LIBERO-Occ: Evaluating and Improving Vision-Language-Action Models under Scene-Induced Occlusion via Viewpoint Imagination

Anonymous ACL submission

Abstract

Vision-Language-Action (VLA) models achieve strong performance on standard manipulation benchmarks, but most evaluations assume that task-relevant objects are fully visible. This assumption often fails in realistic settings, where occlusion makes manipulation partially observable. In this paper, we study *scene-induced occlusion* as a fundamental challenge for VLA models and introduce **LIBERO-Occ**, an occlusion-oriented extension of LIBERO. Experiments show that state-of-the-art VLAs suffer substantial performance degradation under occlusion. To address this issue, we propose **Viewpoint Imagination (VIM)**, which generates a complementary view from an occluded primary observation and conditions action prediction on both observed and imagined evidence. VIM improves robustness across task suites, occlusion types, and severity levels without requiring additional cameras at deployment time, suggesting that viewpoint imagination is an promising mechanism for perception completion in partially observable manipulation.

1 Introduction

Recently, Vision-Language-Action (VLA) models have emerged as a promising paradigm for generalizable robotic manipulation. Some VLAs (Kim et al., 2025; Black et al., 2025; Wang et al., 2025b) have achieved strong performance on standard manipulation benchmarks. However, most existing evaluation protocols implicitly assume that task-relevant objects are fully visible in the input observation. This assumption rarely holds in real-world manipulation environments, where occlusion is a common phenomenon (Guruprasad et al., 2024; Zhang et al., 2026). Under occlusion, manipulation becomes a partially observable decision-making problem (Zhao and Chen, 2021; Xiao et al., 2019), demanding the ability to infer unobserved scene states (Back et al., 2021; Miao et al., 2022). As a result, a critical challenge for deployable VLA systems is whether they can operate when critical task-relevant evidence is not fully observable.

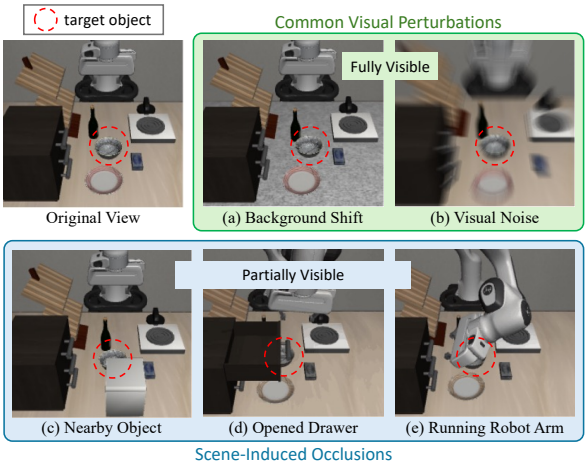


Figure 1: Comparison with common perturbations. Task instruction: *"open the top drawer and put the bowl inside"*. Unlike artificially-crafted perturbations, scene-induced occlusions remove task-relevant evidence from the observation and increase the task difficulty.

We refer to this problem as *scene-induced occlusion*, an intrinsic challenge arising from the physical configuration and dynamics of the manipulation scene itself. As shown in Figure 1, the target object may be initially hidden behind nearby objects, or be covered by manipulated objects or robot arms during operation. Unlike common visual perturbations (Guo et al., 2025; Fei et al., 2025) that artificially alter observation appearance, scene-induced occlusion preserves scene texture while reducing the visibility of task-relevant evidence. Despite its practical importance, scene-induced occlusion remains underexplored in the VLA domain, making it unclear how robust existing models are under such partial observability.

A straightforward way to mitigate occlusion is to increase the physical visibility of the target object, either by introducing additional cameras (Bian

et al., 2025) or by actively adjusting the camera viewpoint (Liu et al., 2026a,b). Though effective, these approaches can be costly in real deployment, as additional cameras require calibration, maintenance, and appropriate placement; and active camera systems introduce extra hardware constraints. This raises a central question: can a VLA model recover missing task-relevant information without adjusting hardware?

Motivated by recent progress in embodied world models and generative modeling (Cen et al., 2025; Wang et al., 2025b), we explore **Viewpoint Imagination (VIM)** as a mechanism for perception completion. The key intuition is that generation can serve as a form of understanding: if a model can imagine a plausible complementary view from an occluded observation, it may recover spatial evidence that is hidden in the primary camera. We therefore propose a viewpoint imagination-based framework for VLA-based manipulation under scene-induced occlusion. Our framework is built upon a world model and aims to transform its generative prior into a perception-completion capability for manipulation. Given an occluded primary observation, the model first generates a complementary viewpoint that may reveal task-relevant information. It then predicts actions conditioned on both the observed image and the imagined view. The framework is trained in two stages: it first learns to generate a complementary viewpoint from the occluded observation, and then jointly optimizes viewpoint imagination and action prediction so that the generated visual evidence is aligned with downstream policy learning. At inference time, the method requires only the original camera observation, avoiding the need for additional hardware.

To systematically evaluate the influence of scene-induced occlusion in VLA manipulation, we construct LIBERO-Occ, an occlusion-oriented extension of the LIBERO benchmark (Liu et al., 2023). LIBERO-Occ introduces controlled scene-induced occlusions while preserving the original task semantics. Moreover, we include multiple occlusion severity levels, enabling reliable evaluation under different degrees of missing task-relevant information. Experiments show that existing VLA models suffer substantial performance degradation under scene-induced occlusion, while our proposed viewpoint imagination framework remians effective, demonstrating the feasibility of using imagined complementary views to improve manipulation robustness under partial observability.

In summary, our contributions are threefold:

- We identify scene-induced occlusion as an important and underexplored challenge for VLA.
- We introduce **LIBERO-Occ**, a benchmark for systematic evaluation under controlled occlusion types and severity levels.
- We propose **VIM**, a viewpoint imagination-based VLA framework that tackles scene-induced occlusion by inferring complementary visual evidence without modifying the hardware setup.

2 LIBERO-Occ Benchmark

To systematically evaluate scene-induced occlusion, we introduce LIBERO-Occ, an occlusion-oriented benchmark built upon LIBERO (Liu et al., 2023). LIBERO is a widely used robotic manipulation benchmark that covers four task suites—Spatial, Object, Goal, and 10—for evaluating generalization across layouts, objects, goals, and long-horizon tasks. LIBERO-Occ extends these suites by introducing physically grounded occlusions while preserving the original task semantics and executability.

