LIBERO-OCC: EVALUATING AND IMPROVING VISION-LANGUAGE-ACTION MODELS UNDER SCENE-INDUCED OCCLUSION VIA VIEWPOINT IMAGINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language-Action (VLA) models have achieved 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 a 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 & 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.

We refer to this problem as *scene-induced occlusion*, an intrinsic challenge arising from the configuration and dynamics of the manipulation scene itself. As shown in Figure 1, the target object may be 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, occlusion preserves scene texture while reducing the visibility of task-relevant

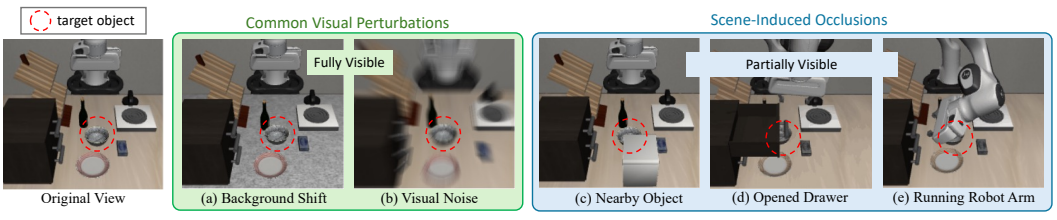


Figure 1: **Comparison with common perturbations.** Task instruction: *"open the top drawer and put the bowl inside"*. Unlike artificially crafted perturbations, scene-induced occlusions remove task-relevant evidence from the observation and increase task difficulty.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

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 and maintenance; and active camera systems introduce extra hardware constraints. This raises a central question: can a VLA model infer task-relevant visual evidence without adjusting hardware?

Motivated by recent progress in 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 evidence that is hidden in the primary camera. We therefore propose a viewpoint imagination-based framework for VLA-based manipulation under occlusion. Our framework is built upon a world model and aims to transform its generative prior into a perception-completion capability. 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 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 occlusions while preserving the original task semantics, and covers different occlusion types and severity levels to enable systematic evaluation under diverse occlusion conditions.

In summary, our contributions are threefold:

- We identify scene-induced occlusion as an underexplored challenge for VLA and introduce **LIBERO-Occ**, a benchmark for systematically evaluating VLA robustness under occlusion.
- We propose **VIM**, a viewpoint imagination-based VLA framework that infers complementary visual evidence without additional cameras.
- Extensive experiments show that existing VLAs suffer substantial degradation under scene-induced occlusion, while VIM consistently improves robustness across task suites, occlusion types, severity levels.

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 Long (LIBERO-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 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 task-relevant entity being hidden: the *manipulated object*, the *receptacle or goal region*, or *both* (dual occlusion). This categorization provides coverage of different sources of task-relevant visual information.

Occlusion Severity. Occlusion is not binary: as more task-relevant evidence becomes hidden, the model must increasingly infer the underlying scene state. We quantify severity using the visibility loss of the occlusion target. Let A_{full} and A_{visible} denote its segmentation areas in the reference and occluded observations, respectively. We define $S_{\text{occ}} = (A_{\text{full}} - A_{\text{visible}})/A_{\text{full}}$. Based on per-suite quartiles, we categorize 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. LIBERO-Occ contains 2,000 task instances spanning different occlusion target types and severity levels. Table 1 summarizes their distribution.

Table 1: **Task distribution.**

Target	Light	Med.	Heavy	Total
Manipulated	257	412	231	900
Receptacle	186	338	226	750
Dual	57	250	43	350
All	**500**	**1000**	**500**	**2000**

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, a vanilla VLA 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 hidden in o_t, making it difficult

[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.

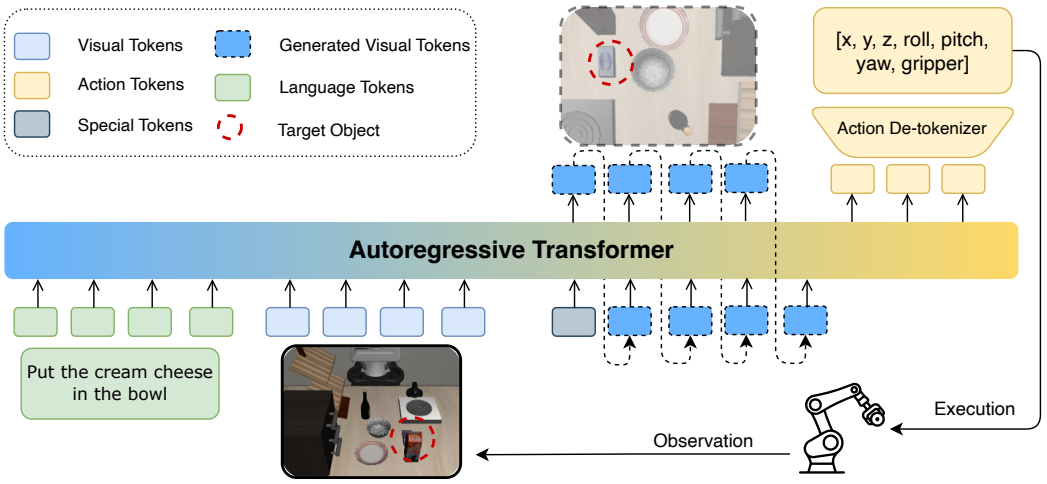


Figure 2: **Viewpoint imagination framework.** Given a primary observation and a language instruction, the model first generates visual tokens of a complementary viewpoint and then predict action tokens. The imagined view may recover task-relevant information that is not visible from the primary camera,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

