SWAN: SCALING LONG-HORIZON WORLD MODELS VIA NEXT-SCALE AUTOREGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robotic world models must faithfully predict the consequences of prescribed actions, including failures, while preserving physical states over extended horizons. However, existing generative models frequently hallucinate success due to strong visual priors, and struggle with error accumulation or prohibitive memory growth during recursive rollouts. To overcome these challenges, we introduce **SWAN**, which **S**cales **W**orld modeling through **A**utoregressive **N**ext-scale prediction by integrating unified action-conditioned prediction with a coarse-to-fine hierarchical streaming memory. First, scale-aligned action representations guide generation across all resolutions to strengthen action-future coupling. Second, a time-scale memory preserves fine recent and coarse distant causal states under a bounded resident-memory budget, coupled with scale-wise dream forcing to bridge the exposure gap. Finally, we align visual dynamics on fresh rollouts via action-following and consistency rewards under fixed controls, without task-success rewards. Experiments demonstrate stable streaming rollouts up to 60 seconds. Across 600 matched sim-to-real rollout pairs with frozen external policies, SWAN achieves a success-rate MAE of 0.062 and a within-task rank correlation of 0.88, while an auxiliary action expert further enables effective real-robot control. Despite residual optimism in contact-rich stages, these results demonstrate the promise of next-scale autoregression for scalable robotic manipulation.

1 INTRODUCTION

Generative world models offer a visual interface for predicting the consequences of robot actions (Yang et al., 2024; Zhu et al., 2024). For robotics, however, visual plausibility is not enough: a useful simulator must follow the prescribed controls, including those that produce unsuccessful outcomes. Otherwise, plausible but overly successful predictions can conceal policy failures. For example, a missed grasp should leave an object on the table rather than attach it to the gripper merely because the instruction implies a successful pick. Subsequent predictions must also preserve the resulting state, including any failed contact, displacement, or corrective action. Extending prediction to continuous interaction introduces a further challenge. Recursive generation can accumulate errors and lose persistent state, while retaining the complete history increases memory use. The central question is therefore how to preserve action-dependent dynamics over long horizons without unbounded growth in historical context.

We introduce SWAN, a next-scale autoregressive world model for action-controllable, long-horizon robotic simulation (Fig. 1). Our key design is to use the same coarse-to-fine hierarchy to construct future observations and retain past interaction states. Initialized from InfinityStar (Liu et al., 2025), SWAN advances through successive temporal intervals and refines each future observation across scales, conditioned on the preceding visual history and intervening controls. Intermediate causal states provide a reusable hierarchy for streaming memory, while explicit visual-token probabilities enable reward-based refinement of the generated dynamics. These requirements are complementary: action grounding specifies the current intervention, memory preserves its accumulated consequences, and rollout adaptation addresses the errors introduced when predicted observations become future context. The common hierarchy links each predicted transition to the state retained for subsequent predictions. Specifically, our core technical contributions are threefold:

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.

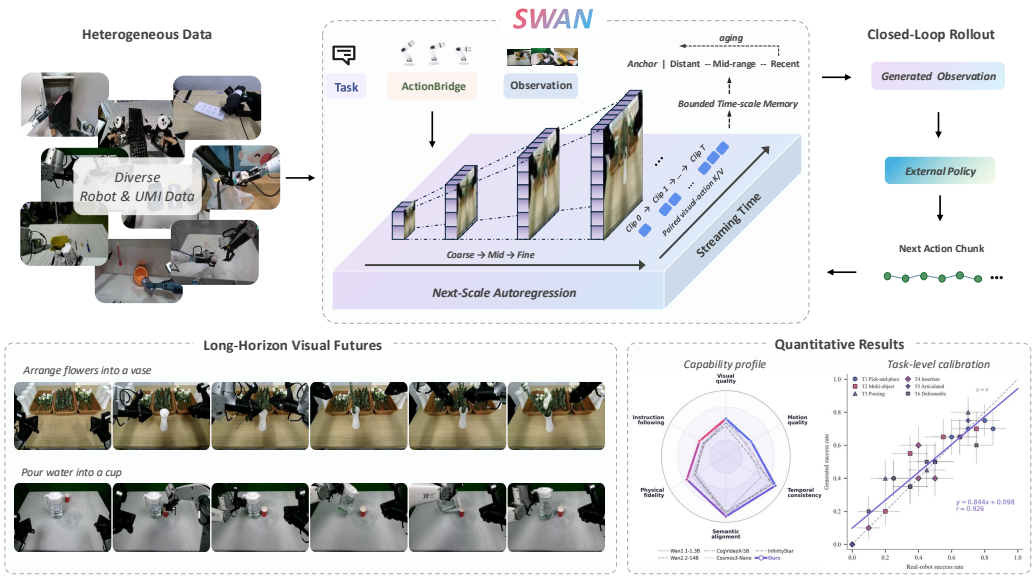


Figure 1: **System-level overview of SWAN.** Robot and UMI data support action-conditioned next-scale prediction with streaming state retention. A frozen external policy consumes generated observations for recursive evaluation. Lower panels show example rollouts and aggregate diagnostics; Fig. 2 details the architecture.

- **Action-conditioned next-scale prediction.** We align action representations with both the visual interval and the prediction scale, making the corresponding controls available throughout coarse-to-fine generation. This targets action–future coupling even when prescribed controls deviate from task or scene priors, rather than treating a likely successful continuation as the desired prediction.
- **Time-scale memory with scale-wise dream forcing.** We select precomputed causal states by age, retaining fine recent records, coarse distant records, and an initial-observation anchor under a bounded memory budget. Each retained visual record remains paired with the actions from the same physical interval; aging changes the selected record rather than recompressing past states online. Scale-wise dream forcing trains prediction on corrupted and model-generated histories to address the mismatch between clean training context and recursive inference.
- **On-policy alignment of visual dynamics.** We optimize fresh visual rollouts under fixed robot-action conditions using action-following and long-term consistency rewards. We update the visual generator, not the robot controller, and exclude task-progress and terminal-success rewards. Autoregressive replay and reference-model regularization constrain drift from the pretrained visual distribution.

Beyond visual prediction, this causal streaming representation naturally supports embodied control. We co-train a lightweight flow-matching action expert on committed world states, providing an auxiliary interface from learned representations to executable robot commands without altering the visual simulation objective.

