Task-Aware Bimanual Affordance Prediction via VLM-Guided Semantic-Geometric Reasoning

Fabian Hahne[1], Vignesh Prasad[1,4,5], Georgia Chalvatzaki[1,4,5], Jan Peters[1,2,3,4,5], Alap Kshirsagar[1]

***Abstract*— Bimanual manipulation requires joint reasoning over object affordances and arm allocation, a challenge for geometry-only planners. To address this, we propose a hierarchical framework leveraging Vision-Language Models (VLMs) for task-aware bimanual affordance prediction without category-specific training. Our approach fuses multi-view RGB-D data to generate global 6-DoF grasps, which the VLM filters to determine task-relevant contact regions and optimal arm assignments. Evaluated on a dual-arm robot across nine real-world tasks—including tool use and human handovers—our approach significantly outperforms existing baselines, demonstrating that VLM-guided semantic reasoning enables highly reliable bimanual manipulation in unstructured environments.**

I. INTRODUCTION

As robots transition into unstructured real-world environments, the ability to manipulate multiple objects becomes critical. A core challenge is identifying task-conditioned object affordances that support stable, functionally appropriate interaction. In bimanual settings, this requires solving the **joint affordance localization and arm allocation problem**: determining where to grasp and which arm to use. Existing grasp synthesis methods [1]–[5] largely focus on single-arm geometry, while current bimanual approaches [6]–[8] often rely on predefined roles or task-specific training, limiting their versatility.

While Vision-Language Models (VLMs) have advanced open-vocabulary grasping [7], [9], their capacity to jointly inform affordance and arm allocation remains unexplored. Robots must simultaneously identify stable configurations, assign arms based on task semantics, and optimize grasp locations to avoid inter-arm conflicts. To address this, we propose a hierarchical bimanual manipulation framework that leverages a VLM for task-aware reasoning without category-specific training. By grounding semantic queries into 3D geometric grasp candidates, our method enables stable, context-aware dual-arm planning.

In summary, our contributions are:

- We reframe bimanual manipulation as a joint affordance localization and arm allocation problem where task semantics drive arm-scene interaction.

[1]Department of Computer Science, Technical University of Darmstadt [2]German Research Center for AI (DFKI) [3]Centre for Cognitive Science, Technical University of Darmstadt [4]Hessian Center for Artificial Intelligence (Hessian.AI), Darmstadt [5]Robotics Institute Germany (RIG)

This work was supported by the German Research Foundation (DFG) Emmy Noether Programme under Grant CH 2676/1-1 and under Germany's Excellence Strategy (EXC 3066/1 "The Adaptive Mind," Project No. 533717223), the EU Horizon Europe Project "ARISE" under Grant 101135959 and the German Federal Ministry of Research, Technology and Space of Germany (BMFTR) Project "RIG" under Grant 16ME1001.

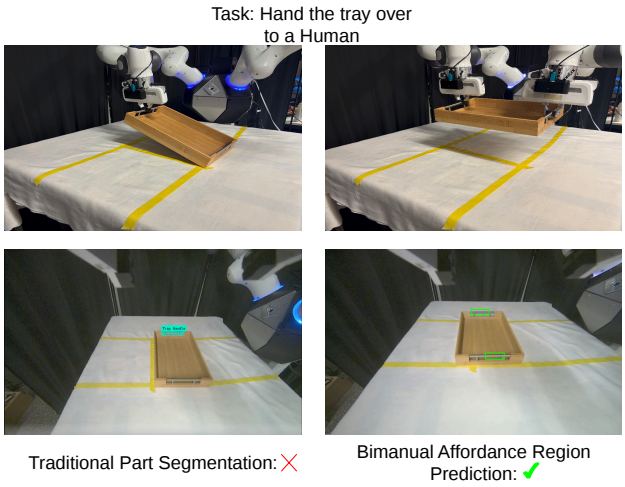


Fig. 1: A motivating example: For the task "Hand the tray over to a human," conventional labeling strategies typically reduce affordance annotation to coarse object-part segmentation for subsequent grasping. Such approaches fail to capture the task-specific requirement of simultaneously grasping both handles to ensure stability and safe transfer. In other words, traditional strategies overlook the coordinated, bimanual nature of the interaction and treat affordances as isolated object parts rather than functionally coupled regions.

