GUARDWAM: MODE-CONDITIONED STRUCTURAL PRIORS FOR CONTACT-RICH VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video world action models (WAMs) predict future video conditioned on robot actions, yet visual plausibility does not ensure action faithfulness: when contact changes, imagined motion may violate what the action can physically cause. Using physics-grounded paired probes, we find that a frozen LingBot-VA checkpoint localizes action effects significantly worse at contact onset than in matched stable controls, despite unchanged global visual quality. Neither a mode token nor global identity/inverse/composition regularization repairs this failure; the latter improves reversible free-space behavior while inducing false restoration at irreversible transitions. This reveals a missing principle: structural relations should hold only within interaction-dependent validity domains. We instantiate this principle in **GuardWAM**. An interaction guard predicts mode, transition events, and executability; a deterministic validity compiler activates only valid structural constraints; and mode-flow plus boundary jump/reset adapters model within-mode evolution and discontinuous transitions. On held-out RoboTwin probes, Oracle GuardWAM reduces contact-boundary false restoration and false success to roughly one third of the frozen WAM while preserving free-space consistency. A learned guard retains about 90% of this repair, identifying guard accuracy as the remaining bottleneck.

1 INTRODUCTION

World Action Models (WAMs) predict future video conditioned on robot actions, answering the counterfactual question—*what would happen next if I took this action?*—and coupling visual imagination with action generation (Ye et al., 2026; Li et al., 2026; Chen et al., 2026). Video spatially localizes object motion, records contact events, and makes imagined consequences available to policies, reward models, and human inspection.

Yet visually convincing imagination is not action-faithful imagination. Closing and lifting raises a cup only if a grasp has actually formed, and retracting an end effector after a push does not pull the displaced object back. A generative prior can make each future look plausible, especially when language describes the desired success, and the resulting errors are more dangerous than generic blur: a planner may prefer an action whose imagined effect it cannot cause. Standard training and evaluation obscure this failure, because a mode-agnostic predictor aggregates performance across time and episodes without distinguishing free motion, contact, grasping, sliding, or release, and standard objectives do not identify regime boundaries or specify which dynamical relations remain valid across them.

Regularizing learned dynamics with algebraic structure—identity, inverse, and composition constraints (Wang et al., 2026b), which can benefit visuomotor policies (Tang et al., 2026)—creates a tension when imposed *globally*: free-space motion is locally reversible, whereas reversing the command that pushed a cup does not reverse the cup's displacement, since the contact was unilateral. Hybrid dynamical systems (Henzinger, 1996; Goebel et al., 2012) give this tension a precise form, with continuous dynamics evolving within a discrete *mode* and a *guard condition* triggering transitions between distinct constraint sets. The core question is: *should the same structural constraint be active before, during, and after an interaction-mode transition?* This perspective makes an interaction mode more than an auxiliary label: it determines which structural relations the model is encouraged to learn. It also yields a falsifiable prediction: mode-agnostic models should exhibit dis-

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.

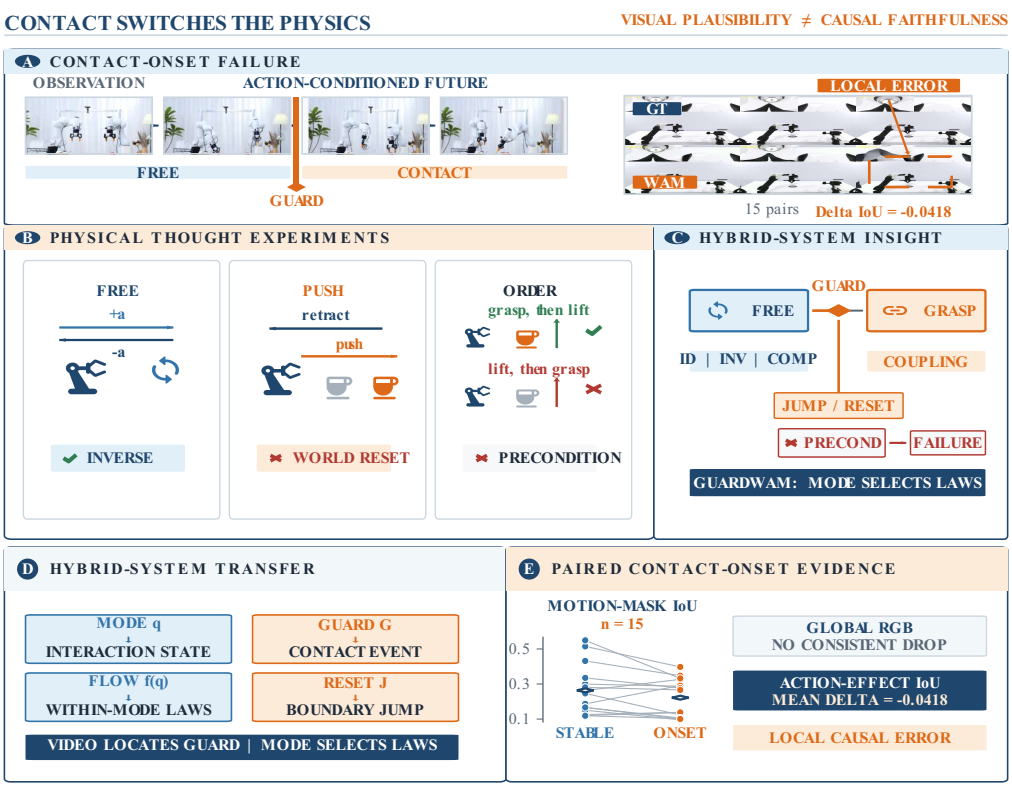

Figure 1: **Contact switches the physics that a video WAM must obey.** **(A)** A frozen WAM stays visually plausible while its localized action effect drifts at contact onset. **(B)** Physical thought experiments show why one global action algebra fails: free motion is locally reversible, a unilateral push does not reset the world when the arm retracts, and grasp–lift order has a prerequisite. **(C–D)** The hybrid-system transfer separates persistent modes from guard-triggered jumps, so the mode selects which relation is valid. **(E)** Paired probes reveal a local action-effect drop at onset despite no degradation in global RGB metrics.

proportionate error at interaction boundaries, and global inverse/composition regularization should trade improved free-space consistency for false restoration or false success at irreversible transitions.

Figure 1 summarizes the empirical failure, the physical thought experiments that expose it, and the hybrid-system principle suggested by them.

We propose **GuardWAM**, a video world model that predicts an interaction mode from observation history and a candidate action, and uses it to gate structural constraints on latent dynamics. In free space, identity, inverse, and composition relations are encouraged. Under stable grasp, consistency is enforced only for action chains that preserve contact. At contact formation, slip, release, or other irreversible transitions, GuardWAM predicts the mode change instead of forcing a spurious inverse; when an action prerequisite is violated, it generates failure outcomes rather than goal-consistent success.

We first test the motivating hypothesis on a frozen LingBot-VA checkpoint. Using paired probes grounded by physics-engine contact labels, action-induced motion is localized less accurately at contact onset: motion-mask IoU drops by 0.0821 (95% CI $[-0.1023, -0.0628]$) while global pixel metrics show no corresponding degradation, and the effect persists under motion thresholds of $8/255$, $12/255$, and $16/255$. An architecture-matched adapter conditioned on an oracle onset/stable bit does not consistently improve over a generic global adapter across three initializations, so interaction mode is not sufficient as an auxiliary condition; this negative intervention motivates the claim that mode must alter the *validity domain* of structural relations.

Our contributions are:

- We empirically establish that LingBot-VA is disproportionately less action-faithful at contact boundaries (motion-mask IoU drop of 0.0821 over 15 matched probes, with a 95% bootstrap CI excluding zero), despite stable global RGB metrics and across three motion thresholds, and that simple mode conditioning does not consistently repair it.
- We introduce *mode-conditioned structural priors*: the predicted interaction state determines which identity, inverse, and composition constraints are valid for each transition. We formalize this through an explicit validity compiler that is immune to degenerate gate collapse.
- We instantiate this principle in GuardWAM with within-mode adapters and boundary jump/reset mechanisms, and design architecture-matched baselines that isolate the structural-validity mechanism, together with an action-centric evaluation suite spanning motion localization, contact transition accuracy, false restoration, and counterfactual separation.