2.1 Occlusion Task Generation

Constructing realistic scene-induced occlusions is non-trivial. Valid occluder placements must effectively block the target from the camera while jointly respecting scene geometry, target object location, and workspace layout, while the resulting task must remain physically executable. To construct LIBERO-Occ at scale, we develop an automatic task generation pipeline that builds occluded task scenes from the original LIBERO tasks.

Step 1: Occlusion Target Identification. Given an original LIBERO task, we first parse its BDDL[1] specification to identify task-relevant entities as occlusion targets. These entities include manipulated objects, receptacles or goal regions. This step determines which objects or regions should be considered valid occlusion targets.

Step 2: View-aware Occluder Placement. To generate physical occlusions rather than image-space masks, we sample occluder placements in 3D scene space. Specifically, for each occlusion target, we calculate candidate placement locations along the camera-to-target ray from the primary

[1] Behavior Domain Definition Language (BDDL) is the symbolic task specification format used by LIBERO. It defines a manipulation task through involved objects, initial scene conditions, and goal predicates.

Occlusion Type	Light	Medium	Heavy	Total
Manipulated object	257	412	231	900
Receptacle	186	338	226	750
Dual	57	250	43	350
All	**500**	**1000**	**500**	**2000**

Table 1: Tasks distribution across occlusion types and severity levels in LIBERO-Occ.

view, so that the inserted occluder is likely to block the target in the observation. We then sample occluder objects from the LIBERO object library and filter out objects that semantically overlap with existing task objects, avoiding ambiguities that could change the intended task semantics.

Step 3: Occlusion Validity Verification. Each feasible candidate occluder is instantiated as an additional object in the scene and rendered in the simulator. To make sure the added occluder is physically plausible, we conduct three verification steps. We retain candidates only if they satisfy three constraints: (1) *Visibility check*: We verify the occlusion effect by rendering both the occluded scene and a reference scene with the occluder removed. We reject candidates that either fail to induce sufficient occlusion or make the target almost completely invisible and thus infeasible. (2) *Physical validity check*: We reject tasks in which the inserted occluder collides with existing objects after reset. (3) *Task executability check*: Finally, we replay the original demonstration in the occluded scene and retain only variants where the task can still be successfully completed. This step ensures that each retained task remains executable under the introduced occlusion, so that model failures can be attributed to occlusion-induced partial observability rather than invalid task construction.

2.2 Benchmark Characterization

LIBERO-Occ characterizes each occluded task along two axes: *what* task-relevant information is hidden and *how much* of that information is missing from the primary view. This two-axis design enables systematic evaluation of model performance across diverse occlusion conditions.

Occlusion Types. We categorize occlusions according to the functional role of the occlusion target in the task. LIBERO-Occ includes three types.

- *Manipulated object occlusion*: the object to be grasped, moved, or otherwise manipulated is occluded. This directly tests whether the model can recognize and localize the manipulation target under incomplete visual evidence.
- *Receptacle occlusion*: the destination region, container, or contact area required for task completion is occluded. This evaluates whether the model can infer where an object should be placed or how it should interact when the goal region is not fully visible.
- *Dual occlusion*: both the manipulated object and the receptacle are simultaneously occluded. This setting tests whether the model can jointly localize the object to act on and infer the partially visible goal region.

Occlusion Severity. Occlusion is not a binary attribute. When only a small region of the target is occluded, the model can rely on partial visual cues for localization. In more challenging scenarios where most of the target object is not visible, however, the model must infer its state from commonsense and spatial reasoning alone. To quantify occlusion severity, we compute the visibility loss of the occlusion target using instance segmentation. Let A_{full} denote the segmentation area of the target object in the original reference scene, and A_{visible} denote its area in the occluded observation. The occlusion severity score is defined as $S_{\text{occ}} = (A_{\text{full}} - A_{\text{visible}})/A_{\text{full}}$, measuring the fraction of the target object rendered invisible by occlusion. Based on S_{occ}, tasks are divided into three difficulty levels using per-suite quartiles: the bottom 25% as light, the middle 50% as medium, and the top 25% as heavy. See Appendix A for details of the split. This partition enables fine-grained analysis of how VLA performance changes as partial observability varies.

Benchmark Statistics. Table 1 reports the distribution of tasks across occlusion types and severity levels. In total, LIBERO-Occ comprises 2,000 tasks spanning different occlusion target types, and severity levels, providing a comprehensive testbed for evaluating VLA robustness under realistic partial observability.

3 Methodology

3.1 Problem Formulation

A standard VLA policy predicts robot actions directly from the current visual observation and language instruction. Given a primary view observation o_t, a task instruction l, and an action chunk a_t,

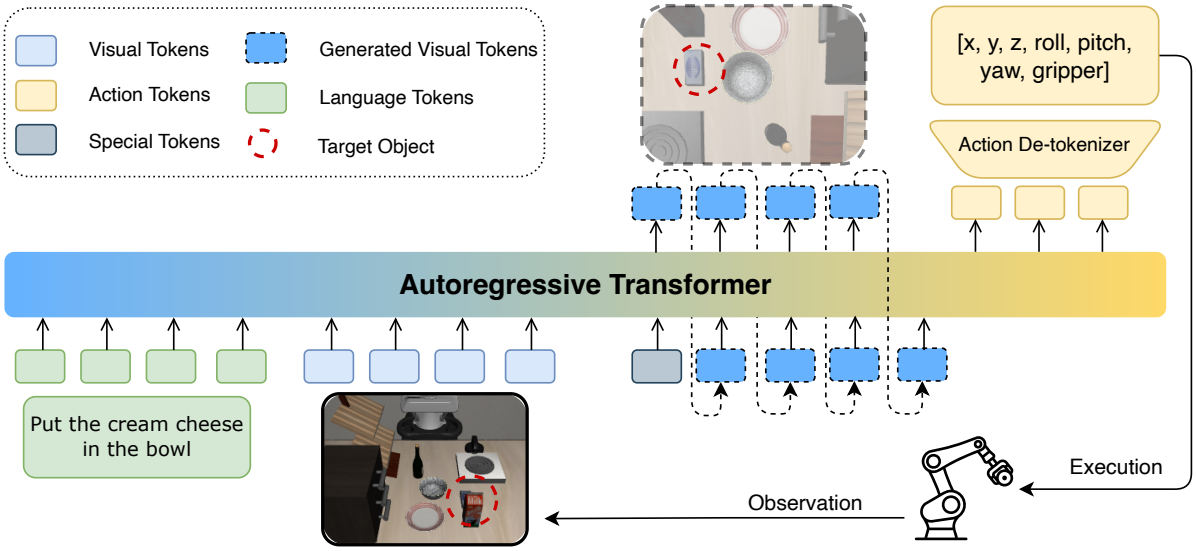


Figure 2: Viewpoint imagination framework. Given a primary observation and a language instruction, the model uses an autoregressive transformer to first generate visual tokens of a complementary viewpoint and then predict action tokens. The imagined view may recover task-relevant information that is not visible from the primary camera, such as the shape, location, or surrounding context of an occluded object.

