

DiNBV-Grasp: Real-Time Distance-Aware Two-Stage Next-Best-View for Robotic Grasping

Zilong Xie^{1*}, Jingyu Gong^{1,2,3*}, Xin Tan^{1,2}, Zhizhong Zhang^{1,3},
Yanyun Qu⁴, Lizhuang Ma¹, and Yuan Xie^{1,2}

¹ School of Computer Science and Technology, East China Normal University, China

² Chongqing Key Laboratory of Precision Optics, Chongqing Institute of East China Normal University, China

³ Shanghai Key Laboratory of Computer Software Evaluating and Testing, China

⁴ School of Informatics, Xiamen University, China

zilong_6037@stu.ecnu.edu.cn yyqu@xmu.edu.cn

{jygong, xtan, zzzhang, lzma, yxie}@cs.ecnu.edu.cn

Abstract. Next-Best-View (NBV) planning is widely adopted for robotic grasping in cluttered and occluded environments. However, existing methods typically assume that a single observation at an arbitrary camera-object distance is sufficient for accurate grasp prediction. This overlooks a critical factor: grasp perception quality is highly sensitive to viewing distance, and the optimal distance varies significantly across object categories and scales. To address this limitation, we propose DiNBV-Grasp, a real-time, distance-aware two-stage NBV framework. In Stage I, we distill a large vision-language model (VLM) into a lightweight planner for global semantic NBV exploration, resolving occlusions and generating an initial grasp hypothesis with low-latency inference. In Stage II, we introduce a local distance-aware refinement mechanism that explicitly models the relationship between viewing distance and grasp success probability. During manipulator approach, a lightweight scoring network evaluates intermediate observations and selects the most informative one for final grasp prediction. Extensive simulation and real-world experiments demonstrate that DiNBV-Grasp consistently improves grasp success rates across diverse objects, particularly for small and occluded targets, while preserving real-time performance, highlighting the importance of distance-aware perception in NBV-based grasping. Our code is available at <https://github.com/XieZilongAI/DiNBV-Grasp>.

Keywords: Robotic grasping · Next-best-view planning · Distance-aware perception

* Equal Contribution ✉ Corresponding Author

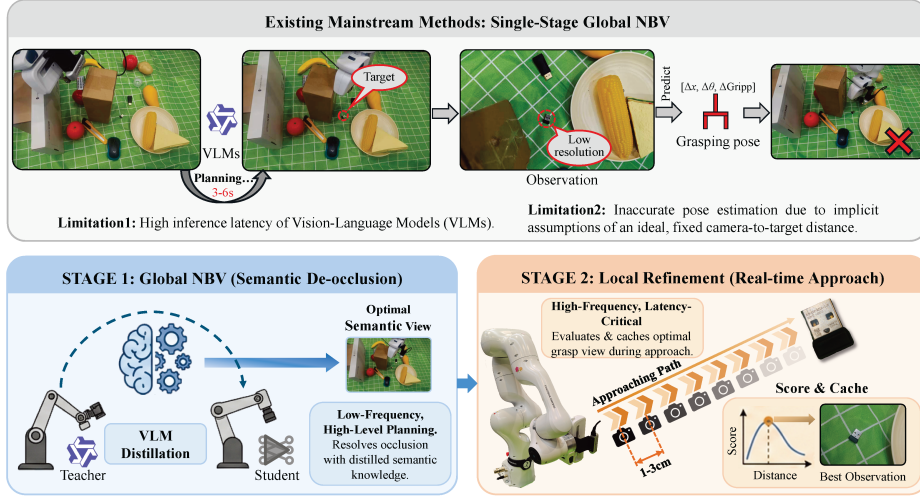


Fig. 1: Motivation of DiNBV-Grasp. Top: existing single-stage NBV pipelines rely on a single globally selected viewpoint, suffering from high VLM latency and distance-agnostic grasp failures. Bottom: the proposed two-stage framework decouples global semantic de-occlusion from local distance-aware refinement, enabling real-time and robust grasping under varying observation distances.

1 Introduction

Reliable robotic grasping in cluttered and unstructured scenes remains challenging due to occlusions and incomplete observations [4, 8, 15]. Next-Best-View (NBV) planning actively moves the sensor to improve visibility and grasp prediction quality. However, most existing NBV approaches [13, 33, 41] simplify the pipeline to selecting a single globally optimal viewpoint followed by grasp prediction from that observation. Similarly, object-centric sparse prediction [20, 21] and keyframe-based methods [7, 55] implicitly assume the existence of a suitable camera-object distance. In practice, grasp prediction quality is highly sensitive to viewing distance, especially for small or geometrically complex objects.

Another limitation lies in computational latency. Recent works [28, 36, 44, 54] incorporate 3D reconstruction [5, 14, 43] or large vision-language models (VLMs) [1, 12, 18, 25, 31] to enhance semantic reasoning. While effective, these approaches [33, 41] typically incur second-level inference delays, making it difficult to support continuous robotic control, as illustrated in Fig. 1 (top). Forcing a single heavy model to handle both global semantic reasoning and high-frequency local decision-making is therefore impractical.

To address these issues, we propose DiNBV-Grasp, a real-time two-stage NBV framework that explicitly models distance-aware grasp perception (Fig. 1, bottom). The first stage performs semantic-aware global exploration. We distill a large VLM into a lightweight student network offline, enabling multi-view eval-

uation within tens of milliseconds per image while preserving semantic scoring behavior.

The second stage operates during manipulator approach. A lightweight distance-aware scorer evaluates intermediate observations using RGB features, depth statistics, viewing distance, and object category information, and selects the most informative observation for final grasp prediction. This decoupling allows global occlusion resolution and local high-frequency refinement to operate under distinct computational constraints.

Extensive experiments validate the effectiveness of this design. In simulation, integrating our distance-aware second stage consistently improves manipulation success rates across all evaluated baselines by 6–9% on both Test Seen and Test Unseen splits. In real-world cluttered scenes, the full system achieves the highest success rate across all five scenarios while reducing the average number of exploration views to 3.10 and maintaining real-time responsiveness. These results demonstrate that explicit distance modeling constitutes a previously overlooked but critical dimension in NBV-based robotic grasping.