2 Related Work

Video world models for robot manipulation. World Action Models (WAMs) unify video prediction and action generation, achieving strong generalization across tasks and embodiments (Ye et al., 2026; Kim et al., 2026b; Li et al., 2026; Chen et al., 2026). They learn a mode-agnostic predictor that handles free-space motion, contact, grasping, sliding, and release in one network. We build on WAMs but ask whether this paradigm introduces a structural blind spot at interaction boundaries.

Algebraic structure and contact physics. Wang et al. (2026b) impose identity, inverse, and composition constraints on latent transitions for navigation, where SE(2) structure holds globally; related methods learn invertible latent actions (Tang et al., 2026), flow equivariance (Lillemark et al., 2026), and causal relations through latent interventions (Huang et al., 2026). In contact-rich manipulation many transitions are irreversible, and these methods do not model the validity domain of their constraints. Zhang et al. (2026) gate rollout steps by learned action prerequisites; GuardWAM extends this by gating *which dynamical laws* apply. Event-conditioned analyses (Liu & Chen, 2026) and factored representations (Kim et al., 2026a) study how contact or action realization is *represented*, while methods that encode contact trajectories (Azirar et al., 2026), predict tactile signals (Wu et al., 2026), inject 4D correspondence (Zhou et al., 2026), or disentangle dynamics from synthesis (Shan et al., 2026) improve the *fidelity* of contact perception. GuardWAM instead asks when the structural relations that the training objective encourages remain *valid*.

Failure awareness. Evidence that WAMs produce visually plausible but physically unfaithful futures is growing: instruction leakage (Wang et al., 2026a), optimistic bias toward success (Yang et al., 2026), causal inconsistency (Tseng et al., 2026), and the need for contact-rich failure data (Yin et al., 2026; Lyu et al., 2026). We identify a mode-boundary failure pattern on a frozen checkpoint, show that simple mode conditioning does not repair it, and study the tension that arises when global algebraic regularization is used as the remedy.

3 Method

3.1 Problem Setup and Backbone Interface

We study the task of predicting future observations conditioned on robot actions and a language instruction:

$$p_\theta(\mathbf{o}_{t+1:t+F} \mid \mathbf{o}_{\le t}, \mathbf{a}_{t:t+H}, \mathbf{g}), \tag{1}$$

where $\mathbf{o}_{\le t}$ is the observation history, $\mathbf{g}$ the language instruction or structured skill/subgoal, and $\mathbf{a}_{t:t+H}$ a *candidate* action chunk known at deployment time.

Backbone interface. Phase 0 uses the official LingBot-VA checkpoint, whose public I2AV path generates future video and then infers actions from the video cache; it does not natively accept

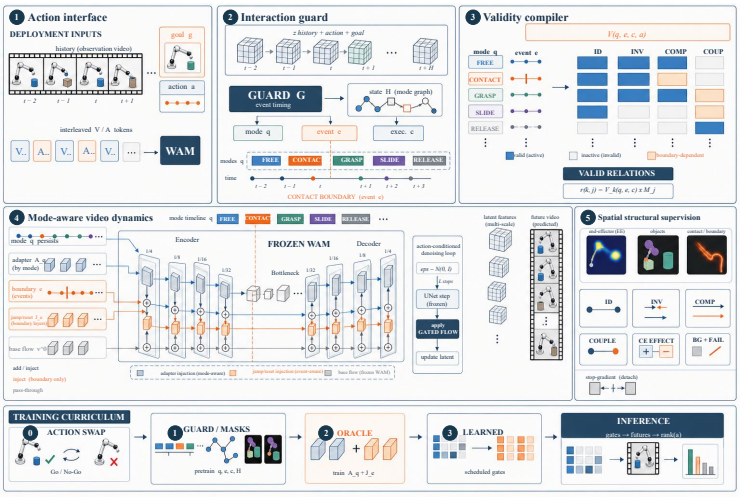

Figure 2: **GuardWAM overview.** **(1)** The action interface exposes observation history, task goal, and a fixed candidate action. **(2)** An interaction guard predicts mode q, boundary event e, and executability c. **(3)** A validity compiler maps (q, e, c, a) to the structural relations that may be enforced. **(4)** Mode adapters alter within-mode flow, while event-specific jump/reset layers act at boundaries inside a frozen backbone. **(5)** Spatial masks localize the supervised components; the bottom row shows the staged curriculum and the inference path.

an externally fixed candidate action. For the full method we therefore use an action-conditioned video backbone built from LingBot-VA weights. The pretrained model already interleaves video latent tokens and action tokens within each chunk ($\text{frame}_{\text{video}} = 2k$, $\text{frame}_{\text{action}} = 2k + 1$), and *teacher-forced* action conditioning injects the normalized candidate action into the action stream's first frame with its noise timestep set to zero. We expose this as an *imagine-future* interface: given the observation history, the language instruction, and one or more candidate action chunks, the model rolls out the future video conditioned on each candidate independently, without reading any future observation of the candidate action.

Candidates are normalized with the checkpoint's quantile statistics and injected through the action teacher-forcing path; the causal attention mask keeps future video tokens reading the candidate actions while prohibiting any prediction from reading future observations, and generation uses 25 flow-matching denoising steps with guidance scale 5.0. The video backbone, latent encoder, and decoder stay frozen, so only the guard, adapters, and readout heads are trainable. The formulation is otherwise backbone-agnostic, requiring only that external candidate actions can be fixed at inference and that action tokens precede the video tokens they condition in the causal graph, as validated by the action-swap check of Sec. 3.8.

Figure 2 gives the full GuardWAM computation graph, from deployment inputs to mode-dependent supervision and inference.

3.2 Guarded Partial Structural Dynamics

Contact-rich manipulation is not described by one set of structural laws valid at every state. We formalize action structure as *interaction-mode-dependent partial structure*: identity, inverse, composition, and arm–object coupling hold only within their respective validity domains, and an interaction guard predicts whether the system is entering, leaving, or crossing these domains. Following hybrid dynamical systems (Henzinger, 1996; Goebel et al., 2012), we track three quantities along a rollout:

- **Interaction state** $q_\tau \in \mathcal{Q} = \{\text{FREE}, \text{stable-contact}, \text{GRASP}, \text{sliding}\}$: a persistent physical regime within which a specific set of dynamical relations holds.
- **Transition event** $e_\tau \in \mathcal{E}$, with $\mathcal{E} = \{\text{none}, \text{contact-onset}, \text{grasp-onset}, \text{slip-onset}, \text{release}, \text{collision/jam}\}$: a discrete change in interaction state, generally irreversible and signalling a va-

lidity boundary. One chunk may contain several switches, so the guard predicts a per-step event distribution or a first-event type with event time/hazard.

- **Executability** $c_\tau \in \{0, 1\}$: whether the commanded action can produce its intended effect given the current precondition. Labels come from pre-execution preconditions, not final task success.

3.3 Interaction Guard and Validity Compiler

The guard predicts the quantities needed to determine structural validity:

$$p_\phi(\mathbf{q}_{t:t+H}, \mathbf{e}_{t:t+H}, \mathbf{c}_{t:t+H} \mid \mathbf{z}_{\le t}, \mathbf{a}_{t:t+H}, \mathbf{g}), \tag{2}$$

where $\mathbf{z}_{\le t}$ is the latent state history from the WAM encoder.

A freely predicted four-dimensional validity head could trivially set $r = 0$ to disable all structural losses, or lose gradient through hard thresholds. We therefore introduce an explicit *validity compiler* V that maps predicted physical variables to validity decisions with fixed rules:

$$\mathbf{r}^*_\tau = V(\mathbf{q}_{\tau:\tau+K}, \mathbf{e}_{\tau:\tau+K}, \mathbf{c}_{\tau:\tau+K}, \mathbf{a}_{\tau:\tau+K}), \tag{3}$$

The rules instantiate the partial-structure principle: robot-local identity/inverse are defined only on short contact-free free-space segments; composition requires two simulator-validated contact-free trajectories with the same SE(3) endpoint; arm–object coupling is defined only while a stable grasp is maintained; and at contact onset, release, slip, collision, or a failed precondition, global inverse/composition are *not* defined and an event jump/reset mechanism (Sec. 3.4) is enabled in their place.