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 The model are trained in two stages. In the first stage, we train the model for viewpoint imagination. The model receives an primary view and the language instruction and is optimized to generate the 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 continues to learn complementary-view generation while being trained to predict actions. During training, we use the ground-truth complementary view o_t^c as the intermediate visual context for action prediction. The overall objective is $\mathcal{L} = \mathcal{L}_{\text{action}} + \lambda \mathcal{L}_{\text{view}}$, where

$$\mathcal{L}_{\text{view}} = -\sum_{i=1}^{N_v} \log p_\theta(v_i \mid v_{<i}, o_t, l), \quad \mathcal{L}_{\text{action}} = -\sum_{j=1}^{N_a} \log p_\theta(a_j \mid a_{<j}, o_t, o_t^c, l).$$

Here, $\mathcal{L}_{\text{view}}$ supervises complementary-view generation, while $\mathcal{L}_{\text{action}}$ trains the policy to use complementary visual evidence for action prediction. We use group-wise loss weighting with λ to balance the two objectives, since the visual and action token sequences can differ substantially in length.

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

We evaluate models on both the original LIBERO benchmark and our proposed LIBERO-Occ benchmark. For the main evaluation, we report the average success rate over 500 rollouts for each task suite. For further analysis, we use 100 random 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$_{\text{OFT}}$ (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. Among them, **UniVLA is the backbone-matched control for VIM**: both are constructed from the same UniVLA world-model backbone, while UniVLA performs action prediction without the proposed viewpoint-imagination process.

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. For fair comparison, all action policies are trained on the same demonstrations replayed in the *original* LIBERO environments, following the preprocessing protocol of OpenVLA$_{\text{OFT}}$ (Kim et al., 2025). The occluded observations used for Stage 1 are generated with the same occlusion-generation procedure as LIBERO-Occ, but are instantiated from training demonstrations rather than evaluation initial states to ensure no overlap between them. Full training-data statistics are provided in Appendix D.1. We use the standard third-person view as the primary observation, and the wrist camera view as the complementary view. More implementation details are provided in appendix C.

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. 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 Main Results

4.2.1 Scene-Induced Occlusion Reveals a Complementary-View Gap

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.

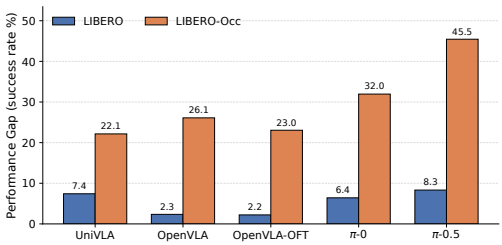


Figure 3: **Complementary-View Gap.**

Table 2: **Success rates on the original LIBERO and LIBERO-Occ benchmarks.** Avg. Drop denotes the decrease from the original LIBERO average to the LIBERO-Occ average. The gray row uses the ground-truth complementary view and is included only as a reference upper bound.

Method	Original LIBERO					LIBERO-Occ					Avg. Drop ↓
	Avg.	Spatial	Goal	Object	Long	Avg.	Spatial	Goal	Object	Long	
UniVLA	88.3	86.0	90.0	95.0	82.0	57.1	61.0	67.4	78.8	21.2	31.2
OpenVLA	92.7	94.0	95.2	96.4	85.0	40.7	39.2	37.4	65.6	20.4	52.0
OpenVLA$_{\text{OFT}}$	95.8	97.0	93.0	97.0	93.0	48.0	48.0	44.8	75.6	23.4	47.8
π-0	89.3	91.0	92.0	95.0	79.0	49.3	52.4	56.0	72.4	16.4	40.0
π-0.5	90.0	90.0	90.0	97.0	83.0	40.6	39.0	40.2	64.0	19.0	49.5
VIM	90.8	94.0	91.0	98.0	80.0	**65.1**	**76.4**	**74.6**	**84.2**	**25.0**	**25.7**
VIM w/ Comp.	93.0	93.0	92.0	98.0	89.0	74.0	83.0	81.0	92.0	40.0	19.0

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. This sharp enlargement of the complementary-view gap indicates that **scene-induced occlusion reveals a hidden dependence on complementary viewpoints**. 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.

4.2.2 Main Results on LIBERO and LIBERO-Occ

Table 2 compares performance on LIBERO and LIBERO-Occ under the complementary-view unavailable setting. While all methods achieve strong performance on the original LIBERO benchmark, most suffer substantial degradation under scene-induced occlusion. In contrast, VIM maintains competitive performance on LIBERO while achieving the highest success rate of 65.1% on LIBERO-Occ, outperforming the backbone-matched UniVLA by 8.0 points with consistent gains across all task suites. Moreover, VIM exhibits the smallest performance drop from LIBERO to LIBERO-Occ. **These results show that VIM preserves strong performance in standard settings while substantially improving robustness under scene-induced occlusion.**

When the ground-truth complementary view is available, VIM further improves its LIBERO-Occ success rate from 65.1% to 74.0%, an 8.9-point gain. **This result not only confirms the value of complementary visual evidence, but also shows that VIM can effectively incorporate either imagined or real complementary views within the same action-generation framework.** Such flexibility allows the proposed method to support both camera-free deployment and scenarios where additional camera observations are available. Qualitative examples in Appendix B further illustrate how complementary views provide task-relevant spatial information that is missing from the primary observation.

