

PASG: A Closed-Loop Framework for Automated Geometric Primitive Extraction and Semantic Anchoring in Robotic Manipulation

Zhihao Zhu* Yifan Zheng* Siyu Pan* Yaohui Jin† Yao Mu†

MoE key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

{zzh2021, yifanzheng, pansiyu0327, jinyh, muyao}@sjtu.edu.cn

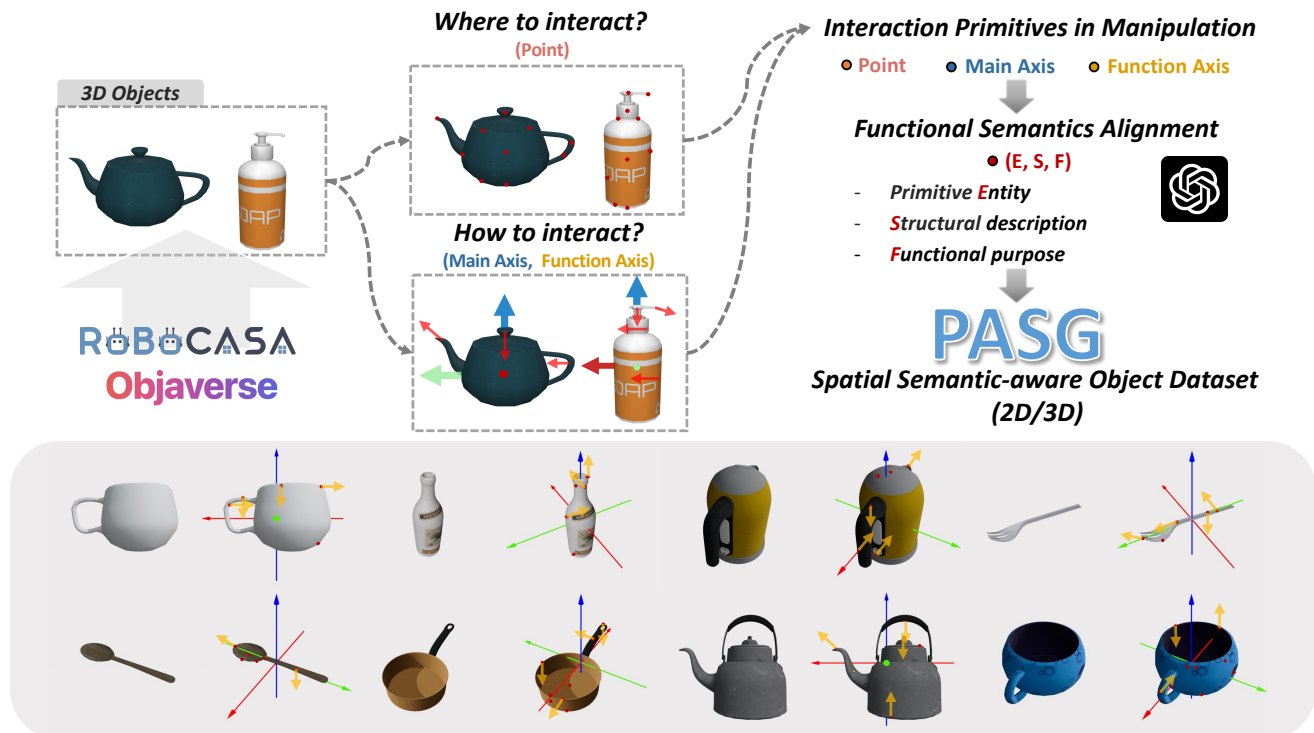


Figure 1. We propose PASG, an automated object-centric spatial-semantic enhancement framework for robotic manipulation. By formalizing interaction primitives and establishing semantic-geometric correspondences, our approach achieves structural coupling between low-level spatial primitives and high-level functional semantics, enabling joint enhancement of manipulation reasoning and semantic-aware 2D/3D object dataset generation.

Abstract

The fragmentation between high-level task semantics and low-level geometric features remains a persistent challenge in robotic manipulation. While vision-language models (VLMs) have shown promise in generating affordance-aware visual representations, the lack of semantic grounding in canonical spaces and reliance on manual annotations severely limit their ability to capture dynamic semantic-affordance relationships. To address these, we propose **Primitive-Aware Semantic Grounding (PASG)**, a closed-

loop framework that introduces: (1) Automatic primitive extraction through geometric feature aggregation, enabling cross-category detection of keypoints and axes; (2) VLM-driven semantic anchoring that dynamically couples geometric primitives with functional affordances and task-relevant description; (3) A spatial-semantic reasoning benchmark and a fine-tuned VLM (Qwen2.5VL-PA). We demonstrate PASG’s effectiveness in practical robotic manipulation tasks across diverse scenarios, achieving performance comparable to manual annotations. PASG achieves a finer-grained semantic-affordance understanding of objects, establishing a unified paradigm for bridging geometric primitives with task semantics in robotic manipulation.

*Equal Contribution.

†Yao Mu and Yaohui Jin are the corresponding authors

1. Introduction

Developing generalizable robotic manipulation in unstructured environments remains challenging due to the semantic asymmetry between low-level interaction primitives (points/axes) and high-level task planning. As large language models (LLMs)[2, 53] and vision-language models (VLMs)[3, 31, 49, 63] demonstrate promise in semantic reasoning and commonsense knowledge, researchers have attempted to integrate these models into robotic manipulation[9, 14, 17, 33]. However, these approaches primarily focus on high-level task decomposition and planning[26, 40, 51, 54, 66]. In contrast, the capacity of semantic reasoning within 3D spatial primitives remains underdeveloped. This limitation stems from insufficient semantic understanding of object canonical spaces—for instance, manually annotated “handle centers” for teapots lack contextual semantics (such as functional descriptions and usage scenarios), leading to inaccurate spatial constraint reasoning—revealing the inherent fragility of strategies that directly map task semantics to canonical spaces devoid of semantic context.