Our main contributions are threefold: (1) We identify and systematically analyze the critical influence of camera–object distance on grasp perception quality, highlighting a previously underexplored dimension in NBV-based manipulation. (2) We propose DiNBV-Grasp, a real-time two-stage framework that decouples global semantic exploration from local distance-aware refinement, enabling adaptive selection of informative observations during manipulator approach. (3) We validate the approach extensively in both simulation and real-world cluttered scenarios, demonstrating consistent improvements in grasp success across diverse object categories, with notable gains on small and partially occluded objects, while maintaining strict real-time performance.

2 Related Work

Object-Centric Grasping. Object-centric grasping methods predict grasp poses directly from visual observations [7, 20, 21, 42, 55]. End-to-end 6-DOF networks such as Contact-GraspNet [37] and VGN [3] generate grasp candidates from RGB-D inputs or voxelized representations in real time. Related 3D pose and shape estimation methods also exploit monocular cues or geometric features to recover object or hand structure under occlusion [24, 52], but they do not address active view selection during grasp execution. These approaches typically assume that sufficiently complete object geometry is observable from a single viewpoint under a suitable camera distance. As a result, their performance degrades under heavy occlusion or extreme near-field observations where geometric evidence becomes incomplete or distorted.

Recent efforts aim to improve robustness through structured or semantic representations. STORM [6] and OmniManip [30] learn object abstractions from primitives or semantic slots to enhance generalization. Interaction-centric frameworks such as FOCUS [10] and AnyDexGrasp [9] further emphasize identifying task-relevant entities in cluttered environments, while SimToolReal [23] extends

grasping to tool-oriented manipulation scenarios. Point-cloud completion methods such as FACNet [50] improve shape recovery from partial observations, but still rely on the available input view rather than optimizing camera motion. Although these methods improve perception and reasoning under clutter, they still rely on pre-acquired observations and generally lack adaptive viewpoint or distance modeling during execution. Keyframe-based strategies partially mitigate occlusion but do not explicitly optimize observation distance throughout the grasping process.

Active Perception for Grasping. Active perception addresses visual uncertainty by planning informative viewpoints [4, 29, 38]. Typical NBV strategies optimize information gain, entropy reduction, or grasp prediction uncertainty [33]. For example, GraspView [41] employs VLM-based scoring to evaluate rendered candidate views, while Neural Graspness Field [29] predicts spatial grasp distributions to guide camera motion via grasp inconsistency. Other systems, such as ActiveUMI [51] and RASP [27], incorporate joint viewpoint planning and depth completion inspired by human exploration strategies. More recent cognitive frameworks like CoMe-VLA [17] formulate active perception as a non-Markovian decision process for long-horizon tasks.

Despite these advances, most NBV-based methods focus primarily on global exploration to resolve occlusion before executing a final grasp from a single selected observation. The camera-object distance during the approach phase is typically treated as incidental rather than an explicit optimization variable, leaving its influence on grasp prediction quality insufficiently modeled.

Foundation Models in Robotic Perception and Planning. Foundation Models, particularly Vision-Language Models (VLMs), have been increasingly integrated into robotic systems for open-vocabulary perception, semantic reasoning, and task planning [26, 32, 47]. VISO-Grasp [33] leverages VLMs to reason about spatial relationships between targets and occluders for semantic-driven NBV planning. Halo [34] and InstructVLA [49] introduce embodied reasoning chains to align perception with downstream actions.

Related foundation-model advances also include adaptive CLIP prompting for dense scene parsing [16], depth foundation models for generalizable visual depth estimation [48], lifelong skill scaling [53], action-aligned affordance planning [40], and semantic-geometric mapping for embodied navigation [19].

However, large-scale VLM inference incurs substantial computational overhead [41], limiting real-time applicability in closed-loop control. Consequently, recent research has explored distillation and model compression. BitVLA [39] proposes a native 1-bit architecture for edge deployment, while Drive-KD [22] and Shallow- π [11] adopt multi-teacher distillation and architectural pruning to improve inference efficiency. Nevertheless, most FM-based grasping systems operate primarily on 2D image inputs without explicitly incorporating 3D spatial priors or depth-aware modeling [35].

These observations suggest that while foundation models enhance semantic reasoning, their integration into real-time robotic grasping requires lightweight designs and tighter coupling with geometry-aware active perception mechanisms.

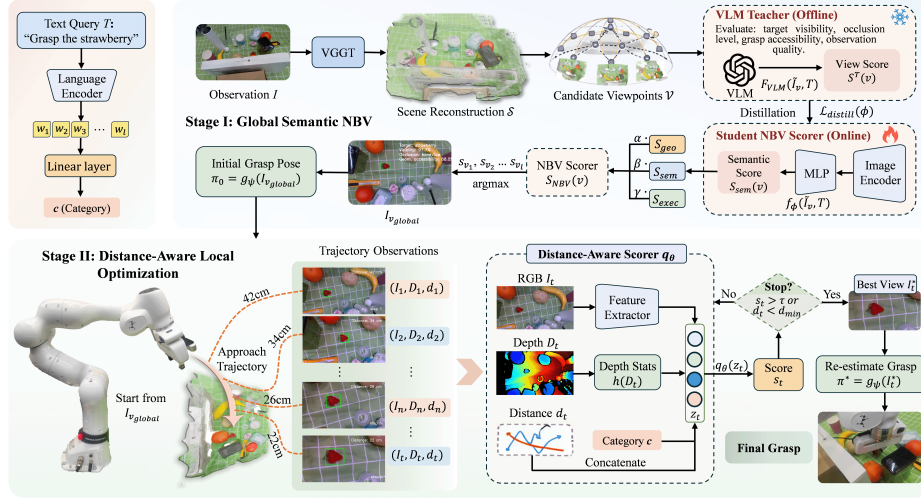


Fig. 2: Framework of DiNBV-Grasp. Stage I performs global semantic NBV planning via VLM distillation to resolve occlusions and generate an initial grasp hypothesis. Stage II conducts high-frequency distance-aware view scoring during manipulator approach and selects the most informative observation for final grasp prediction.

3 Method

3.1 Problem Formulation

Let $\mathcal{V} = \{v_k\}_{k=1}^N$ denote a discrete set of candidate camera viewpoints in the robot workspace, where each $v_k \in SE(3)$ represents a 6-DoF camera pose. The robot incrementally observes the environment and constructs a global scene representation $\mathcal{S}$.