4.2.3 Robustness across Occlusion Conditions.

Table 3: **Success rates on LIBERO-Occ grouped by occlusion target type.**

Method	Manip.	Recep.	Dual	Avg.
UniVLA	47.8	81.9	28.0	57.1
OpenVLA	28.9	67.5	13.4	40.7
OpenVLA$_{\text{OFT}}$	35.2	77.9	16.6	48.0
π-0	40.8	70.8	25.1	49.3
π-0.5	29.8	67.9	9.7	40.6
VIM	**54.7**	**91.3**	**35.4**	**65.1**

Analysis by occlusion target type. As shown in Table 3, performance varies substantially with the occluded target. Dual occlusion is generally the most challenging, as both the manipulated object and the receptacle are partially hidden. Manipulated-object occlusion is more difficult than receptacle occlusion, reflecting the greater precision required to localize and grasp a manipulation target, while placement can tolerate greater uncertainty in the receptacle location. VIM consistently achieves the best performance across all three occlusion types, indicating that viewpoint imagination improves robustness across different forms of task-relevant occlusion.

Analysis by occlusion severity. Figure 4 analyzes model performance across different occlusion severity levels. Stronger occlusion generally leads to lower success rates, although the trend is not

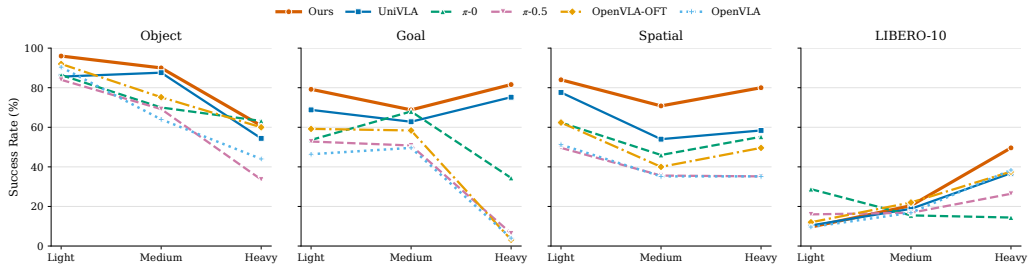


Figure 4: **Success rates under different occlusion severity levels on LIBERO-Occ.**

strictly monotonic across all settings. This is because our severity score measures only the fraction of the target that is visually occluded, whereas overall task difficulty also depends on factors such as instruction complexity and scene configuration. Consequently, a higher occlusion severity does not necessarily imply a more difficult manipulation task. Despite this variation, VIM maintains consistently strong performance across severity levels. **These results suggest that VIM is less sensitive to the loss of primary-view evidence, as viewpoint imagination provides complementary visual information when the primary observation becomes incomplete.**

4.2.4 Generalization Beyond LIBERO-Occ

We evaluate whether the robustness gains of VIM generalize beyond the standard LIBERO-Occ. We consider two settings: (1) an OOD occlusion split containing occluders and occluder–target combinations that are absent from the training data, and (2) LIBERO-Plus (Fei et al., 2025), an external benchmark for evaluating robustness under common perturbations like in Fig.1. Details of the OOD split construction are provided in Appendix E.

Results are summarized in Table 4. On the OOD occlusion split, VIM outperforms the strongest baseline by 10.2 points, indicating that it generalizes to novel occluder–target configurations. We observe a similar trend on LIBERO-Plus, where VIM achieves the best performance across all task suites and improves the average success rate by 6.1 points over the strongest baseline, π-0.5, and by 7.7 points over the backbone-matched UniVLA. Together, these results suggest that the robustness benefits of viewpoint imagination extend beyond LIBERO-Occ to unseen occlusion configurations and broader visual perturbations.

Table 4: **Generalization evaluation.**

Method	Avg.	Spa.	Goal	Obj.	Long
OOD Occlusion					
UniVLA	59.1	60.4	65.4	74.6	36.1
$\text{OpenVLA}_{\text{OFT}}$	58.3	67.4	57.5	68.9	39.4
π-0.5	52.6	45.8	48.6	74.6	41.1
VIM	**69.3**	**79.9**	**72.1**	**78.5**	**46.7**
LIBERO-Plus					
UniVLA	48.0	51.6	45.0	50.6	44.8
$\text{OpenVLA}_{\text{OFT}}$	48.5	52.4	43.4	56.0	42.2
π-0.5	49.6	44.2	52.4	51.8	49.8
VIM	**55.7**	**56.8**	**54.2**	**61.2**	**50.6**

Table 5: **Transferability and modularity.**

View	Policy	Avg.	Spa.	Goal	Obj.	Long
Transfer to External Policies						
–	$\text{OpenVLA}_{\text{OFT}}$	45.8	47.0	**39.0**	74.0	23.0
VIM	$\text{OpenVLA}_{\text{OFT}}$	**52.3**	**57.0**	36.0	**92.0**	**24.0**
–	UniVLA	55.3	61.0	59.0	82.0	19.0
VIM	UniVLA	**62.0**	**72.0**	**69.0**	**84.0**	**23.0**
External View Generator						
Qwen	VIM	53.8	61.0	63.0	72.0	19.0
VIM	VIM	**65.0**	**78.0**	**73.0**	**84.0**	**25.0**
Qwen	π-0.5	46.6	56.0	54.0	**59.4**	17.1
VIM	π-0.5	**54.7**	**67.0**	**73.0**	58.0	**20.8**

4.3 Further Analysis

4.3.1 Generated-View Fidelity and Hallucination Analysis

We evaluate the fidelity of imagined complementary views using LPIPS and SSIM, together with a task-oriented human evaluation. The human study assesses whether generated views preserve manipulation-relevant visual information. Metric definitions and human evaluation criteria are provided in Appendix D.2.