To endow robots with spatial primitive understanding, current approaches typically fine-tune VLMs on large-scale manipulation demonstrations to enhance spatial semantic reasoning. However, such methods depend on manually annotated geometric primitives (e.g., keypoints, axes), which leads to high annotation costs and inherently limits generalizability. Recent pioneering work leverages pre-trained large vision models (LVM) [3, 50] to detect interaction features, followed by VLM-based semantic filtering to identify task-relevant primitives [23]. Nevertheless, such frameworks exhibit two systemic weaknesses: (1) Automated detection methods (e.g., SAM [28], DINOv2 [43]) lack verification mechanisms, propagating errors from undetected or misaligned primitives and drastically degrade success rates; (2) Incomplete canonical space definitions - overlooking essential orientation features of object like main axes while only include keypoints and directions - result in manipulation failures during grasping or transportation tasks [59]. These limitations underscore the need for a unified framework that integrates automated primitive extraction with semantic-task contextualization.

To address these challenges, as shown in Fig 1, we propose PASG, a closed-loop framework establishing the mapping between spatial primitives and functional semantics. It offers several key innovations: First, our geometry-aware feature extraction module automatically detects interaction primitives (keypoints, directions, and principal axes) through visual foundation models (VFMs) integration with geometric topology analysis, without any manual annota-

tion. Second, our dynamic semantic anchoring mechanism employs VLMs to contextualize primitives with multi-granularity semantics - from low-level descriptions (“edge of the neck”) to high-level intents (“Crucial for aligning the bottle during pouring or filling.”) - while implementing self-corrective feature extraction loops. Third, we validate the practicality of PASG through extensive manipulation experiments conducted on diverse tasks within a simulated environment, demonstrating competitive or superior performance compared to human annotations. Lastly, we provide Robocasa-PA, an extensive benchmark that supports scalable, object-based evaluations of functional primitive understanding in manipulation scenarios, and develop Qwen2.5VL-PA through parameter-efficient LoRA fine-tuning, achieving 77.8% overall accuracy (+33.9% absolute improvement) with minimal cross-domain variance.

Our contributions are as follows:

- We propose a novel framework that automatically annotates hierarchical semantics for object interaction primitives, bridging the gap between low-level geometric features and high-level task semantics.
- We introduce Robocasa-PA, featuring 8,343 validated visual questions across three task evaluations, providing the first benchmark for assessing functional primitive understanding in manipulation.
- We demonstrate PASG’s effectiveness in real-world manipulation scenarios, achieving competitive performance relative to human annotations and enhancing diversity and flexibility in grasp and interaction primitives.

2. Related Work

2.1. Language-Grounded Manipulation

Natural language has emerged as a critical interface for robotic manipulation. Existing approaches fall into two categories: (1) End-to-end models: Cross-modal Transformers [16, 37, 46, 62], and Vision-Language-Action (VLA) models trained on large-scale robotic datasets (e.g., RT-X, Open X-Embodiment) [7, 8, 35, 44], unify perception [22, 32], planning [15, 64], and action cross-modally through latent space alignment [20, 36, 56], demonstrating strong generalization. However, their dependence on domain-specific robotic data introduces scalability bottlenecks. (2) Decoupled Language-Planning Architectures: This approach separates low-level motion control from high-level task reasoning, leveraging VLMs for instruction-based subgoal decomposition [10, 13, 18, 19, 55] and formulate constraint optimization problems based on geometric primitives [21, 24, 38]. This method effectively harnesses VLMs’ multimodal semantic reasoning capabilities to enhance interpretability. However, these methods suffer from coarse coupling mechanisms that lead to primitive-semantic misalignment. Contemporaneous work SOFAR

Method	Geometric Primitives	Primitive Extraction	Semantic Coupling	Adaptive Refinement [†]
ReKep [24]	Keypoints	VFM + VLM Detection (Task-Level)	No	No
CoPa [21]	Keypoints, Function Axes	VFM + VLM Detection (Task-Level)	No	No
OmniManip [45]	Keypoints, Main Axes	VFM + VLM Detection (Task-Level)	No	Yes
FUNCTO [52]	Keypoints, Function Axes	VFM + VLM Detection (Task-Level)	Predefined Semantics	No
Robotwin [38]	Keypoints, Function Axes, Main Axes	Human-Annotated (Object-Level)	Predefined Semantics	No
SoFar [48]	Main Axes	Predefined Orientation (Object-Level)	Automatic Spatial Semantic Anchoring	No
PASG (Ours)	Keypoints, Function Axes, Main Axes	VFM + VLM Detection (Object-Level)	Automatic Spatial Semantic Anchoring	Yes

Table 1. Normative interaction primitive and semantic coupling across different frameworks in robotic manipulation tasks: PASG as the first automated closed-loop framework with primitive extraction, semantic anchoring, and self-refinement. [†] Adaptive refinement refers to the mechanism that self-corrects erroneous or omitted geometric primitives. OmniManip employs computational constraint optimization and scene rendering for VLM validation, while our method directly detects annotation-primitive misalignment for efficient self-correction.

addresses this limitation by proposing direction-aware spatial understanding module PointSO, which connects geometric reasoning with functional semantics [48]. Unlike SOFAR’s predefined directional priors, PASG framework focuses on fine-grained keypoints and functional vectors to construct hierarchical semantic anchoring, achieving deeper integration between task semantics and spatial primitives.

2.2. Spatial Reasoning for Manipulation

Spatial reasoning in manipulation involves inferring interaction constraints from object’s spatial primitives to guide robot actions. Keypoints are widely adopted for constructing spatial constraints due to their semantic interpretability and ease of representation [29, 47, 60]. However, manual annotation of keypoints limits scalability. Recent advances address this by leveraging visual prompts (scene images) fed directly into vision-language models (VLMs) to autonomously identify task-relevant points, achieving zero-shot consistency [1, 27, 52]. Nevertheless, the discrete nature of keypoints and detection instability degrade reasoning performance while failing to provide effective guidance for end-effector pose optimization. To address this, researchers integrate 6D pose estimation with positional awareness, significantly enhancing end-effector pose stability [5, 45, 57]. However, traditional orientation representations suffer from sparsity and static template-based definitions that lack semantic foundations, limiting adaptability in open-world manipulation. PASG resolves these by augmenting geometric primitives with function-aware directional vectors (e.g., ”pressing direction” for buttons, ”grasping direction” for handles), enabling flexible yet stable directional constraints. Through automated extraction

of primitives using VFMs and hierarchical semantic alignment with VLMs, PASG achieves closed-loop optimization of spatial reasoning paradigms in open-world scenarios.