Given a target object specified by a textual query T , the goal is to actively select informative viewpoints and execute a successful grasp.

Our system operates in two stages: (1) global NBV planning, which iteratively selects an informative viewpoint $v_{\text{global}} \in \mathcal{V}$ and predicts an initial grasp hypothesis π_0 , and (2) distance-aware local view selection, which evaluates observations collected during manipulator approach and selects the most informative one to refine the final grasp π^* .

3.2 Scene Reconstruction and Multi-View Rendering

Following [41], the system reconstructs a global scene model $\mathcal{S}$ from accumulated RGB-D observations, forming a unified 3D representation that supports rendering from arbitrary viewpoints.

Candidate viewpoints. A discrete set $\mathcal{V}$ is generated by sampling camera poses on a hemispherical shell centered around the target, with discretized radii

and viewing angles. Each candidate pose is validated via inverse kinematics and collision checking.

View rendering and evaluation. For each $v \in \mathcal{V}$, a synthetic observation is rendered:

$$\tilde{I}_v = \mathcal{R}(\mathcal{S}, v). \quad (1)$$

These virtual observations are evaluated without moving the camera, and the robot executes motion only to the selected next-best-view.

3.3 Global NBV via VLM Distillation

At exploration step t , the next-best-view is selected by maximizing a visibility-aware utility function that evaluates geometric visibility, semantic relevance, and execution feasibility:

$$S_{\text{NBV}}(v, t) = \alpha_t S_{\text{geo}}(v) + \beta_t S_{\text{sem}}(v) + \gamma_t S_{\text{exec}}(v), \quad (2)$$

where $(\alpha_t, \beta_t, \gamma_t)$ are visibility-adaptive weights conditioned on the current target visibility V_t . All three score terms are normalized over the feasible candidate set at step t : S_{geo} estimates geometric information gain from the expected newly visible target surface, S_{sem} is the distilled teacher-VLM score measuring target visibility, semantic relevance, and graspability, and S_{exec} encodes rule-based execution feasibility, including inverse-kinematics validity, collision checking, and safety constraints. When the target is heavily occluded, the utility emphasizes geometric exploration; as visibility improves, semantic and execution terms receive higher influence for selecting grasp-relevant and physically reachable views. The selected viewpoint is:

$$v_{\text{global}} = \arg \max_{v \in \mathcal{V}} S_{\text{NBV}}(v, t). \quad (3)$$

Directly computing semantic utility with large vision-language models is computationally expensive. We therefore distill the scoring behavior of a teacher VLM into a lightweight student predictor. A teacher VLM produces semantic scores:

$$S^T(v) = F_{\text{VLM}}(\tilde{I}_v, T), \quad (4)$$

and a compact model f_ϕ is trained to approximate this mapping:

$$\mathcal{L}_{\text{distill}}(\phi) = \mathbb{E}_{v \sim \mathcal{V}} \left[\ell(S^T(v), f_\phi(\tilde{I}_v, T)) \right]. \quad (5)$$

At runtime, the distilled predictor directly instantiates the semantic utility term:

$$S_{\text{sem}}(v) = f_\phi(\tilde{I}_v, T). \quad (6)$$

Global NBV planning is executed in a closed-loop manner. At each step, the system renders candidate views from the current scene representation, evaluates

their utilities, and selects the next-best-view v_{global} . The robot moves to the selected viewpoint, captures a new observation $I_{v_{\text{global}}}$, and updates the scene representation $\mathcal{S}$.

This process terminates deterministically when either the number of executed exploration views reaches N_{max} or the target visibility V_t exceeds a threshold η_v . If no valid observation or grasp candidate is found by termination, the system flags the target as ungraspable and safely aborts.

After exploration terminates, the system selects the final global observation and predicts an initial grasp hypothesis:

$$\pi_0 = g_\psi(I_{v_{\text{global}}}), \quad (7)$$

where g_ψ is a grasp prediction network.

3.4 Distance-Aware Local View Selection During Approach

After the global exploration stage, the manipulator approaches the target along a collision-free trajectory. During this motion, the system continuously evaluates intermediate observations and selects the most informative one for final grasp prediction.

Observation representation. At each approach step t , the robot acquires an RGB-D observation and measures the current camera-object distance d_t . Each observation is encoded into a compact feature vector:

$$z_t = [\phi_{\text{rgb}}(I_t), h(D_t), \hat{d}_t, c], \quad (8)$$

where ϕ_{rgb} is a visual feature extractor, $h(D_t)$ denotes low-dimensional depth statistics, $\hat{d}_t$ is the normalized distance, and category encoding c is obtained using a pre-trained CLIP text encoder, which converts the category name into a 512-dimensional text embedding. Only the text encoder is used in our implementation, enabling the representation to generalize to unseen categories without relying on a fixed vocabulary. The network uses normalized depth arrays and low-dimensional depth statistics; colorized depth maps are used only for visualization.

Distance-aware success scoring. We model view quality using a lightweight distance discriminator:

$$p_t = q_\theta(z_t) \in [0, 1], \quad (9)$$

which predicts the probability that a grasp generated from the current observation will succeed. Unlike heuristic distance selection, this model directly estimates action success likelihood conditioned on visual evidence, object category, and viewing distance.

Online selection. During the approach, the system maintains the best observation encountered so far:

$$t_T^* = \arg \max_{1 \leq t \leq T} p_t. \quad (10)$$

The approach terminates when either the predicted success probability exceeds a task-dependent threshold τ or a minimum safe distance $d_{\min}$ is reached:

$$\text{stop if } \left(\max_{1 \leq t \leq T} p_t \geq \tau \right) \text{ or } d_T \leq d_{\min}. \quad (11)$$

The final grasp pose is estimated from the most promising observation:

$$\pi^* = g_\psi(I_{t_T^*}). \quad (12)$$

This mechanism enables continuous view optimization during motion and explicitly accounts for the dependence of grasp success on viewing distance.

3.5 Training Objective

The global semantic student f_ϕ is trained using the distillation loss defined previously. Here we describe the learning objective for the distance-aware discriminator q_θ .