a vanilla VLA model parameterized by θ models

$$p_\theta(a_t \mid o_t, l).$$

This formulation implicitly assumes that the observed image contains sufficient task-relevant spatial evidence for action prediction. However, under scene-induced occlusion, important entities may be partially hideen in o_t, making it difficult to infer well-grounded actions. We therefore introduce a viewpoint imagination formulation. The model first predicts a complementary view $\hat{o}_t^c$ from the occluded primary observation, and then conditions action prediction on both the observed and imagined views:

$$p_\theta(\hat{o}_t^c, a_t \mid o_t, l) = p_\theta(\hat{o}_t^c \mid o_t, l) \cdot p_\theta(a_t \mid o_t, \hat{o}_t^c, l).$$

Here, o_t is the available primary-view observation and $\hat{o}_t^c$ is a generated complementary view, such as a wrist or gripper view that may expose information occluded from the primary camera.

3.2 Viewpoint Imagination Framework

Figure 2 illustrates our method. To jointly model imagined views and actions, the framework uses a unified autoregressive model that is capable of understanding and generating both textual and visual tokens. It consists of two coupled processes: viewpoint imagination and action prediction. First, given the occluded primary observation o_t and the language instruction l, the model performs viewpoint imagination to generate the visual tokens of a complementary view, which serves as an intermediate representation of the potentially missing scene information. The model then predicts action tokens conditioned on both the original observation and the generated complementary-view tokens. In this way, the generated complementary view serves not merely as an auxiliary reconstruction target, but as an intermediate visual reasoning step that bridges perception under occlusion and policy generation.

3.3 Training and Inference

Training We train the model in two stages. In the first stage, we train the model for viewpoint imagination. The model is given the occluded front view and the language instruction, and is optimized to generate the corresponding complementary view. This stage equips the model with the ability to infer hidden information from the occluded observation by learning cross-view spatial correspondence.

In the second stage, we jointly optimize viewpoint imagination and action prediction. The model predicts action tokens while continuing to receive supervision on complementary-view generation. The training objective is

$$\mathcal{L} = \mathcal{L}_{action} + \lambda \mathcal{L}_{view}$$

where $\mathcal{L}_{view}$ is the token-level loss for complementary view generation and $\mathcal{L}_{\text{action}}$ is the loss for action prediction. We use group-wise loss weighting with λ to balance the visual generation objective and the action prediction objective, as the two to-

ken groups can differ substantially in length. This stage teaches the model to use the inferred evidence for policy decisions.

Inference At inference time, the robot receives only the current primary observation o_t and language instruction l. The model first generates an imagined complementary view $\hat{o}_t^c$. It then predicts the low-level action sequence a_t conditioned on both the primary observation and the generated view. Since the complementary view is generated internally, the method can be deployed without additional cameras or active viewpoint control. If a real complementary view o_t^c is available, it can directly replace the generated view. Therefore, our VIM framework supports both camera-free viewpoint imagination and camera-assisted deployment within a unified action-generation interface.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate models on both the original LIBERO benchmark and our proposed LIBERO-Occ benchmark. The original LIBERO benchmark provides a standard in-domain evaluation setting. LIBERO-Occ, in contrast, introduces scene-induced occlusions while preserving the original task semantics and executable trajectory family. This enables us to isolate the effect of occlusion and examine whether VLA models can act reliably with missing information. For the main evaluation, we report the average success rate over 500 rollouts for each task suite. For ablation studies, we use 100 rollouts per task suite to reduce evaluation cost.

Baselines. We compare VIM with representative state-of-the-art VLA models, including OpenVLA (Kim et al., 2024), OpenVLA-OFT (Kim et al., 2025), π-0 (Black et al., 2024), π-0.5 (Black et al., 2025), and UniVLA (Wang et al., 2025b). These baselines cover both autoregressive and diffusion-based VLA policies and have shown strong performance on standard manipulation benchmarks.

Implementation Details We implement our viewpoint imagination policy on top of the world model trained by UniVLA (Wang et al., 2025b), using an Emu3-MoE (Wang et al., 2024) autoregressive backbone for unified image-token and action-token generation. Robot actions are tokenized by the FAST (Pertsch et al., 2025) action tokenizer.

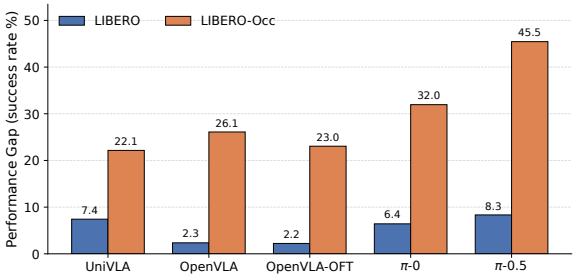


Figure 3: Performance gap of VLA models when the complementary view is unavailable. Larger drops indicate stronger dependence on additional visual evidence and weaker robustness under partial observability. Scene-induced occlusion substantially amplifies the benefit of complementary views, revealing that existing VLA models depend on additional visual evidence when task-relevant information is partially missing.

For fair comparison, all methods are finetuned on the same demonstration data, preprocessed following (Kim et al., 2025). Training is conducted on 8 H100 GPUs with a batch size of 192. We use a cosine learning-rate schedule with a peak learning rate of 8×10^{-5} for the first stage and 4×10^{-5} for the second stage. We use the standard third-person view as the primary observation, and the wrist/gripper-camera view as the complementary view. We set λ to 0.5 in stage two. More implementation details are provided in appendix.

Observation Settings. We evaluate two observation settings. In the *complementary-view available* setting, models receive both the primary view and the ground-truth complementary view, providing an upper-bound estimate when additional visual evidence is physically available. In the *complementary-view unavailable* setting, models receive only the primary view, corresponding to the standard fixed-camera deployment scenario where task-relevant information may be occluded.