3. Method

In this section, we explore the following research questions: (1) How to define spatial geometric primitives for objects in manipulation tasks? (2) How to automatically extract geometric primitives from objects? (3) How does PASG achieve dynamic alignment between geometric primitives and task semantics? (4) How to leverage PASG to enhance reasoning in manipulation tasks?

3.1. Semantic Primitives in Robotic Manipulation

In robotic manipulation tasks, spatial primitives of objects serve as fundamental building blocks for planning and executing actions. Traditionally, these primitives are defined by geometric entities E (e.g., points, axes, orientations). However, beyond pure geometry, each primitive often carries both semantic S and functional information F . For instance, a cup’s handle represents not merely a spatial protrusion but provides an affordance for grasping while implicitly suggesting proper manipulation strategies. By augmenting geometric primitives with such attributes, we can better capture an object’s intended usage and operational constraints. To formally integrate these aspects, we define an *interaction primitive* as a triplet (E, S, F) that combines geometric, structural, and functional properties:

- E – Geometric primitive entity
- S – Structural description of the object component
- F – Functional role in manipulation tasks

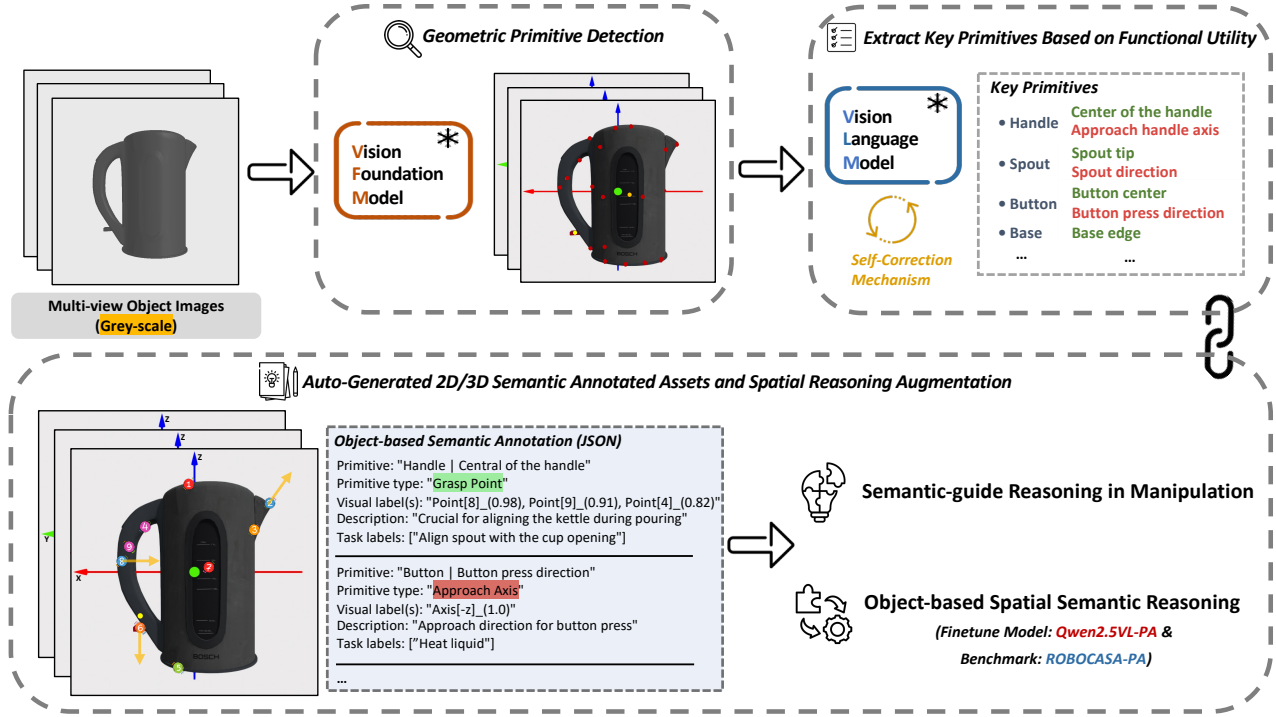


Figure 2. Overview of PASG

To further integrate operational task semantics, we categorize interaction primitives into two functionally distinct classes based on manipulation requirements: *point-based* (\mathcal{P}) and *axis-based* (\mathcal{A}) primitives.

Point Interaction Primitives (\mathcal{P}) denote specific object locations critical for manipulation.

- **Anchor Point** (p_a): A reference position that determines how an object should be placed or aligned in the workspace. (e.g. the spout tip over a cup for pouring)
- **Grasp Point** (p_g): A location on the object optimized for a secure hold by a robot’s end-effector. (e.g. the center of a mug’s handle)
- **Actuation Point** (p_{act}): The specific spot that triggers a mechanism or function when pressed or manipulated. (e.g. the power button on a microwave)

Axis Interaction Primitives (\mathcal{A}) encode directional information derived from geometric properties and functional requirements. These specify object orientation and motion constraints:

- **Primary Axis** (a_p): The principal orientation axis of the object, usually dictated by its geometry or symmetry. (e.g. vertical primary axis z in a teapot)
- **Functional Axis** (a_f): The axis aligned with the object’s intended action or function – essentially the direction along which the object exerts its effect. (e.g. the hammerhead direction in a hammer)
- **Approach Axis** (a_{app}): The direction from which a robot’s end-effector should approach the object to inter-

act with a specific point. (e.g. approaching a handle from the side along the handle’s orientation)

3.2. Geometry Primitive Extraction

To obtain accurate spatial primitive detection, we employ a VFM for fine-grained region segmentation, followed by geometric-topological processing for hierarchical keypoint detection and filtering.

VFM-Based Region Segmentation We utilize a pre-trained VFM (Semantic SAM [30]) for fine-grained semantic segmentation. To enable this, we first acquire multi-view RGB images ($\mathcal{I} = \{I_1, \dots, I_n\}$) from the object’s 3D mesh data, which are then resized and fed into the segmentation model. Inspired by SoM [58], we preprocess the obtained segmentation masks ($\mathcal{M} = \{M_1, \dots, M_n\}$) through connected component analysis to extract the largest foreground region while eliminating small areas and background noise.

