ASK THE WORLD: CORRECTION-GUIDED ACTIVE PERCEPTION FOR VLAS CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language-action (VLA) policies can fail when the spatial relation needed for an action decision is not visible from the current observations. Inference-time correction offers a way to improve frozen policies, but a proposed correction can itself be uncertain: the available evidence may be insufficient to determine whether applying it would actually help. We introduce **Ask the World**, a correction-guided active perception framework that uses a tentative correction to determine what the robot should observe before intervening. Our key idea is to treat the correction as a decision-directed visual query: it identifies the spatial relation that must be resolved to accept or reject the intervention. When the current evidence is insufficient, Ask the World selects a viewpoint that exposes this relation, acquires a new observation, and reassesses the correction before execution. We instantiate the framework with **LingBot** as the frozen base VLA and **Astra** as an inference-time visual reasoning assistant. LingBot provides the nominal action chunk, while Astra proposes bounded local corrections, assesses whether the native observations provide sufficient evidence, and acquires additional views only when needed. The newly acquired evidence is used to reassess the correction against the same nominal action before execution.

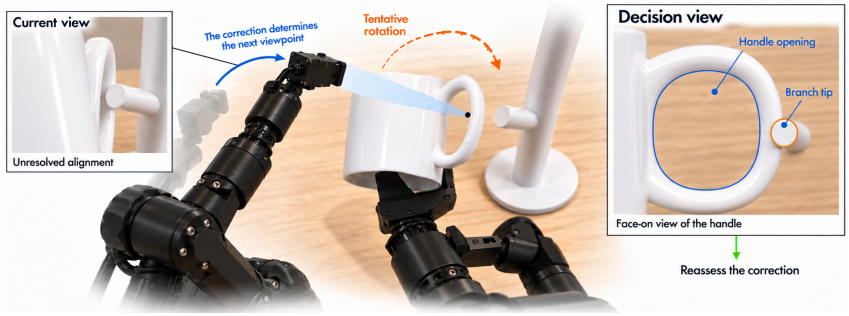

Figure 1: **Ask the World.** When a tentative correction cannot be judged from the current view, the correction defines what the robot needs to see next. The robot acquires a decision view that reveals the relevant spatial relation, then reassesses whether the correction should be applied.

1 INTRODUCTION

Consider a person hanging a mug on a hook. They suspect that the mug should move slightly farther forward, but the current view does not reveal the front-to-back relation between the handle opening and the hook. The difficulty is not knowing what "move forward" means, but knowing whether that adjustment is actually necessary. Rather than executing the tentative correction immediately, the person can first lean sideways to inspect the missing relation and then decide whether to intervene. This simple example illustrates an important capability of embodied intelligence: when an action judgment depends on evidence that is not currently visible, an agent can actively change what it observes before deciding how to act.

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.

Vision-language-action (VLA) policies map visual observations and task instructions to robot actions, providing a reusable foundation for general-purpose manipulation (Kim et al., 2025; Wu et al., 2026a). We study a deployment setting in which the underlying VLA remains frozen while an inference-time assistant can inspect its predicted actions and propose bounded local corrections. Such assistance can improve execution without retraining the base policy. However, identifying a plausible correction does not imply that sufficient evidence exists to justify it: the spatial relation that determines whether a correction is beneficial may be occluded, visually ambiguous, or poorly represented by the available viewpoints. Existing inference-time assistance and VLM-based control methods largely reason about how to act from the observations currently available. We instead ask a complementary question: *when the current observations are insufficient to determine whether a proposed correction is beneficial, what should the robot observe next?*

Our key observation is that a tentative correction itself contains information about what needs to be observed. A local correction changes a particular component of the predicted action—for example, translating the end effector along one direction or adjusting its orientation—and whether that change is beneficial often depends on a specific spatial relation in the scene. In mug hanging, for instance, whether the mug should move farther forward depends on the front-to-back relation between the handle opening and the hook tip. We therefore treat the tentative correction as a *decision-directed visual query*: it identifies the relation that must be resolved before the intervention can be accepted or rejected. When that relation is not observable from the current views, the correction guides the acquisition of a targeted observation that can provide the missing evidence. In this sense, a correction is not only a candidate intervention; it is also a question to the environment.

We introduce **Ask the World** to realize this correction-guided active perception process. We instantiate the framework with **LingBot** as the frozen base VLA and **Astra** as an inference-time visual reasoning assistant. LingBot first generates a nominal action chunk from its native observations, which remains fixed throughout the subsequent correction process. Given this action and the three native views, Astra proposes a bounded local correction and assesses whether the available evidence is sufficient to determine whether the correction is warranted. If the evidence is sufficient, the system directly verifies whether to retain or modify the nominal action. Otherwise, Astra uses the pending correction to identify the spatial relation that remains unresolved and selects a new *decision view* expected to clarify it. The additional observation is provided to Astra, which reassesses the correction against the same nominal action and may retain, revise, or withdraw the proposal. Only the final evidence-supported correction is mapped back to the original action chunk before execution.