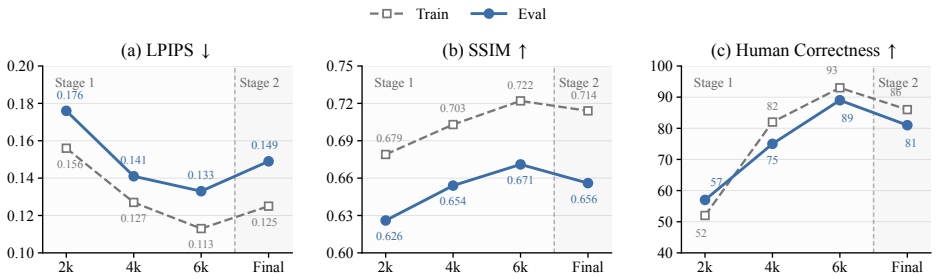


Figure 5: **Generated-view fidelity across training.** Train serves as an in-distribution reference, while Eval measures generalization to held-out observations. View quality improves steadily during Stage 1 and remains high after joint action–view optimization in Stage 2.

Training-set results serve as an in-distribution reference, while evaluation-set results measure generalization to held-out observations. As shown in Figure 5, view quality improves steadily during Stage 1, with the evaluation set closely tracking the training trend. The consistent trends across metrics indicate improved perceptual and task-relevant fidelity. After joint optimization with action prediction in Stage 2, reconstruction fidelity decreases slightly but remains high, suggesting that two-stage training does not simply maximize image reconstruction fidelity. Overall, the results suggest that VIM generates visually coherent complementary views that preserve task-relevant scene information.

4.3.2 VIM AS A STANDALONE VIEW GENERATOR

We investigate whether viewpoint imagination can serve as a standalone view generator for other VLA models. Specifically, we feed VIM-generated complementary views to external VLA policies, evaluating whether the generated views provide useful visual evidence independent of our action model. We further compare VIM with an external generator by fine-tuning Qwen-Image-Edit-2511 (Wu et al., 2025) to generate wrist-view images.

Results are shown in Table 5. View denotes the complementary-view source; "–" denotes primary-view-only input. VIM-generated views improve OpenVLA$_{\text{OFT}}$ and UniVLA by 6.5 and 6.7 points on average, respectively, showing that they provide useful task-relevant evidence beyond the VIM policy itself. Replacing VIM's internal imagination with Qwen-generated views make the performance decrease 11.2 points. This suggests that effective viewpoint imagination requires not only generating plausible complementary views, but also aligning the imagined evidence with downstream action prediction. Overall, VIM-generated views are transferable across policies, while joint viewpoint imagination and action learning yields the strongest performance.

4.3.3 ABLATION STUDY ON TWO-STAGE TRAINING

We 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 6: **Result of ablation on the two-stage training strategy.** S1: Stage-1 and S2: Stage-2.

Variant	Avg.	Spatial	Goal	Object	Long
w/o S2 view loss	0.00	0.00	0.00	0.00	0.00
w/o S1 view train.	36.3	42.0	39.0	61.0	3.00
Full model	**65.0**	**78.0**	**73.0**	**84.0**	**25.0**

Table 6 shows that both stages are essential for effective viewpoint imagination and manipulation. Removing Stage-1 training causes a substantial performance drop. This indicates that directly learning action prediction without a 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 is interleaved with non-visual tokens rather than forming a valid visual-token grid. This collapse prevents reliable action generation. Thus, the Stage-2 view loss acts not merely as an auxiliary reconstruction objective, but as a structural regularizer that maintains a valid imagination-to-action generation interface.

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 are usually visible from the observations. Recent robustness studies further examine several 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 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. To address this challenge, we propose VIM, a viewpoint imagination framework that generates complementary visual evidence from a primary observation for action prediction. Experiment results show that VIM 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 under occlusion.

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 relies on the model's visual prior to infer a complementary view, which may be inaccurate when the primary observation provides insufficient evidence, particularly under severe dual occlusion. Moreover, generating complementary-view tokens before action prediction introduces additional inference latency and computation. Future work could improve both reliability and efficiency through uncertainty-aware or more lightweight imagination mechanisms.

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.

AI USE DISCLOSURE

We used generative AI tools, to assist with manuscript writing and editing, improve clarity and organization, provide feedback on the presentation of experiments and results, and assist with LaTeX formatting and occasional code debugging. The research problem, benchmark, proposed method, experimental design, implementation, experiments, and scientific conclusions were developed and validated by the authors. All AI-assisted content and technical suggestions were reviewed and revised by the authors, and experimental results and technical descriptions were verified against the actual implementation and evaluation outputs. The authors take full responsibility for the final content of this work.

8 ETHICS STATEMENT

This work studies the robustness of vision-language-action models under scene-induced occlusion and is evaluated primarily in simulated robotic manipulation environments. LIBERO-Occ is constructed from existing benchmark tasks and does not involve the collection of personal or sensitive user data. The human evaluation in this work is limited to assessing the visual correctness of generated complementary views.

9 REPRODUCIBILITY STATEMENT.

We provide detailed information for reproducibility in the main paper and appendix, including the construction of LIBERO-Occ, the VIM architecture and training procedure, baseline configurations, implementation details, training-data statistics, and evaluation protocols. We also document the construction of the OOD evaluation splits. Upon acceptance, we will publicly release the LIBERO-Occ benchmark together with the training and evaluation code required to reproduce our results.