Keypoint Extraction For geometric keypoint (\mathcal{K}_{raw}) detection, we extract representative geometric positions including centers ($\mathcal{C} = \{c_1, \dots, c_n\}$) and corner feature points ($\mathcal{F} = \{f_1, \dots, f_n\}$) in segmentation masks M . The centroid (c) is calculated through mask moments. Corner detection employs polygon approximation (cv2.approxPolyDP()) and curvature analysis. Additionally, Principal Component Analysis (PCA) calculates two orthogonal axes from the mask, where their intersections with the boundary are extracted as supplementary feature points.

Keypoint Filtering To reduce redundancy in detected key-

points (\mathcal{K}_{raw}), we implement a two-stage filtering mechanism: DBSCAN clustering removes locally dense points, while optimized farthest point sampling globally selects distinctive features, yielding final keypoints ($\mathcal{K}_{\text{filter}}$) including center point set (\mathcal{C}) and feature points set (\mathcal{F}).

Principal Axis Calibration For standardized axis representation, most 3D objects in datasets provide pre-aligned main axes. We rectify significant deviations by defining the Z-axis as the line connecting geometric centroids of top/bottom view masks. Orthogonal X/Y axes are subsequently generated. To ensure cross-view consistency, we enforce color standardization for axis visualization across different objects and viewpoints.

3.3. Task-Oriented Semantic Annotation

To generate task-oriented primitive annotations, we establish a mapping from task intentions to spatial primitives through three core stages: task-driven semantic primitive identification, semantic-geometric alignment, and dynamic self-verification.

Object-Centric Semantic Primitive Identification We initiate with task scenario construction and subgoal decomposition to identify semantically critical primitives in an object o . Specifically, we use VLMs to analyze geometric and physical features from multi-view images (\mathcal{I}) to infer potential manipulation tasks ($\mathcal{T} = \{t_1, \dots, t_m\}$). Each task t_i is decomposed into sub-stages with explicit operation goals per stage:

$$\mathcal{G}_i = \{g_{i1}, \dots, g_{ik}\} \quad (1)$$

To identify task-relevant spatial primitives, we establish primitive constraints for each subgoal:

$$\mathcal{R}_{ij}^o(P_{ij}^o, A_{ij}^o) \implies g_{ij} \quad (2)$$

where $P_{ij}^o \subseteq \mathcal{P}^o$ denotes point primitives and $A_{ij}^o \subseteq \mathcal{A}^o$ axis primitives for object o . The unified primitive set is obtained through:

$$\mathcal{E}^o = \left(\bigcup_{i,j} P_{ij}^o, \bigcup_{i,j} A_{ij}^o \right) \quad (3)$$

Notably, natural language descriptions of constraints (\mathcal{R}_{ij}^o) and primitives (\mathcal{E}^o) enable parallel semantic-geometric processing.

Visual-Semantic Primitive Alignment To align task-driven geometric primitives with visual annotated primitives from Sec 3.2, we leverage VLM’s multimodal geometric reasoning for mapping. For semantic keypoints, we employ a multi-candidate matching strategy: when multiple geometric keypoints ($p_i \in \mathcal{K}_{\text{filter}}^o$) match semantic descriptions in \mathcal{E}^o in object o (e.g., "handle center" may correspond to multiple points detected on the handle), we record all candidates with confidence scores ($s \in [0, 1]$) to ensure operational robustness and handle occlusions. Unmatched cases

return *NONE*. For semantic orientations, we adopt dual representation: 1) symbolic axis descriptions (e.g., +Z-axis for vertical orientation), 2) flexible point-pair directions (e.g., spout direction from spout base to tip). Through cross-view consistency, we ensure point label persistence across perspectives, generating robust spatial-semantic mappings with 2D/3D annotations as shown in Fig. 2.

Algorithm 1 Dynamic Self-Refine Matching Mechanism

Input:

- Object \mathcal{O}_i
- Segmentation granularity levels $\Gamma = \{\gamma_1, \dots, \gamma_N\}$
- Geometric primitives $\mathcal{K}_i = \{\mathcal{P}_i^{(1)}, \mathcal{A}_i^{(1)}, \dots\}$
- Semantic primitives $\mathcal{E}_i = \{\mathcal{P}_i^{\text{sem}}, \mathcal{A}_i^{\text{sem}}\}$
- Correspondence set $\mathcal{C}_i = (\mathcal{E}_i, \mathcal{K}_i^{(N)}, s_i)$

Output: Refined correspondence set $\hat{\mathcal{C}}_i$ or Matching Failure

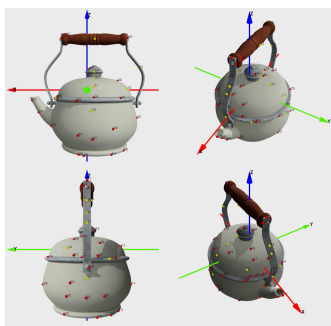
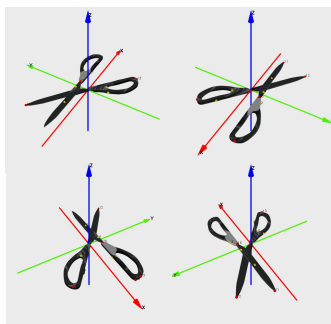
```

1: Initialize  $\gamma \leftarrow \gamma_1, t \leftarrow 0, \tau_{max} \leftarrow 5$ 
2: while  $t < \tau_{max}$  do
3:    $t \leftarrow t + 1$ 
4:   Match: Evaluate confidence scores  $s_i$  in correspondence set  $\mathcal{C}_i$ 
5:    $\mathcal{L} \leftarrow \{e_j \in \mathcal{E}_i \mid s_j < 0.5 \vee \text{unmatched}\}$ 
6:   if  $\mathcal{L} = \emptyset$  then
7:     return  $\hat{\mathcal{C}}_i = (\mathcal{E}_i, \mathcal{K}_i^{(\gamma)}, s)$ 
8:   else
9:     if  $t = \tau_{max}$  or  $\gamma = \gamma_{max}$  then
10:      return Matching Failure
11:     end if
12:      $\gamma \leftarrow \gamma^+$  ▷ Update to the next (finer)
13:     Resample:  $\mathcal{K}_i^{(\gamma)} \leftarrow \text{SemanticSAM}(\mathcal{O}_i, \gamma)$ 
14:     Align:  $\mathcal{C}_i^{(\gamma)} \leftarrow \text{VLM}(\mathcal{L}, \mathcal{K}_i^{(\gamma)})$ 
15:     Refine:  $\hat{\mathcal{C}}_i \leftarrow \text{ReplaceLowConf}(\mathcal{C}_i^{(\gamma)}, \mathcal{L})$ 
16:   end if
17: end while
18: return Matching Failure

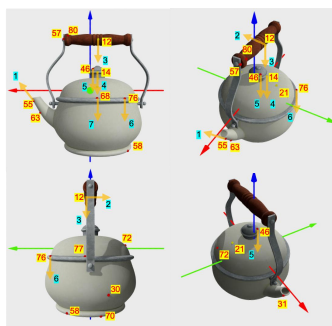
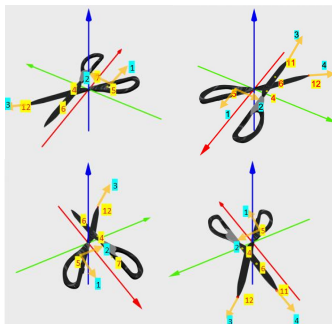
```