Experiments show improved action following and reduced long-horizon inconsistency, with streaming rollouts evaluated up to 60 seconds. To test whether the learned dynamics preserve policy-relevant outcomes, we execute the same frozen policy checkpoints in SWAN and on the real robot. Across 600 matched generated–real rollout pairs, SWAN achieves a success-rate MAE of 0.062 and mean within-task Spearman correlation of 0.88, with residual optimism at contact-rich stages. These results support its use for policy evaluation within the tested setting. Separately, we deploy a co-trained action expert on the physical robot to test whether the learned world states support executable control. While the matched-rollout evaluation benchmarks simulator fidelity under external policies, this auxiliary deployment confirms that our learned causal states directly transfer to effective physical execution. Together, these results demonstrate the promise of next-scale autoregression for scalable robotic simulation and control.

2 Related Work

Action-conditioned world models. UniSim and IRASim render interaction outcomes conditioned on instructions or robot controls (Yang et al., 2024; Zhu et al., 2024). Unified World Models and Cosmos Policy further connect visual prediction with executable actions and planning (Zhu et al., 2025; Kim et al., 2026). Our primary objective is forward simulation under prescribed controls, including unsuccessful actions; direct control is an auxiliary extension. The central evaluation therefore asks whether predicted dynamics preserve the outcomes of frozen robot policies, rather than equating realistic video with a reliable simulator.

Next-scale generation. VAR replaces token-wise image generation with coarse-to-fine token-map prediction (Tian et al., 2024). Infinity scales this approach with bitwise representations, and InfinityStar extends autoregression across space and time for image and video generation (Han et al., 2025; Liu et al., 2025). We initialize from InfinityStar rather than introduce a new visual tokenizer or claim the underlying generation hierarchy as our contribution. Our additions align control information with that hierarchy, reuse its causal states as streaming memory, and optimize its conditional visual dynamics on generated rollouts.

Memory and rollout adaptation. FramePack assigns different context lengths to historical frames, while Self Forcing trains on self-generated histories to reduce exposure bias (Zhang et al., 2025; Huang et al., 2025). These address related but distinct questions: which past information to retain, and how to predict from imperfect history. We combine age-dependent selection of precomputed next-scale states with supervised dream forcing. For post-training, RLIR derives action-following rewards by recovering controls from generated video (Ye et al., 2025). Our visual alignment combines control–transition agreement with long-horizon state, boundary, and cross-view consistency, while fixing robot controls and excluding task-progress and success rewards. This objective refines the visual simulator rather than optimizing a robot controller for task completion.

3 Method

SWAN predicts synchronized visual futures from an initial observation $\mathbf{O}_0$, task instruction T_g, and control chunks $A_{1:C}$ (Fig. 2). The visual generator is initialized from InfinityStar (Liu et al., 2025); its tokenizer remains frozen. We first learn action-conditioned prediction, then adapt to generated histories, and finally align visual dynamics with rollout feedback. The following sections define these three components and the auxiliary action expert.

3.1 Action-Grounded Next-Scale Prediction

Representation and factorization. Let c index temporal clips, ℓ the coarse-to-fine scale, and r a residual refinement repeat. The bundle $\mathbf{Z}_{c,\ell,r}$ contains the visual codes for all available synchronized views. Each prediction refines the cumulative reconstruction from preceding decisions, denoted $\prec (\ell, r)$.

With historical memory $\mathcal{M}_c$ and aligned action condition $\mathbf{X}^A_{c,\kappa(\ell)}$, the visual sequence factorizes as

$$p_\theta(\mathbf{Z}_{1:C} \mid \mathbf{O}_0, A_{1:C}, T_g) = \prod_{c=1}^{C} \prod_{(\ell,r)} p_\theta\Big(\mathbf{Z}_{c,\ell,r} \mid \mathbf{Z}_{c,\prec(\ell,r)}, \mathcal{M}_c, \mathbf{X}^A_{c,\kappa(\ell)}, T_g\Big). \tag{1}$$

Action grounding and causal access. For each camera, a deterministic renderer projects the plan-side trajectory initialized from the rollout state, using only prescribed controls and calibration—never future measured states or target video. A shared causal encoder processes this action stream continuously; features are sliced at the same clip boundaries as RGB and projected to retained action scales $\mathcal{S}_A$. The map $\kappa(\ell)$ selects the finest retained action scale not exceeding ℓ, without pooling controls across time or clips.

Action features enter as condition prefixes. Each visual query directly reads only the action block for its current clip, mapped scale, and camera; adjacent clips, future controls, and other cameras' action blocks are masked. A residual cross-view adapter exchanges synchronized causal visual states

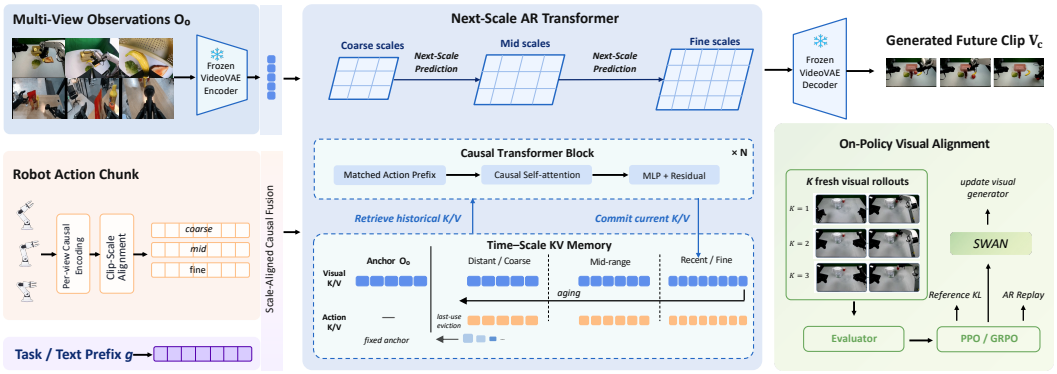


Figure 2: **SWAN framework.** Scale-aligned action prefixes condition next-scale visual generation. Precomputed causal states form bounded time–scale memory, with an initial-observation anchor. On-policy alignment updates the visual generator under fixed controls using fidelity rewards, reference-model regularization, and real-trajectory replay.

without changing this direct action mask or imposing a camera-wise generation order. The initial-observation anchor contains no future action. Additional grounding details appear in Appendix B.