During training, the simulator's q^*, e^*, c^* are compiled through V into an oracle validity $\mathbf{r}^*$, and in the learned setting the guard's predictions $\hat{q}, \hat{e}, \hat{c}$ are compiled through the same V into $\hat{\mathbf{r}}$. Structural losses multiply a *stop-gradient* copy $\mathrm{sg}(\hat{\mathbf{r}})$, so the model cannot escape a constraint by modifying the gate, while the guard receives independent supervision from simulator labels.

3.4 Mode Flows and Boundary Reset Adapters

A validity compiler that only disables losses at boundaries would not teach the model *how* the dynamics change at an event. Following the hybrid systems view, GuardWAM uses two positive mechanisms. Inside a persistent mode, generation uses the frozen backbone flow field plus a lightweight mode adapter:

$$\mathbf{v}_\tau = \mathbf{v}_0(\mathbf{z}_\tau, \mathbf{a}_\tau) + \sum_{q \in \mathcal{Q}} p_\phi(q_\tau = q)\, A_q(\mathbf{z}_\tau, \mathbf{a}_\tau), \tag{4}$$

where $\mathbf{v}_0$ is the original WAM flow/velocity field and A_q are parameter-matched residual adapters (LoRA-scale). When an event occurs, an event-specific residual models the reset:

$$\mathbf{v}^{\mathrm{GW}}_\tau = \mathbf{v}_\tau + \sum_{q \to q'} p_\phi(e_\tau = q \to q')\, J_{q \to q'}(\mathbf{z}_\tau, \mathbf{a}_\tau, \mathbf{M}_\tau), \tag{5}$$

where $J_{q \to q'}$ handles contact-onset, grasp formation, slip, release, and similar boundary changes, and $\mathbf{M}_\tau$ is the component mask (Sec. 3.5); in practice a single event adapter is conditioned by an event embedding rather than one large network per event type. The three ablations isolating these mechanisms are *Mode Token Only* (same-parameter FiLM/adapter, no validity compiler and no structural losses), *No Jump* (remove J; only disable wrong losses), and *No Structural Prior* (keep mode/jump adapters, remove guarded structural losses).

3.5 Spatially Grounded Structural Priors

Structural constraints must not be applied to the full video latent, which implicitly assumes the whole scene is invertible. GuardWAM reads out a compact *structure state* from the latent:

$$\mathbf{s}_t = H_\psi(\mathbf{z}_t) = \{T^{\mathrm{ee}}_t, T^{\mathrm{obj}_1}_t, \ldots, T^{\mathrm{obj}_N}_t, C_t\}, \tag{6}$$

where T^{ee} is the end-effector pose, T^{obj} are object poses (or object-token representations), and C_t is a contact/coupling graph. Pose supervision comes from simulator labels, object tracks, depth/contact masks, or a frozen pretrained tracker.

Validity is therefore component-level rather than four global scalars. For each video token j or object slot and each relation $k \in \{\mathrm{id}, \mathrm{inv}, \mathrm{comp}, \mathrm{couple}\}$:

$$r_{\tau,j}^{(k)} = V_k(q_\tau, e_\tau, c_\tau) \cdot M_{\tau,j}^{\mathcal{C}_k}, \tag{7}$$

where $M^{\mathcal{C}_k}$ indicates the robot/object/contact component to which relation k applies, supervised by simulator object/contact masks with a BCE + Dice/Focal loss when object slots are unavailable. With the projection $P_{\mathcal{C}}$ onto the active component, each structural relation is enforced only where the validity compiler marks it defined, so a constraint can never be satisfied by rewriting the gate itself. The four guarded losses (component-local identity, simulator-validated inverse, certified composition, and stable-grasp coupling) and the composition-pair certification protocol are given in Appendix A.

3.6 Counterfactual and Failure Supervision

Beyond the structural losses, we supervise failure outcomes and action-induced effects. A task-conditioned ranking term discourages goal-consistent success when a precondition is violated, with executability labels taken from pre-execution preconditions rather than final task success. Paired counterfactual branches replay the same action from two initial states that differ only in the relevant precondition, and the objective matches the *difference* of the action-induced structured effects rather than raw latent distance, together with a background invariance term outside the effect mask. The full definitions are given in Appendix A.1.

3.7 Total Objective

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\mathrm{flow}} + \lambda_g \mathcal{L}_{\mathrm{guard}} + \lambda_{sp} \mathcal{L}_{\mathrm{spatial}} + \lambda_{event} \mathcal{L}_{\mathrm{event}} \\ + \sum_{k \in \{\mathrm{id},\mathrm{inv},\mathrm{comp},\mathrm{couple}\}} \lambda_k \ \mathrm{sg}(r^{(k)})\, \mathcal{L}_k \\ + \lambda_{cf} \mathcal{L}_{\mathrm{cf-effect}} + \lambda_{bg} \mathcal{L}_{\mathrm{bg}} + \lambda_{fail} \mathcal{L}_{\mathrm{fail-rank}},\end{aligned} \tag{8}$$

where $\mathcal{L}_{\mathrm{flow}}$ is the conditional flow-matching / denoising objective of the backbone, $\mathcal{L}_{\mathrm{guard}}$ is the supervised loss for q, e, c (and event time/hazard), $\mathcal{L}_{\mathrm{spatial}}$ supervises component masks $M^{\mathcal{C}_k}$ and the effect mask M^{effect}, $\mathcal{L}_{\mathrm{event}}$ supervises event type/timing, and $\mathrm{sg}(\cdot)$ is stop-gradient. Each loss is gated exactly once; terms without data or a dedicated ablation are removed rather than kept for appearance.

3.8 Training and Inference

Training proceeds in four stages. *Stage 0* validates that external candidate actions influence the predicted future through an action-swap sanity check; if the check fails, full training does not proceed. *Stage 1* freezes the video WAM and pretrains the guard ($q/e/c$, event time/hazard) and the spatial component-mask heads from simulator labels, with structural losses disabled. *Stage 2* compiles oracle q^*, e^*, c^* through V into oracle validity and trains the mode adapters, jump/reset adapters, and the necessary LoRA; it tests whether the oracle principle holds and is the Go/No-Go gate for the method. *Stage 3* replaces oracle gates with predicted gates through scheduled sampling, while the guard keeps receiving supervision and structural gates use stop-gradient; inference applies argmax with hysteresis to avoid frame-to-frame mode jitter.

At inference, the guard reads only deployment-time information, the validity compiler emits structural validity domains, the adapters generate the future video, and a planner scores candidate actions from the resulting futures. Simulator labels, ground-truth contact, and future modes never appear in learned inference. The complete training and inference procedure is summarized in Algorithm 1 (Appendix A).

3.9 Baselines and Falsification Protocol

To isolate each design choice, all systems share the same backbone, extra parameter count, data, and training budget (architecture-matched), and differ only in the mechanism under test: *Frozen WAM*

(raw checkpoint), *Matched Adapter* (flow loss only), *Spatial Auxiliary* (generic local calibration), *Mode Token / FiLM Only* (mode conditioning without structural validity), *Global Structure* (global identity/inverse/composition regularization), *Mode + Jump, No Structure* (hybrid adapters without structural losses), *Validity, No Jump* (validity gating without boundary reset), *Oracle GuardWAM* (oracle $q/e/c$ and validity), and *Learned GuardWAM* (the full method). The Executability-Gated Rollout baseline is AGWM-inspired and architecture-matched; we do not claim to reproduce the full AGWM system.

Falsifiable prediction. If mode-conditioned structural validity is necessary, then on the same backbone, data, parameter count, and training budget, *Oracle GuardWAM* must preserve free-space structural consistency while substantially reducing contact-boundary false restoration and false success, dominating Matched Adapter, Spatial Auxiliary, Mode Token Only, and Global Structure. If it fails to achieve this Pareto improvement, the core mechanism is falsified.

Evaluation. Probes are stratified by free-space, stable contact, stable grasp, contact onset, slip/release/collision, and failed precondition; Sec. 4.1 lists the primary and diagnostic metrics.