REFERENCES

Seunghyeok Back, Joosoon Lee, Taewon Kim, Sangjun Noh, Raeyoung Kang, Seongho Bak, and Kyoobin Lee. Unseen object amodal instance segmentation via hierarchical occlusion modeling. *2022 International Conference on Robotics and Automation (ICRA)*, pp. 5085–5092, 2021. URL https://api.semanticscholar.org/CorpusID:237605342.

Jinyue Bian, Zhaoxing Zhang, Zhengyu Liang, Shiwei Zheng, Shengtao Zhang, Rong Shen, Chenyu Yang, and Anzhou Hou. Vla-lpaf: Lightweight perspective-adaptive fusion for vision-language-action to enable more unconstrained robotic manipulation. *ArXiv*, abs/2509.18183, 2025. URL https://api.semanticscholar.org/CorpusID:281496707.

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

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, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In Joseph Lim, Shuran Song, and Hae-Won Park (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 27–30 Sep 2025. URL https://proceedings.mlr.press/v305/black25a.html.

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. Worldvla: Towards autoregressive action world model, 2025. URL https://arxiv.org/abs/2506.21539.

Yixiang Dai, Siang Chen, Kaiqin Yang, Dingchang Hu, Pengwei Xie, Guosheng Li, Yuan Shen, and Guijin Wang. Active-perceptive language-oriented grasp policy for heavily cluttered scenes. *IEEE Robotics and Automation Letters*, 10:11094–11101, 2025. URL https://api.semanticscholar.org/CorpusID:281045887.

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

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

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. Libero-plus: In-depth robustness analysis of vision-language-action models. *ArXiv*, abs/2510.13626, 2025. URL https://api.semanticscholar.org/CorpusID:282102298.

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. On robustness of vision-language-action model against multi-modal perturbations. *ArXiv*, abs/2510.00037, 2025. URL https://api.semanticscholar.org/CorpusID:281706065.

Pranav Guruprasad, Harshvardhan Digvijay Sikka, Jaewoo Song, Yangyue Wang, and Paul Pu Liang. Benchmarking vision, language, & action models on robotic learning tasks. *ArXiv*, abs/2411.05821, 2024. URL https://api.semanticscholar.org/CorpusID:273962883.

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

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *ArXiv*, abs/2502.19645, 2025. URL https://api.semanticscholar.org/CorpusID:276647709.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qian Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *ArXiv*, abs/2306.03310, 2023. URL https://api.semanticscholar.org/CorpusID:259089508.

Mengzhen Liu, Enshen Zhou, Cheng Chi, Yi Han, Shanyu Rong, Liming Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Sapave: Towards active perception and manipulation in vision-language-action models for robotics. 2026a. URL https://api.semanticscholar.org/CorpusID:286493172.

Zhenyang Liu, Yongchong Gu, Yikai Wang, Xiangyang Xue, and Yanwei Fu. Activevla: Injecting active perception into vision-language-action models for precise 3d robotic manipulation. *ArXiv*, abs/2601.08325, 2026b. URL https://api.semanticscholar.org/CorpusID:284704513.

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

Yinglong Miao, Rui Wang, and Kostas E. Bekris. Safe, occlusion-aware manipulation for online object reconstruction in confined spaces. In *International Symposium of Robotics Research*, 2022. URL https://api.semanticscholar.org/CorpusID:249017771.

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

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

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, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023.

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

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, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024. URL https://arxiv.org/abs/2409.18869.

Yu-Quan Wang, Xinghang Li, Wenxuan Wang, Junbo Zhang, Yingyan Li, Yuntao Chen, Xinlong Wang, and Zhaoxiang Zhang. Unified vision-language-action model. *ArXiv*, abs/2506.19850, 2025b. URL https://api.semanticscholar.org/CorpusID:279999860.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

Yuchen Xiao, Sammie Katt, Andreas ten Pas, Shengjian Chen, and Chris Amato. Online planning for target object search in clutter under partial observability. *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8241–8247, 2019. URL https://api.semanticscholar.org/CorpusID:146805205.

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

Kaidi Zhang, Heng Zhang, Zhengtong Xu, Zhiyuan Zhang, Md. Rakibul Islam Prince, Xiang Li, Xiaojing Han, Yuhao Zhou, Arash Ajoudani, and Yu She. Tacvla: Contact-aware tactile fusion for robust vision-language-action manipulation. 2026. URL https://api.semanticscholar.org/CorpusID:286517672.

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

Wenrui Zhao and Weidong Chen. Hierarchical pomdp planning for object manipulation in clutter. *Robotics Auton. Syst.*, 139:103736, 2021. URL https://api.semanticscholar.org/CorpusID:232369667.

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

Xueyang Zhou, Yangming Xu, Guiyao Tie, Yongchao Chen, Guowen Zhang, Duanfeng Chu, Pan Zhou, and Lichao Sun. Libero-pro: Towards robust and fair evaluation of vision-language-action models beyond memorization. *ArXiv*, abs/2510.03827, 2025. URL https://api.semanticscholar.org/CorpusID:281843691.