Supervision. Training samples are collected from approach trajectories in diverse scenes. Each observation feature z_t is assigned a binary label $y_t \in \{0, 1\}$ indicating whether the executed manipulation succeeds. The discriminator is trained to predict the probability of successful grasp execution conditioned on the observation.

Sample-level success prediction. We first optimize a binary classification objective using cross-entropy loss:

$$\mathcal{L}_{\text{bce}} = \mathbb{E}_t \left[-y_t \log q_\theta(z_t) - (1 - y_t) \log(1 - q_\theta(z_t)) \right]. \quad (13)$$

This term ensures that the predicted success probability is calibrated at the individual observation level.

Scene-level distance ranking. For observations collected from the same scene, successful viewing distances should receive higher scores than unsuccessful ones. We therefore impose a ranking constraint:

$$\mathcal{L}_{\text{rank}} = \max(0, m - (\bar{p}_{\text{succ}} - \bar{p}_{\text{fail}})), \quad (14)$$

where $\bar{p}_{\text{succ}}$ and $\bar{p}_{\text{fail}}$ denote the mean predicted success probabilities over successful and failed distances within the same scene, and m is a margin.

Pairwise contrastive separation. To further enforce discriminability between good and poor viewing distances, we apply a pairwise margin ranking loss on sampled success–failure pairs from the same scene:

$$\mathcal{L}_{\text{ctr}} = \mathbb{E}_{(i,j)} \max(0, m - (p_i - p_j)), \quad (15)$$

where i and j index successful and failed observations respectively.

Overall objective. The final training objective for the local scorer is:

$$\mathcal{L}_{\text{scorer}}(\theta) = \mathcal{L}_{\text{bce}} + \lambda_1 \mathcal{L}_{\text{rank}} + \lambda_2 \mathcal{L}_{\text{ctr}}, \quad (16)$$

where λ_1 and λ_2 balance scene-level ordering and pairwise discrimination.

Joint training. The complete learning objective of the framework is:

$$\mathcal{L}(\phi, \theta) = \mathcal{L}_{\text{distill}}(\phi) + \mathcal{L}_{\text{scorer}}(\theta) + \mathcal{R}(\phi, \theta), \quad (17)$$

where $\mathcal{R}$ denotes weight regularization and optional auxiliary supervision.

4 Experiments

We evaluate the proposed method in both large-scale simulations and real-world robotic experiments to validate its algorithmic effectiveness and practical robustness.

4.1 Simulation Setup

Experiments are conducted in interactive environments built on the SAPIEN physics simulator with PartNet-Mobility assets [45]. Manipulation is performed using a Franka Panda arm with a suction gripper. We collected approximately 40,000 successful trajectories across 20 categories for training. Evaluation is split into *Test Seen* (novel instances of observed categories) and *Test Unseen* (novel categories).

4.2 Real-World Experimental Setup

The physical platform consists of a 7-DoF Franka Emika Panda manipulator and an Intel RealSense D435i RGB-D camera. The system is implemented in Python/PyTorch, utilizing ROS 2 for middleware and MoveIt2 for motion planning. Grasp execution is powered by Contact-GraspNet [37]. Each object is evaluated over 10 independent trials.

4.3 Evaluation Metrics

To evaluate our framework comprehensively, we employ a consistent set of metrics across both simulation and real-world experiments. We report Average Success Rate (ASR, %) for simulation experiments and Success Rate (SR, %) for real-world experiments. In simulation, ASR is defined as the percentage of successful grasps where the target displacement exceeds $\delta = 0.1$ after execution. In real-world experiments, SR is defined as the ratio of successful grasps over 10 trials. Abort Rate (AR, %) quantifies the percentage of trials where the system fails to identify a valid observation within the maximum number of exploration steps N_{max} . Average Views (#Views) denotes the average number of observations, including the initial view, acquired before transitioning to the local approach stage. Regarding temporal performance, we report the Average Frame Time (ms) in simulation to measure inference latency per frame, and the Average Total Time (s) in real-world experiments to capture the full duration of perception, planning, and execution.

4.4 Implementation Details

The global NBV exploration is configured with a maximum of $N_{\max} = 5$, with early transition to Stage II triggered when target visibility V_t exceeds 0.75. The local distance-aware scorer uses $\tau = 0.8$ as the success-probability threshold during approach. The distance-aware discriminator processes a 1035-dimensional feature composed of RGB features (512D), depth statistics (10D), normalized distance (1D), and a category encoding (512D). The minimum camera-object distance is set to $d_{\min} = 15$ cm, which approximately corresponds to the lower bound of reliable depth sensing for the Intel RealSense camera, where depth measurements remain stable without severe near-field noise. The training objective combines a scene-level ranking loss ($\lambda_1 = 1.0$) and a pairwise contrastive loss ($\lambda_2 = 0.5$, margin $m = 0.2$). Optimization is performed using Adam with learning rate 10^{-4} , batch size 128, for 50 epochs. The RGB encoder remains frozen during training, and distances are normalized to $[0, 1]$ within the sampled range. For VLM distillation, Qwen3-VL-32B-Instruct [2] is used as the teacher model to generate semantic supervision for candidate rendered views. The student network adopts a lightweight DINOv3-based visual encoder followed by a small MLP scoring head.

4.5 Baselines and Integration Protocol

In simulation, we compare against Flowbot3D [7], Implicit3D [55], ManipLLM [21], CrayonRobo [20], and RobMRAG [46], which span geometric, learning-based, and vision-language paradigms. All assume manipulation prediction from a single observation at a suitable distance. Only the proposed second stage is activated and integrated as a plug-in module, while baseline pipelines remain unchanged. During execution, observations acquired at different distances are evaluated and the one with highest predicted manipulation relevance is selected for final pose estimation. Additional strategies include *random* (random observation selection), *nearest* (closest-distance selection), *Baseline* (original setting), and *heuristic*. The heuristic baseline selects the observation distance using a weighted combination of a Gaussian prior centered at a category-dependent preferred distance and a depth-quality score measuring depth consistency and valid sensing range.

