engel, N., Belagiannis, V., Dietmayer, K.: Point transformer. *IEEE Access* **9**, 134826–134840 (2021)
 15. Fan, G., Qi, Z., Shi, W., Ma, K.: Point-gcc: Universal self-supervised 3d scene pre-training via geometry-color contrast. In: Cai, J., Kankanhalli, M.S., Prabhakaran, B., Boll, S., Subramanian, R., Zheng, L., Singh, V.K., César, P., Xie, L., Xu, D. (eds.) Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024. pp. 4709–4718. ACM (2024)
 16. Fang, K., Liu, F., Abbeel, P., Levine, S.: MOKA: Open-World Robotic Manipulation through Mark-Based Visual Prompting. In: Proceedings of Robotics: Science and Systems. Delft, Netherlands (July 2024)
 17. Garosi, M., Tedoldi, R., Boscaini, D., Mancini, M., Sebe, N., Poiesi, F.: 3d part segmentation via geometric aggregation of 2d visual features. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3257–3267. IEEE (2025)
 18. Gibson, J.J.: The theory of affordances:(1979). In: The people, place, and space reader, pp. 56–60. Routledge (2014)
 19. Hamdi, A., Giancola, S., Ghanem, B.: MVTN: multi-view transformation network for 3d shape recognition. In: ICCV. pp. 1–11. IEEE (2021)
 20. He, J., Li, D., Yu, X., Qi, Z., Zhang, W., Chen, J., Zhang, Z., Yi, L., Wang, H.: Dexvlg: Dexterous vision-language-grasp model at scale. *CoRR* **abs/2507.02747** (2025)
 21. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. In: CVPR (2022)
 22. Huang, H., Lin, F., Hu, Y., Wang, S., Gao, Y.: Copa: General robotic manipulation through spatial constraints of parts with foundation models. *CoRR* **abs/2403.08248** (2024)
 23. Huang, S., Chang, H., Liu, Y., Zhu, Y., Dong, H., Boularias, A., Gao, P., Li, H.: A3VLM: actionable articulation-aware vision language model. In: Agrawal, P.,

- Kroemer, O., Burgard, W. (eds.) Conference on Robot Learning, 6-9 November 2024, Munich, Germany. Proceedings of Machine Learning Research, vol. 270, pp. 1675–1690. PMLR (2024)
24. Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., Fei-Fei, L.: Voxposer: Composable 3d value maps for robotic manipulation with language models (2023)
 25. Ichter, B., Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., Ho, D., Ibarz, J., Irpan, A., Jang, E., Julian, R., Kalashnikov, D., Levine, S., Lu, Y., Parada, C., Rao, K., Sermanet, P., Toshev, A., Vanhoucke, V., Xia, F., Xiao, T., Xu, P., Yan, M., Brown, N., Ahn, M., Cortes, O., Sievers, N., Tan, C., Xu, S., Reyes, D., Rettinghouse, J., Quiambao, J., Pastor, P., Luu, L., Lee, K., Kuang, Y., Jesmonth, S., Joshi, N.J., Jeffrey, K., Ruano, R.J., Hsu, J., Gopalakrishnan, K., David, B., Zeng, A., Fu, C.K.: Do as I can, not as I say: Grounding language in robotic affordances. In: Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand. Proceedings of Machine Learning Research, vol. 205, pp. 287–318. PMLR (2022)
 26. Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M.K., Chen, L.Y., Ellis, K., Fagan, P.D., Hejna, J., Itkina, M., Lepert, M., Ma, Y.J., Miller, P.T., Wu, J., Belkhale, S., Dass, S., Ha, H., Jain, A., Lee, A., Lee, Y., Memmel, M., Park, S., Radosavovic, I., Wang, K., Zhan, A., Black, K., Chi, C., Hatch, K.B., Lin, S., Lu, J., Mercat, J., Rehman, A., Sanketi, P.R., Sharma, A., Simpson, C., Vuong, Q., Walke, H.R., Wulfe, B., Xiao, T., Yang, J.H., Yavary, A., Zhao, T.Z., Agia, C., Baijal, R., Castro, M.G., Chen, D., Chen, Q., Chung, T., Drake, J., Foster, E.P., et al.: DROID: A large-scale in-the-wild robot manipulation dataset. CoRR **abs/2403.12945** (2024)
 27. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E.P., Lam, G., Sanketi, P., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., Finn, C.: Openvla: An open-source vision-language-action model. CoRR **abs/2406.09246** (2024)