Supervised learning. We optimize the balanced negative log-likelihood

$$\mathcal{L}_{\mathrm{AR}} = -\mathbb{E}_{\mathcal{D}}\mathbb{E}_{(c,\ell,r)\sim \mathrm{Bal}}\left[\frac{1}{N_{c,\ell,r}}\log p_\theta\Big(\mathbf{Z}_{c,\ell,r} \mid \mathbf{Z}_{c,\prec(\ell,r)}, \mathcal{M}_c, \mathbf{X}^A_{c,\kappa(\ell)}, T_g\Big)\right], \tag{2}$$

where Bal averages clips, scales, and refinement repeats uniformly; $N_{c,\ell,r}$ counts valid visual decisions after unavailable camera sites are masked. Bundle log-probabilities sum these decisions. Samples without reliable actions retain visual supervision with $\mathbf{X}^A = \varnothing$.

3.2 Time–Scale Memory and Dream Forcing

Selecting historical states. We reuse states already formed during next-scale generation rather than encoding a separate compressed history. At retained pre-prediction boundary s, clip u provides causal state $\mathbf{H}^{\mathrm{pre}}_{u,s}$ before observing that boundary's target. A larger s exposes a finer cumulative prefix. Each record $\mathbf{B}_{u,s} = (\mathbf{H}^{\mathrm{pre}}_{u,s}, \mathsf{A}_u)$ pairs this state with the action bank from the same physical interval. Before clip c, we read

$$\mathcal{M}_c = \{\mathbf{H}_{\mathrm{anc}}\} \cup \{\mathbf{B}_{c-\delta,s(\delta)} : 1 \le \delta \le \min(W_M, c-1)\}, \qquad \delta_1 < \delta_2 \Rightarrow s(\delta_1) \ge s(\delta_2), \tag{3}$$

where W_M is the fixed retention horizon and $\mathbf{H}_{\mathrm{anc}}$ encodes only $\mathbf{O}_0$. The anchor supplies an identity/layout reference; rolling records carry changing world state. Aging selects a precomputed boundary rather than converting KV states online. Visual-state selection does not imply temporal compression of the paired controls.

The resident bank $\mathcal{R}^{\mathrm{mat}}_c$ includes both active records and precomputed records awaiting later reads. Records are evicted after their last scheduled use. A fixed age–scale schedule enforces $\sup_c \mathrm{mem}(\mathcal{R}^{\mathrm{mat}}_c) \le B_{\mathrm{KV}}$; with fixed clip size, history storage and working memory do not grow with the number of completed clips. Synchronized views commit together before records are aged and evicted. Appendix C specifies the record lifecycle and memory accounting.

Learning from imperfect history. Memory controls context cost but does not remove the exposure gap. We mix clean teacher-forced histories, age-dependent corrupted histories, and detached suffixes sampled from a lagged model $p_{\bar\theta}$ using the same streaming transition $\mathcal{U}$. For a generated suffix of d clips,

$$\begin{aligned} \widehat{\mathbf{Z}}_{c:c+d-1} &\sim p^{\mathrm{stream}}_{\bar\theta}(\cdot \mid \mathcal{M}_c, \mathbf{X}^A_{c:c+d-1}, T_g), \\ \mathcal{L}_{\mathrm{self}} &= -\log p_\theta\Big(\mathbf{Z}^{\mathrm{gt}}_{c+d} \mid \mathrm{sg}[\mathcal{U}^{(d)}(\mathcal{M}_c, \widehat{\mathbf{Z}}_{c:c+d-1}, \mathbf{X}^A_{c:c+d-1})], \mathbf{X}^A_{c+d}, T_g\Big). \end{aligned} \tag{4}$$

Here sg stops gradients through the generated context. We combine this loss with clean and corrupted-history likelihoods as $\mathcal{L}_{\text{long}} = \mathcal{L}_{\text{AR}} + \lambda_{\text{cor}}\mathcal{L}_{\text{cor}} + \lambda_{\text{self}}\mathcal{L}_{\text{self}}$. The target remains the clean recorded future in every branch. Dream forcing is supervised adaptation to imperfect context, not reward optimization.

3.3 On-Policy Alignment of Visual Dynamics

Visual policy and fidelity rewards. A fixed sampling transform defines $\pi_\theta^V(\cdot \mid \xi) = \mathcal{S}_\eta[p_\theta(\cdot \mid \xi)]$, where ξ contains the visual prefix, conditions, mask, and memory addresses. A freshly frozen behavior copy samples $K \geq 2$ rollouts with the same initial observation, instruction, and prescribed controls. "On-policy" refers to this visual distribution, not the robot controller. Two frozen scorers assign clip rewards:

$$\begin{aligned} r_{i,c} = {} & \lambda_{\text{act}} R_{\text{act}}(\widehat{\mathbf{V}}^{\,i}_{c-1:c}, A_c) \\ & + \lambda_{\text{long}} R_{\text{long}}(\mathbf{O}_0, \widehat{\mathbf{V}}^{\,i}_{b_c-1:c}, \mathbf{A}_{b_c:c}), \qquad b_c = \max(1, c - W + 1). \end{aligned} \tag{5}$$

Here $\widehat{\mathbf{V}}^{\,i}$ is decoded rollout i, with $\widehat{\mathbf{V}}^{\,i}_0 = \mathbf{O}_0$. The action scorer contrasts matched controls with shifted, reversed, or cross-trajectory actions; the long-horizon scorer contrasts contiguous real windows with spliced, drifted, or view-desynchronized sequences. They assess control–transition agreement and state, boundary, and cross-view consistency, respectively. Task progress and terminal success are excluded from the rewards and reserved for evaluation.

Prior-preserving optimization. Discounted clip returns are normalized within each rollout group to obtain detached advantages. A clipped group-relative objective (Schulman et al., 2017; Shao et al., 2024) $\mathcal{L}_{\text{PG}}$ applies them to sampled visual-token log-probabilities without a learned critic. We optimize

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{PG}} + \beta\, \mathbb{E}_{\xi \sim \mathcal{D}^{\text{bal}}_{\text{roll}}} D_{\text{KL}}\big(\pi_\theta^V(\cdot \mid \xi) \| \pi_{\text{ref}}^V(\cdot \mid \xi)\big) + \lambda_{\text{AR}} \mathcal{L}^{\text{real}}_{\text{AR}}. \tag{6}$$

The reference penalty limits distributional drift; real-trajectory replay retains visual supervision. All next-scale factors participate with balanced normalization. Only the visual generator and conditioning adapters update; the tokenizer, scorers, reference model, and action-producing modules remain fixed. Appendix D gives the complete optimization objective and separates alignment from frozen-policy closed-loop evaluation.