In real-world experiments, we compare against *initial-view*, which predicts grasp from a predefined initial camera pose only; *fixed-traj.*, where the robot follows fixed circular trajectories centered on the target object at a 30° depression angle and uniformly samples up to $N_{\max}$ views, terminating early upon confident detection; and the method of Breyer et al. [4], a closed-loop next-best-view strategy based on geometric information gain for grasp planning. All methods share identical hardware configuration, planning stack, and execution interface to ensure fair comparison.

4.6 Simulation Results

Table 1 reports the simulation results on both the Test Seen and Test Unseen splits. Across all evaluated object-centric grasping methods, integrating our

Table 1: Simulation results (ASR, %) with our distance-aware second-stage plug-in on object-centric grasping methods.

Test Seen Split					
	Random	Nearest	Baseline	Heuristic	Ours
Flowbot3D [7]	38.88	29.76	37.00	40.09	45.16
Implicit3D [55]	45.90	38.37	46.00	48.58	52.62
ManipLLM [21]	54.90	41.37	56.00	61.58	65.62
CrayonRobo [20]	72.52	59.94	74.00	79.92	83.01
RobMRAG (<i>zero-shot</i>) [46]	45.24	43.16	49.25	52.61	57.69
Test Unseen Split					
	Random	Nearest	Baseline	Heuristic	Ours
Flowbot3D [7]	31.92	25.23	32.00	34.03	38.53
Implicit3D [55]	41.25	30.24	41.00	45.60	47.42
ManipLLM [21]	50.25	31.24	51.00	55.60	58.42
CrayonRobo [20]	70.56	55.44	72.00	75.88	79.67
RobMRAG (<i>zero-shot</i>) [46]	44.01	34.52	43.53	46.24	51.76

distance-aware second-stage module consistently improves manipulation success rates.

On the Test Seen split, our method outperforms each baseline’s default setting by a clear margin, with gains of approximately 6–9 percentage points across all methods. Notably, the improvement is consistent across diverse paradigms, including geometry-based (Flowbot3D), learning-based (Implicit3D), and vision-language-based approaches (ManipLLM, CrayonRobo, and RobMRAG), indicating strong generality. On the more challenging Test Unseen split, where generalization to novel object categories is required, our method maintains similarly robust gains. The success rates achieved under the default and random distance settings of the baseline methods are comparable. This demonstrates that relying solely on fixed observation distances or handcrafted rules is insufficient to capture the intricate relationship between perception quality and manipulation success.

Overall, these results validate that explicitly modeling observation distance and performing local view selection during the robot’s approach can serve as an effective plug-in for object-centric manipulation pipelines, significantly improving robustness without altering their core architectures or incurring additional computational overhead.

4.7 Real-world Results

We evaluate the proposed framework in five highly cluttered real-world scenarios (Scenes a–e). As shown in Fig. 3, Stage I progressively resolves occlusions and converges to an informative global viewpoint. Stage II then refines the obser-



Fig. 3: Stage I: Global View Exploration. Columns correspond to five cluttered real-world scenes (a–e) in Table 2. Rows show the progression from the initial arbitrary view (top), to intermediate exploration (middle), and the final selected Next-Best-View v_{global} (bottom), which exposes the target for initial grasp hypothesis.

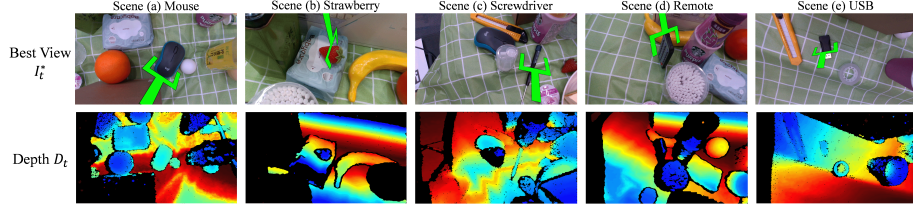


Fig. 4: Stage II: Distance-Aware Local Refinement. Starting from the global view selected in Stage I, the robot evaluates intermediate observations during approach. The top row shows the selected optimal observation I_t^* , and the bottom row presents the corresponding depth map D_t used for final grasp synthesis.

vation during approach, selecting the optimal view I_t^* and corresponding depth map D_t for grasp execution (Fig. 4).

Quantitative results in Table 2 show that *Ours (Full)* consistently achieves the highest Success Rate (SR) across all scenes. Compared to geometric NBV [4], our method improves SR while requiring substantially fewer exploration views (≤ 3.10 vs. ≥ 3.82) and significantly shorter execution time.

Importantly, the comparison between *Ours (w/o Stage II)* and *Ours (Full)* demonstrates the contribution of the proposed distance-aware refinement. While Stage I alone already outperforms existing baselines, incorporating Stage II further improves grasp success in four out of five scenes. Notably, SR increases from 2/10 to 4/10 in Scene (c) and from 3/10 to 5/10 in Scene (e), both involving small or slender objects. These gains validate that distance-aware local view selection effectively enhances grasp reliability beyond global occlusion resolution.

Performance remains relatively lower in Scene (c), where the screwdriver’s thin geometry limits grasp affordances and requires high precision from the parallel-jaw gripper. This indicates that residual failures are primarily due to physical constraints rather than perceptual deficiencies.

Table 2: Real-world experimental results in highly cluttered and occluded scenes.