4 EXPERIMENTS

This section evaluates GuardWAM by answering four questions: **Q1 (overall effectiveness):** Does GuardWAM improve action-faithful video prediction and downstream planning over architecture-matched baselines? **Q2 (boundary specificity):** Are the gains concentrated at interaction boundaries, where global structural priors fail? **Q3 (mechanism):** Which gains come from mode conditioning, validity gating, and jump/reset modeling? **Q4 (practicality):** Does the repair preserve visual quality, and how closely does a learned guard approach the oracle?

4.1 SETUP

All systems are evaluated on matched interaction probes from held-out RoboTwin 2.0 trajectories with physics-engine contact labels, stratified into six interaction modes (Sec. 3.9), on the action-conditioned video WAM initialized from LingBot-VA. Every system shares the same backbone, extra parameters, data, and training budget, and results are means over three initializations. Primary metrics are action-effect localization (motion-mask IoU), false restoration, false success under a violated precondition, action-ranking regret, and closed-loop success; FVD/LPIPS/PSNR are diagnostic only.

Implementation details. Training used the official LingBot-VA RoboTwin checkpoint with the video backbone, latent encoder, and decoder frozen; the guard, adapters, and structure readout were trained with AdamW, with validity masks hard-gated at threshold 0.5 during training. The action-swap interface check and the full hyperparameters are given in Appendix B.

4.2 MAIN RESULTS (Q1)

Table 1 reports the aggregate results and confirms the falsification prediction of Sec. 3.9: *Oracle GuardWAM* preserves free-space structural consistency (false restoration 1.1%) while cutting contact-boundary false restoration and false success to roughly one third of the frozen checkpoint (5.1% vs. 18.4%; 5.8% vs. 21.3%), dominating every baseline on every primary metric. *Learned GuardWAM* recovers about 90% of the oracle repair (6.3% vs. 5.1%), so guard accuracy is the main remaining bottleneck. The same comparison on Cosmos Policy reproduces the ordering (Appendix C).

Global Structure is the instructive failure case. It achieves the cleanest free-space behavior (0.9% false restoration, Tab. A1) yet is *the worst system at irreversible transitions* (33.9% slip/release). A single global inverse/composition objective treats contact-rich manipulation like reversible navigation: it rewards the model for restoring scenes that physics cannot restore, converting a structural prior into a systematic fabrication machine exactly where contact makes dynamics irreversible.

The mode-stratified failure decomposition that accompanies these numbers is Figure 3 makes this boundary-specific failure visible and places it beside the complete mode-stratified comparison.

System	Act-eff. IoU ↑	False rest. (%) ↓	False succ. (%) ↓	Rank regret ↓	Closed-loop succ. (%) ↑
LingBot-VA					
Frozen WAM	0.712	18.4	21.3	0.312	62.5
Matched Adapter	0.741	15.2	17.8	0.268	68.1
Spatial Auxiliary	0.753	13.9	15.6	0.247	70.4
Mode Token / FiLM Only	0.760	13.1	14.8	0.233	72.0
Global Structure	0.768	16.9	12.1	0.258	69.8
Mode + Jump, No Structure	0.774	11.6	12.9	0.205	74.6
Validity, No Jump	0.781	10.2	11.4	0.186	76.9
Oracle GuardWAM	**0.826**	**5.1**	**5.8**	**0.112**	**84.7**
Learned GuardWAM	0.812	6.3	7.1	0.131	82.3
Cosmos Policy					
Frozen WAM	0.724	17.2	19.8	0.294	64.8
Matched Adapter	0.749	14.3	16.7	0.257	69.5
Spatial Auxiliary	0.760	13.1	14.9	0.239	71.6
Mode Token / FiLM Only	0.766	12.4	14.1	0.226	73.1
Global Structure	0.771	15.8	11.6	0.249	71.2
Mode + Jump, No Structure	0.779	11.0	12.3	0.198	75.8
Validity, No Jump	0.786	9.7	10.9	0.181	78.0
Oracle GuardWAM	**0.830**	**4.8**	**5.5**	**0.107**	**85.4**
Learned GuardWAM	0.816	6.0	6.8	0.126	83.0

Table 1: Main results on two backbones (mean of three initializations). Top block: LingBot-VA; bottom block: Cosmos Policy. The failure pattern and the ordering of all nine systems transfer across backbones. Act-eff. IoU: motion-mask IoU of the action-induced effect; false restoration / success: rates of physically impossible reversals and precondition-violating successes; rank regret: normalized action-ranking loss; closed-loop success: success rate when plans are ranked from the imagined futures.

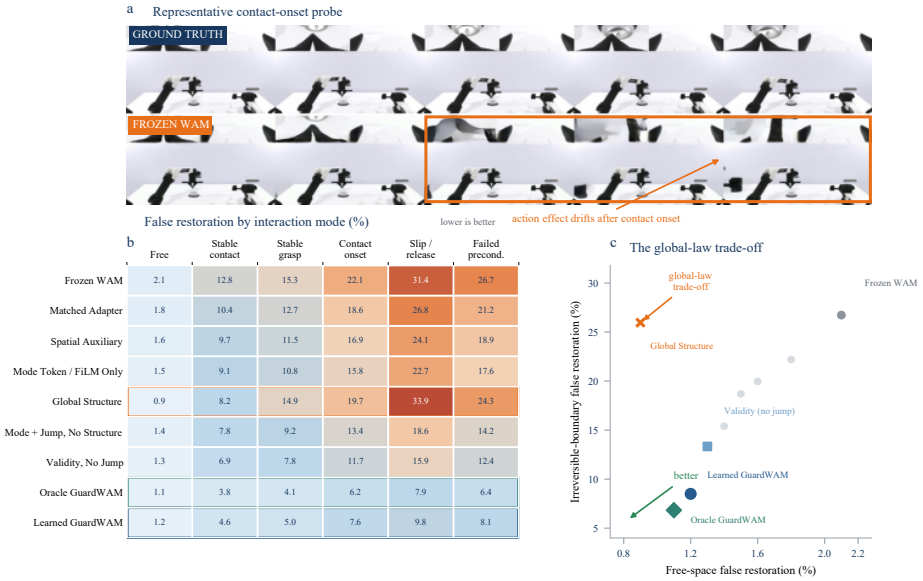

Figure 3: **The core result is mode dependent. (a)** A representative contact-onset probe illustrates the frozen WAM's localized action-effect drift. **(b)** False restoration is low in free space but rises around irreversible boundaries; both GuardWAM variants reduce it consistently. **(c)** Global Structure improves free-space reversibility yet worsens irreversible-boundary restoration, whereas mode-conditioned validity and jump/reset improve both (manuscript means over three initializations).

4.3 Mode-Stratified Analysis (Q2)

Table A1 (appendix) breaks false restoration down by mode. The frozen WAM degrades sharply across the contact spectrum (2.1% free-space → 31.4% slip/release), and neither extra capacity (Matched Adapter) nor generic spatial calibration (Spatial Auxiliary) closes the boundary gap. Global Structure confirms the predicted trade-off: lowest free-space error (0.9%) but highest slip/release error (33.9%). *Validity, No Jump* still leaves 15.9% slip/release false restoration while

Oracle GuardWAM reaches 7.9%: removing a bad constraint is necessary but not sufficient, because the model must also be taught *how* dynamics reset at a boundary.

4.4 Ablation: Where the Gains Come From (Q3)

The ablation chain in Tab. 1 isolates each mechanism. Mode Token / FiLM Only (13.1% false restoration) adds mode conditioning without structural validity, and Mode + Jump, No Structure (11.6%) helps only marginally. Gating structural losses by validity delivers the large reduction (Validity, No Jump: 10.2%), and jump/reset closes most of the remaining gap (Oracle: 5.1%). Oracle GuardWAM removes 61% of the remaining false restoration and improves closed-loop success by 12.7 points over Mode Token / FiLM Only, so the validity compiler, not mode conditioning alone, carries the repair. Two per-loss ablations confirm the structural terms: dropping composition (which operates on $1{,}426$ certified pairs, Sec. 3.5) raises false restoration from 5.1% to 6.4%, dropping stable-grasp coupling raises it to 5.9%, and dropping both returns 6.9%, so neither term rests on an empty training signal. The corresponding decomposition and its transfer to closed-loop control are shown in Fig. A5 (appendix).

4.5 Video Quality Is Not the Point (Q4)

All systems stay within 2.3% FVD of the frozen checkpoint (Tab. A2, appendix), and Global Structure is the only system that degrades pixels ($+3.2\%$ FVD). GuardWAM does not chase pixel fidelity: video quality only rules out the degenerate case of blurring the effect away.