3.4 Auxiliary Action Expert

A co-trained expert uses flow matching (Lipman et al., 2022) to predict executable action chunks from committed visual memory, current proprioceptive or plan-side state, and the instruction. It cannot read the current visual target, current action condition, or future records. Its predicted controls use the same grounding interface as external controls; the visual generator never reads private expert states. Paired teacher-forced transitions supervise both action prediction and the corresponding visual future; unpaired histories retain only visual losses. This auxiliary world-to-action extension does not change the visual alignment objective. Appendix H gives its interface and joint training loss.

4 Experiments

4.1 Experimental Setup

Training data. We combine the public AgiBotWorld-Beta (Bu et al., 2025) and ManipArena (Sun et al., 2026) datasets with additional recordings from a bimanual tabletop platform and a handheld-gripper interface (Chi et al., 2024), including policy interventions and recovery. Semantic segmentation and caption validation of AgiBotWorld-Beta yield 987,508 captioned clips from 165,560 source videos, counted before latent-cache validation. Public data broaden scene and task coverage, while the additional recordings match the target robot and observation geometry. Intervention data retain missed grasps, misalignment, retries, and corrections rather than filtering by success. Rollback branches are locally matched examples, not exact physical counterfactuals.

Model	Image quality ↑	Motion smooth. ↑	Subject consist. ↑	Interaction quality ↑	Trajectory accuracy ↑	Action following ↑
InfinityStar	0.602	0.601	0.803	0.484	0.251	–
Wan2.1-1.3B	0.679	0.672	0.608	0.452	0.197	–
Wan2.2-14B	0.690	**0.793**	0.803	0.476	0.159	–
CogVideoX-5B	**0.718**	0.686	0.776	0.400	0.177	–
Cosmos3-Nano	0.693	0.701	0.861	0.516	0.202	0.044
SWAN w/o action conditioning	0.683	0.687	0.856	0.501	0.418	0.019
SWAN	0.697	0.715	**0.873**	**0.546**	**0.539**	**0.290**

Table 1: **Generation quality and action conditioning.** External baselines (upper block) and our variants (lower block); cross-model inputs/adaptation may differ. Appendix F supplies the other eight metrics. Bold marks maxima; dashes are unreported, not zero.

We standardize retained video at 10 FPS and enable action conditioning only when synchronized controls and calibration are reliable. Valid visual streams without reliable actions remain in a video-only pool; corrupted recordings are excluded. Both pools supervise visual prediction, whereas only paired teacher-forced transitions provide action targets to the auxiliary expert. Outcome and intervention annotations are retained when available, without treating a takeover as evidence of successful recovery. Appendix A details curation and provenance.

Training. We initialize the visual generator from InfinityStar and freeze the visual tokenizer. Training comprises action-grounded likelihood learning, long-horizon adaptation with dream forcing, and on-policy visual alignment; the objectives and update boundaries are given in Sec. 3. Appendix D describes the training and rollout procedures.

Evaluation settings. The generation benchmark contains 200 in-distribution (ID) and 100 out-of-distribution (OOD) tasks. Synchronized control trajectories are available for 80% of evaluation samples; the remainder use text and an initial image. The OOD set covers novel object–verb pairings, paraphrased instructions, unseen scene configurations, and task compositions. Following the WorldArena protocol (Shang et al., 2026), we assess visual, temporal, and interaction quality, with Action Following and Trajectory Accuracy for action-conditioned evaluation. Appendix E describes the benchmark and metrics. We additionally test streaming up to 60 seconds and compare matched closed-loop executions of frozen external policy checkpoints in the model and on the real robot. This policy-evaluation setting is distinct from physical control by the auxiliary action expert.

4.2 Action-Conditioned Prediction

Generation quality and control fidelity. Table 1 compares visual quality and control fidelity. SWAN obtains Trajectory Accuracy 0.539 and Action Following 0.290, versus 0.202 and 0.044 for Cosmos3-Nano. Image Quality and Motion Smoothness remain below the best respective baselines, so these gains do not imply uniform visual superiority. Cross-model input and training differences limit component attribution; pooled scores do not establish OOD-specific gains.

Effect of action conditioning. Without action conditioning, Image Quality remains 0.683, but Action Following falls from 0.290 to 0.019 and Trajectory Accuracy from 0.539 to 0.418. Interaction Quality also falls from 0.546 to 0.501. This contrast supports the role of action information beyond visual plausibility, not scale alignment over alternative injection designs; Appendix F.1 details its scope.

Changing controls from a shared state. Figure 3 varies controls with the initial observation and instruction fixed. Distinct paths illustrate control sensitivity; Trajectory Accuracy additionally measures reference-motion agreement. Neither metric is task success. These examples do not isolate intervention-data effects; further visualizations appear in Appendix F.

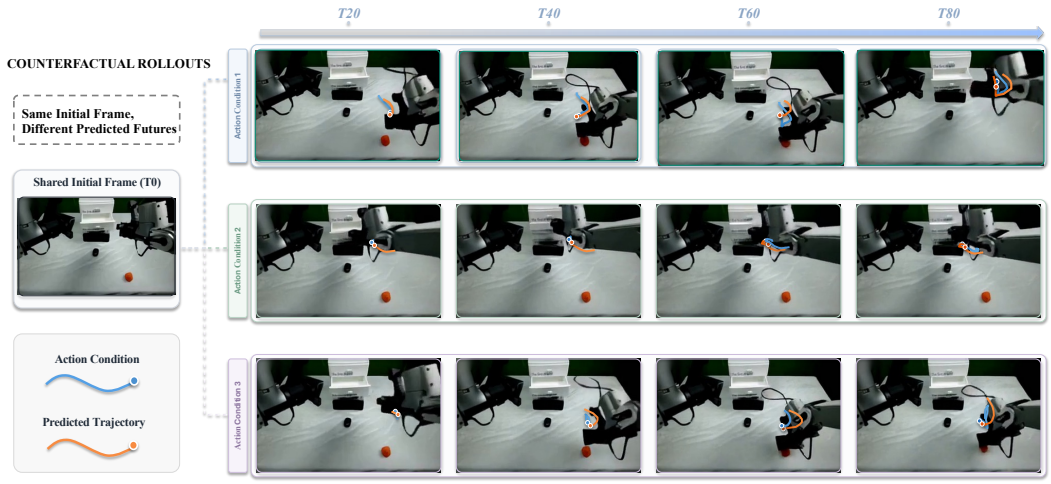


Figure 3: **Action-dependent futures from a shared initial observation.** Blue: prescribed controls; orange: recovered generated trajectory. Branches are model-side action interventions, not exact physical rollback pairs.