4.2 Scene-Induced Occlusion Exposes Hidden Dependence on Complementary Views

We first examine whether current VLA models rely on complementary visual information to maintain robust manipulation performance under occlusion. We report the performance gap between the complementary-view available setting and unavailable setting. As shown in Figure 3, the gain from providing the true complementary view is relatively modest on the original LIBERO benchmark, ranging from 2.2 to 8.3 percentage points across models. In contrast, the same gap increases substantially on LIBERO-Occ, ranging from 22.1 to 45.5 points.

Method	Original LIBERO					LIBERO-Occ					Avg. Drop ↓
	Avg.	Spatial	Goal	Object	LIBERO-10	Avg.	Spatial	Goal	Object	LIBERO-10	
UniVLA	88.25	86.00	90.00	95.00	82.00	57.10	61.00	67.40	78.80	21.20	31.15
OpenVLA	92.65	94.00	95.20	96.40	85.00	40.65	39.20	37.40	65.60	20.40	52.00
OpenVLA-OFT	95.75	97.00	93.00	97.00	93.00	47.95	48.00	44.80	75.60	23.40	47.80
π-0	89.25	91.00	92.00	95.00	79.00	49.30	52.40	56.00	72.40	16.40	39.95
π-0.5	90.00	90.00	90.00	97.00	83.00	40.55	39.00	40.20	64.00	19.00	49.45
Ours	90.75	94.00	91.00	98.00	80.00	**65.05**	**76.40**	**74.60**	**84.20**	**25.00**	**25.70**
Ours w/ GT Comp.	93.00	93.00	92.00	98.00	89.00	74.00	83.00	81.00	92.00	40.00	19.00

Table 2: Success rates on the original LIBERO and LIBERO-Occ benchmarks. We report results on each suite and the average over suites. Avg. Drop denotes the decrease from the original LIBERO average to the LIBERO-Occ average. Bold numbers indicate the best results under the complementary-view unavailable setting. The gray row uses the ground-truth complementary view and is included only as a reference upper bound.

This sharp enlargement of the complementary-view gap indicates that scene-induced occlusion reveals a hidden dependence on viewpoints that expose task-relevant evidence. While standard LIBERO observations often contain sufficient information for action prediction, LIBERO-Occ turns manipulation into a partially observable problem, in which existing VLA policies still have limited ability to infer hidden scene states.

4.3 Main Results on LIBERO and LIBERO-Occ

Table 2 reports the results on both the original LIBERO and LIBERO-Occ benchmarks under complementary-view unavailable setting. On the original LIBERO benchmark, all methods perform strongly, and OpenVLA-OFT achieves the best average success rate. However, this advantage does not transfer to occluded scenes: OpenVLA-OFT drops to 47.95% on LIBERO-Occ, and other baselines exhibit similarly large declines. These results show that occlusion is a challenging and underexplored failure mode for current VLA systems. By contrast, our method achieves the highest LIBERO-Occ average success rate of 65.05%, outperforming the strongest baseline by 7.95 percentage points. The improvement is consistent across different task suites, indicating that viewpoint imagination benefits different types of manipulation tasks rather than only a narrow subset. This result also supports the intuition that generative visual modeling can provide useful intermediate representations for recovering task-relevant evidence under occlusion. Moreover, our method shows the smallest average drop from original LIBERO to LIBERO-Occ, suggesting that viewpoint imagination improves robustness to scene-induced occlusion without relying on additional cameras at deployment time. When the ground-truth complementary view is available, our model further improves the LIBERO-Occ average success rate from 65.05% to 74.00%, confirming that complementary visual evidence provides useful missing information. Qualitative examples in Appendix B illustrate how the imagined complementary view recovers task-relevant spatial evidence in the occluded observation.

Overall, the main results show that LIBERO-Occ exposes a substantial robustness gap in current VLA models, while the proposed viewpoint imagination framework effectively mitigates this gap without requiring additional hardware at deployment time.

4.4 Analysis across Occlusion Severity.

We further analyze model performance under different occlusion severity levels, as shown in Figure 4. Overall, stronger occlusion generally leads to larger performance degradation for existing VLA baselines. This confirms that LIBERO-Occ does not merely introduce superficial visual changes, but progressively hides task-relevant evidence needed for reliable action prediction. In the Object suite, most baselines drop sharply from light to heavy occlusion, indicating that directly hiding the object substantially weakens object recognition and localization. In the Goal suite, several baselines collapse under heavy occlusion, suggesting that inferring the destination region is particularly difficult when the target receptacle or placement area is no longer visible. In contrast, our method remains consistently competitive across severity levels and achieves the strongest overall performance in almost all settings. Notably, it preserves high success rates under heavy occlusions, where complementary spatial evidence is especially important. These results suggest that viewpoint imagination improves robustness under

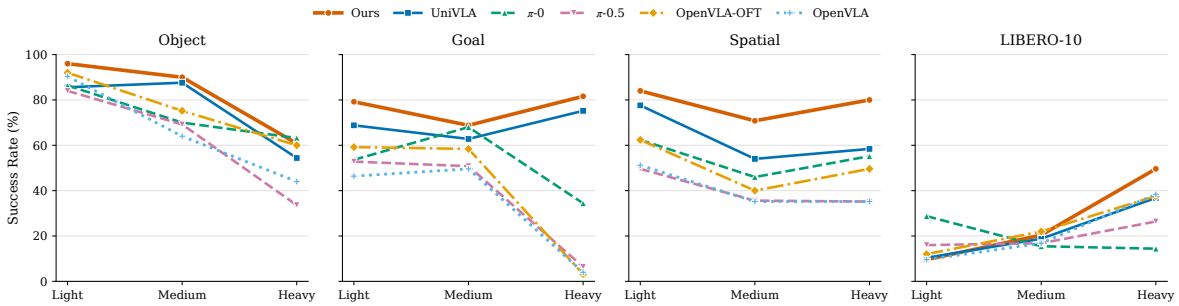


Figure 4: Success rates under different occlusion severity levels on LIBERO-Occ. We report results for Object, Goal, Spatial, and LIBERO-10 suites. As occlusion becomes more severe, existing VLA baselines generally suffer larger performance degradation, especially when task-relevant objects or goal regions are heavily hidden. Our method maintains stronger performance across severity levels, demonstrating improved robustness under increasing partial observability.

Method	Manipulated	Receptacle	Dual	Overall
UniVLA	47.78	81.87	28.00	57.10
OpenVLA	28.89	67.47	13.43	40.65
OpenVLA-OFT	35.22	77.87	16.57	47.95
π-0	40.78	70.80	25.14	49.30
π-0.5	29.78	67.87	9.71	40.55
Ours	**54.67**	**91.33**	**35.43**	**65.05**

Table 3: Success rates on LIBERO-Occ grouped by occlusion target type. Results are aggregated across all suites. The best result in each column is shown in bold.

different levels of partial observability, by helping the policy recover missing information.