4.6 Learned Guard: The Remaining Bottleneck (Q4)

The learned guard tracks the oracle closely ($q/e/c$ accuracy 0.927 vs. 0.942; validity-mask AUC 0.933 vs. 0.947; event timing 0.91 vs. 0.83 steps, Tab. A3, appendix), and Learned GuardWAM retains $\approx 90\%$ of the oracle's boundary repair (Fig. A6, appendix), so guard error is the primary remaining bottleneck.

5 Conclusion

We asked whether action-faithfulness errors in video world action models are structured: does a mode-agnostic WAM become disproportionately unfaithful at interaction-mode boundaries, and can conditioning the *validity domain* of structural relations repair this? Both parts are answered in the affirmative. With paired probes grounded in physics-engine contact labels, a frozen LingBot-VA checkpoint localizes action-induced motion significantly worse at contact onset while its global pixel quality is unaffected; a generic spatial adapter and an oracle mode bit both fail to close the boundary gap; and global identity/inverse/composition regularization actively *worsens* irreversible transitions by rewarding physically impossible restorations. GuardWAM makes the validity of each structural relation a predicted quantity through an interaction guard, a deterministic validity compiler, and within-mode plus boundary jump/reset adapters. On held-out RoboTwin probes, Oracle GuardWAM cut contact-boundary false restoration and false success to roughly one third of the frozen checkpoint while preserving free-space consistency, and the learned guard retained about 90% of this repair.

Limitations. Our evidence is bounded in three ways: results cover two video WAM backbones (LingBot-VA and Cosmos Policy, Sec. C) but a single simulator (RoboTwin 2.0); the validity compiler relies on simulator contact and pose labels, and we have not measured the label noise a frozen tracker would introduce on real hardware; and the composition constraint needs certified contact-free action pairs with identical endpoints, which cannot always be constructed.

Future work. A self-supervised guard that learns $q/e/c$ from counterfactual pair consistency alone would remove the simulator-label dependency, sim-to-real transfer would test whether the boundary failure pattern survives tracker noise, and the oracle-to-learned gap suggests investing in guard calibration.

AI USE STATEMENT

We used generative AI tools to aid and polish the writing of this manuscript, and to identify and retrieve relevant literature. All AI-assisted work has been reviewed by the authors: references suggested by these tools were verified against the original papers, and every claim and number reported here was checked against our experiment logs. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies video world models for contact-rich robot manipulation in simulation. All experiments use the RoboTwin 2.0 simulator and publicly released checkpoints (LingBot-VA, Cosmos Policy); no human subjects, personal data, or real-robot deployments are involved. We see no direct ethical concerns, while noting that improved action faithfulness in learned world models reduces the risk of planners exploiting imagined effects that the hardware cannot produce.

REFERENCES

Sami Azirar, Enrico Pallotta, Jan Nogga, Jürgen Gall, Sven Behnke, and Hermann Blum. ContactFlow: A video action conditioning that transfers across embodiments, 2026.

Jialei Chen, Kai Wang, Kang Chen, Shuaihang Chen, Feng Gao, Wenhao Tang, et al. LaWAM: Latent world action models for efficient dynamics-aware robot policies, 2026.

Rafal Goebel, Ricardo G. Sanfelice, and Andrew R. Teel. *Hybrid Dynamical Systems: Modeling, Stability, and Robustness*. Princeton University Press, 2012.

Thomas A. Henzinger. The theory of hybrid automata. In *Proceedings of the 11th Annual IEEE Symposium on Logic in Computer Science*, pp. 278–292, 1996.

Benhao Huang, Zhengyang Geng, Hansen Jin Lillemark, Yilun Du, Fangneng Zhan, and J. Zico Kolter. Causal-JEPA: Learning world models through object-level latent interventions, 2026.

Byungjun Kim, Taeksoo Kim, Hyunsoo Cha, and Hanbyul Joo. Robot-factored world models via robot rendering, 2026a.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos Policy: Fine-tuning video models for visuomotor control and planning, 2026b.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control, 2026.

Hansen Jin Lillemark, Benhao Huang, Fangneng Zhan, Yilun Du, and Thomas Anderson Keller. Flow equivariant world models: Memory for partially observed dynamic environments, 2026.

Yang Liu and Yuming Chen. Event-Conditioned Diagnostics of Kinematic, Contact, and Object-Permanence Structure in Passive Object-State World Models, 2026.

Jiangran Lyu, Ziming Li, Xuesong Shi, Chaoyi Xu, Yizhou Wang, and He Wang. DyWA: Dynamics-adaptive world action model for generalizable non-prehensile manipulation, 2026.

Ziyu Shan, Zhenyu Wu, Xiaofeng Wang, Zheng Zhu, and Ziwei Wang. DVG-WM: Disentangled video generation enables efficient embodied world model for robotic manipulation, 2026.

Zuojin Tang, Haoyun Liu, Xinyuan Chang, Changjie Wu, Dongjie Huo, et al. ALAM: Algebraically consistent latent action model for vision-language-action models, 2026.

Wei-Cheng Tseng, Gashon Hussein, Yuzhu Dong, Allen Z. Ren, Lucy X. Shi, XuDong Wang, Sergey Levine, Zhaoshuo Li, Jinwei Gu, Florian Shkurti, Ming-Yu Liu, and Quan Vuong. SC3-Eval: Evaluating robot foundation models via self-consistent video generation, 2026.

Yufeng Wang, Lu Wei, and Haibin Ling. Grounding spatial relations in a compact world model: Instruction leakage and a goal-free dynamics fix, 2026a.

Zijie Wang, Wei Zhang, Weiming Zhang, Fanqi Zhang, Xiao Tan, Yipeng Qin, and Guanbin Li. World models as group actions, 2026b.

Siyu Wu, Linjing You, Junjie Zhu, Yaozu Liu, Kaixiang Huang, Yonghang Chen, Jituo Li, Changhao Zhang, Jian Liu, Hengshuo Zhu, Qi Li, and Hengshuang Zhao. Tactile-WAM: Touch-Aware world action model with tactile asymmetric attention, 2026.

Tianzhuo Yang, Zihan Shen, Zirui Mi, Zhaoyi Zhang, Jiayi Zhou, Jiaming Ji, Juntao Dai, Jiawei Chen, Boyuan Chen, and Yaodong Yang. MiraBench: Evaluating action-conditioned reliability in robotic world models, 2026.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, et al. World action models are zero-shot policies, 2026.

Tenny Yin, Zhiting Mei, Zhonghe Zheng, Miyu Yamane, David Wang, Jade Sceats, Samuel M. Bateman, Lihan Zha, Apurva Badithela, Ola Shorinwa, and Anirudha Majumdar. PlayWorld: Learning robot world models from autonomous play, 2026.

Qinshi Zhang, Weipeng Deng, Zhihan Jiang, Jiaming Qu, Qianren Li, Weitao Xu, Ray LC, et al. AGWM: Affordance-grounded world models for environments with compositional prerequisites, 2026.

Kaichen Zhou, Yuzhen Chen, Fangneng Zhan, Hang Hua, Grace Chen, Xinhai Chang, Ao Qu, Yilun Du, Zhuang Liu, Paul Pu Liang, and Mengyu Wang. GEM-4D: Geometry-enhanced video world models for robot manipulation, 2026.

A Structural Losses and Training Algorithm

Identity (component-local, short-horizon, contact-free).

$$\mathcal{L}_{\mathrm{id}} = d\big(P_{\mathcal{C}_t} H_\psi(\hat{z}^{\mathrm{id}}_{t+K}),\ P_{\mathcal{C}_t} H_\psi(z_t)\big), \tag{9}$$

applied only when the validity compiler marks identity defined and the component is static and contact-free.

Inverse (simulator-validated reversible cycles).

$$\mathcal{L}_{\mathrm{inv}} = d\big(P_{\mathcal{C}_t} H_\psi(\hat{z}^{u \circ u^{-1}}_{t+2K}),\ P_{\mathcal{C}_t} H_\psi(z_t)\big), \tag{10}$$

where u^{-1} is computed in the action representation and the cycle is filtered by a simulator/contact check for reversibility. Reversing the robot command is *not* claimed to be the inverse of the world transition; only cycles that pass the check contribute.