This formulation differs from both direct correction and unconditional multi-view observation. Ask the World first determines whether the native observations are sufficient to resolve the pending correction decision, and acquires an additional view only when unresolved evidence could affect that decision. Moreover, the selected view is not intended to be generally informative: it is chosen to clarify the specific spatial relation that determines whether the proposed correction should be retained, revised, or rejected. We evaluate Ask the World on 16 bimanual manipulation tasks in the RoboTwin(Chen et al., 2026) simulator. Under the same evaluation protocol, Ask the World achieves an average success rate of 93.0% and outperforms the frozen LingBot policy as well as RoboDawnGuo et al. (2026) and RoboProbeZhang et al. (2026), with particularly pronounced gains on tasks in which the relevant spatial relation is difficult to infer from the native views, such as mug hanging. These results suggest that selectively acquiring correction-relevant evidence before intervention can improve the execution of frozen VLA policies.

Our contributions are threefold:

- We introduce *correction-guided active perception*, treating a tentative action correction as a decision-directed visual query that identifies what evidence should be acquired before an intervention is accepted or rejected.
- We develop **Ask the World**, an inference-time framework that couples correction proposal, evidence-sufficiency assessment, targeted viewpoint selection, and post-observation reassessment while keeping the underlying VLA frozen, and instantiate it with LingBot and Astra.
- We evaluate Ask the World across 16 bimanual manipulation tasks in RoboTwin against the frozen base policy and alternative VLM-based manipulation methods, demonstrating the benefit of acquiring correction-relevant evidence before intervention.

2 Related Work

Active Perception and Task-Directed View Selection. Active perception studies how an agent can improve downstream decisions by actively changing what it observes. Early work in robotic grasping showed that selecting a viewpoint conditioned on an approximate desired grasp can improve grasp localization and confidence Gualtieri & Platt (2017). More recent approaches couple view selection more tightly with manipulation objectives. ACE-NBV selects new observations according to predicted grasp affordances under candidate viewpoints Zhang et al. (2023), while Vision in Action learns task-relevant searching, tracking, and focusing behaviors from human demonstrations using an actively controlled camera Xiong et al. (2025). ActiveArena further highlights active evidence acquisition as a distinct challenge for robotic manipulation, with tasks that require multi-round visual exploration rather than relying on passive observations alone Li et al. (2026). Closest to our setting, INSPECT learns viewpoint preferences for verifying part presence and installation state, conditioning observation selection on the evidence required by a specific inspection claim Wen et al. (2026). These works establish that perception can be actively conditioned on downstream task or verification objectives. Ask the World considers a different source of conditioning: a tentative action correction generated online during policy execution. Rather than selecting a view to improve general task visibility or answer a predefined inspection query, we use the pending correction to identify the spatial relation whose resolution is required to decide whether that intervention should be executed.

Inference-Time Correction and Verification for VLAs. A growing body of work improves pretrained robot policies at inference time without retraining the underlying policy from scratch. Residual approaches augment a frozen VLA with an additional policy that modifies its predicted actions, for example by learning object-centric residual corrections Kim et al. (2026). Other methods focus on runtime verification and adaptive execution. Do What You Say predicts the outcomes of candidate action sequences and uses a VLM-based verifier to steer execution toward actions consistent with the policy's reasoning Wu et al. (2026b). Continue or Replan? keeps the base VLA frozen while learning when an action chunk should continue and when it should be replanned Xu et al. (2026). CheckVLA instead uses an action-conditioned world model to detect deviations during execution and revise the remaining action sequence Liu et al. (2026), while CARE uses physically grounded 3D monitoring to trigger local corrective actions or re-operations after execution failures Xiao et al. (2026). These methods primarily address how, when, or whether to modify policy execution using evidence that is already available, predicted, or reconstructed. Our focus is complementary: when a plausible correction cannot yet be justified by the current visual evidence, Ask the World actively changes the observation itself. The unresolved correction specifies what spatial relation must be inspected, and the newly acquired evidence is then used to reassess whether the intervention remains warranted.