4.5 Analysis by Occlusion Target Types.

We also analyze performance according to which task-relevant entity is occluded, as shown in Table 3. The results reveal clear differences across occlusion types. Most models achieve higher success rates under receptacle occlusion. In contrast, manipulated object occlusion causes a larger degradation. This trend is also consistent with the asymmetric difficulty of manipulation behaviors: grasping a partially occluded object requires precise recognition and localization of the manipulated object, whereas placing an object into an occluded receptacle can often tolerate larger spatial uncertainty. Dual occlusion is the most challenging setting, as both the object to manipulate and the target region may be partially missing; all baselines drop substantially in this case. Our method achieves the best performance across all three occlusion types. These consistent gains indicate that viewpoint imagination is not merely exploiting an easier occlusion subset, but helps recover missing visual evidence and infer spatial relations across different forms of partial observability.

Variant	Avg.	Spatial	Goal	Object	10
w/o S2 view loss	0.00	0.00	0.00	0.00	0.00
w/o S1 view train.	36.25	42.00	39.00	61.00	3.00
Full model	**65.00**	**78.00**	**73.00**	**84.00**	**25.00**

Table 4: Ablation study on the two-stage training strategy on LIBERO-Occ. S1: Stage-1 and S2: Stage-2.

4.6 Ablation Study

4.6.1 Ablation on Two-Stage Training

We first study the role of the proposed two-stage training strategy. We compare three variants: (1) *Ours w/o Stage-1 View Training*, which removes the viewpoint-imagination training stage and directly trains the model for action prediction with the Stage-2 objective; (2) *Ours w/o Stage-2 View Loss*, which keeps Stage-1 but removes the auxiliary viewpoint-imagination loss during Stage-2; and (3) the full model.

Table 4 shows that both stages are essential for effective viewpoint imagination and downstream manipulation. Removing Stage-1 view training causes a substantial performance drop. This indicates that directly learning action prediction without a dedicated viewpoint-imagination pretraining stage is insufficient for acquiring reliable cross-view spatial correspondence.

Removing the Stage-2 view loss causes complete failure across all suites. Inspection of the generated sequences shows that, the output tokens are interleaved with non-visual text and special tokens rather than forming a valid visual-token grid. Because action prediction is autoregressively conditioned on this intermediate visual segment, this format collapse prevents reliable action generation. Thus, the Stage-2 view loss acts not merely as an auxiliary reconstruction objective, but as a struc-

Method	Avg.	Spatial	Goal	Object	10
Ours → UniVLA	62.00	72.00	69.00	**84.00**	23.00
Ours	**65.00**	**78.00**	**73.00**	**84.00**	**25.00**

Table 5: Ablation on unified viewpoint imagination and action prediction. "Ours → UniVLA": the separated pipeline.

tural regularizer that maintains a valid imagination-to-action generation interface.

4.6.2 Ablation on Unified Viewpoint Imagination and Action Prediction

We further evaluate whether unifying viewpoint imagination and action prediction is beneficial. To this end, we compare our full model with a separated pipeline, where our model first generates the complementary-view image and UniVLA then uses it for action prediction.

As shown in Table 5, the separated pipeline achieves 62.00% average success rate on LIBERO-Occ, which is close to but weaker than our model. This indicates that the generated view contains meaningful task-relevant evidence that can also benefit an external VLA policy. At the same time, the remaining gap suggests that the unified autoregressive formulation helps the policy better exploit imagined visual tokens by optimizing viewpoint imagination and action prediction within the same sequence-generation process.

5 Related Work

Vision-Language-Action Models and Robust Manipulation Evaluation. Vision-Language-Action (VLA) models have shown strong potential for generalizable robotic manipulation by mapping visual observations and language instructions to robot actions (Zitkovich et al., 2023; Driess et al., 2023; Kim et al., 2025; Team et al., 2024; Black et al., 2025). They are commonly evaluated on benchmarks such as LIBERO (Liu et al., 2023), Bridge-V2 (Walke et al., 2023), and CALVIN (Mees et al., 2022), where task-relevant objects and goal regions are usually visible from the input observation. Recent robustness studies further examine visual or environmental perturbations, including background changes, lighting variations, and image noise (Fei et al., 2025; Zhou et al., 2025; Guo et al., 2025). However, these settings often preserve task-relevant visual evidence. In contrast, scene-induced occlusion hides such evidence and makes manipulation partially observable. Our work isolates occlusion as a physically grounded evaluation factor and introduces LIBERO-Occ to systematically measure VLA robustness under occlusion.

Perception Completion for Manipulation under Occlusion. Occlusion is often addressed by acquiring additional visual evidence through multi-view perception or active camera control. Multi-camera robot policies use complementary viewpoints to provide useful spatial cues for manipulation (Bian et al., 2025; Deng et al., 2025; Xue et al., 2025), while active perception selects viewpoints that reduce uncertainty or reveal occluded objects (Wang et al., 2025a; Dai et al., 2025). Recent VLA-oriented methods further incorporate active viewpoint selection or task-aware view exploration into manipulation policies (Liu et al., 2026b,a). Although effective, these approaches usually require additional cameras and calibration, increasing deployment cost and hardware complexity. In parallel, embodied world models and generative VLA models suggest that generative visual modeling can provide useful intermediate representations for control (Zhou et al., 2024; Cen et al., 2025; Wang et al., 2025b; Zhao et al., 2025). VIM builds on this insight but uses generation for perception completion: it imagines a complementary view from the occluded primary observation and uses it as intermediate evidence for action prediction.

6 Conclusion

In this work, we study scene-induced occlusion as a practical partial-observability challenge for vision-language-action models. We introduce LIBERO-Occ, an occlusion-oriented extension of LIBERO. Experiments show that strong VLA models suffer substantial degradation under occlusion and rely more heavily on complementary views when task-relevant evidence is missing.

To address this challenge, we propose VIM, a viewpoint imagination framework that generates complementary visual evidence from an occluded primary observation and uses it for action prediction. Results on LIBERO-Occ show that VIM consistently improves robustness across task suites, severity levels, and occlusion target types without requiring additional cameras at inference time. These findings suggest that viewpoint imagination is a promising mechanism for perception completion in partially observable manipulation.