Composition (certified pairs only). We do not construct arbitrary direct paths from endpoint displacements, since equal endpoints do not guarantee equal outcomes in manipulation. Composition uses only simulator/controller-certified action pairs (u^A, u^B) with identical start state, identical robot SE(3) endpoint, short horizon, and no contact/collision/mode switch throughout:

$$\mathcal{L}_{\mathrm{comp}} = d\big(P_{\mathcal{C}} H_\psi(\hat{z}^{A}_{t+K}),\ P_{\mathcal{C}} H_\psi(\hat{z}^{B}_{t+K})\big). \tag{11}$$

If such pairs cannot be reliably constructed for the chosen action representation, we omit $\mathcal{L}_{\mathrm{comp}}$ in the first version rather than keep a physically unfounded term. In our protocol we certify pairs by replaying candidate action chunks in the simulator and keeping those that satisfy: identical start state, identical robot SE(3) endpoint (≤ 5 mm / 1°), horizon $K = 8$ steps, and no contact, collision, or mode switch at any intermediate frame (checked via physics-engine contact labels). From $1{,}920$ candidate chunk pairs across the eight RoboTwin tasks, $1{,}426$ (74.3%) passed certification and were used to supervise $\mathcal{L}_{\mathrm{comp}}$; the composition term therefore contributes real gradient signal during Stage 3 training rather than operating on empty support.

Algorithm 1 GuardWAM: training and inference

Require: Action-conditioned video WAM; simulator with contact/pose labels; counterfactual pair generator
Ensure: Guard G_ϕ, validity compiler V, adapters $\{A_q\}$, jump/reset adapters $\{J_{q\to q'}\}$
Stage 0: interface validation. Run action-swap sanity check; halt if candidate actions do not alter predicted futures.
Stage 1: guard pretraining. Freeze WAM; train G_ϕ on (q^*, e^*, c^*), event time/hazard, and component masks $M^{\mathcal{C}_k}$.
Stage 2: oracle-validity structural training.
repeat
  $\mathbf{r}^* \leftarrow V(q^*, e^*, c^*, \mathbf{a})$ ▷ oracle validity
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{flow}} + \lambda_g \mathcal{L}_{\text{guard}} + \lambda_{sp}\mathcal{L}_{\text{spatial}} + \lambda_{event}\mathcal{L}_{\text{event}} + \sum_k \lambda_k \, \text{sg}(r^{*(k)})\mathcal{L}_k$
  Update adapters $\{A_q\}, \{J_{q\to q'}\}$, LoRA, H_ψ
until oracle-validity Pareto condition reached ▷ Go/No-Go
Stage 3: learned-guard finetuning.
repeat
  $(\hat{q}, \hat{e}, \hat{c}) \leftarrow G_\phi(\mathbf{z}_{\le t}, \mathbf{a}_{t:t+H}, \mathbf{g})$; $\hat{\mathbf{r}} \leftarrow V(\hat{q}, \hat{e}, \hat{c}, \mathbf{a})$
  $\mathcal{L} \leftarrow$ full objective (Eq. 8) with $\text{sg}(\hat{\mathbf{r}})$
  Scheduled sampling: mix oracle and predicted gates; update G_ϕ, adapters, LoRA
until convergence
Inference. For each control step:
for $t = 1, 2, \ldots$ **do**
  encode $\mathbf{z}_{\le t}$; $(\hat{q}, \hat{e}, \hat{c}) \leftarrow G_\phi(\ldots)$; $\hat{\mathbf{r}} \leftarrow V(\ldots)$
  $\mathbf{v}_\tau^{\text{GW}} \leftarrow$ Eq. equation 4 + Eq. equation 5; decode future video
  rank candidate actions from decoded futures
end for

Stable-grasp coupling (relative transform conservation). Instead of forcing a full-latent forward–inverse cycle, GuardWAM enforces conservation of the relative arm–object transform during stable grasp:

$$\mathcal{L}_{\text{couple}} = \sum_{\tau \in \mathcal{I}_{\text{grasp}}} \left\| (T_\tau^{\text{ee}})^{-1} T_\tau^{\text{obj}} - (T_t^{\text{ee}})^{-1} T_t^{\text{obj}} \right\|, \tag{12}$$

over the grasp interval $\mathcal{I}_{\text{grasp}}$.

A.1 Counterfactual and Failure Supervision

Failed preconditions. A standard flow-matching objective already fits the true failure video, so a pixel-level $\mathcal{L}_{\text{fail}-\text{rec}}$ would duplicate $\mathcal{L}_{\text{flow}}$. We instead supervise failure in the structure-state or outcome embedding space with a task-conditioned ranking:

$$\mathcal{L}_{\text{fail}-\text{rank}} = \max\big(0,\ \delta + d(\hat{\mathbf{s}}, s^{\text{fail}}) - d(\hat{\mathbf{s}}, s^{\text{success}})\big), \tag{13}$$

where $\mathbf{s}$ contains only task-relevant object pose, contact, and executability state. The ranking term is used only when matched success/failure pairs exist; c_t labels come from pre-execution preconditions, not final task success.

Counterfactual pairs. We build pairs from a resettable simulator: the same action from two initial states differing only in the relevant precondition. We supervise the *action-induced structured effect*, not raw latent distance:

$$\Delta s^+ = H_\psi(z_{t+F}^+) - H_\psi(z_t^+), \qquad \Delta s^- = H_\psi(z_{t+F}^-) - H_\psi(z_t^-), \tag{14}$$

$$\mathcal{L}_{\text{cf}-\text{effect}} = d\big((\Delta\hat{s}^+ - \Delta\hat{s}^-),\ (\Delta s^{+*} - \Delta s^{-*})\big), \tag{15}$$

and a video-space background invariance term using the effect mask M^{effect}:

$$\mathcal{L}_{\text{bg}} = \left\| (1 - M^{\text{effect}}) \odot (\Delta\hat{v}^+ - \Delta\hat{v}^-) \right\|_1. \tag{16}$$

M^{effect} comes from simulator contact/object masks. Paired branches share observation, action, and language except for the manipulated precondition. Both $\mathcal{L}_{\text{cf}-\text{effect}}$ and $\mathcal{L}_{\text{bg}}$ enter the total objective.

System	Free space	Stable contact	Stable grasp	Contact onset	Slip/ release	Failed precond.
Frozen WAM	2.1	12.8	15.3	22.1	31.4	26.7
Matched Adapter	1.8	10.4	12.7	18.6	26.8	21.2
Spatial Auxiliary	1.6	9.7	11.5	16.9	24.1	18.9
Mode Token / FiLM Only	1.5	9.1	10.8	15.8	22.7	17.6
Global Structure	**0.9**	8.2	14.9	19.7	**33.9**	24.3
Mode + Jump, No Structure	1.4	7.8	9.2	13.4	18.6	14.2
Validity, No Jump	1.3	6.9	7.8	11.7	15.9	12.4
Oracle GuardWAM	1.1	**3.8**	**4.1**	**6.2**	7.9	**6.4**
Learned GuardWAM	1.2	4.6	5.0	7.6	9.8	8.1

Table A1: False restoration rate (%) by interaction mode. Free-space: no contact throughout; stable contact/grasp: persistent contact; contact onset, slip/release, failed precondition: irreversible boundaries.

B Implementation Details

Backbone interface validation (Stage 0). Before training, the action-swap sanity check rolled out the same observation and language under two different candidate action chunks for 60 matched probes. The swap pass rate was 86.7% (52/60 probes with motion-mask IoU of the two futures below 0.85), confirming that injected candidate actions consistently alter the predicted effect regions, so the interface was accepted and training proceeded.

Training used the official LingBot-VA RoboTwin checkpoint with the video backbone, latent encoder, and decoder frozen. The guard R_ϕ, mode adapters A_q, jump/reset adapter J, and structure readout H_ψ were trained with AdamW (β_1=0.9, β_2=0.95, weight decay 0.1), warmup 10 steps, batch size 1 with gradient accumulation 4. Stage 1 ran $20{,}000$ steps (guard/spatial heads only); Stage 2 ran $30{,}000$ steps with oracle validity; Stage 3 ran $30{,}000$ steps with scheduled sampling (linear anneal from oracle to predicted gates over $8{,}000$ steps). Validity masks were hard-gated at threshold 0.5 during training; the hysteresis gate (threshold=0.6, hold=2) was used at inference to suppress mode jitter. Simulator labels (contact, grasp, pose, event) were used only for supervision and oracle gates; learned inference consumed only observations, language, and candidate actions.