 28. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.B.: Segment anything. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 3992–4003. IEEE (2023)
 29. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
 30. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022)
 31. Li, X., Zhang, M., Geng, Y., Geng, H., Long, Y., Shen, Y., Zhang, R., Liu, J., Dong, H.: Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 18061–18070. IEEE (2024)
 32. Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., Zeng, A.: Code as policies: Language model programs for embodied control. In: IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023. pp. 9493–9500. IEEE (2023)
 33. Liu, H., Lee, L., Lee, K., Abbeel, P.: Instruction-following agents with jointly pre-trained vision-language models. CoRR **abs/2210.13431** (2022)

34. Liu, M., Zhu, Y., Cai, H., Han, S., Ling, Z., Porikli, F., Su, H.: Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21736–21746 (2023)
35. Liu, M., Zhu, Y., Cai, H., Han, S., Ling, Z., Porikli, F., Su, H.: Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21736–21746 (2023)
36. Liu, Z., Zhang, Z., Cao, Y., Hu, H., Tong, X.: Group-free 3d object detection via transformers. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 2929–2938. IEEE (2021)
37. Lu, D., Kong, L., Huang, T., Lee, G.H.: Geal: Generalizable 3d affordance learning with cross-modal consistency. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1680–1690 (2025)
38. Ma, Z., Yue, Y., Gkioxari, G.: Find any part in 3d. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7818–7827 (2025)
39. Mao, J., Xue, Y., Niu, M., Bai, H., Feng, J., Liang, X., Xu, H., Xu, C.: Voxel transformer for 3d object detection. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 3144–3153. IEEE (2021)
40. Mao, W., Zhong, W., Jiang, Z., Fang, D., Zhang, Z., Lan, Z., Jia, F., Wang, T., Fan, H., Yoshie, O.: Robomatrix: A skill-centric hierarchical framework for scalable robot task planning and execution in open-world. CoRR **abs/2412.00171** (2024)
41. Maturana, D., Scherer, S.A.: Voxnet: A 3d convolutional neural network for real-time object recognition. pp. 922–928. IEEE (2015)
42. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 909–918. Computer Vision Foundation / IEEE (2019)
43. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 909–918 (2019)
44. Nasiriany, S., Kirmani, S., Ding, T., Smith, L., Zhu, Y., Driess, D., Sadigh, D., Xiao, T.: Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. CoRR **abs/2411.02704** (2024)
45. Pang, Y., Wang, W., Tay, F.E.H., Liu, W., Tian, Y., Yuan, L.: Masked autoencoders for point cloud self-supervised learning. In: ECCV (2022)
46. Peng, S., Genova, K., Jiang, C.M., Tagliasacchi, A., Pollefeys, M., Funkhouser, T.A.: Openscene: 3d scene understanding with open vocabularies. In: CVPR (2023)
47. Peng, Y., Cui, Y., Tang, H., Qi, Z., Dong, R., Bai, J., Han, C., Ge, Z., Zhang, X., Xia, S.T.: Dreambench++: A human-aligned benchmark for personalized image generation. CoRR **abs/2406.16855** (2024)
48. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR. pp. 77–85 (2017)
49. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. pp. 5099–5108 (2017)

50. Qi, Z., Dong, R., Fan, G., Ge, Z., Zhang, X., Ma, K., Yi, L.: Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In: ICML (2023)
51. Qi, Z., Dong, R., Zhang, S., Geng, H., Han, C., Ge, Z., Yi, L., Ma, K.: Shapellm: Universal 3d object understanding for embodied interaction. In: Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLIII. Lecture Notes in Computer Science, vol. 15101, pp. 214–238. Springer (2024)