7 Limitations

This work has several limitations. First, LIBERO-Occ is constructed in simulation based on LIBERO. Although each occlusion is physically instantiated and verified through demonstration replay, the benchmark cannot fully capture the sensing noise, object variability, and dynamics of real-world manipulation environments. Thus, our results should be interpreted as a controlled study of scene-induced occlusion rather than a substitute for real-robot evaluation.

Second, VIM generates a complementary view from the primary observation, which makes it deployable without additional cameras but also constrains it by the model's learned visual prior. When the primary view contains very limited evidence, the imagined view may be inaccurate, especially under severe dual occlusion. Future work could improve reliability by modeling uncertainty or integrating temporal evidence across multiple frames.

Finally, our experiments mainly use the wrist/gripper view as the complementary viewpoint and require paired complementary-view observations during training. Other viewpoints, such as side, bird, or task-adaptive views, may provide different benefits, and such paired data may not always be available. Learning viewpoint imagination from weaker supervision or less structured video data remains an important direction for future work.

References

Seunghyeok Back, Joosoon Lee, Taewon Kim, Sangjun Noh, Raeyoung Kang, Seongho Bak, and Kyoobin Lee. 2021. Unseen object amodal instance segmentation via hierarchical occlusion modeling. *2022 International Conference on Robotics and Automation (ICRA)*, pages 5085–5092.

Jinyue Bian, Zhaoxing Zhang, Zhengyu Liang, Shiwei Zheng, Shengtao Zhang, Rong Shen, Chenyu Yang, and Anzhou Hou. 2025. Vla-lpaf: Lightweight perspective-adaptive fusion for vision-language-action to enable more unconstrained robotic manipulation. *ArXiv*, abs/2509.18183.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, brian ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, and 16 others. 2025. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 17–40. PMLR.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, and 1 others. 2024. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. 2025. Worldvla: Towards autoregressive action world model. *Preprint*, arXiv:2506.21539.

Yixiang Dai, Siang Chen, Kaiqin Yang, Dingchang Hu, Pengwei Xie, Guosheng Li, Yuan Shen, and Guijin Wang. 2025. Active-perceptive language-oriented grasp policy for heavily cluttered scenes. *IEEE Robotics and Automation Letters*, 10:11094–11101.

Shengliang Deng, Mi Yan, Yixin Zheng, Jiayi Su, Wenhao Zhang, Xiaoguang Zhao, Heming Cui, Zhizheng Zhang, and He Wang. 2025. Stereovla: Enhancing vision-language-action models with stereo vision. *arXiv preprint arXiv:2512.21970*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, and 1 others. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Senyu Fei, Siyin Wang, Junhao Shi, Z. G. Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. 2025. Libero-plus: In-depth robustness analysis of vision-language-action models. *ArXiv*, abs/2510.13626.

Jiani Guo, Zhen Wu, Changhe Tu, Yiyao Ma, Xiangqi Kong, Zhiqian Liu, Jiaming Ji, Shuning Zhang, Yuanpei Chen, Kai Chen, Qi Dou, Yaodong Yang, Xianglong Liu, Huijie Zhao, Weifeng Lv, and Simin Li. 2025. On robustness of vision-language-action model against multi-modal perturbations. *ArXiv*, abs/2510.00037.

Pranav Guruprasad, Harshvardhan Digvijay Sikka, Jaewoo Song, Yangyue Wang, and Paul Pu Liang. 2024. Benchmarking vision, language, & action models on robotic learning tasks. *ArXiv*, abs/2411.05821.

Moo Jin Kim, Chelsea Finn, and Percy Liang. 2025. Fine-tuning vision-language-action models: Optimizing speed and success. *ArXiv*, abs/2502.19645.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, and 1 others. 2024. Openvla: An open-source vision-language-action model, 2024. *URL https://arxiv.org/abs/2406.09246*, 1(2):4.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qian Liu, Yuke Zhu, and Peter Stone. 2023. Libero: Benchmarking knowledge transfer for lifelong robot learning. *ArXiv*, abs/2306.03310.

Mengzhen Liu, Enshen Zhou, Cheng Chi, Yi Han, Shanyu Rong, Liming Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. 2026a. Sapave: Towards active perception and manipulation in vision-language-action models for robotics.

Zhenyang Liu, Yongchong Gu, Yikai Wang, Xiangyang Xue, and Yanwei Fu. 2026b. Activevla: Injecting active perception into vision-language-action models for precise 3d robotic manipulation. *ArXiv*, abs/2601.08325.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. 2022. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334.

Yinglong Miao, Rui Wang, and Kostas E. Bekris. 2022. Safe, occlusion-aware manipulation for online object reconstruction in confined spaces. In *International Symposium of Robotics Research*.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. 2025. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, and 1 others. 2024. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*.

Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, and 1 others. 2023. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR.

Guokang Wang, Hang Li, Shuyuan Zhang, Di Guo, Yanhong Liu, and Huaping Liu. 2025a. Observe then act: Asynchronous active vision-action model for robotic manipulation. *IEEE Robotics and Automation Letters*.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, and 6 others. 2024. Emu3: Next-token prediction is all you need. *Preprint*, arXiv:2409.18869.

Yu-Quan Wang, Xinghang Li, Wenxuan Wang, Junbo Zhang, Yingyan Li, Yuntao Chen, Xinlong Wang, and Zhaoxiang Zhang. 2025b. Unified vision-language-action model. *ArXiv*, abs/2506.19850.

Yuchen Xiao, Sammie Katt, Andreas ten Pas, Shengjian Chen, and Chris Amato. 2019. Online planning for target object search in clutter under partial observability. *2019 International Conference on Robotics and Automation (ICRA)*, pages 8241–8247.

Han Xue, Jieji Ren, Wendi Chen, Gu Zhang, Yuan Fang, Guoying Gu, Huazhe Xu, and Cewu Lu. 2025. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. *arXiv preprint arXiv:2503.02881*.

Kaidi Zhang, Heng Zhang, Zhengtong Xu, Zhiyuan Zhang, Md. Rakibul Islam Prince, Xiang Li, Xiaojing Han, Yuhao Zhou, Arash Ajoudani, and Yu She. 2026. Tacvla: Contact-aware tactile fusion for robust vision-language-action manipulation.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, and 1 others. 2025. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713.

Wenrui Zhao and Weidong Chen. 2021. Hierarchical pomdp planning for object manipulation in clutter. *Robotics Auton. Syst.*, 139:103736.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. 2024. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*.

Xueyang Zhou, Yangming Xu, Guiyao Tie, Yongchao Chen, Guowen Zhang, Duanfeng Chu, Pan Zhou, and Lichao Sun. 2025. Libero-pro: Towards robust and fair evaluation of vision-language-action models beyond memorization. *ArXiv*, abs/2510.03827.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, and 1 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.