Dynamic Self-Refine Matching Mechanism To address the inaccuracies in primitives detection and information loss caused by feature extractors, we adopt a dynamic self-refine matching algorithm, as shown in Algorithm 1. First, lexical pattern matching validates annotation mappings. When low-confidence predictions (< 0.5) or missing primitives (labeled as *NONE*) are detected, the framework triggers hierarchical segmentation-based incremental annotation. Leveraging Semantic SAM’s multi-granularity segmentation, this process refines geometric primitives via adaptive resampling, forming a closed-loop dynamic optimization workflow: *segment-align-detect-resample*. Experiments demonstrate that our method achieves a 98% matching success rate on our dataset and effectively mitigates error propagation from poor segmentation.

Keypoint Detection Results



Semantic Filtering Results



Semantically Annotated Output

```

Take the scissor as an example
"Central area of the handle": {
  "primitive_type": "Point : Grasp Point",
  "primitives": [
    "point[5]_(0.92)", # Located at the center of the left handle with clear visibility
    "point[7]_(0.90)" # Located at the center of the right handle, ensuring a
    secure grip),
    "description": "Secure grip on the handle for various actions.",
    "task_labels": ["Grasp Scissors"]
  }
  "Perpendicular to handle (approach handle)": {
    "primitive_type": "Axis : Approach Axis",
    "primitives": [
      "axis[+x]_(0.95)", # Horizontal approach perpendicular to the handle for
      grasping
      "axis[+y]_(0.90)" # Vertical approach perpendicular to the handle for grasping),
      "description": "Approach direction for handle grasp.",
      "task_labels": ["Grasp Scissors"]
    }
    "Blade tip": {
      "primitive_type": "Point : Anchor Point",
      "primitives": [
        "point[12]_(0.93)", # Clear tip endpoint of the right blade
        "point[11]_(0.90)" # Clear tip endpoint of the left blade],
        "description": "Align blade with cutting or opening line.",
        "task_labels": ["Align Scissors with Plant", "Align Scissors with Package"]
      }
      "Blade edge": {
        "primitive_type": "Axis : Functional Axis",
        "primitives": [
          "point[6][12]_(0.88)", # Point 6 (blade edge) to point 12 (blade tip) for right
          blade
          "point[6][11]_(0.85)" # Point 6 (blade edge) to point 11 (blade tip) for left
          blade],
          "description": "Ensure proper cutting, trimming, or opening direction.",
          "task_labels": ["Execute Cutting Action", "Execute Trimming Action", "Execute
          Opening Action"]
        }
        "Pivot point": {
          "primitive_type": "Point : Hinge Point",
          "primitives": [
            "point[4]_(0.95)" # Central pivot point for rotation],
            "description": "Rotation point for cutting, trimming, or opening action.",
            "task_labels": ["Execute Cutting Action", "Execute Trimming Action", "Execute
            Opening Action"]
          }
  }

```

Figure 3. Keypoints and Axes Annotated Output. This framework demonstrates the process of detecting, filtering, and semantically annotating functional keypoints and axes on 3D objects. The visualization progresses from initial keypoint detection (left column) to semantic filtering (middle column), culminating in rich semantic annotation (right column).

3.4. Semantic-guide Reasoning in Manipulation

Beyond generating geometrically annotated object datasets, our framework facilitates the integration of spatial semantics into manipulation tasks. By leveraging multi-modal inputs—including geometrically annotated objects and their corresponding hierarchical semantics—PASG provides semantic-aware visual cues and textual guidance to downstream reasoning modules.

Specifically, through our manipulation experiments (in Section 4.2), we validate that the PASG pipeline can reliably identify interaction primitives across diverse object categories. Compared to manually labeled counterparts, PASG-generated annotations exhibit greater semantic diversity (e.g., multiple meaningful grasp or actuation points). This diversity contributes to flexible downstream usage and competitive task success rates across varied manipulation tasks, demonstrating the practical effectiveness of PASG in generating semantically grounded geometric annotations for real-world deployment.

4. Experiment

4.1. Semantic-aware Object Dataset

Data Sources and Scales We collect raw datasets sourced from RoboCasa [39] and Objaverse [11, 12]. RoboCasa

provides over 2,500 high-quality 3D objects covering more than 150 categories in everyday tasks, whereas Objaverse is a large-scale open dataset containing over 800,000 annotated 3D objects. Since our focus is on objects with meaningful spatial semantics, we filtered the 3D assets based on their tags and titles, excluding objects with narrowly defined functionalities or limited manipulability potential (e.g., food items). Leveraging texture detection, we further refined our selection to obtain a high-quality dataset of 5,231 objects. To ensure effective object segmentation, all selected objects are rescaled to appropriate sizes and rendered from multiple viewpoints, including four horizontal orthographic views and four 45-degree oblique orthographic projections. In total, we acquired a 5,231 object 3D dataset as well as an 41,848 image 2D dataset for subsequent semantic annotation.