General-Purpose Vision-Language Models for Robot Control. Recent work has also explored whether general-purpose vision-language models can participate directly in robot control. RoboDawn exposes a compact interface of translation, rotation, and gripper commands to an agentic VLM, enabling closed-loop control through repeated observation, reasoning, and action Guo et al. (2026). RoboProbe goes further by studying an "LLM-as-policy" setting in which general-purpose models such as Astra directly generate robot motion targets without relying on a learned manipulation policy Zhang et al. (2026). Ask the World assigns the visual reasoning model a different role. We retain a pretrained VLA as the manipulation policy and use the VLM as an inference-time assistant that proposes candidate corrections, assesses whether the available observations are sufficient to evaluate them, and seeks additional evidence when they are not. Our framework therefore does not replace the learned policy with a general-purpose VLM; instead, it connects inference-time correction with active perception by making the tentative correction itself the query that determines what should be observed next.

3 Problem Setup

We consider pre-execution intervention for a frozen vision–language–action (VLA) policy under incomplete visual evidence. At action-chunk step k, the policy receives a task instruction ℓ, robot state s_k, and three native RGB observations from the head, left-wrist, and right-wrist cameras,

$$I_k^{\mathrm{nat}} = \{I_k^{\mathrm{head}}, I_k^{\mathrm{left}}, I_k^{\mathrm{right}}\}. \tag{1}$$

The frozen policy π_θ predicts a nominal action chunk

$$A_k = \pi_\theta(I_k^{\text{nat}}, s_k, \ell) = (a_{k,0}, \ldots, a_{k,H-1}) \in \mathbb{R}^{H\times 14}, \tag{2}$$

where $H = 50$. Each action contains six joint targets for each arm and one gripper command per arm. We instantiate π_θ with LingBot and keep its parameters fixed throughout inference.

Before execution, forward kinematics converts A_k into nominal end-effector trajectories

$$\mathcal{T}_k = \{(p_{b,k,i}, R_{b,k,i})\}_{b\in\{L,R\},\, i=0}^{H-1}. \tag{3}$$

Together with the native images and auxiliary calibration and robot geometry Ξ_k, these trajectories form the initial evidence

$$\mathcal{E}_k^0 = \{I_k^{\text{nat}}, \mathcal{T}_k, \Xi_k\}. \tag{4}$$

This lets the reasoning module relate visible scene geometry to what the nominal action is predicted to do.

The environment additionally provides an independently queryable virtual camera that returns an RGB observation I_k^v of the same paused scene from viewpoint v. Such observations can expose geometry that is occluded or ambiguous in the native views, but do not physically probe future contact outcomes.

Within one correction cycle, A_k remains the reference plan while the available evidence may grow through additional observations. New views are used by the reasoning assistant rather than fed back to LingBot to regenerate the current action chunk. The problem is therefore to decide whether the same nominal action should be retained or modified as decision-relevant evidence becomes available.

Importantly, uncertainty alone does not justify intervention. A correction may be plausible but insufficiently supported, while visual ambiguity does not imply that the nominal action is wrong. Ask the World therefore first asks whether the available evidence is sufficient to judge a pending intervention and acquires additional evidence only when it is not.

4 Ask the World

4.1 Overview

Ask the World separates proposing a correction from committing to it. Given A_k, Astra proposes a tentative correction, determines what evidence is required to judge it, and checks whether the current observations already provide that evidence. If not, the pending correction guides additional observation toward the unresolved relation. Astra then reassesses the proposal using the accumulated evidence and chooses to apply, revise, or retain before execution.

Proposal generation, evidence assessment, observation selection, and reassessment are different structured-prompt roles of the same frozen Astra model; no additional model training is required.

4.2 Correction as a Decision-Directed Visual Query

Given the nominal action and available evidence, Astra jointly proposes a candidate correction and a decision query:

$$(C_k^r, Q_k^r) = \mathcal{A}_{\text{prop}}(A_k, \mathcal{E}_k^j, \ell), \tag{5}$$

where r indexes the correction version and j the number of additional observations. C_k^r specifies *how the nominal action may need to change*, whereas Q_k^r specifies *what must be resolved before deciding whether that change is warranted.*

We represent the query as

$$Q_k^r = (\mathcal{R}_k^r, \mathcal{S}_k^{r,+}, \mathcal{S}_k^{r,-}, \mathcal{U}_k^r), \tag{6}$$

where $\mathcal{R}_k^r$ contains decision-relevant relations, $\mathcal{S}_k^{r,+}$ and $\mathcal{S}_k^{r,-}$ specify evidence that supports or refutes the correction, and $\mathcal{U}_k^r$ specifies how these outcomes affect the intervention decision. Thus, Q_k^r defines alternatives that must be distinguished rather than merely describing the scene.