- A hierarchical framework integrating VLM-based reasoning with geometric grasp generation for task-aware execution without task-specific training.
- An arm allocation strategy derived from affordance regions that outperforms geometry-only and semantic baselines across nine real-world bimanual tasks.

II. RELATED WORK

The integration of Large Language Models (LLMs) and Vision-Language Models (VLMs) into robotic grasping has enabled open-ended semantic understanding and zero-shot generalization [10]. Pioneering works like Lan-grasp [9] chain these models with planners to achieve task reasoning without specific training. This paradigm has been extended through large-scale datasets [11], visual grounding [12], and advanced reasoning for implicit intent [13] and strategic decluttering [14]. While some methods explore geometric decomposition [15] or dexterous multi-fingered grasping [16], [17], they remain largely focused on single-arm manipulation.

A significant limitation in current literature is the reduction of grasp identification to coarse *object part selection*, which

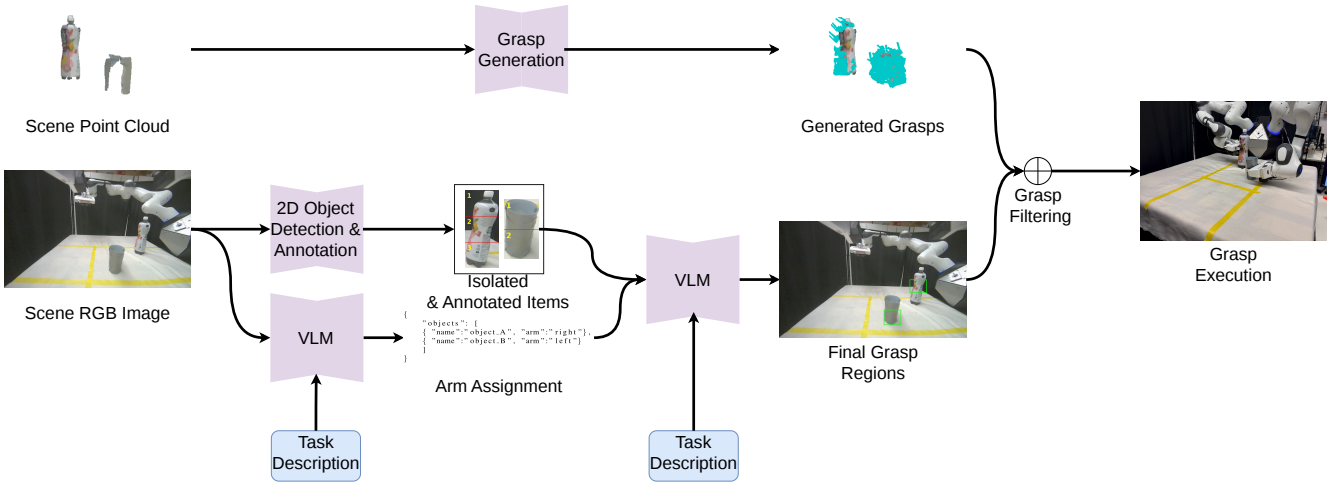


Fig. 2: Overview of our proposed approach. Given a scene RGB image, a vision–language model (VLM) performs 2D object detection and annotation, followed by arm assignment and identification of final grasp regions based on the task description. In parallel, the RGB-D input is converted into a scene point cloud, from which candidate grasps are generated. These grasps are then filtered according to the grasp regions specified by the VLM, removing proposals that fall outside the designated areas. The selected grasps which align with the task-relevant parts of each object are then executed.

fails to capture the precise localized regions required for functional contact [18]. Bimanual manipulation exacerbates this by requiring joint optimization of contact regions and division of labor between hands to satisfy spatial and task constraints.

The most relevant work, 2HandedAfforder [18], utilizes a VLM-based model trained on human activity videos. However, this necessitates extensive data curation and training. In contrast, our approach leverages the zero-shot reasoning of large VLMs. By combining region-based visual prompting [6] with semantic reasoning [9], we extend existing paradigms to predict precise bimanual affordance regions and actionable grasp configurations, providing a lightweight, deployable solution for unstructured environments.