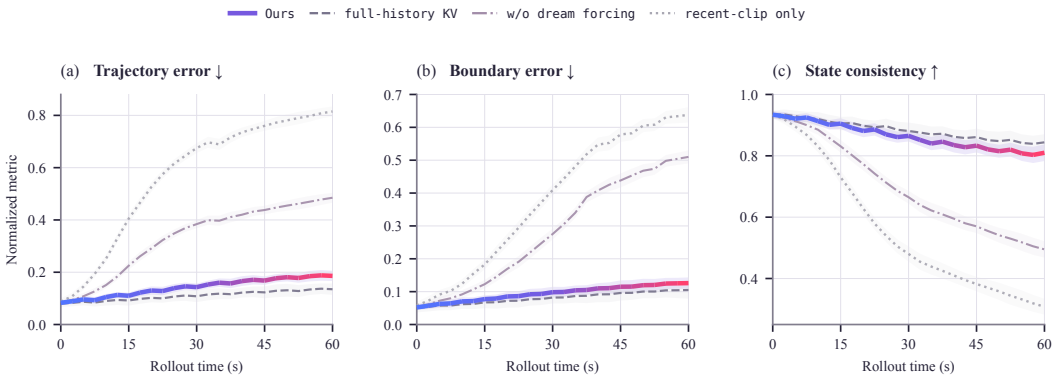


Figure 4: **Streaming dynamics over 60 seconds.** Trajectory, boundary, and state metrics distinguish control, continuity, and retention failures. Full-history KV is a quality reference, not bounded-memory inference.

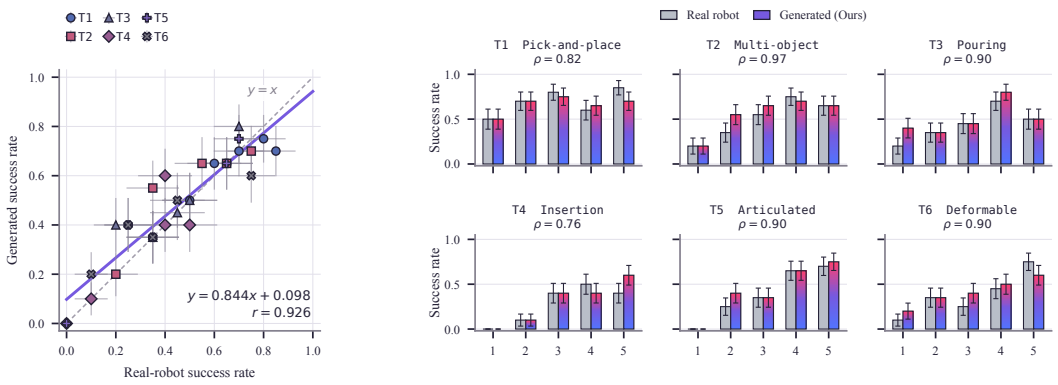
4.3 Long-Horizon Rollout and Memory

Protocol and variants. We evaluate 60-second streaming under matched initial observations and controls, without real-future feedback. We compare time–scale memory with recent-clip-only context, full-history KV, and *w/o dream forcing*, which removes self-generated-history training but retains inference memory. Trajectory error uses normalized dynamic time warping of end-effector paths; boundary error measures commit discontinuity; state consistency evaluates robot, object, and contact state.

Quality under bounded historical context. The complete model accumulates less error than recent-only context or training without generated histories (Fig. 4). Full-history KV remains stronger: bounded retention trades some quality for history allocation independent of rollout length. Actual latency and peak-memory benefits require matched measurements including all resident records, even those awaiting later reads. Appendix F adds windowed PSNR/FID/FVD, continuous-rollout and cross-view diagnostics, complementing action/state checks.

Checkpoint	Action following ↑	Trajectory accuracy ↑	Boundary error ↓
Supervised	0.264	0.512	0.118
Reward-aligned	**0.290**	**0.539**	**0.104**

Table 2: **Effect of visual alignment.** Supervised/aligned checkpoint results under matched initial observations, controls, and sampling settings. This stage comparison does not separately ablate the reward terms, KL penalty, or replay.



(a) Task-level success-rate calibration

(b) Within-task checkpoint ranking

Figure 5: **Frozen-policy evaluation.** Calibration across 30 task–checkpoint cells (left); within-task comparison of five checkpoints (right), with binomial standard errors and per-task Spearman correlations. Cross-task correlation cannot substitute for within-task ranking.

4.4 Effect of Visual Alignment

We compare supervised and reward-aligned checkpoints with identical held-out initial observations, controls, and sampling settings; both recurse on generated history. Alignment improves Action Following, Trajectory Accuracy, and boundary error (Table 2), supporting post-training refinement without separating individual rewards, KL regularization, or replay.

Excluding task-progress and success rewards does not establish scorer/metric independence. The frozen-policy test below probes evaluation utility; Sec. 4.6 separately reports physical control and the no-RL ablation.

4.5 Closed-Loop Policy Evaluation

Matched policy executions. Unlike fixed-action replay, prediction errors in closed loop can change the policy's next action. Five frozen manipulation policy checkpoints run in SWAN and on the robot across six tasks and 20 matched initial configurations: $6 \times 5 \times 20 = 600$ pairs. The model-side policy receives only generated observations after initialization. Blinded annotators label outcomes and subtask completion; unrecoverable artifacts count as failures. Appendix G specifies tasks, labels, and statistics.

Calibration, ranking, and failure detection. Across 30 task–checkpoint cells, success-rate MAE is 0.062 (95% CI $[0.06, 0.11]$), signed bias $+0.028$, and calibration slope 0.84. Mean within-task Spearman correlation is 0.88. The model selects the best real-robot checkpoint on four of six tasks, with mean regret 0.025 (Fig. 5), supporting policy comparison within the tested setting.

Episode-level balanced accuracy is 0.88 (95% CI $[0.83, 0.93]$), and failure recall is 0.86. However, 45 of 332 real failures are predicted as successes: FPR is 0.136 (95% CI $[0.09, 0.19]$). Contact/insertion optimism (Appendix G) can favor weak checkpoints, so physical validation remains necessary. With 20 trials per cell, we emphasize aggregate trends rather than small cell-level differences.