52. Qi, Z., Yu, M., Dong, R., Ma, K.: VPP: efficient conditional 3d generation via voxel-point progressive representation. In: NeurIPS (2023)
53. Qian, G., Li, Y., Peng, H., Mai, J., Hammoud, H.A.A.K., Elhoseiny, M., Ghanem, B.: Pointnext: Revisiting pointnet++ with improved training and scaling strategies. In: NeurIPS (2022)
54. Qian, S., Chen, W., Bai, M., Zhou, X., Tu, Z., Li, L.E.: Affordancellm: Grounding affordance from vision language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7587–7597 (2024)
55. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20051–20060 (2024)
56. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021)
57. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
58. Shridhar, M., Manuelli, L., Fox, D.: Perceiver-actor: A multi-task transformer for robotic manipulation. In: Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand. Proceedings of Machine Learning Research, vol. 205, pp. 785–799. PMLR (2022)
59. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
60. Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E.G.: Multi-view convolutional neural networks for 3d shape recognition. In: ICCV (2015)
61. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
62. Uy, M.A., Pham, Q., Hua, B., Nguyen, D.T., Yeung, S.: Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 1588–1597. IEEE (2019)
63. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. pp. 5998–6008 (2017)
64. Wang, C., Fang, H., Gou, M., Fang, H., Gao, J., Lu, C.: Graspness discovery in clutters for fast and accurate grasp detection. CoRR **abs/2406.11142** (2024)
65. Wu, H., Yang, P., Asano, Y.M., Snoek, C.G.: Segment any 3d-part in a scene from a sentence. arXiv preprint arXiv:2506.19331 (2025)

66. Wu, H., Yang, P., Asano, Y.M., Snoek, C.G.: Segment any 3d-part in a scene from a sentence. arXiv preprint arXiv:2506.19331 (2025)
67. Wu, T., Yang, G., Li, Z., Zhang, K., Liu, Z., Guibas, L.J., Lin, D., Wetzstein, G.: Gpt-4v(ision) is a human-aligned evaluator for text-to-3d generation. In: CVPR (2024)
68. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 1912–1920. IEEE Computer Society (2015)
69. Xie, S., Gu, J., Guo, D., Qi, C.R., Guibas, L.J., Litany, O.: Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In: ECCV. Lecture Notes in Computer Science, vol. 12348, pp. 574–591. Springer (2020)
70. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. CoRR **abs/2310.11441** (2023)
71. Yang, S., Du, Y., Ghasemipour, S.K.S., Tompson, J., Kaelbling, L.P., Schuurmans, D., Abbeel, P.: Learning interactive real-world simulators. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024)
72. Yi, L., Kim, V.G., Ceylan, D., Shen, I., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., Guibas, L.J.: A scalable active framework for region annotation in 3d shape collections. ACM Trans. Graph. **35**(6), 210:1–210:12 (2016)
73. Yi, L., Su, H., Guo, X., Guibas, L.J.: Syncspeccnn: Synchronized spectral CNN for 3d shape segmentation. In: CVPR (2017)
74. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J.: Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In: CVPR (2022)
75. Yuan, W., Duan, J., Blukis, V., Pumacay, W., Krishna, R., Murali, A., Mousavian, A., Fox, D.: Robopoint: A vision-language model for spatial affordance prediction for robotics. CoRR **abs/2406.10721** (2024)