A Details about LIBERO-Occ

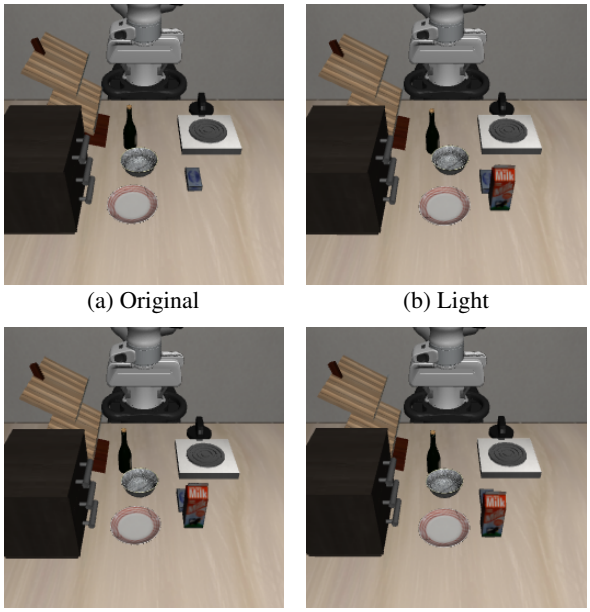
Following the original LIBERO protocol, each suite contains 10 task templates, and each task template is evaluated with 50 initial states that differ slightly in object configurations. In LIBERO-Occ, we instantiate occluders for these initial states. Because different occluder placements can induce different occlusion targets and severity levels, we treat each occluded initial state as a distinct occluded task instance. Thus, each LIBERO-Occ suite contains 500 occluded task instances, yielding 2,000 instances in total across the four suites.

A.1 Severity Distribution

To characterize the occlusion severity distribution in each LIBERO-Occ suite, we report the first and third quartiles of the occlusion severity in Table 6. The first quartile, Q_1, corresponds to the 25th percentile, meaning that 25% of episodes have an occlusion ratio no larger than this value. The third quartile, Q_3, corresponds to the 75th percentile, meaning that 75% of episodes have an occlusion ratio no larger than this value. These quartiles summarize the spread of occlusion severity across suites and define the central range of occlusion ratios used in the benchmark.

Table 6: Occlusion severity quartiles for each LIBERO-Occ suite. Severity is measured by the occlusion ratio in the manifest.

Suite	N	Q_1	Q_3
Spatial	500	0.5979	0.8874
Goal	500	0.3663	0.8239
Object	500	0.5771	0.8013
LIBERO-10	500	0.4278	0.8006



(a) Original (b) Light

(c) Medium (d) Heavy

Figure 5: Examples of different occlusion severity levels in LIBERO-Occ. The original scene is shown for reference. As severity increases, more target-related visual evidence becomes unavailable from the primary view, making the task increasingly partially observable.

A.2 Examples

To provide an intuitive visualization of the severity split, Figure 5 shows representative examples from the same task under different occlusion levels. Compared with the original scene, light occlusion only removes a small portion of the task-relevant visual evidence, while medium and heavy occlusions progressively hide larger parts of the target. These examples illustrate that the proposed severity metric reflects meaningful changes in partial observability rather than superficial appearance variations.

B Qualitative Examples

We provide qualitative examples in Figures 6 and 7. Each figure shows five uniformly sampled timesteps from the rollout. The first three columns show the rollouts of our model, with columns correspond to the primary view, the ground-truth complementary view, the imagined complementary view generated. The fourth column is the primary-view rollout of π-0.5 on the same task. These examples illustrate that the imagined view provides task-relevant spatial evidence that is weakly visible or hidden from the primary camera, while the baseline policy fails under the same scene-induced occlusion.

Setting	Value
Base Model	UniVLA World Model
Action tokenizer	FAST
Hardware	8 H100 GPUs
Global batch size	192
Precision	bfloat16 with TF32 enabled
Optimizer	AdamW
AdamW β_1, β_2	0.9, 0.95
Weight decay	0.1
LR schedule	Cosine decay
Stage-1 peak learning rate	8×10^{-5}
Stage-2 peak learning rate	4×10^{-5}
Stage-1 training steps	4,000
Stage-2 training steps	6,000
Stage-2 weight λ	0.5
Action chunk length	10
Random seed	42
Camera resolution	200

Table 7: Implementation details for VIM training.

C Implementation Details

We summarize the main implementation details in Table 7. The model uses the third-person camera

as the primary observation and the wrist/gripper camera as the complementary view.