Scene	Instruction	Method	SR	FR	AR	#Views	Time/s
a	Grasp the mouse	<i>initial-view</i>	0/10	10/10	0/10	1.00	8.25
		<i>fixed-traj.</i>	2/10	3/10	5/10	4.20	13.52
		<i>Breyer et al.</i>	4/10	3/10	3/10	4.31	24.81
		<i>Ours (w/o Stage II)</i>	6/10	4/10	0/10	2.15	10.82
		<i>Ours (Full)</i>	6/10	4/10	0/10	2.15	12.45
b	Take the strawberry	<i>initial-view</i>	1/10	9/10	0/10	1.00	6.82
		<i>fixed-traj.</i>	1/10	3/10	6/10	4.56	11.45
		<i>Breyer et al.</i>	3/10	2/10	5/10	4.82	28.34
		<i>Ours (w/o Stage II)</i>	4/10	4/10	2/10	2.85	12.15
		<i>Ours (Full)</i>	5/10	3/10	2/10	2.85	13.72
c	Catch the screwdriver	<i>initial-view</i>	0/10	10/10	0/10	1.00	8.05
		<i>fixed-traj.</i>	0/10	3/10	7/10	5.02	12.10
		<i>Breyer et al.</i>	1/10	5/10	4/10	4.55	26.52
		<i>Ours (w/o Stage II)</i>	2/10	6/10	2/10	2.60	11.42
		<i>Ours (Full)</i>	4/10	4/10	2/10	2.60	12.98
d	Pick up the remote controller	<i>initial-view</i>	0/10	10/10	0/10	1.00	7.55
		<i>fixed-traj.</i>	3/10	2/10	5/10	4.51	13.25
		<i>Breyer et al.</i>	3/10	3/10	4/10	3.82	22.44
		<i>Ours (w/o Stage II)</i>	5/10	5/10	0/10	2.30	9.95
		<i>Ours (Full)</i>	6/10	4/10	0/10	2.30	11.40
e	Retrieve the USB drive	<i>initial-view</i>	0/10	10/10	0/10	1.00	6.52
		<i>fixed-traj.</i>	0/10	2/10	8/10	5.45	11.88
		<i>Breyer et al.</i>	0/10	4/10	6/10	5.21	30.12
		<i>Ours (w/o Stage II)</i>	3/10	3/10	4/10	3.10	13.82
		<i>Ours (Full)</i>	5/10	1/10	4/10	3.10	15.28

4.8 Ablation Studies

To thoroughly evaluate the individual contributions of our proposed components and justify our architectural choices, we conduct extensive ablation studies in simulation. We analyze the design trade-offs in both the first-stage global NBV planner and the second-stage local view scorer.

Table 3 compares the teacher VLM and distilled students. Directly using the VLM requires the fewest views (2.15) but incurs prohibitive latency (693 ms). A lightweight ResNet-18 reduces latency to 17 ms but increases required views (3.19). DINOv3-Base achieves a near-teacher view efficiency (2.26) with practical latency (28 ms), offering the best trade-off between semantic capability and real-time performance.

Table 4 evaluates feature design. Adding depth statistics and distance improves ASR from 47.92% to 49.58%. Incorporating object category prior further

Table 3: Ablation of Stage I Distilled Models. Evaluating the trade-off between search efficiency (#Views) and inference latency (Avg Frame Time).

Distilled Model	#Views ↓	Avg Frame Time (ms) ↓
VLM Teacher (No Distill)	2.15	693
ResNet-18	3.19	17
ViT-Base	2.60	23
DINOv3-Base (Ours)	2.26	28

Table 4: Ablation of Stage II Input Features (Baseline: ManipLLM). *Img*: Image, *Dep*: Depth, *Dist*: Distance, *Cat*: Category, *Aff*: Affordance.

Scorer Configuration	Test Seen ASR (%)	Test Unseen ASR (%)	Avg Frame Time (ms)
ResNet-18 (<i>Img</i>)	47.92	44.09	18.82
+ <i>Dep</i> + <i>Dist</i>	49.58	46.77	26.10
+ <i>Dep</i> + <i>Dist</i> + <i>Cat</i>	59.30	54.06	27.97
+ <i>Dep</i> + <i>Dist</i> + <i>Cat</i> + <i>Aff</i>	65.91	59.31	210.68
DINOv3-Tiny			
+ <i>Dep</i> + <i>Dist</i> + <i>Cat</i> (Ours)	65.62	58.42	47.82

boosts performance (+9.72% Seen, +7.29% Unseen), indicating the importance of scale-aware distance modeling. Although adding affordance features slightly increases ASR (65.91%), it raises latency above 210 ms, violating real-time constraints. Using DINOv3-Tiny without heavy affordance features achieves competitive performance (65.62% / 58.42%) under 48 ms latency, satisfying high-frequency refinement requirements.

Fig. 5(a) shows that increasing τ reduces final distance (73.5 cm $\rightarrow$ 22.2 cm) and improves ASR until saturation at $\tau \approx 0.8$. Excessive proximity introduces sensing artifacts and kinematic constraints.

Fig. 5(b) illustrates scale-dependent behavior: small objects require near-field observations ($d < 30$ cm), whereas large objects achieve peak confidence at intermediate distances ($d \approx 62$ cm). This validates the necessity of explicit distance-aware modeling.

5 Conclusion

We presented DiNBV-Grasp, a distance-aware two-stage Next-Best-View framework that explicitly models the impact of camera-object distance on grasp success. Unlike conventional NBV methods that rely on a single global viewpoint, our approach extends active perception into the manipulator approach phase through continuous distance-aware view evaluation. By decoupling global semantic exploration from lightweight local refinement, the framework balances VLM-based reasoning capability with strict real-time constraints and can be

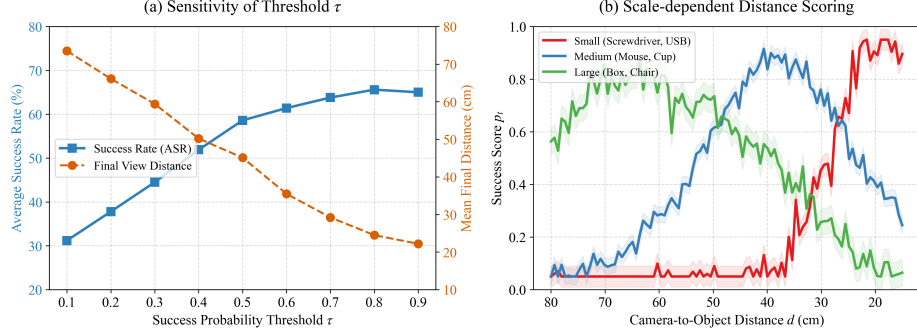


Fig. 5: Ablation on success threshold and object scale. (a) Increasing the success threshold τ forces closer observation, improving ASR until saturation at 65.6%. (b) Learned success score p_t versus camera-to-object distance d across object scales. Small objects require near-field observations, whereas large objects peak at intermediate distances and degrade under extreme proximity.