76. Yuan, Z., Wei, T., Cheng, S., Zhang, G., Chen, Y., Xu, H.: Learning to manipulate anywhere: A visual generalizable framework for reinforcement learning. In: Agrawal, P., Kroemer, O., Burgard, W. (eds.) Conference on Robot Learning, 6-9 November 2024, Munich, Germany. Proceedings of Machine Learning Research, vol. 270, pp. 1815–1833. PMLR (2024)
77. Zhang, J., Liu, H., Li, D., Yu, X., Geng, H., Ding, Y., Chen, J., Wang, H.: Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes. In: 8th Annual Conference on Robot Learning (2024)
78. Zhang, J., Dong, R., Ma, K.: CLIP-FO3D: learning free open-world 3d scene representations from 2d dense CLIP. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023. pp. 2040–2051. IEEE (2023)
79. Zhang, S., Qi, Z., Dong, R., Bai, X., Wei, X.: Positional prompt tuning for efficient 3d representation learning. CoRR **abs/2408.11567** (2024)
80. Zhang, W., Liu, H., Li, B., He, J., Qi, Z., Wang, Y., Zhao, S., Yu, X., Zeng, W., Jin, X.: Hybrid-grained feature aggregation with coarse-to-fine language guidance for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6678–6692 (October 2025)
81. Zhang, W., Liu, H., Qi, Z., Wang, Y., Yu, X., Zhang, J., Dong, R., He, J., Wang, H., Zhang, Z., Yi, L., Zeng, W., Jin, X.: Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge. CoRR **abs/2507.04447** (2025)

82. Zhao, T.Z., Kumar, V., Levine, S., Finn, C.: Learning fine-grained bimanual manipulation with low-cost hardware. In: *Robotics: Science and Systems XIX*, Daegu, Republic of Korea, July 10-14, 2023 (2023)
83. Zhen, H., Qiu, X., Chen, P., Yang, J., Yan, X., Du, Y., Hong, Y., Gan, C.: 3d-vla: A 3d vision-language-action generative world model. In: *Forty-first International Conference on Machine Learning, ICML 2024*, Vienna, Austria, July 21-27, 2024. OpenReview.net (2024)
84. Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging llm-as-a-judge with mt-bench and chatbot arena. In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, New Orleans, LA, USA, December 10 - 16, 2023 (2023)
85. Zhou, E., Su, Q., Chi, C., Zhang, Z., Wang, Z., Huang, T., Sheng, L., Wang, H.: Code-as-monitor: Constraint-aware visual programming for reactive and proactive robotic failure detection. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 6919–6929 (2025)
86. Zhou, Y., Gu, J., Chiang, T.Y., Xiang, F., Su, H.: Point-sam: Promptable 3d segmentation model for point clouds. arXiv preprint arXiv:2406.17741 (2024)
87. Zhou, Y., Gu, J., Chiang, T.Y., Xiang, F., Su, H.: Point-sam: Promptable 3d segmentation model for point clouds. arXiv preprint arXiv:2406.17741 (2024)
88. Zhu, X., Zhang, R., He, B., Guo, Z., Zeng, Z., Qin, Z., Zhang, S., Gao, P.: Pointclip V2: prompting CLIP and GPT for powerful 3d open-world learning. In: *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, Paris, France, October 1-6, 2023. pp. 2639–2650. IEEE (2023)
89. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., Vuong, Q., Vanhoucke, V., Tran, H.T., Soricut, R., Singh, A., Singh, J., Sermanet, P., Sanketi, P.R., Salazar, G., Ryoo, M.S., Reymann, K., Rao, K., Pertsch, K., Mordatch, I., Michalewski, H., Lu, Y., Levine, S., Lee, L., Lee, T.E., Leal, I., Kuang, Y., Kalashnikov, D., Julian, R., Joshi, N.J., Irpan, A., Ichter, B., Hsu, J., Herzog, A., Hausman, K., Gopalakrishnan, K., Fu, C., Florence, P., Finn, C., Dubey, K.A., Driess, D., Ding, T., Choromanski, K.M., Chen, X., Chebotar, Y., Carbajal, J., Brown, N., Brohan, A., Arenas, M.G., Han, K.: RT-2: vision-language-action models transfer web knowledge to robotic control. In: *Conference on Robot Learning, CoRL 2023*, 6-9 November 2023, Atlanta, GA, USA. *Proceedings of Machine Learning Research*, vol. 229, pp. 2165–2183. PMLR (2023)