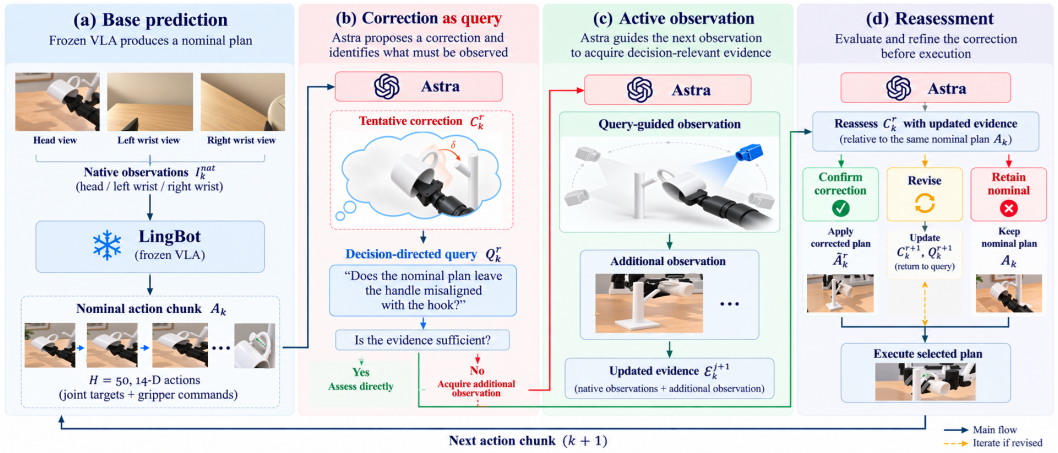

Figure 2: **Ask the World pipeline.** (a) A frozen VLA predicts a nominal action chunk A_k from native observations. (b) Astra proposes a tentative correction C_k^r and converts it into a decision-directed visual query Q_k^r specifying what must be resolved to judge the correction. (c) When the native evidence is insufficient, Astra guides additional observation to acquire decision-relevant evidence. (d) Astra reassesses the correction against the accumulated evidence and either applies it, revises the proposal, or retains the nominal plan before execution.

These relations are defined relative to the nominal plan. For example, if Astra proposes additional forward motion, observing that a target lies in front of the current gripper is insufficient: the nominal trajectory may already move far enough later in the chunk. The relevant question is whether the planned trajectory falls short of, reaches, or passes the target along the proposed correction direction. Ask the World therefore uses the correction to define the evidence needed to decide whether the intervention should be made.

4.3 Correction-Guided Evidence Acquisition

The system first evaluates whether the current evidence is sufficient to resolve the decision query:

$$R_k^{r,j} = \mathcal{A}_{\text{ready}}(A_k, C_k^r, Q_k^r, \mathcal{E}_k^j) \in \{0, 1\}. \tag{7}$$

Here $j = 0$ denotes the native observations. READY measures evidence coverage, not support for the correction: it is true only when the decision-relevant alternatives in Q_k^r can be distinguished from the available observations. A ready candidate may still be rejected.

When $R_k^{r,j} = 0$, Astra guides an additional observation toward the unresolved relation. The objective is not generic visual coverage or evidence that merely confirms the proposal, but evidence that can distinguish interpretations leading to different intervention decisions.

In RoboTwin, we instantiate this principle through virtual viewpoint selection:

$$v_{k,j+1} = \mathcal{A}_{\text{obs}}(C_k^r, Q_k^r, \mathcal{E}_k^j), \tag{8}$$

followed by

$$\mathcal{E}_k^{j+1} = \mathcal{E}_k^j \cup \{(v_{k,j+1}, I_k^{v_{k,j+1}})\}. \tag{9}$$

The viewpoint implementation is therefore a specific realization of a broader principle: the pending correction determines what relation remains decision-relevant, and observation is directed toward resolving it.

4.4 Evidence-Based Reassessment

Once the evidence is ready, Astra reassesses the candidate relative to the same nominal action:

$$D_k^{r,j} = \mathcal{A}_{\text{verify}}(A_k, C_k^r, Q_k^r, \mathcal{E}_k^j, \mathcal{H}_k), \tag{10}$$

where $\mathcal{H}_k$ records previous candidate versions and associated feedback.

Reassessment asks whether the accumulated evidence justifies modifying A_k as proposed by C_k^r. It produces three outcomes: APPLY, when the evidence supports the correction; REVISE, when the evidence motivates a new (C_k^{r+1}, Q_k^{r+1}); and RETAIN, when the evidence supports the identity candidate C_k^{id} and the nominal plan is kept. Importantly, failure to establish a correction is not by itself evidence for RETAIN.

Existing observations can be reused after revision, but evidence coverage and support are reassessed for the new query. Observation and revision budgets bound the loop. If neither a supported correction nor a supported nominal plan can be established within these budgets, the system returns UNCERTAIN and falls back to A_k; this fallback does not certify the nominal plan as correct or safe.