III. PROPOSED APPROACH

Our framework integrates high-level semantic reasoning with low-level geometric sampling to bridge the gap between abstract task descriptions and precise bimanual execution. The pipeline operates across three distinct stages: multi-view representation, hierarchical reasoning via Vision-Language Models (VLMs), and candidate filtering for motion synthesis.

A. Multi-View Scene Representation and Global Sampling

To mitigate self-occlusion and environmental blind spots inherent in single-view systems, we fuse synchronized RGB-D data from two strategically placed sensors into a unified 3D point cloud $\mathcal{P}$ in the robot's base frame $\mathcal{F}_{base}$. This dense representation is processed using a voxel-downsampling filter to maintain computational efficiency while preserving surface topology.

This fused point cloud serves as the input for a training-free, off-the-shelf grasp generator (e.g., GPD [19] or Contact-GraspNet [20]). The generator samples the entire workspace to identify a global set of 6-DOF candidate poses $\mathcal{G} = \{g_1, g_2, \ldots, g_n\}$, where each $g_i \in SE(3)$. Candidates are initially ranked based on:

- **Surface Normals:** Local curvature alignment with the gripper's closing axis.
- **Collision Geometry:** Voxel-based checking to ensure the gripper mesh does not intersect with the scene $\mathcal{P}$ or the robot's own kinematic chain.

B. Hierarchical Semantic Reasoning

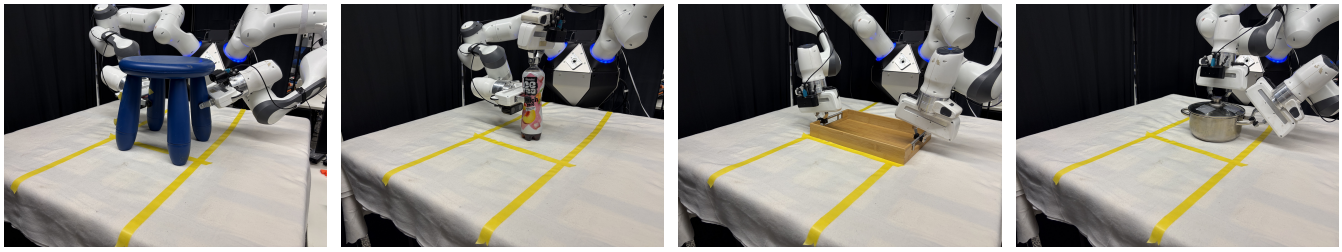
While geometric sampling provides feasible grasps, it lacks the task-context necessary for complex bimanual manipulation (e.g., knowing to hold a bottle by the base rather than the cap). We utilize a VLM to refine the global pool $\mathcal{G}$ through a top-down hierarchical process.

1) Step 1: Arm Allocation and Strategy Synthesis: The VLM evaluates the natural language task description T and a detected object list O. It determines the mechanical requirements—specifically whether an object requires a `left`, `right`, or `bimanual` strategy based on its scale, mass distribution, and the intended interaction (e.g., stabilization vs. manipulation).

```
Output Format: {"objects":
[{"name": "<obj>", "arm":
"left/right/bimanual"}]}
```

2) Step 2: Region-Based Affordance Selection: To bridge the modality gap between 2D reasoning and 3D execution, we isolate objects using 2D instance segmentation masks. We overlay a size-adaptive grid $\mathcal{M}$ over the segmented object, where each cell $c_{i,j}$ has dimensions proportional to the robot's gripper width. The VLM acts as a zero-shot affordance selector, identifying specific grid cells that represent optimal contact zones for the assigned task.

```
Output Format: {"cell_robot_right":
["<cell_id>"], "cell_robot_left":
["<cell_id>"]}
```



(a) Stool handover: Symmetric bimanual grasp for stable lifting and human access.

(b) Bottle opening: Asymmetric roles with one arm stabilizing and the other unscrewing.

(c) Tray handover: Coordinated dual-arm grasping for balanced transfer.

(d) Pot opening: Complementary roles for body stabilization and lid removal.

Fig. 3: Successful executions demonstrating symmetric stabilization, complementary manipulation, and fine-grained affordance selection through our integrated framework.