Data Annotation As mentioned in Section 3.2, we first utilize VFM combined with topological methods to extract and visualize key primitives from multi-view object images. To ensure consistent and effective visual labeling, we perform the projection of 2D keypoints onto their corresponding 3D objects and the subsequent disambiguation and filtering through Open3D [65]. Each annotated point is assigned a visual index to facilitate identification. Leveraging the strong semantic reasoning capabilities of VLM, we feed

the annotated multi-view images into GPT-4o [42] as a visual prompt for spatial semantic annotating, as detailed in Section 3.3. Based on annotations, we filter out primitives that lack critical semantic information, producing the final set of semantic annotations for each object.

Dataset Validation To evaluate the effectiveness of our dataset generation pipeline, we randomly sample 50 annotated objects and conduct a manual inspection. Our experiment shows that, during the semantic identification phase, GPT-4o achieves an accuracy of 91.6%. During the geometric-semantic alignment phase, it demonstrates strict alignment accuracy of 75.8% (requiring full primitive correctness) and alignment effectiveness of 91.5% (ensuring at least one valid primitive-semantic match). These findings highlight GPT-4o’s strong spatial reasoning abilities under structured prompts, confirming the reliability of our annotated dataset. However, complex spatial semantics of the objects remain a challenge for GPT-4o. To further improve annotation quality, we employ a multi-round validation strategy where GPT-4o iteratively self-corrects its annotations, enhancing the robustness of the dataset’s semantic labels. Additional validation experiments are provided in the supplemental material.

4.2. Manipulation Task Evaluation

To validate the effectiveness of PASG in robotic manipulation, we conduct comprehensive evaluations using the RoboTwin[38] simulation platform, open-source environment designed to emulate realistic robotic manipulation scenarios. RoboTwin provides standardized benchmarks that ensure both reproducibility and practical relevance.

Task Setting We evaluate PASG’s performance across six representative manipulation tasks, including tasks requiring dual-arm coordination and single-arm manipulations, as well as interactions within cluttered environments. Specifically, the tasks are: (1) Block Hammer Beat, (2) Container Place, (3) Dual Bottles Pick, (4) Empty Cup Place, (5) Pick Apple, and (6) Messy Shoe Place. Detailed descriptions of each setup are provided in the supplemental material.

Quantitative Results We quantitatively compare the task success rates of PASG against a baseline involving manual annotations. While human annotations are manually curated and serve as the gold standard, PASG annotations are generated automatically with minimal manual filtering. Each task is executed 100 times using randomly initialized seeds to ensure robustness of the evaluation. Results of this comparison are summarized in Table 2, the PASG-based policy achieves competitive performance compared to manual annotations, and even outperforms them in tasks such as “Block Hammer Beat” and “Empty Cup Place”. These results demonstrate the ability of our pipeline to generate functional and accurate geometric primitives for downstream manipulation.

Method	BHB	CP	DBP	ECP	PAM	SP	Avg.
Human Annotation	79.0	93.0	95.0	73.0	85.0	83.0	84.67
PASG	82.0	89.0	70.0	76.0	81.0	69.0	77.83

Table 2. Task success rates (%) for different manipulation scenarios. Bold highlights where PASG outperforms human annotations.

Qualitative Results A key advantage of PASG is its ability to generate a richer and more diverse set of interaction primitives compared to manual annotation. Human labelers, constrained by cost and effort, tend to identify only a few optimal points. In contrast, our automated framework identifies a wider array of semantically meaningful points, as illustrated in Figure 4. This diversity provides the manipulation policy with greater flexibility and enhances robustness to variations in task execution. For instance, the availability of multiple valid grasp points on a shoe or a cup allows successful execution even in cases of occlusion or restricted access.

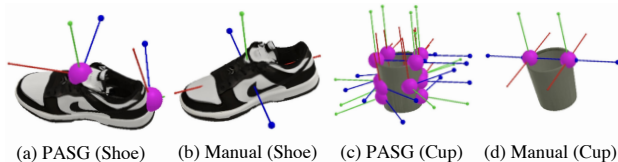


Figure 4. Compared to manual annotation, PASG tends to generate a more diverse and semantically accurate set of interaction points.

4.3. Object-based Spatial-Semantic Reasoning

Benchmark To evaluate whether our framework effectively captures spatial primitives (keypoints and axes), we developed Robocasa-PA, a visual question-answer benchmark derived from the Robocasa dataset using PASG. This spatial-aware benchmark evaluates models’ understanding of functional geometric primitives in robotic manipulation scenarios. The dataset comprises three question categories: Task 1 *Type Identification* (determining the functional category of spatial primitives from visual features), Task 2 *Task Association* (linking detected primitives to specific manipulation tasks), and Task 3 *Task-to-Primitive Mapping* (identifying primitives required to accomplish given tasks). These three question types collectively assess both geometric perception (e.g., structural recognition of primitives) and task-aware reasoning (e.g., understanding the functional significance of primitives in context).

The dataset is divided into three parts for evaluating model generalization. We first generate 6,979 questions from a designated pool of base objects, allocating 80% (5,583 questions) as the fine-tuning training set to establish a foundational understanding of primitive structures. The remaining 20% (1,396 questions) of the same object pool formed the in-distribution test set. To rigorously assess cross-domain adaptability, we introduce an out-of-distribution test set comprising 1,364 questions exclusively derived from unseen objects, ensuring strict isolation from

Model	in-distribution test set				out-of-distribution test set			
	Task1	Task2	Task3	Overall	Task1	Task2	Task3	Overall
GPT-4V [41]	32.92	38.14	46.1	39.04	37.56	36.35	43.93	39.00
GPT-4O [42]	37.87	40.07	52.48	43.19	38.29	42.99	43.69	41.79
GPT-4O-mini [42]	37.32	30.26	29.85	32.26	37.87	27.59	42.08	34.96
LLaVA-1.5 [34]	30.20	33.04	32.15	31.95	31.95	25.09	36.89	30.72
Claude-3.5 [4]	29.95	35.68	34.28	33.6	25.12	34.13	36.17	32.04
Qwen-2.5VL [6]	50.00	36.56	47.99	43.91	46.10	43.73	40.05	43.33
SpaceMantis [25]	37.13	10.54	46.57	29.15	33.66	8.86	26.70	21.70
RoboPoint [61]	32.67	0	16.31	14.40	34.63	0	16.26	15.32
Qwen2.5VL-PA	86.63	83.48	61.70	77.79	89.02	81.73	67.72	79.69