4.5 ACTION REALIZATION AND EXECUTION

A supported correction is represented as sparse end-effector translation and rotation residuals, together with optional gripper and timing edits. After expanding these residuals over the action chunk, the corrected end-effector poses are

$$\begin{aligned} \widetilde{p}_{b,k,i}^r &= p_{b,k,i} + \Delta p_{b,k,i}^r, \\ \widetilde{R}_{b,k,i}^r &= \exp([\omega_{b,k,i}^r]_\times) R_{b,k,i}. \end{aligned} \tag{11}$$

The action mapper

$$\widetilde{A}_k^r = \mathcal{M}(A_k, C_k^r) \tag{12}$$

uses inverse kinematics to recover modified joint trajectories and combines them with the candidate gripper commands; unchanged components are inherited from A_k.

Before execution, the mapped action is checked for numerical validity, joint limits, FK reconstruction error, and command consistency. Failed checks can trigger revision when budget remains. Only the final selected chunk, either $\widetilde{A}_k^r$ or A_k, is executed. Intermediate candidates are never physically tested. If the task continues, LingBot then produces A_{k+1} and a new correction cycle begins.

5 EXPERIMENTS

We evaluate Ask the World from three complementary perspectives. First, we ask whether correction-guided active perception improves the end-to-end performance of a frozen VLA across diverse manipulation tasks. Second, we examine whether the observed gains are consistent with the mechanism proposed in Section 4: additional observation should be most useful when a pending correction depends on spatial evidence that is missing or ambiguous in the native views. Finally, we use diagnostic experiments to separate the effects of action correction, additional observation, and correction-conditioned evidence acquisition.

5.1 EXPERIMENTAL SETUP

We evaluate on 16 bimanual manipulation tasks in the RoboTwin simulator. The tasks cover a range of manipulation behaviors, including object placement, handover, and interactions that require precise relative position and orientation reasoning. LingBot serves as the frozen base VLA, while Astra is used as the visual reasoning model in Ask the World.

We compare against the unmodified LingBot policy and two VLM-based manipulation approaches, RoboDawn and RoboProbe. All methods are evaluated on the same task set and with the same task-success criterion. We report task-level success rates and the macro-average across all 16 tasks.

Ask the World does not update the parameters of either LingBot or Astra. LingBot receives only the native head and wrist observations described in Section 3. Additional observations are queried only by Astra when the available evidence is insufficient to judge a pending correction and are not fed back to LingBot to regenerate the current action chunk. This isolates inference-time evidence acquisition and correction from multi-view replanning of the underlying VLA.

Table 1: Complete-task success rates (%) on 16 RoboTwin 2.0 tasks. ATW combines Astra with the frozen LingBot policy. The macro-average weights all tasks equally. Bold marks the highest rate in each row, including ties. † denotes locally evaluated implementations.

Task	RoboDawn†	RoboProbe†	LingBot	ATW (ours)
Grab Roller	85.00	80.00	**100.00**	95.00
Handover Block	13.33	0.00	99.00	**100.00**
Handover Mic	66.67	35.71	99.00	**100.00**
Hanging Mug	5.00	5.26	49.00	**70.00**
Lift Pot	20.00	13.33	**100.00**	**100.00**
Pick Diverse Bottles	36.84	5.26	86.00	**95.00**
Pick Dual Bottles	20.00	0.00	**96.00**	90.00
Place Bread Basket	75.00	30.00	92.00	**95.00**
Place Burger Fries	75.00	35.00	**99.00**	95.00
Place Can Basket	40.00	25.00	85.00	**90.00**
Place Cans Plasticbox	5.00	15.00	**95.00**	**95.00**
Place Dual Shoes	14.29	0.00	**96.00**	90.00
Place Object Basket	37.50	12.50	87.00	**95.00**
Put Object Cabinet	35.29	23.53	92.00	**95.00**
Scan Object	47.06	5.88	**98.00**	95.00
Put Bottles Dustbin	0.00	0.00	89.00	**90.00**
Macro-average	36.00	17.90	91.38	**93.13**

For each action chunk, Ask the World operates on the nominal action predicted by LingBot and either retains it or applies an evidence-supported correction. Intermediate candidates are never executed. When the system cannot establish a supported intervention within its reasoning budget, it falls back to the nominal action as described in Section 4.4.

5.2 Overall Manipulation Performance

Table 1 reports success rates for all 16 manipulation tasks. Ask the World achieves a **93.0% macro-average success rate** and outperforms LingBot, RoboDawn, and RoboProbe under the shared evaluation protocol. Importantly, this improvement is obtained without fine-tuning the frozen VLA or training an additional action policy.