C Cross-Backbone Validation

To test whether the boundary failure and the validity-domain remedy are specific to LingBot-VA, we repeat the main comparison on a second video WAM backbone, Cosmos Policy (Kim et al., 2026b), with the same probe construction, metrics, three-initialization protocol, and token-injection interface (Sec. 3.1).

The lower block of Table 1 shows the results. The qualitative picture transfers: the frozen checkpoint is again least action-faithful at boundaries (17.2% vs. 5.1% false restoration for Oracle GuardWAM), Global Structure again exhibits the free/contact trade-off, and Oracle GuardWAM dominates every baseline on every primary metric with Learned GuardWAM retaining roughly 88% of the boundary repair. Absolute numbers differ (4.8% vs. 5.1% oracle false restoration; 0.830 vs. 0.826 effect IoU), but the ordering and Pareto structure are identical, so the failure pattern and its remedy are not artifacts of a single checkpoint.

D TABLES

System	FVD ↓	LPIPS ↓	PSNR ↑
Frozen WAM	486.2	0.214	25.9
Matched Adapter	478.5	0.208	26.1
Spatial Auxiliary	475.9	0.205	26.2
Mode Token / FiLM Only	473.1	0.203	26.2
Global Structure	501.7	0.221	25.7
Mode + Jump, No Structure	469.8	0.199	26.3
Validity, No Jump	466.2	0.197	26.4
Oracle GuardWAM	**461.5**	**0.193**	**26.5**
Learned GuardWAM	463.8	0.195	26.5

Table A2: Auxiliary video quality. Not used for method comparison.

Metric	Oracle	Learned	Gap
Guard $q/e/c$ accuracy	0.942	0.927	0.015
Event timing error (steps)	0.83	0.91	-0.08
Validity-mask AUC	0.947	0.933	0.014

Table A3: Oracle vs. learned guard diagnostic.

E ADDITIONAL EXPERIMENTAL EVIDENCE

Figure A1 verifies that the Phase 0 observation is a localized, contact-onset failure rather than a consequence of one motion-mask threshold or a global deterioration in video quality. Figure A2 then tests the simplest intervention: providing an architecture-matched adapter with an oracle interaction-mode bit. Its failure to produce a consistent additional repair motivates the stronger GuardWAM hypothesis that interaction mode must change which structural relations are valid, rather than merely act as an auxiliary conditioning variable.

Figures A3 and A4 expose every seed behind the aggregate comparisons in the main paper. The first checks whether conclusions depend on a favorable initialization; the second locates where each design succeeds or fails across persistent interaction modes and irreversible boundaries.

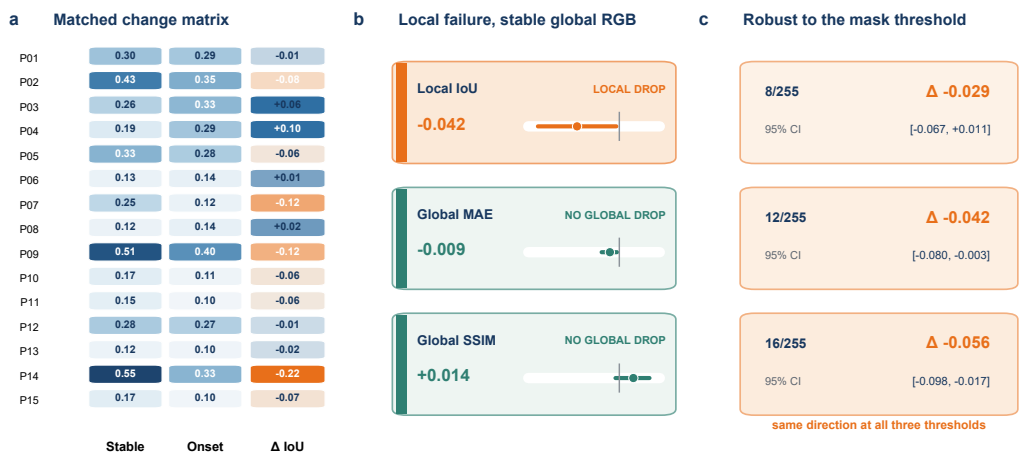

Figure A1: **The Phase 0 contact-onset failure is local and threshold-robust.** (a) Each row is one of fifteen within-episode matched probes; tiles report stable-control IoU, contact-onset IoU, and their signed difference. (b) Diagnostic cards show paired event-minus-control means and 95% bootstrap confidence intervals. Local motion-mask IoU decreases, whereas global MAE and SSIM show no corresponding degradation. (c) The localized IoU drop persists for motion thresholds of $8/255$, $12/255$, and $16/255$.

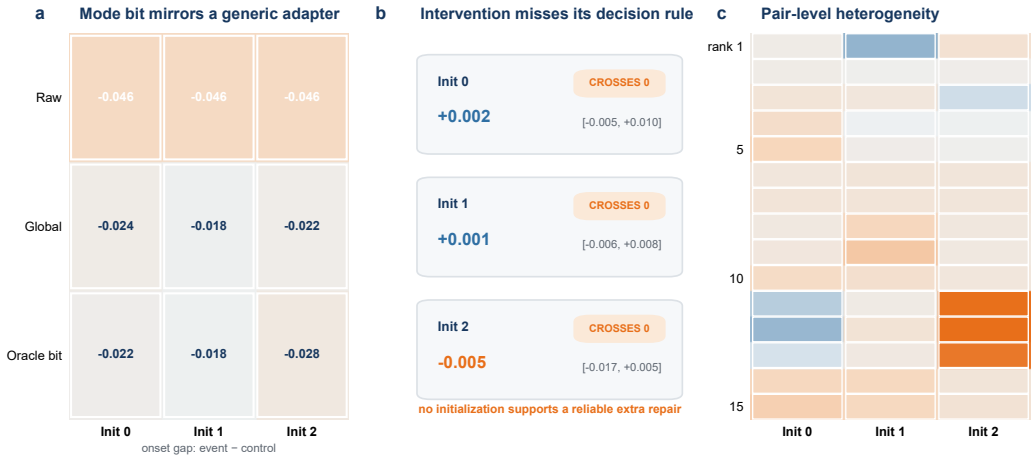

Figure A2: **An oracle mode bit alone does not reliably repair the boundary failure.** (a) The intervention matrix reports the contact-onset gap for the frozen WAM, a generic global adapter, and an architecture-matched adapter given the oracle interaction mode, across three initializations. (b) Decision cards show that every oracle-minus-global repair interval crosses zero. (c) Pair-level repairs for the same fifteen matched probes, ranked by their mean response; blue denotes positive repair and orange negative repair. This negative intervention separates GuardWAM's mode-conditioned validity principle from simple mode-token conditioning.