C. Candidate Filtering and Bimanual Execution

The selected 2D cells are back-projected into the 3D workspace using the camera intrinsics K to create semantic volumes V_{left} and V_{right}. We filter the global grasp pool $\mathcal{G}$ such that a candidate g_i is retained only if its contact center $p_i \in V_{arm}$. Final execution is governed by two coordination modes:

- **Parallel Mode (Decoupled):** For tasks involving independent objects, the system selects the highest-scoring collision-free candidate g^* for each arm within its respective volume V.
- **Coordinated Mode (Coupled):** For single-object bimanual grasps (e.g., lifting a heavy tray), the system performs a joint optimization. We seek a pair (g_L, g_R) that maximizes a composite quality score $Q_{total} = Q(g_L) + Q(g_R)$ subject to:

$$\|p_L - p_R\|_2 \geq d_{min} \tag{1}$$

where d_{min} is a safety threshold to prevent inter-gripper collision during the approach and closing phases.

This hierarchical approach ensures that the resulting motions are not only geometrically reachable but also semantically appropriate for the specific functional requirements of the task.

IV. EXPERIMENTAL EVALUATION

We evaluate whether modeling language-conditioned bimanual affordances produces manipulation that is both geometrically feasible and semantically consistent. Unlike metrics focusing solely on grasp stability, our experiments measure **strategy alignment**: whether the robot uses the correct arm(s) and contacts the correct regions for a given task.

A. Experimental Setup

The platform consists of two Franka Emika Panda arms with overlapping workspaces.

- **Perception:** Two ZED X Mini cameras provide fused 3D point clouds, reducing occlusions.
- **Grasp Generation:** AnyGrasp [21] generates 6-DoF candidates from the fused cloud.
- **Filtering:** GPT-5 performs arm allocation and grid-based region selection to filter the global grasp pool.

B. Experimental Tasks and Baselines

We tested nine tasks across four categories: **Parallel/Coordination** (Stir, Pick/Open Pot), **Affordance-Sensitive** (Open Bottle, Pour, Sweep, Drill), and **Large Object/Handover** (Tray, Stool). We compared our framework against:

- **Geometry-Only:** Heuristic arm assignment and no region filtering.
- **Arm Only / Region Only:** Ablations removing region selection or arm allocation, respectively.
- **VLPart:** Filtering grasps using open-vocabulary part segmentation [22].

C. Results and Analysis

Performance is measured by **strategy alignment rate (%)**—matching human-defined criteria for arm choice, hand count, and contact regions.

Our method achieves a **mean alignment of 88.9%**, significantly outperforming all baselines.

- **Coordination:** Arm allocation is vital for tasks like "Pick Pot" (80%), while region selection is critical for "Stir" (80%). Our framework integrates both to reach 90% and 100% respectively.
- **Affordance:** For tools (Drill/Sweep), arm selection is the primary success driver due to workspace constraints. For functional tasks (Bottle/Pour), region selection dominates.
- **Handovers:** Coordinated reasoning ensures symmetric, human-accessible grasps on large objects where heuristic baselines fail (30% vs 100% on Stool).

V. DISCUSSION

Our results indicate that evaluating manipulation through *grasp strategy alignment* effectively isolates semantic intent from physical stability. Geometry-only baselines fail under this criterion, confirming that collision-free 6-DOF poses are insufficient for task-appropriate execution without semantic grounding.

Ablation studies reveal a functional decomposition of the bimanual problem. **Arm allocation** excels in tasks dominated

TABLE I: Strategy Alignment Rate (%) across tasks.

Method	**Parallel / Coord.**			**Affordance-Sensitive**				**Large Object**		**Mean**
	Stir	Pick	Open	Bottle	Pour	Sweep	Drill	Tray	Stool	
Geometry-Only	10	20	0	0	0	10	0	10	30	9.0
Arm Only	20	80	20	20	10	80	70	30	80	45.6
Region Only	80	20	80	80	80	20	30	80	30	55.6
VLPart	60	50	60	50	60	50	50	60	60	55.6
Ours	**100**	**80**	**90**	**90**	**80**	**100**	**70**	**90**	**100**	**88.9**

by global coordination (e.g., *Sweep*, *Drill*), while **region-based reasoning** is critical for functional contact selection (e.g., *Open Bottle*, *Pour*). Neither alone is sufficient; tasks like *Stool Handover* or *Pick Pot* require both global arm assignment and local spatial symmetry to ensure stability and human accessibility.