The improvement is not merely an argument for replacing the VLA with a stronger reasoning model. In Ask the World, LingBot remains responsible for producing the nominal manipulation action, while Astra reasons about whether intervention is necessary and what visual evidence would be useful before committing to that intervention. The comparison with direct VLM-based manipulation baselines therefore supports a complementary role for general-purpose visual reasoning: rather than directly replacing the manipulation policy, it can serve as an evidence-seeking assistant around a frozen policy.

The task-level results are also consistent with the intended operating regime of Ask the World. The most substantial benefits arise on tasks in which the correctness of a local correction depends on spatial relations that are difficult to resolve from the native observations. Mug hanging is a representative example. A nominal trajectory can appear plausible while the depth and orientation between the mug-handle opening and the hook remain ambiguous. In this case, deciding whether to translate or rotate the trajectory requires evidence that is not necessarily contained in the original views.

By contrast, when the geometry needed for action selection is already well exposed, additional sensing has less opportunity to change the decision. This distinction is central to our formulation: Ask the World does not assume that more observations are always useful; it invokes active perception when the evidence needed for a pending intervention is unresolved.

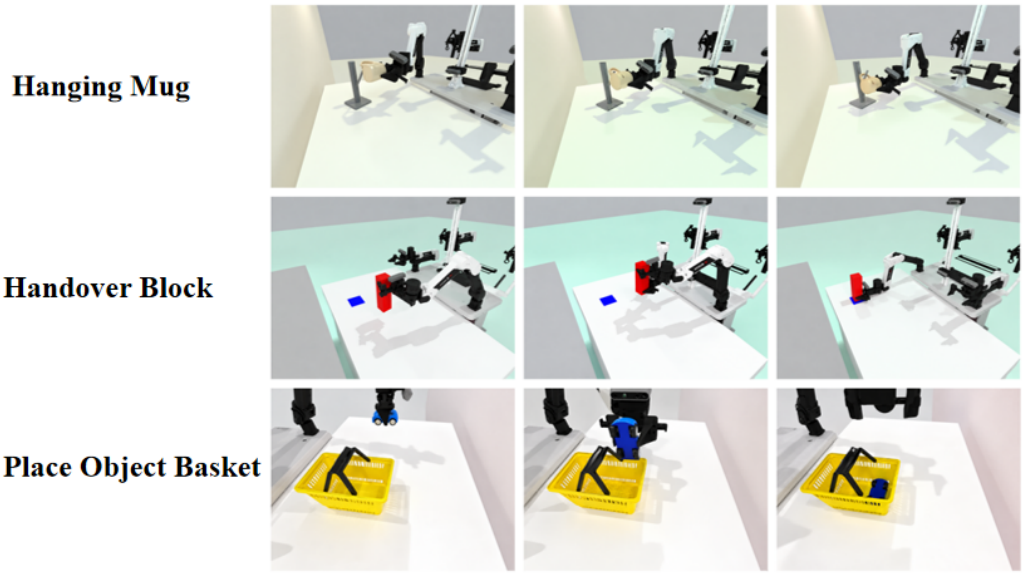

Figure 3: **Representative manipulation cases.** Ask the World is evaluated on tasks involving different spatial relations, including mug–hook alignment, object handover, and object placement. These examples illustrate settings in which a nominal action can appear plausible from the native observations while deciding whether to intervene requires more precise geometric evidence.

5.3 Where Does Additional Evidence Help?

Figure 3 shows representative manipulation cases from the evaluation. The examples cover different forms of geometric reasoning: mug–hook alignment in *Hanging Mug*, relative object and gripper configuration during *Handover Block*, and object–container alignment in *Place Object Basket*. In each case, the objects themselves may already be visible from the native observations, while the relation required to judge a local correction can remain uncertain.

This observation motivates an important distinction between *object visibility* and *decision sufficiency*. A clearer image of the manipulated object is not necessarily useful if it does not reveal the relation that distinguishes competing action choices. Ask the World therefore conditions observation on the pending correction: the correction specifies which relation must be resolved, and Astra seeks an observation informative about that relation.

The newly acquired evidence is not restricted to confirming the initial proposal. It can support the correction, indicate that the nominal plan is already adequate, or reveal that a different modification is required. This non-confirmatory behavior is important because an assistant that always executes its first correction could improve failed actions while simultaneously damaging actions that were already correct. Our reassessment stage instead requires the intervention to remain supported after the new evidence is incorporated.

5.4 Diagnostic Decomposition of Observation and Correction

The 16-task benchmark establishes the end-to-end performance of Ask the World, but success rate alone does not identify the source of the gain. In particular, three simpler explanations must be considered: (i) Astra may improve execution through correction even without new observations; (ii) any additional view may help regardless of how it is selected; and (iii) some failures may reflect a systematic action bias that can be addressed by a fixed residual.