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

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 7. The occlusion severity of dual occlusion is the union of the manipulated-object and receptacle occlusion ratios. 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 7: 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
Long	500	0.4278	0.8006

A.2 Examples

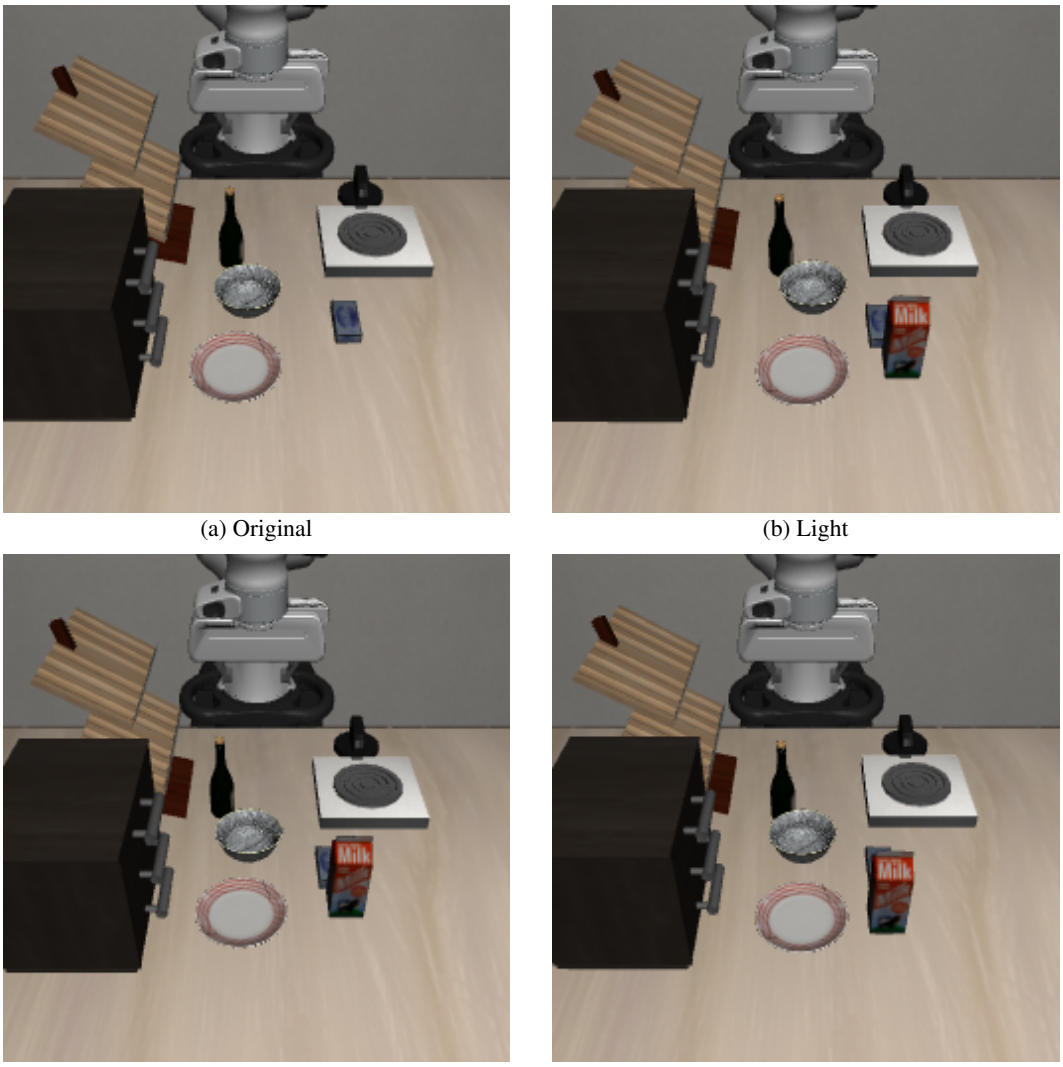
To provide an intuitive visualization of the severity split, Figure 6 shows representative examples from the same task under different occlusion levels. Compared with the original scene, light occlusion only removes a small portion of the task-relevant visual evidence, while medium and heavy occlusions 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 7 and 8. Each figure shows five uniformly sampled timesteps from the rollout. The first three columns show VIM rollouts and correspond to the primary view, ground-truth complementary view, and imagined complementary view, respectively. 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.

C Implementation Details

We summarize the main implementation details in Table 8. The model uses the third-person camera as the primary observation and the wrist/gripper camera as the complementary view.



(a) Original (b) Light

(c) Medium (d) Heavy

Figure 6: 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.

D Training and Evaluation Protocol

D.1 Details of the training and evaluation data

We provide additional details on the training data, supervision signals, and separation between training and evaluation instances. Table 9 summarizes the data used by each stage of VIM and by the action-policy baselines.

Action supervision. We replay the original LIBERO demonstrations and retain 1,722 successful episodes for action-policy training. VIM and all baselines use exactly these episodes for action supervision. No method is trained with actions from LIBERO-Occ or other occluded evaluation scenes.