4.6 Real-Robot Control with the Action Expert

Controller and execution protocol. Unlike the frozen external-policy test in Sec. 4.5, here SWAN's co-trained action expert controls the tabletop bimanual robot. At each chunk boundary, it reads synchronized head/wrist observations, the instruction, and measured proprioception, predicts controls, then receives new physical observations after execution.

Tasks and metric. We compare against $\pi_{0.5}$ (Physical Intelligence et al., 2025), LingBot-VA (Li et al., 2026), and DreamZero (Ye et al., 2026) on seven tasks: cup arrangement, spoon placement, ring insertion, basket placement, headphone sorting, shape classification, and ordered button pressing. Task Progress credits partial completion on a 0–100 scale, not binary success. Table 3 gives unweighted task means; Fig. 6 compares the full model with external baselines. This seven-task benchmark is distinct from the six-task simulator test.

Metric	$\pi_{0.5}$	LingBot-VA	DreamZero	SWAN w/o RL	SWAN
Mean Task Progress ↑	49.6	34.0	44.1	67.4	**69.1**

Table 3: **Real-robot control and RL ablation.** Unweighted seven-task mean on a 0–100 progress scale, not success rate. Per-task ablation scores appear in Table 7.

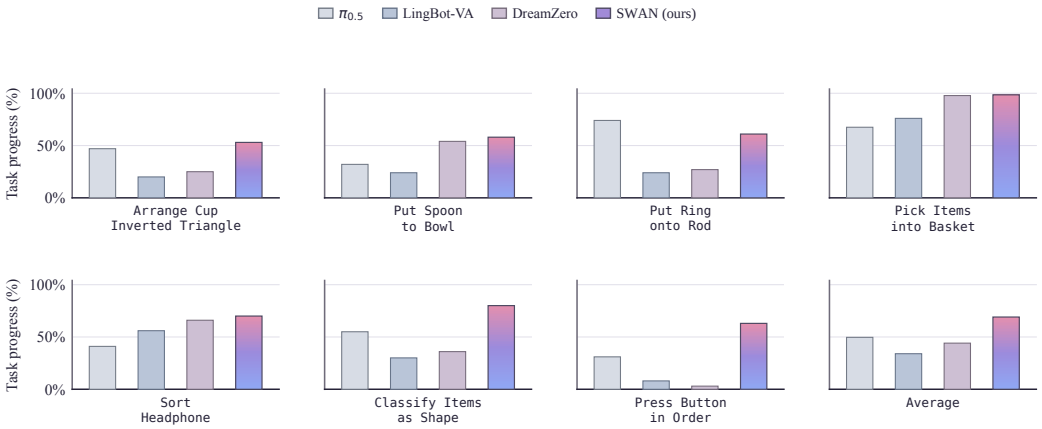


Figure 6: **Real-robot control across seven tasks.** Task Progress for the full model and external baselines; last panel: task mean. $\pi_{0.5}$ is stronger on ring insertion. The no-RL variant is tabulated separately.

Results and RL ablation. The full model reaches mean Task Progress 69.1, versus 49.6 for $\pi_{0.5}$, although $\pi_{0.5}$ remains stronger on ring insertion. The no-RL variant scores 67.4: the full model is higher on all seven tasks, by 1–3 points per task and approximately 1.7 points on average (Table 3). This is a modest additional gain, not an explanation of the full cross-model advantage. Per-task values are in Appendix H.2; causal attribution to visual alignment requires matched expert-training histories and evaluation settings, and these means alone do not establish statistical significance.

5 Conclusion and Limitations

SWAN connects action-grounded next-scale prediction, time–scale memory, and visual alignment, with an auxiliary control expert. Evidence covers 60-second streaming, six-task policy evaluation, and seven-task physical control. Full-history KV retains a quality advantage; practical memory and latency benefits remain unmeasured. Heterogeneous baselines limit component attribution, while optimistic contact predictions can favor weak policies. The model supports evaluation in the tested setting, not replacement of physical validation.

AI USE STATEMENT

Generative AI tools were used for drafting, condensation, structural editing, and copy-editing. The authors are responsible for the final manuscript, including verification of AI-assisted claims, citations, and numerical consistency. Writing assistance is not a substitute for validation of experimental results.

ETHICS STATEMENT

The study concerns robot manipulation in a controlled laboratory setting. The training corpus combines public robot datasets with additional robot and handheld-gripper recordings, including human-operated interventions. Relevant considerations include collection permissions, dataset terms, operator privacy, and safe execution of unsuccessful actions. The world model's optimistic failure predictions should not be treated as safety guarantees or replace physical validation.

REPRODUCIBILITY STATEMENT

The model and objectives are given in Sec. 3. Data curation, action grounding, memory, and optimization details appear in Appendices A–D. Evaluation protocols and extended results appear in Appendices E, F, and G; action-expert training and deployment are described in Appendix H. The reported experiments require the corresponding configurations, split manifests, evaluator versions, and checkpoint provenance for independent reproduction.

REFERENCES

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xu Huang, Shu Jiang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025. URL https://arxiv.org/abs/2503.06669.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *arXiv preprint arXiv:2402.10329*, 2024. URL https://arxiv.org/abs/2402.10329.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15733–15744, 2025. URL https://arxiv.org/abs/2412.04431.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.08500.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025. URL https://arxiv.org/abs/2506.08009.

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. *arXiv preprint arXiv:2601.16163*, 2026. URL https://arxiv.org/abs/2601.16163.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026. URL https://arxiv.org/abs/2601.21998.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. URL https://arxiv.org/abs/2210.02747.

Jinlai Liu, Jian Han, Bin Yan, Hui Wu, Fengda Zhu, Xing Wang, Yi Jiang, Bingyue Peng, and Zehuan Yuan. InfinityStar: Unified spacetime autoregressive modeling for visual generation. *arXiv preprint arXiv:2511.04675*, 2025. URL https://arxiv.org/abs/2511.04675.