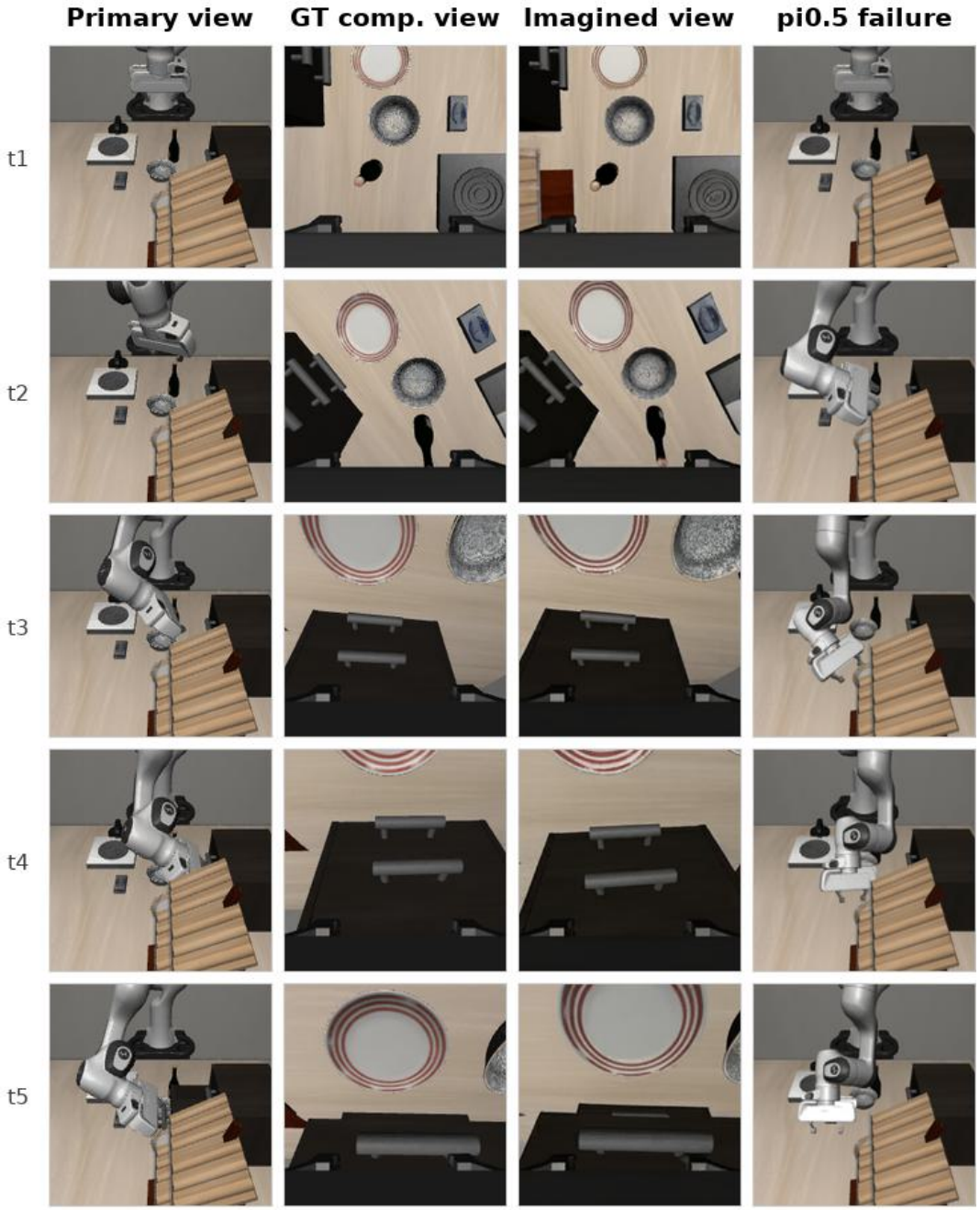


Figure 6: Qualitative example on LIBERO-Occ. Task instruction: *"open the middle drawer of the cabinet"*. The drawer is partially occluded by scene objects. The imagined complementary view closely matches the ground-truth complementary view and provides additional spatial evidence for action prediction. In contrast, π-0.5 fails under the same primary-view setting.

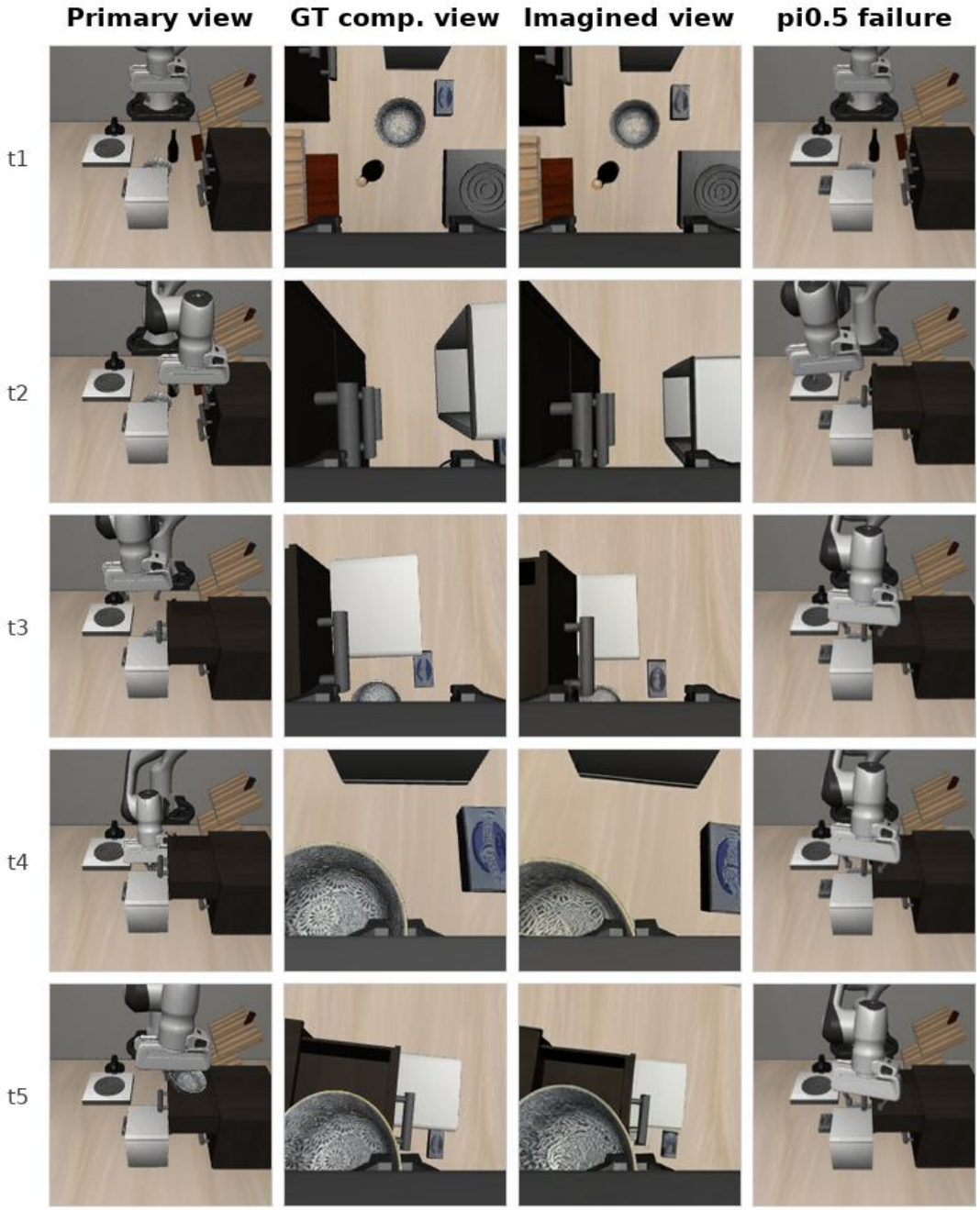


Figure 7: Qualitative example on LIBERO-Occ. Task instruction: *"open the top drawer and put the bowl inside"*. The manipulated object becomes difficult to locate from the primary camera as the drawer, occluder, and robot arm block it. Our imagined complementary view recovers a viewpoint that exposes the drawer and bowl more clearly, supporting successful manipulation. π-0.5 fails on the same task when only the primary view is available.