integrated into existing object-centric grasping pipelines. Extensive simulation and real-world experiments demonstrate consistent improvements in grasp success across diverse object categories, with notable gains on small and partially occluded objects. Ablation results further show that distance cues complement RGB-D observations by helping select grasp-relevant views under varying object scales and sensor quality. These results underscore observation distance as a critical yet previously underexplored factor in NBV-based robotic manipulation, and suggest that future active perception systems should jointly consider visibility, geometry, execution feasibility, and distance-dependent sensing quality.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62502159, W2521174, 62476090, U23A20343, 62302167, 62222602), Fundamental Research Funds for the Central Universities, Natural Science Foundation of Shanghai (Grant No. 25ZR1402135), the Shanghai Committee of Science and Technology (Grant No. 25511103300, 25511104302, 25511102700), “Chen Guang” project supported by Shanghai Municipal Education Commission and Shanghai Education Development Foundation (Grant No. 25CGA22), Natural Science Foundation of Chongqing (Grant No. CSTB2025NSCQ-GPX0445), Young Elite Scientists Sponsorship Program by CAST (Grant No. YESS20240780), Open Project Program of the State Key Laboratory of CAD&CG (Grant No. A2501), Zhejiang University, Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, ECNU.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* **35**, 23716–23736 (2022)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631* (2025)
3. Breyer, M., Chung, J.J., Ott, L., Siegwart, R.Y., Nieto, J.I.: Volumetric grasping network: Real-time 6 dof grasp detection in clutter. In: *Conference on Robot Learning* (2021)
4. Breyer, M., Ott, L., Siegwart, R., Chung, J.J.: Closed-loop next-best-view planning for target-driven grasping. In: *IROS*. pp. 1411–1416. IEEE (2022)
5. Cabon, Y., Stoffl, L., Antsfeld, L., Csurka, G., Chidlovskii, B., Revaud, J., Leroy, V.: Must3r: Multi-view network for stereo 3d reconstruction. In: *CVPR* (2025)
6. Chapin, A., Dellandréa, E., Chen, L.: Storm: Slot-based task-aware object-centric representation for robotic manipulation. *arXiv preprint arXiv:2601.20381* (2026)
7. Eisner, B., Zhang, H., Held, D.: Flowbot3d: Learning 3d articulation flow to manipulate articulated objects. In: *RSS* (2022)
8. Fang, H.S., Wang, C., Gou, M., Lu, C.: Graspnet-1billion: A large-scale benchmark for general object grasping. In: *CVPR*. pp. 11441–11450 (2020)
9. Fang, H.S., Yan, H., Tang, Z., Fang, H., Wang, C., Lu, C.: Anydexgrasp: General dexterous grasping for different hands with human-level learning efficiency. *arXiv preprint arXiv:2502.16420* (2025)
10. Ferraro, S., Mazzaglia, P., Verbelen, T., Dhoedt, B.: Focus: object-centric world models for robotic manipulation. *Frontiers in Neurorobotics* **19**, 1585386 (2025)
11. Jeon, B., Choi, Y., Kim, T.: Shallow- π : Knowledge distillation for flow-based vlas. *arXiv preprint arXiv:2601.20262* (2026)
12. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: *International conference on machine learning*. pp. 4904–4916. PMLR (2021)
13. Jin, L., Chen, X., Rückin, J., Popović, M.: Neu-nbv: Next best view planning using uncertainty estimation in image-based neural rendering. In: *IROS*. pp. 11305–11312. IEEE (2023)
14. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**(4), 139–1 (2023)
15. Lauri, M., Pajarinen, J., Peters, J., Frintrop, S.: Multi-sensor next-best-view planning as matroid-constrained submodular maximization. *IEEE Robotics and Automation Letters* **5**(4), 5323–5330 (2020)
16. Li, J., Huang, Y., Wu, M., Zhang, B., Ji, X., Zhang, C.: Clip-sp: Vision-language model with adaptive prompting for scene parsing. *Computational Visual Media* **10**(4), 741–752 (2024)
17. Li, J., Qiao, Y., Guo, Y., Chen, C., Lian, W.: Act, sense, act: Learning non-markovian active perception strategies from large-scale egocentric human data. *arXiv preprint arXiv:2602.04600* (2026)
18. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: *International conference on machine learning*. pp. 19730–19742. PMLR (2023)

19. Li, K., Qian, T., Yang, L., Fu, Y., Gong, J., Wang, X., He, L.: Bridging the 2d-3d gap: A hierarchical semantic-geometric map for vision language navigation. In: CVPR. pp. 15243–15252 (2026)
20. Li, X., Xu, J., Zhang, M., Liu, J., Shen, Y., Ponomarenko, I., Xu, J., Heng, L., Huang, S., Zhang, S., et al.: Object-centric prompt-driven vision-language-action model for robotic manipulation. In: CVPR. pp. 27638–27648 (2025)
21. 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: CVPR. pp. 18061–18070 (2024)
22. Lian, W., Tang, Z., Li, H., Gao, T., Wang, Y., Wang, Z., Meng, L., Ru, T., Cui, Z., Zhu, Y., et al.: Drive-kd: Multi-teacher distillation for vlms in autonomous driving. arXiv preprint arXiv:2601.21288 (2026)
23. Lin, T., Sachdev, K., Fan, L., Malik, J., Zhu, Y.: Sim-to-real reinforcement learning for vision-based dexterous manipulation on humanoids. arXiv preprint arXiv:2502.20396 (2025)
24. Liu, C., Chen, F., Deng, L., Yi, R., Zheng, L., Zhu, C., Wang, J., Xu, K.: 6dof pose estimation of a 3d rigid object based on edge-enhanced point pair features. Computational Visual Media **10**(1), 61–77 (2024)
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
26. 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: CVPR. pp. 21736–21746 (2023)
27. Liu, Y., Jia, T., Zhang, H., Yang, G., Wang, H., Chen, D.: Rasp: Robot active scene perception with joint viewpoint planning and depth completion in cluttered environments. IEEE Transactions on Automation Science and Engineering (2025)
28. Liu, Z., Gu, Y., Wang, Y., Xue, X., Fu, Y.: Activevla: Injecting active perception into vision-language-action models for precise 3d robotic manipulation. arXiv preprint arXiv:2601.08325 (2026)
29. Ma, H., Shi, M., Gao, B., Huang, D.: Active perception for grasp detection via neural graspness field. Advances in Neural Information Processing Systems **37**, 38122–38141 (2024)
30. Pan, M., Zhang, J., Wu, T., Zhao, Y., Gao, W., Dong, H.: Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. In: CVPR. pp. 17359–17369 (2025)
31. 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)
32. Radosavovic, I., Xiao, T., James, S., Abbeel, P., Malik, J., Darrell, T.: Real-world robot learning with masked visual pre-training. In: Conference on Robot Learning. pp. 416–426. PMLR (2023)
33. Shi, Y., Wen, D., Chen, G., Welte, E., Liu, S., Peng, K., Stiefelwagen, R., Rayyes, R.: Viso-grasp: Vision-language informed spatial object-centric 6-dof active view planning and grasping in clutter and invisibility. arXiv preprint arXiv:2503.12609 (2025)
34. Shou, Q., Zhu, F., Chen, S., Yan, P., Yan, Z., Miao, Y., Pang, X., Hong, Z., Shi, R., Huang, H., et al.: Halo: A unified vision-language-action model for embodied multimodal chain-of-thought reasoning. arXiv preprint arXiv:2602.21157 (2026)
35. Sripada, V., Carter, S., Guerin, F., Ghahramani, A.: Ap-vlm: Active perception enabled by vision-language models. arXiv preprint arXiv:2409.17641 (2024)