NVIDIA. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026. URL https://arxiv.org/abs/2606.02800.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. URL https://arxiv.org/abs/2504.16054.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, Chen Gao, Wei Wu, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Wenwu Zhu, and Yong Li. WorldArena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026. URL https://arxiv.org/abs/2602.08971.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Yu Sun, Meng Cao, Yang Ping, Kaidong Zhang, Qingxuan Chen, Rongtao Xu, Liangwang Ruan, Xuecheng Chen, Dongxiu Liu, Yunxiao Yan, Zunnan Xu, Runze Xu, Charles Yang, Peilun Zhang, Xiaofan Li, Ruyi Gan, Liang Ma, Yuehao Yin, Jincheng Yu, Lufang Chen, Yuxin Liang, Peng Zhai, Hao Wang, Ivan Laptev, Ian Reid, Qian Wang, and Xiaodan Liang. ManipArena: Comprehensive real-world evaluation of reasoning-oriented generalist robot manipulation. *arXiv preprint arXiv:2603.28545*, 2026. URL https://arxiv.org/abs/2603.28545.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2404.02905.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. URL https://arxiv.org/abs/1812.01717.

Wan Team. Wan2.2: Open and advanced large-scale video generative models. Official release repository, 2025. URL https://github.com/Wan-Video/Wan2.2. Inference code and model weights released July 28, 2025.

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06114.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng,

Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.06072.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026. URL https://arxiv.org/abs/2602.15922.

Yang Ye, Tianyu He, Shuo Yang, and Jiang Bian. Reinforcement learning with inverse rewards for world model post-training. *arXiv preprint arXiv:2509.23958*, 2025. URL https://arxiv.org/abs/2509.23958.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *arXiv preprint arXiv:2504.12626*, 2025. URL https://arxiv.org/abs/2504.12626.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025. URL https://arxiv.org/abs/2504.02792.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A fine-grained world model for robot manipulation. *arXiv preprint arXiv:2406.14540*, 2024. URL https://arxiv.org/abs/2406.14540.

A Training Data and Curation

A.1 Sources and Counting Units

The public sources are AgiBotWorld-Beta (Bu et al., 2025) and ManipArena (Sun et al., 2026). Semantic segmentation and caption validation of AgiBotWorld-Beta produce 987,508 captioned clips from 165,560 source videos. These counts describe preprocessing outputs before latent-cache validation; the final training collection additionally depends on cache validity and split assignment. Clip counts and source-video counts are reported separately because segmentation can produce multiple clips from one video.

Recordings from the bimanual tabletop platform contain motion, gripper states, and multi-view video in the target observation geometry. The portable handheld-gripper UMI subset expands interaction coverage without occupying a robot platform; its motion supplies control conditioning only when synchronization and the geometric calibration chain are valid. Policy-intervention and recovery records retain source and intervention identities rather than being merged anonymously with nominal demonstrations.

A.2 Interventions and Recovery

We retain unsuccessful but physically valid transitions so that training includes outcomes not captured by nominal successful demonstrations. In *immediate intervention*, the operator takes over from the current policy state and continues with corrective controls. In *rollback-and-replay*, the erroneous execution is retained, the system is brought back toward the pre-error state, and the operator records a corrective branch. These records preserve missed contacts, misalignment, re-grasping, retries, and subsequent corrections.

Physical rollback need not reproduce object pose, contact, or compliance exactly. We therefore treat the branches as locally matched intervention examples, not exact counterfactuals. Direct branch comparisons use synchronized states and visual context around the rollback point to check comparability. Failure, takeover, recovery-completion, and episode-outcome labels are kept separately when available: a local failure is not necessarily terminal, and intervention alone does not establish successful recovery. The visual objective predicts the recorded outcome rather than relabeling a failed interaction as a successful one.

A.3 Integrity, Synchronization, and Supervision Routing

We check frame decodability, resolution and frame-rate metadata, timestamp monotonicity, multi-view synchronization, video/control length consistency, and available robot states. Corrupted or unusable visual records are rejected. A usable visual stream is retained without action conditioning when its control timing, kinematics, or calibration is unreliable. This distinction prevents invalid action metadata from contaminating the conditioned pool while preserving valid video supervision. Physically meaningful waiting, holding, missed grasps, drops, and retries are not discarded merely because motion is small or the task fails.

Long recordings are segmented at semantic task or subtask boundaries and standardized to 10 FPS. Each manifest entry stores source identity, the annotated interval, original frame rate and resolution, language description, and links to available action/calibration metadata. Intervention clips retain pre-error and corrective context when available. For each retained visual sample, let m_i^{cal} indicate that action timing, intrinsics, extrinsics, and kinematics pass their checks. The pools are

$$\begin{aligned} \mathcal{D}_{\mathrm{ac}} &= \{(\mathbf{v}_i, \mathbf{a}_i, \mathbf{l}_i) : m_i^{\mathrm{cal}} = 1\}, \\ \mathcal{D}_{\mathrm{vid}} &= \{(\mathbf{v}_i, \mathbf{l}_i) : m_i^{\mathrm{cal}} = 0\}. \end{aligned} \tag{7}$$

Both pools train visual prediction. Only $\mathcal{D}_{\mathrm{ac}}$ supplies an aligned control condition; action-expert supervision further requires a paired teacher-forced transition, as in Appendix H.

A.4 Action Geometry and Cache Provenance

For calibrated robot and UMI records, end-effector poses and gripper opening are converted to a common representation. To make the projection convention explicit, let $\mathbf{x}_e \in \mathbb{R}^3$ be an end-effector

keypoint, $\mathbf{R}_{v\leftarrow e}, \mathbf{t}_{v\leftarrow e}$ its rigid transform into a calibrated camera v, and $\mathbf{K}_v$ the intrinsic matrix after the RGB resize/crop. Then

$$\begin{aligned} \mathbf{y}_v &= \mathbf{K}_v(\mathbf{R}_{v\leftarrow e}\mathbf{x}_e + \mathbf{t}_{v\leftarrow e}), \\ \mathbf{u}_v &= (y_{v,1}/y_{v,3},\ y_{v,2}/y_{v,3}). \end{aligned} \tag{8}$$

Invalid or behind-camera projections are masked. The rendered condition represents end-effector centers, orientation axes, and gripper state; RGB and action inputs use matched temporal indices and spatial transforms. As required by Appendix B, rollout conditioning must be constructed from prescribed controls and causally available state, not future realized telemetry.

The frozen visual tokenizer caches RGB latents. Cache keys record source video, annotated interval, target frame rate, resize/crop, and representation version; training checks cache existence and tensor compatibility. Each example retains calibration status, quality flags, and available intervention/outcome labels for source-aware sampling and audit.

B Additional Action-Grounding Details

Camera-aligned control streams. For camera v, the deterministic renderer $\mathcal{R}^{(v)}$ maps the plan-side trajectory initialized from rollout state p_0 onto the camera timeline. Its inputs are prescribed controls, the available initialization state, and calibration; realized future telemetry and target RGB are never used. The action encoder processes the rendered stream causally and continuously. Its features are sliced by the visual clip intervals $\mathcal{J}_c$, preserving the correspondence between each control interval and its predicted visual consequence.