Viewpoint-imagination supervision. Stage 1 uses 3,539 replay episodes collected from original and automatically generated occluded LIBERO environments. Both successful and failed rollouts

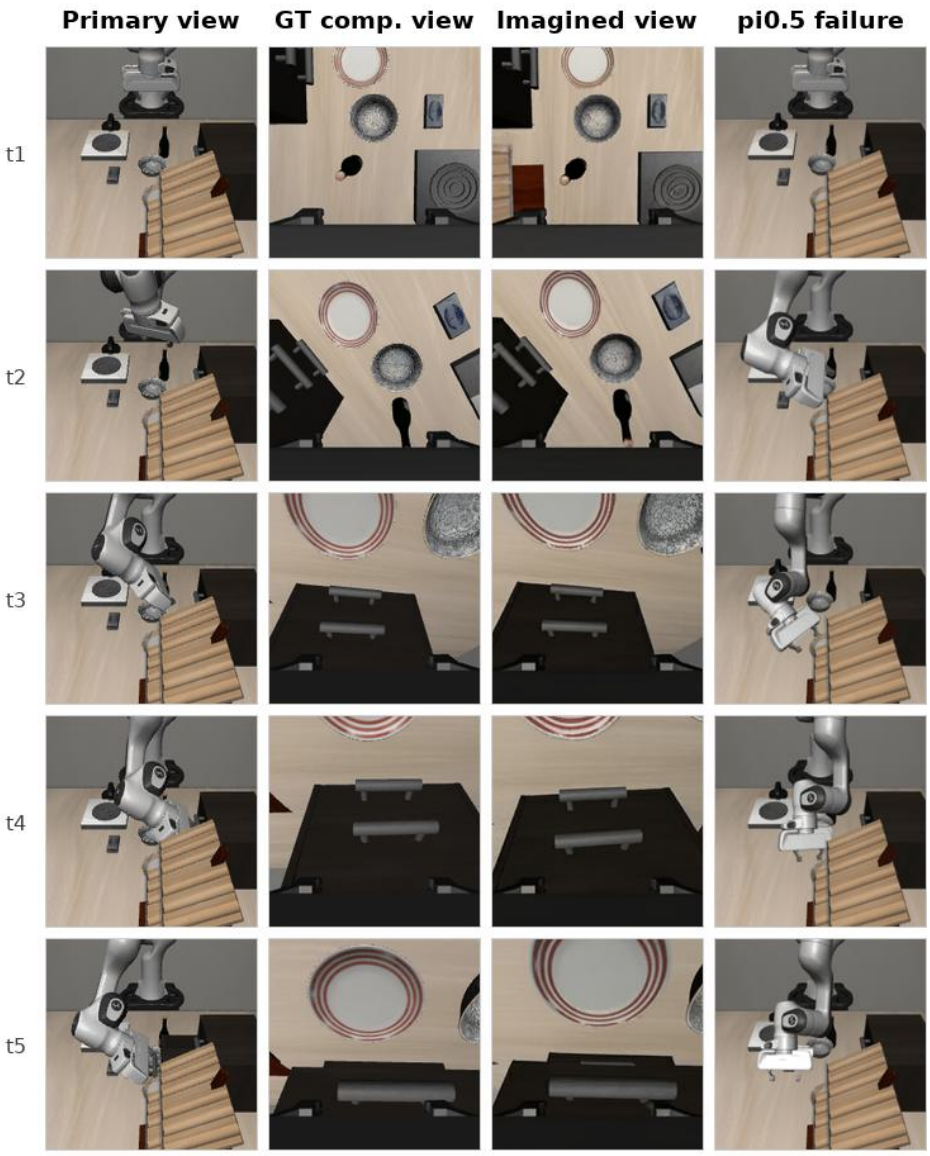


Figure 7: Qualitative example on LIBERO-Occ. Task instruction: *"open the middle drawer of the cabinet"*. The drawer is partially occluded by scene objects. The imagined complementary view closely matches the ground-truth complementary view and provides additional spatial evidence for action prediction. In contrast, π-0.5 fails under the same primary-view setting.

are included because this stage learns primary-to-complementary view prediction rather than action imitation. These data provide additional visual supervision to VIM but no additional action labels.

Backbone-matched control. VIM and UniVLA use the same UniVLA world-model backbone and the same action-supervised demonstrations. UniVLA predicts actions directly from the primary observation, while VIM additionally learns and uses complementary-view imagination.

Train–evaluation separation. Stage 1 and LIBERO-Occ use the same general occlusion-generation pipeline but different instantiated scenes. Following the original LIBERO, evaluation uses initial states distinct from the demonstration data, resulting in different object configurations, occluder placements, trajectories, primary-view sequences, and complementary-view sequences.