Table 3. Spatial comprehension evaluation on our visual question-answer benchmark. Numbers represent accuracy (%).

training instances at both object and primitive levels. All images are validated to ensure visibility of the referenced primitives. Each question follows a single-choice format, and accuracy is used to evaluate performance across all sets. **Finetune** We fine-tuned Qwen-2.5VL [6] using Low-Rank Adaptation (LoRA) to assess whether the VQA benchmark supports knowledge transfer in primitive compositional reasoning. By constraining parameter updates to low-rank decomposition matrices, performance improvements can be causally attributed to knowledge distillation from the benchmark. We selected several VLMs as baselines, including general-purpose large-scale vision language models and models with spatial awareness capabilities proposed in prior works, as shown in Table 3.

The fine-tuned Qwen2.5VL-PA shows significant improvements over the baselines, validating our framework’s ability to distill geometric-semantic knowledge and task-aware reasoning. Notably, the model exhibits consistent robustness in out-of-distribution tests, showcasing enhanced cross-domain adaptability and task-transfer potential.

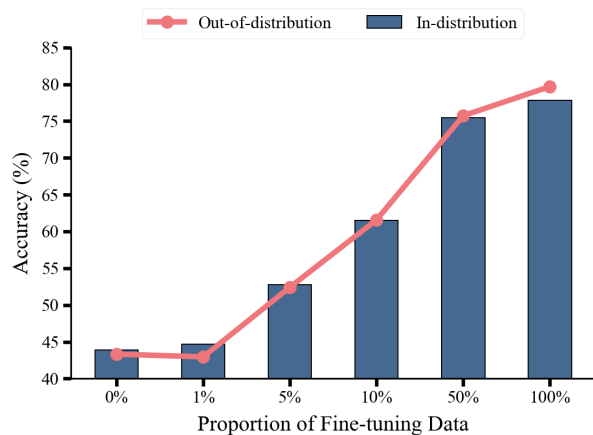


Figure 5. Data Effectiveness Study

Data Effectiveness To evaluate the effectiveness of fine-tuning data, we conducted a progressive scaling experi-

ment: fine-tune the model with randomly sampled subsets of 1% (55 samples), 5% (279 samples), 10% (558 samples), and 50% (2,791 samples) from the original training set. As shown in Fig 5, with only 5% data, the model achieved an absolute accuracy improvement of approximately 10% on both in-distribution and out-of-distribution test sets, corresponding to a 20.6% relative improvement over the original non-fine-tuned model. When increasing the subset to 10%, the absolute accuracy further improved by 20%, achieving a 41.12% relative improvement over the baseline. Notably, the performance gap between out-of-distribution and in-distribution test sets remained stable within $\pm 2\%$. These results demonstrate that our proposed data filtering pipeline can efficiently extract high-value data, significantly reducing annotation costs while ensuring robust cross-distribution generalization.

5. Conclusion

This paper presents PASG, a closed-loop framework that bridges task semantics and geometric primitives in robotic manipulation. By combining automated geometric primitive extraction with VLM-driven semantic anchoring, PASG enables spatial-semantic reasoning in unstructured environments. It overcomes key limitations in existing systems through geometry-aware feature aggregation, dynamic coupling of primitives with functional affordances, and self-corrective mechanisms to reduce error propagation. Evaluations on the RoboTwin platform across diverse manipulation tasks show PASG performs competitively with human annotations, even outperforming them in certain tasks. PASG’s ability to generate diverse interaction primitives enhances task flexibility and robustness, making it suitable for real-world applications. Additionally, the Robocasa-PA benchmark and fine-tuned Qwen2.5VL-PA model demonstrate the framework’s effectiveness, providing an automated pipeline for generating high-quality annotated data that improves generalization and cross-domain adaptability in robotic manipulation.

Acknowledgement This work was supported by the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and the Fundamental Research Funds for the Central Universities.

References

- [1] Ahmed Abdelreheem, Ivan Skorokhodov, Maks Ovsjanikov, and Peter Wonka. Satr: Zero-shot semantic segmentation of 3d shapes, 2023. 3
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2
- [4] Anthropic. Claude 3.5: Technical overview, 2024. 8
- [5] Philipp Ausrerlechner, David Haberger, Stefan Thalhammer, Jean-Baptiste Weibel, and Markus Vincze. Zs6d: Zero-shot 6d object pose estimation using vision transformers. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 463–469, 2024. 3
- [6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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, 2025. 8
- [7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 2
- [8] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 2
- [9] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465, 2024. 2
- [10] Yongchao Chen, Jacob Arkin, Charles Dawson, Yang Zhang, Nicholas Roy, and Chuchu Fan. Autotamp: Autoregressive task and motion planning with llms as translators and checkers. In *2024 IEEE International conference on robotics and automation (ICRA)*, pages 6695–6702. IEEE, 2024. 2
- [11] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Anirudha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. In *Advances in Neural Information Processing Systems*, pages 35799–35813. Curran Associates, Inc., 2023. 6
- [12] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Anirudha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13142–13153, 2023. 6
- [13] Yan Ding, Xiaohan Zhang, Chris Paxton, and Shiqi Zhang. Task and motion planning with large language models for object rearrangement. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2086–2092. IEEE, 2023. 2
- [14] Jiafei Duan, Wentao Yuan, Wilbert Pumacay, Yi Ru Wang, Kiana Ehsani, Dieter Fox, and Ranjay Krishna. Manipulate-anything: Automating real-world robots using vision-language models, 2024. 2
- [15] Junming Fan and Pai Zheng. A vision-language-guided robotic action planning approach for ambiguity mitigation in human-robot collaborative manufacturing. *Journal of Manufacturing Systems*, 74:1009–1018, 2024. 2
- [16] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers, 2021. 2
- [17] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12462–12469. IEEE, 2024. 2
- [18] Caelan Reed Garrett, Rohan Chitnis, Rachel Holladay, Beomjoon Kim, Tom Silver, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Integrated task and motion planning. *Annual review of control, robotics, and autonomous systems*, 4(1):265–293, 2021. 2
- [19] Huihui Guo, Fan Wu, Yunchuan Qin, Ruihui Li, Keqin Li, and Kenli Li. Recent trends in task and motion planning for robotics: A survey. *ACM Computing Surveys*, 55(13s):1–36, 2023. 2
- [20] Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. Multimodal latent space learning for chain-of-thought reasoning in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18180–18187, 2024. 2
- [21] Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9488–9495. IEEE, 2024. 2, 3
- [22] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023. 2