Table 2: **Diagnostic decomposition on the alarm task.** The experiment separately tests additional observation and inference-time correction. These runs use a development protocol distinct from the final 16-task benchmark and are therefore not pooled with Table 1.

Variant	Success	Rate
Base policy	5/12	41.7%
Additional observation	6/12	50.0%
Correction from native observations	10/12	83.3%
Observation + correction	12/12	100.0%

We therefore retain controlled development experiments that separately vary observation and correction. Table 2 reports one such decomposition on the alarm task. These experiments were collected during method development under a different protocol from the final 16-task benchmark, so we report them as diagnostic evidence and do not combine them with the main quantitative results.

Several observations follow from this diagnostic. First, additional observation alone changes performance only modestly in this batch, suggesting that simply exposing the system to another image does not explain the full gain. Second, correction using the native observations already produces a substantial improvement, confirming that inference-time action adjustment is itself a strong component. Finally, combining observation with correction yields the strongest result in this development setting, motivating the central question studied in the final method: *when a correction remains uncertain, can targeted evidence acquisition improve the intervention decision?*

The development experiments also reveal important alternative explanations. Fixed residual corrections were a strong baseline in separate runs, indicating that some failures can arise from systematic action bias rather than missing visual evidence. Likewise, viewpoint-selection experiments varied across development batches, showing that a visually clearer or closer observation is not automatically more useful for downstream manipulation. These observations motivated our decision-directed formulation: the goal is not to find a universally "better" camera view, but to acquire evidence relevant to the particular correction under consideration.

We therefore interpret the 93.0% result conservatively. It demonstrates that the complete Ask the World system improves end-to-end manipulation under our evaluation protocol, while the diagnostic experiments clarify that correction, observation, and viewpoint choice are distinct factors. A stronger causal claim about the independent contribution of query-guided observation requires observation-matched controls in which the correction model, modification permissions, candidate-view pool, and observation budget are held fixed.

6 Limitations

Ask the World targets failures in which a local correction depends on spatial evidence that is missing or ambiguous in the native observations; it does not address missing task skills, low-level control errors, or contact dynamics that cannot be inferred from the current scene. Our evaluation is currently limited to RoboTwin with a queryable virtual camera, abstracting away the physical cost and constraints of embodied viewpoint acquisition. We also evaluate one frozen VLA (LingBot) with one reasoning model (Astra), so broader generalization across policies, reasoning models, and real robots remains to be established.

7 Conclusion

We introduced **Ask the World**, a correction-guided active perception framework for frozen VLA policies. Our key idea is to treat a tentative correction not only as a candidate intervention, but also as a query that identifies what spatial evidence must be resolved before acting. Across 16 bimanual RoboTwin tasks, Ask the World achieves a 93.0% macro-average success rate and outperforms LingBot, RoboDawn, and RoboProbe under our evaluation protocol. These results suggest that visual reasoning can complement a frozen policy by seeking evidence for uncertain interventions rather than replacing the policy itself.

AI USE STATEMENT

In this work, we used generative AI tools to improve the readability, grammar, and wording of the manuscript, assist with literature search and reference formatting, and improve the organization and structure of the paper. We did not use generative AI tools for tasks requiring mandatory disclosure, such as developing the methodology, designing experiments, or interpreting experimental results. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies inference-time visual reasoning for robotic manipulation and is evaluated in simulation on RoboTwin 2.0. The reported evaluation does not involve human-subject experiments or the collection of personal data. Incorrect visual judgments or action corrections could cause collisions, object damage, or injury when deployed on physical robots. The checks used in our framework do not constitute formal safety guarantees. Deployment on physical robots therefore requires additional hardware-specific validation, independent safety mechanisms, and appropriate human oversight.

REPRODUCIBILITY STATEMENT

We describe the proposed framework, experimental setup, and evaluation procedures in the paper. The methodological descriptions and reported experimental settings are intended to support the reproducibility of our results and facilitate future comparisons.

REFERENCES

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan-ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=itonej9GIV.

Marcus Gualtieri and Robert Platt. Viewpoint selection for grasp detection. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 258–264. IEEE, 2017. doi: 10.1109/IROS.2017.8202166. URL https://doi.org/10.1109/IROS.2017.8202166.

Meng-Hao Guo, Zhe-Han Mo, Jia-Jun Wang, Yi Zhang, Kejin Wang, Yi-Xuan Deng, Jia-Peng Zhang, Yongming Rao, and Shi-Min Hu. Transferring the intelligence of VLMs to robotic control. *arXiv preprint arXiv:2609.22966*, 2026. URL https://arxiv.org/abs/2609.22966.