Comparisons with VLPart highlight the limitations of object-centric part segmentation. While segmentation identifies plausible components, it lacks an *interaction-centric* model. It frequently fails to enforce the symmetric or complementary arm configurations required for large-scale coordination. In contrast, our hierarchical framework explicitly models these affordances by conditioning region selection on global arm allocation. This approach achieves superior strategy alignment across coordination-heavy and tool-oriented tasks, suggesting that reasoning over integrated task intent and spatial symmetry is essential for robust, zero-shot bimanual manipulation.

VI. CONCLUSION

We presented a hierarchical framework that reframes dual-arm manipulation as a joint problem of semantic affordance localization and arm allocation. By grounding VLM-based reasoning in geometric grasp candidates, our approach achieves zero-shot, task-aware grasping without specialized training. Experiments across nine real-world tasks demonstrate that our method substantially outperforms geometry-only and purely semantic baselines. Ablation results confirm that bimanual reliability requires both global arm assignment and local region selection, as neither component alone generalizes across all task categories.

Despite these gains, limitations persist: the pipeline remains sensitive to heavy occlusions and lacks explicit kinematic feasibility checking during the allocation phase. Furthermore, our current execution model assumes a quasi-static environment. Future work will focus on integrating kinematic feedback directly into the VLM reasoning loop and extending the framework to dynamic, mobile manipulation scenarios.

REFERENCES

[1] L. Wang, *et al.*, "Goal-auxiliary actor-critic for 6d robotic grasping with point clouds," in *Conference on Robot Learning*, 2022.

[2] J. Urain, *et al.*, "Se(3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion," *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[3] S. Jauhri, *et al.*, "Learning any-view 6dof robotic grasping in cluttered scenes via neural surface rendering," *Robotics: Science and Systems (R:SS)*, 2024.

[4] M. F. Karim, *et al.*, "Dagdiff: Guiding dual-arm grasp diffusion to stable and collision-free grasps," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2026.

[5] ——, "Dg16m: A large-scale dataset for dual-arm grasping with force-optimized grasps," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[6] J. Liu, *et al.*, "Leveraging semantic and geometric information for zero-shot robot-to-human handover," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[7] C. Tang, *et al.*, "Graspgpt: Leveraging semantic knowledge from a large language model for task-oriented grasping," *IEEE Robotics and Automation Letters*, 2023.

[8] G. Singh, *et al.*, "Constrained 6-dof grasp generation on complex shapes for improved dual-arm manipulation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[9] R. Mirjalili, *et al.*, "Lan-grasp: Using large language models for semantic object grasping," *arXiv preprint arXiv:2310.05239*, 2023.

[10] Y. Kim, *et al.*, "A survey on integration of large language models with intelligent robots," *Intelligent Service Robotics*, 2024.

[11] A. D. Vuong, *et al.*, "Language-driven grasp detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[12] Y. Lu, *et al.*, "Vl-grasp: a 6-dof interactive grasp policy for language-oriented objects in cluttered indoor scenes," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[13] S. Jin, *et al.*, "Reasoning grasping via multimodal large language model," in *Conference on Robot Learning*. PMLR, 2025.

[14] Y. Qian, *et al.*, "Thinkgrasp: A vision-language system for strategic part grasping in clutter," in *Conference on Robot Learning (CoRL)*, 2024.

[15] S. Li, *et al.*, "Shapegrasp: Zero-shot task-oriented grasping with large language models through geometric decomposition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[16] Y.-L. Wei, *et al.*, "Grasp as you say: Language-guided dexterous grasp generation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[17] Z. Li, *et al.*, "Language-guided dexterous functional grasping by llm generated grasp functionality and synergy for humanoid manipulation," *IEEE Transactions on Automation Science and Engineering*, 2025.

[18] M. Heidinger, *et al.*, "2handedafforder: Learning precise actionable bimanual affordances from human videos," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[19] A. Ten Pas, *et al.*, "Grasp pose detection in point clouds," *The International Journal of Robotics Research*, 2017.

[20] M. Sundermeyer, *et al.*, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[21] H.-S. Fang, *et al.*, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, 2023.

[22] P. Sun, *et al.*, "Going denser with open-vocabulary part segmentation," in *IEEE/CVF International Conference on Computer Vision*, 2023.