36. Strong, M., Lei, B., Swann, A., Jiang, W., Daniilidis, K., Kennedy, M.: Next best sense: Guiding vision and touch with fisherrf for 3d gaussian splatting. In: ICRA. pp. 3204–3210. IEEE (2025)
37. Sundermeyer, M., Mousavian, A., Triebel, R., Fox, D.: Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In: ICRA. pp. 13438–13444 (2021)
38. Vutetakis, D., Xiao, J.: Active perception network for non-myopic online exploration and visual surface coverage. *The International Journal of Robotics Research* **44**(2), 247–272 (2025)
39. Wang, H., Xiong, C., Wang, R., Chen, X.: Bitvla: 1-bit vision-language-action models for robotics manipulation. arXiv preprint arXiv:2506.07530 (2025)
40. Wang, R., Fu, Y., Li, Y., Lin, T., Qian, T., Elhoseiny, M., Zhao, B., Fu, Y., Jiang, Y.G., Xue, X.: Afford-vla: Action-aligned visual planning via internalized affordance. arXiv preprint arXiv:2605.24203 (2026)
41. Wang, S., Dai, M., Su, J., Liu, L., Chen, C., Wu, X., Lin, L.: Graspview: Active perception scoring and best-view optimization for robotic grasping in cluttered environments. arXiv preprint arXiv:2511.04199 (2025)
42. Wei, Z., Xu, Z., Guo, J., Hou, Y., Gao, C., Cai, Z., Luo, J., Shao, L.: $\mathcal{D}(\mathcal{R}, \mathcal{O})$ grasp: A unified representation of robot and object interaction for cross-embodiment dexterous grasping. In: ICRA. pp. 4982–4988. IEEE (2025)
43. Wewer, C., Raj, K., Ilg, E., Schiele, B., Lenssen, J.E.: Latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) *Computer Vision – ECCV 2024*. pp. 456–473. Springer Nature Switzerland, Cham (2025)
44. Wilson, J., Almeida, M., Mahajan, S., Labrie, M., Ghaffari, M., Ghasemalizadeh, O., Sun, M., Kuo, C.H., Sen, A.: Pop-gs: Next best view in 3d-gaussian splatting with p-optimality. In: CVPR. pp. 3646–3655 (2025)
45. Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Jiang, H., Yuan, Y., Wang, H., et al.: Sapien: A simulated part-based interactive environment. In: CVPR. pp. 11097–11107 (2020)
46. Xie, Z., Gong, J., Tan, X., Zhang, Z., Xie, Y.: Zero-shot robotic manipulation via 3d gaussian splatting-enhanced multimodal retrieval-augmented generation. arXiv preprint arXiv:2603.00500 (2026)
47. Xu, J., Jin, S., Lei, Y., Zhang, Y., Zhang, L.: Rt-grasp: Reasoning tuning robotic grasping via multi-modal large language model. In: IROS. pp. 7323–7330. IEEE (2024)
48. Xu, Z., Zhou, H., Peng, S., Lin, H., Guo, H., Shao, J., Yang, P., Yang, Q., Miao, S., He, X., et al.: Towards depth foundation models: Recent trends in vision-based depth estimation. *Computational Visual Media* (2026)
49. Yang, S., Li, H., Chen, Y., Wang, B., Tian, Y., Wang, T., Wang, H., Zhao, F., Liao, Y., Pang, J.: Instructvla: Vision-language-action instruction tuning from understanding to manipulation. arXiv preprint arXiv:2507.17520 (2025)
50. Yu, X., Li, J., Wong, C.C., Vong, C.M., Liang, Y.: Facnet: Feature alignment fast point cloud completion network. *Computational Visual Media* **11**(1), 141–157 (2025)
51. Zeng, Q., Li, C., John, J.S., Zhou, Z., Wen, J., Feng, G., Zhu, Y., Xu, Y.: Activeumi: Robotic manipulation with active perception from robot-free human demonstrations. arXiv preprint arXiv:2510.01607 (2025)
52. Zhang, F., Zhao, L., Li, S., Su, W., Liu, L., Tao, W.: 3d hand pose and shape estimation from monocular rgb via efficient 2d cues. *Computational Visual Media* **10**(1), 79–96 (2024)

- 53. Zhang, H., Gong, J., Zhang, Z., Tan, X., Qu, Y., Xie, Y.: T2s: Tokenized skill scaling for lifelong imitation learning. arXiv preprint arXiv:2508.01167 (2025)
- 54. Zhang, X., Wang, D., Han, S., Li, W., Zhao, B., Wang, Z., Duan, X., Fang, C., Li, X., He, J.: Affordance-driven next-best-view planning for robotic grasping. arXiv preprint arXiv:2309.09556 (2023)
- 55. Zhong, C., Zheng, Y., Zheng, Y., Zhao, H., Yi, L., Mu, X., Wang, L., Li, P., Zhou, G., Yang, C., et al.: 3d implicit transporter for temporally consistent keypoint discovery. In: ICCV. pp. 3869–3880 (2023)