Kinam Kim, Namiko Saito, Heecheol Kim, Katsushi Ikeuchi, Jaegul Choo, and Yasuyuki Matsushita. Object-centric residual RL for zero-shot sim-to-real VLA enhancement. In *Conference on Robot Learning (CoRL)*, 2026. URL https://arxiv.org/abs/2606.18953.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025. URL https://proceedings.mlr.press/v270/kim25c.html.

Yibo Li, Enshen Zhou, Rui Chen, Yanjun Ding, Mengzhen Liu, Yi Han, Jiabo Zhan, Lipeng Wang, Shanghang Zhang, and Lu Sheng. ActiveArena: Benchmarking and understanding active perception in robotic manipulation. *arXiv preprint arXiv:2609.24124*, 2026. URL https://arxiv.org/abs/2609.24124.

Yushan Liu, Peibo Sun, Xintao Chao, Zhenyang Yang, Yifan Xie, Lingfeng Zhang, Shoujie Li, Chenyu Tang, Fang Chen, Xiao-Ping Zhang, and Wenbo Ding. CheckVLA: Execution-time

verification with action-conditioned world model for long-horizon mobile manipulation. *arXiv preprint arXiv:2607.26789*, 2026. URL https://arxiv.org/abs/2607.26789.

Di Wen, Kailun Yang, Wenhao Guo, Yitian Shi, Junwei Zheng, Yufan Chen, Ruiping Liu, Jiale Wei, Rania Rayyes, and Kunyu Peng. INSPECT: Learning robot view selection from assistant use. *arXiv preprint arXiv:2609.20615*, 2026. URL https://arxiv.org/abs/2609.20615.

Wei Wu, Fan Lu, Yunnan Wang, Shuai Yang, Shi Liu, Fangjing Wang, Qian Zhu, He Sun, Yong Wang, Shuailei Ma, Yiyu Ren, Kejia Zhang, Hui Yu, Jingmei Zhao, Shuai Zhou, Zhenqi Qiu, Houlong Xiong, Ziyu Wang, Zechen Wang, Ran Cheng, Yong-Lu Li, Yongtao Huang, Xing Zhu, Yujun Shen, and Kecheng Zheng. A pragmatic VLA foundation model. *arXiv preprint arXiv:2601.18692*, 2026a. URL https://arxiv.org/abs/2601.18692.

Yilin Wu, Anqi Li, Tucker Hermans, Fabio Ramos, Andrea Bajcsy, and Claudia Pérez-D'Arpino. Do what you say: Steering vision-language-action models via runtime reasoning-action alignment verification. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, 2026b. URL https://arxiv.org/abs/2510.16281.

Junlan Xiao, Junwei Jiang, Zaibin Zhang, Yifan Wang, Zhongbo Zhang, Huchuan Lu, and Lijun Wang. CARE: Experience-guided atomic corrective execution for vision-language-action policies. *arXiv preprint arXiv:2609.24118*, 2026. URL https://arxiv.org/abs/2609.24118.

Haoyu Xiong, Xiaomeng Xu, Jimmy Wu, Yifan Hou, Jeannette Bohg, and Shuran Song. Vision in action: Learning active perception from human demonstrations. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 5450–5463. PMLR, 2025. URL https://proceedings.mlr.press/v305/xiong25a.html.

Weichen Xu, Zhenhua Liu, Lin Luo, Yaobo Liang, Chengtang Yao, Qingyu Mei, Jian Cao, Xixin Cao, Xing Zhang, Jiaolong Yang, and Baining Guo. Continue or replan? bernoulli-continuation policy learning for adaptive horizon execution. *arXiv preprint arXiv:2608.03483*, 2026. URL https://arxiv.org/abs/2608.03483.

Wenbo Zhang, Kaixuan Wang, Yutao Ouyang, Xiaoyu Huang, Liyang Li, Kailun Su, Weiyang Jin, Wenhao Chai, Haotian Liang, Zhiyang Dou, Yue Chen, and Tianxing Chen. An unexpected robot policy: Early evaluations of GPT-6 Astra on RoboDojo and beyond. *arXiv preprint arXiv:2609.24170*, 2026. URL https://arxiv.org/abs/2609.24170.

Xuechao Zhang, Dong Wang, Sun Han, Weichuang Li, Bin Zhao, Zhigang Wang, Xiaoming Duan, Chongrong Fang, Xuelong Li, and Jianping He. Affordance-driven next-best-view planning for robotic grasping. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2849–2862. PMLR, 2023. URL https://proceedings.mlr.press/v229/zhang23i.html.

A Appendix

You may include other additional sections here.