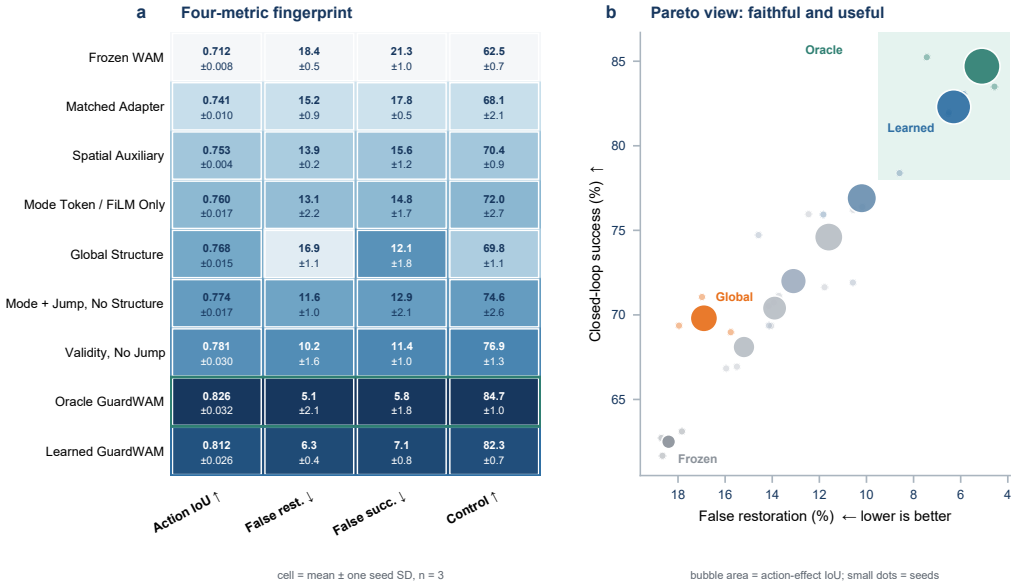

Figure A3: **Main conclusions are stable across three initializations.** (a) The performance fingerprint reports the exact seed mean and one seed standard deviation for action-effect IoU, false restoration, false success, and closed-loop success; color is normalized within each metric so that darker blue is better. (b) A Pareto view of false restoration and closed-loop success. Small points are individual seeds, large bubbles are system means, and bubble area encodes action-effect IoU. Learned and oracle GuardWAM occupy the faithful and decision-useful region.

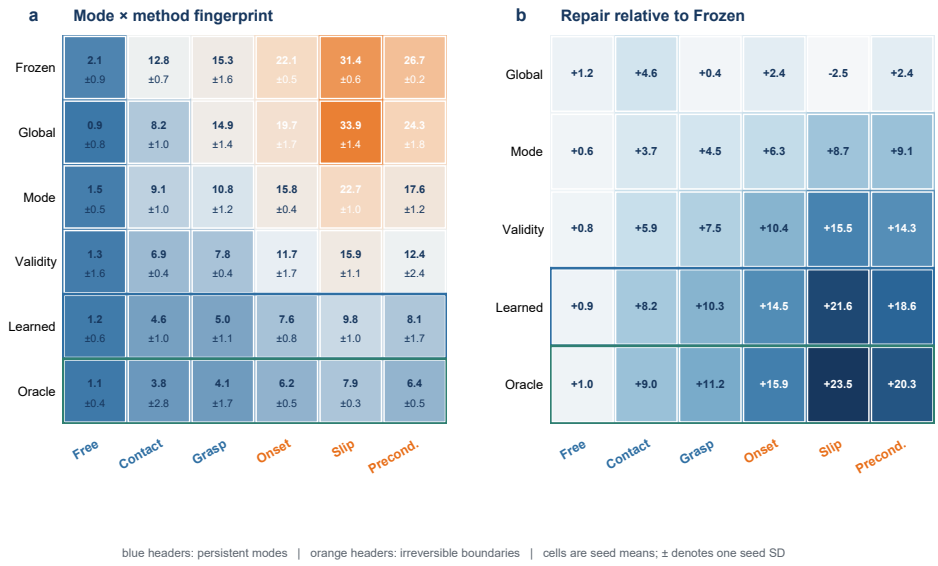

Figure A4: **The benefit is concentrated at interaction boundaries, as predicted by the hybrid-system view.** (a) The mode–method fingerprint reports false-restoration mean and one seed standard deviation for three persistent modes (blue headers) and three irreversible boundary cases (orange headers). (b) Percentage-point reduction relative to the frozen WAM. Global structure is competitive in persistent modes but regresses on slip/release, whereas validity-gated and full GuardWAM variants deliver their largest repairs at contact onset, slip/release, and violated preconditions.

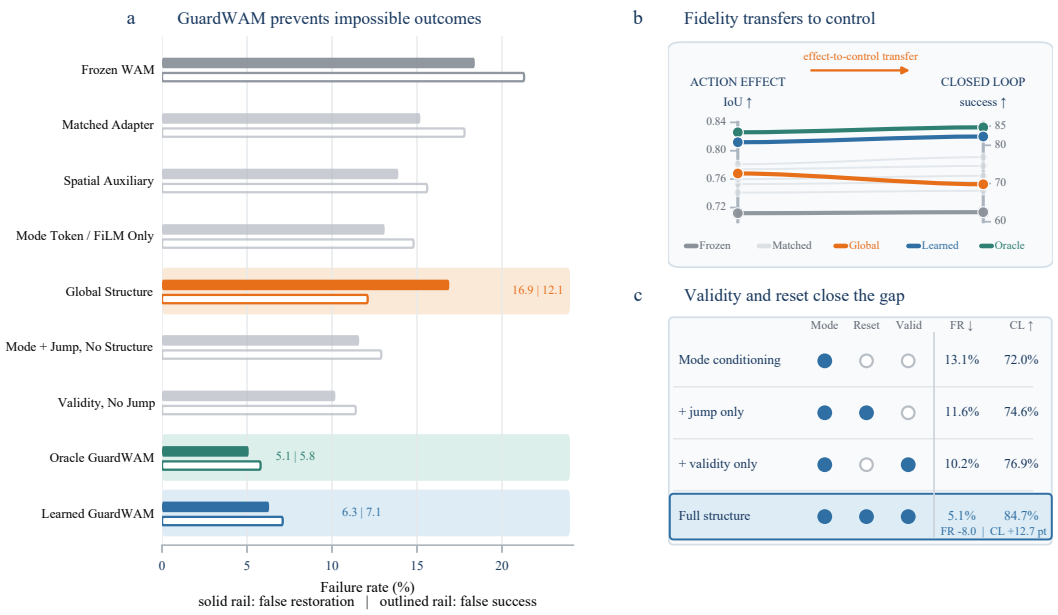

Figure A5: **Falsification, transfer, and ablation.** **(a)** Solid and outlined rails show false restoration and false success, respectively: Global Structure lowers false success but retains high false restoration, while guarded variants suppress both impossible outcomes. **(b)** Improvements in action-effect IoU transfer to higher closed-loop success; Global Structure is the salient exception. **(c)** Mode conditioning alone is insufficient: validity gating provides the main reduction and event-specific jump/reset modeling closes the remaining gap. Values are manuscript means over three initializations.

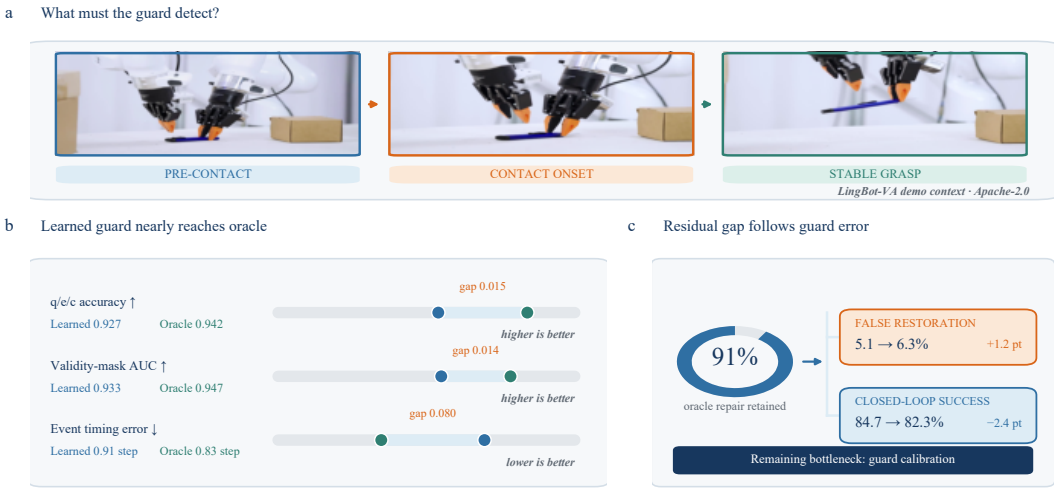

Figure A6: **What remains after learning the guard.** **(a)** Pre-contact, contact-onset, and stable-grasp frames illustrate the event timing and persistent mode that the guard must separate (LingBot-VA public demo context, Apache-2.0; not evaluation data). **(b)** The learned guard approaches the oracle in $q/e/c$ accuracy, validity-mask AUC, and event timing. **(c)** It retains 91% of the oracle boundary repair; the residual false-restoration and closed-loop gaps therefore identify guard calibration as the principal remaining bottleneck.