- [23] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models, 2023. 2
- [24] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024. 2, 3
- [25] Dongfu Jiang, Xuan He, Huaye Zeng, Con Wei, Max Ku, Qian Liu, and Wenhui Chen. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024. 8
- [26] Shyam Sundar Kannan, Vishnunandan LN Venkatesh, and Byung-Cheol Min. Smart-llm: Smart multi-agent robot task planning using large language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12140–12147. IEEE, 2024. 2
- [27] Hyunjin Kim and Minhyuk Sung. Partstad: 2d-to-3d part segmentation task adaptation, 2024. 3
- [28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2
- [29] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Openpipaf: Composite fields for semantic keypoint detection and spatio-temporal association. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):13498–13511, 2022. 3
- [30] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. 4
- [31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2
- [32] Dingning Liu, Xiaoshui Huang, Yuenan Hou, Zhihui Wang, Zhenfei Yin, Yongshun Gong, Peng Gao, and Wanli Ouyang. Uni3d-llm: Unifying point cloud perception, generation and editing with large language models, 2024. 2
- [33] Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-vocabulary robotic manipulation through mark-based visual prompting. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024. 2
- [34] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 8
- [35] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2
- [36] Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. *Advances in Neural Information Processing Systems*, 36:55394–55414, 2023. 2
- [37] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(4):1–23, 2021. 2
- [38] Yao Mu, Tianxing Chen, Shijia Peng, Zanxin Chen, Zeyu Gao, Yude Zou, Lunkai Lin, Zhiqiang Xie, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins (early version), 2024. 2, 3, 7, 16
- [39] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots, 2024. 6
- [40] Kazuma Obata, Tatsuya Aoki, Takato Horii, Tadahiro Taniguchi, and Takayuki Nagai. Lip-llm: Integrating linear programming and dependency graph with large language models for multi-robot task planning. *IEEE Robotics and Automation Letters*, 2024. 2
- [41] OpenAI. Gpt-4v(ision) system card, 2023. 8
- [42] OpenAI. Gpt-4o system card, 2024. 7, 8
- [43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 2
- [44] Abby O’Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024. 2
- [45] Mingjie Pan, Jiyao Zhang, Tianshu Wu, Yinghao Zhao, Wenlong Gao, and Hao Dong. Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints, 2025. 3
- [46] Seonghyun Park, An Gia Vien, and Chul Lee. Cross-modal transformers for infrared and visible image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2):770–785, 2024. 2
- [47] Georgios Pavlakos, Xiaowei Zhou, Aaron Chan, Konstantinos G. Derpanis, and Kostas Daniilidis. 6-dof object pose from semantic keypoints. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2011–2018, 2017. 3
- [48] Zekun Qi, Wenyao Zhang, Yufei Ding, Runpei Dong, Xinqiang Yu, Jingwen Li, Lingyun Xu, Baoyu Li, Xialin He, Guofan Fan, Jiazhaoh Zhang, Jiawei He, Jiayuan Gu, Xin Jin, Kaisheng Ma, Zhizheng Zhang, He Wang, and Li Yi. So-far: Language-grounded orientation bridges spatial reasoning and object manipulation, 2025. 3
- [49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

- Amanda Aspell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. [2](#)
- [50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Aspell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. [2](#)
- [51] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE, 2023. [2](#)
- [52] Chao Tang, Anxing Xiao, Yuhong Deng, Tianrun Hu, Wenlong Dong, Hanbo Zhang, David Hsu, and Hong Zhang. Functo: Function-centric one-shot imitation learning for tool manipulation, 2025. [3](#)
- [53] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. [2](#)
- [54] Ruoyu Wang, Zhipeng Yang, Zinan Zhao, Xinyan Tong, Zhi Hong, and Kun Qian. Llm-based robot task planning with exceptional handling for general purpose service robots. In *2024 43rd Chinese Control Conference (CCC)*, pages 4439–4444. IEEE, 2024. [2](#)
- [55] Shu Wang, Muzhi Han, Ziyuan Jiao, Zeyu Zhang, Ying Nian Wu, Song-Chun Zhu, and Hangxin Liu. Llm³: Large language model-based task and motion planning with motion failure reasoning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12086–12092. IEEE, 2024. [2](#)
- [56] Wei Wang, Shuo Ren, Yao Qian, Shujie Liu, Yu Shi, Yanmin Qian, and Michael Zeng. Optimizing alignment of speech and language latent spaces for end-to-end speech recognition and understanding. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7802–7806. IEEE, 2022. [2](#)
- [57] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17868–17879, 2024. [3](#)
- [58] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v, 2023. [4](#)
- [59] Qiaojun Yu, Ce Hao, Junbo Wang, Wenhai Liu, Liu Liu, Yao Mu, Yang You, Hengxu Yan, and Cewu Lu. Manipose: A comprehensive benchmark for pose-aware object manipulation in robotics, 2024. [2](#)
- [60] Haocheng Yuan, Chen Zhao, Shichao Fan, Jiayi Jiang, and Jiaqi Yang. Unsupervised learning of 3d semantic keypoints with mutual reconstruction. In *Computer Vision – ECCV 2022*, pages 534–549, Cham, 2022. Springer Nature Switzerland. [3](#)
- [61] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024. [8](#)
- [62] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12): 14679–14694, 2023. [2](#)
- [63] Jingyi Zhang, Jiaying Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. [2](#)
- [64] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model, 2024. [2](#)
- [65] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing, 2018. [6](#)
- [66] Zhehua Zhou, Jiayang Song, Kunpeng Yao, Zhan Shu, and Lei Ma. Isr-llm: Iterative self-refined large language model for long-horizon sequential task planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2081–2088. IEEE, 2024. [2](#)