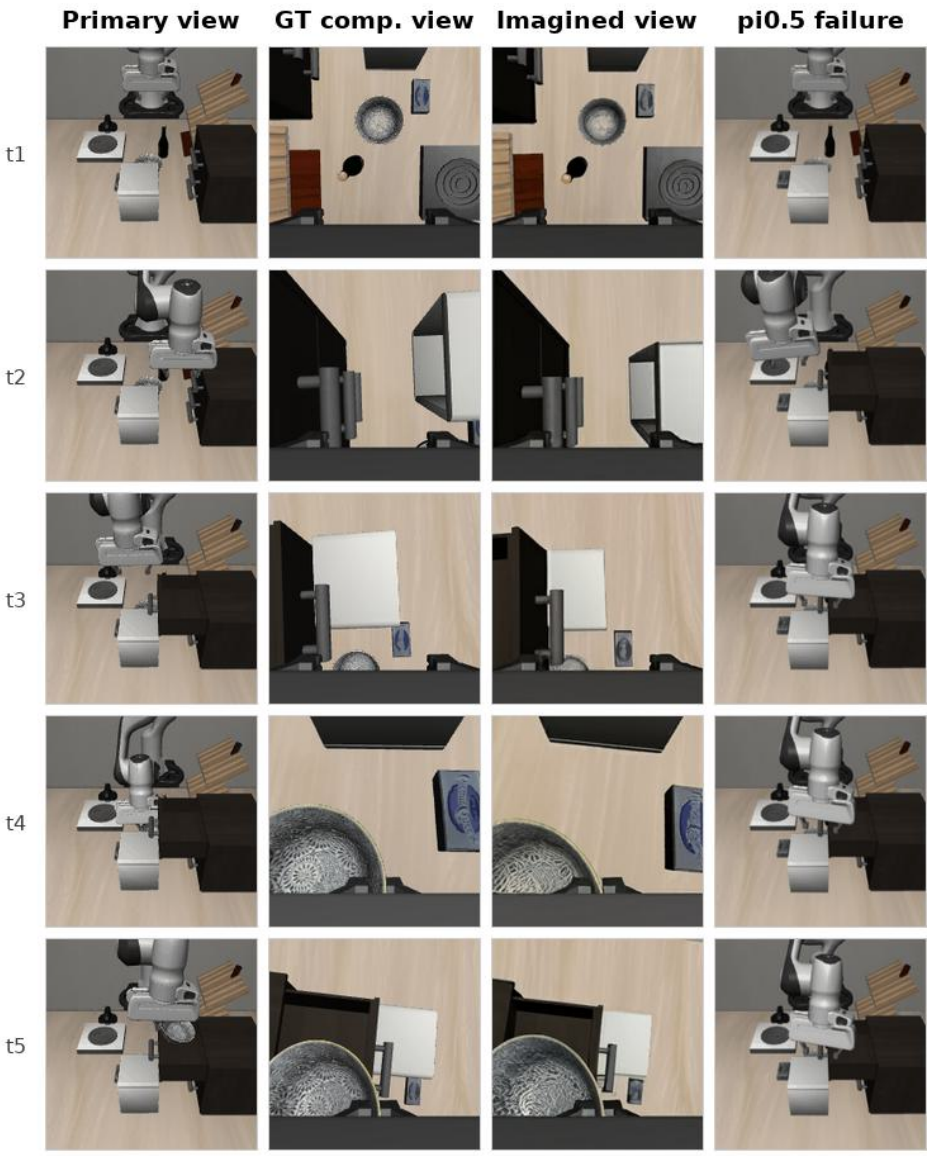


Figure 8: Qualitative example on LIBERO-Occ. Task instruction: *"open the top drawer and put the bowl inside"*. The manipulated object becomes difficult to locate from the primary camera as the drawer, occluder, and robot arm block it. Our imagined complementary view recovers a viewpoint that exposes the drawer and bowl more clearly, supporting successful manipulation. π-0.5 fails on the same task when only the primary view is available.

Thus, there is no instance-level overlap between Stage 1 training data and LIBERO-Occ evaluation instances.

D.2 Generated-View Evaluation Protocol

We evaluate generated complementary views using both automatic metrics and task-oriented human evaluation. For the automatic evaluation, each generated view is compared with the corresponding ground-truth complementary-view observation. We report LPIPS and SSIM. LPIPS measures perceptual similarity between images, with lower values indicating greater perceptual fidelity, while SSIM measures structural similarity, with higher values indicating better preservation of image structure.

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 8: Implementation details for VIM training.

Data	Environment	Episodes	Rollouts	View supervision	Action supervision	Used by
Stage 1	Original + occluded LIBERO	3,539	Success + failure	Primary $\rightarrow$ complementary	No	VIM
Stage 2	Original LIBERO	1,722	Success only	Complementary-view loss	Yes	VIM
Baseline fine-tuning	Original LIBERO	1,722	Success only	No	Yes	All baselines

Table 9: **Training-data composition and supervision signals.** Stage 1 uses paired primary- and complementary-view observations for cross-view generation and does not use action supervision. Stage 2 and all action-policy baselines use the same 1,722 successful demonstrations from the original LIBERO environments for action supervision. No method is trained with action labels from LIBERO-Occ.

We additionally conduct a task-oriented human evaluation on 100 randomly sampled episodes, since automatic image-similarity metrics may not fully capture errors that affect manipulation. A generated view is considered correct only if it can be successfully decoded, preserves the shapes and spatial positions of task-relevant objects, and correctly represents the robot arm and gripper state. The reported human score is the percentage of generated views satisfying all of these criteria.

E Details of the OOD Occlusion Split

To evaluate generalization to unseen occlusion configurations, we construct an OOD split based on *unseen occluder–target combinations*. Specifically, we retain only occluder–target pairings that do not appear in the viewpoint-imagination training data. The task generation procedure otherwise follows the same pipeline as LIBERO-Occ. Each candidate instance is filtered to ensure that the inserted occluder induces meaningful target occlusion while avoiding physical collisions, obstruction of task execution, or ambiguities that alter the original task semantics. These constraints substantially reduce the number of valid OOD instances that can be generated.

The OOD split also exhibits slightly lower occlusion severity than the standard LIBERO-Occ benchmark. Some occluders available for unseen combinations are relatively short and have large footprints, making it difficult to produce severe occlusion without colliding with task-relevant objects or interfering with task execution. As a result, the average occlusion severity of the OOD split is 0.57, compared with 0.66 for LIBERO-Occ. This difference partly explains why both VIM and the baseline models achieve higher absolute success rates on the OOD split than on the standard benchmark. Therefore, the OOD experiment is intended primarily to evaluate generalization to unseen occluder–target combinations rather than increased occlusion difficulty. Importantly, VIM maintains a clear

advantage over the baselines under these unseen configurations, indicating that its robustness gains are not limited to occlusion combinations observed during training.