Scale selection. For each interval, features are projected at the retained action scales $\mathcal{S}_A$ and collected in $\mathbf{X}^A_{c,j}$. The selector $\kappa(\ell)$ reads the finest retained action scale no finer than visual scale ℓ. This is a choice of representation scale, not a temporal pooling operation: it neither merges controls across clips nor replaces a control trajectory by its time average.

Direct and cross-view conditioning. Action tokens are condition prefixes in the visual Transformer's sequence. A query indexed by (c, ℓ, r, v) reads text and the direct action block $(c, \kappa(\ell), v)$. Direct attention to other views' action blocks, neighboring clips, or future controls is masked. The cross-view adapter exchanges only synchronized causal visual states; it does not relax the direct action mask. Unavailable camera sites are masked from the visual likelihood. The identity anchor is derived from $\mathbf{O}_0$ alone and has no future-action condition. These restrictions distinguish access to a control from access to its as-yet-unobserved visual outcome.

The data-side geometric preprocessing is described in Sec. A.4. Source-specific calibration availability controls whether an example receives action-conditioned or video-only supervision.

C Memory Records and Dream-Forcing Details

C.1 Precomputed Records and Resident Storage

Let $\mathcal{S}_M$ be the set of retained pre-prediction boundaries and $\mathcal{S}_A$ the retained action scales. For clip u, we materialize

$$\mathsf{A}_u = \{\mathbf{X}^A_{u,j}\}_{j\in\mathcal{S}_A}, \qquad \mathbf{B}_{u,s} = (\mathbf{H}^{\mathrm{pre}}_{u,s}, \mathsf{A}_u), \qquad s \in \mathcal{S}_M. \tag{9}$$

$\mathbf{H}^{\mathrm{pre}}_{u,s}$ is captured before the target at boundary s is observed. Larger s corresponds to a finer cumulative visual prefix. The fixed monotone schedule in Eq. (3) selects a record according to clip age. Aging changes the selected boundary address; it does not apply a learned compressor or an online conversion to cached KV states. Each selected visual record stays paired with the action bank for its own physical interval.

The active read set $\mathcal{M}_c$ is not the full resident allocation. Storage also includes precomputed records that will be selected only at later ages. The resident bank and its bound are

$$\begin{aligned} \mathcal{R}^{\mathrm{mat}}_c &= \{\mathbf{H}_{\mathrm{anc}}\} \cup \{\mathbf{B}_{u,s} : (u,s) \text{ is resident before clip } c\}, \\ \sup_c \mathrm{mem}(\mathcal{R}^{\mathrm{mat}}_c) &\le B_{\mathrm{KV}}. \end{aligned} \tag{10}$$

Records are evicted after their last scheduled read. The bound accounts for resident visual records, paired action data, and the persistent anchor, including records awaiting future use. For fixed clip dimensions, retained boundaries, and retention horizon, both this allocation and the working memory for one clip are independent of the number of completed clips. This storage bound does not assert that bounded memory matches the prediction quality of a full-history model.

C.2 Streaming Record Lifecycle

Before predicting clip c, the model assembles $\mathcal{M}_c$, causally encodes the current action block, and generates the synchronized visual bundle in next-scale order. States captured at retained boundaries become persistent only after the complete bundle commits. The transition then ages records, changes active cutoffs, and evicts expired entries before clip $c+1$. The initial anchor remains fixed, while rolling records propagate the changing world state.

C.3 Clean, Corrupted, and Self-Generated Histories

For age-dependent corruption, sample $\widetilde{\mathcal{M}}_c \sim \mathcal{C}_{\mathrm{age}}(\mathcal{M}_c)$ and optimize $\mathcal{L}_{\mathrm{cor}}$, the balanced visual likelihood with $\widetilde{\mathcal{M}}_c$ replacing the clean history. For self-generated history, Eq. (4) samples a detached suffix from a lagged streaming model and supervises the next clean recorded clip. $\mathcal{U}^{(d)}$ is the composition of d streaming memory transitions, including the same commit and eviction operations used at inference. The full objective is

$$\mathcal{L}_{\mathrm{long}} = \mathcal{L}_{\mathrm{AR}} + \lambda_{\mathrm{cor}}\mathcal{L}_{\mathrm{cor}} + \lambda_{\mathrm{self}}\mathcal{L}_{\mathrm{self}}. \tag{11}$$

Only the conditioning history changes across these branches. The prediction target remains a clean recorded future, and gradients do not pass through the sampled suffix. Dream forcing is therefore supervised adaptation to recursive context, not reward optimization.

D Training, Visual Alignment, and Rollout Details

D.1 Supervised World-Model Training

Starting from InfinityStar, we train the autoregressive generator and its action/view-conditioning modules on the curated latent corpus of Appendix A, while keeping the visual tokenizer frozen. Training first uses clean ground-truth contexts and the next-scale likelihood in Eq. (2). The initial observation supplies an action-free anchor; each future clip is paired with its own control chunk. Unreliable action streams are omitted as conditions without removing valid visual supervision, and unavailable views are masked from the likelihood.

Long-horizon adaptation continues on contiguous multi-clip windows with bounded time–scale memory. Clean histories are mixed with age-aware corrupted histories and detached suffixes sampled from a lagged streaming model. Supervision remains the clean recorded future, as specified in Appendix C. When the auxiliary action branch is trained, only paired teacher-forced transitions contribute action-target supervision (Appendix H).

D.2 Scorers and Visual-Policy Optimization

Before visual-policy optimization, the action-following scorer is trained by contrasting matched controls with temporally shifted, reversed, or cross-trajectory actions. The long-horizon scorer contrasts real contiguous windows with spliced, drifted, or view-desynchronized sequences. Both are then frozen. They assess visual-dynamics fidelity; task progress and terminal success are not reward targets.

At each alignment iteration, a just-frozen behavior copy samples $K \geq 2$ fresh visual rollouts from a common initial observation or committed stream state under the same instruction and controls. The fixed sampling transform $\mathcal{S}_\eta$ defines the policy probabilities used for generation and optimization. Decoded rollouts receive the clip rewards in Eq. (5). Returns and detached group-relative advantages are

$$G_{i,c} = \sum_{u=c}^{C} \gamma^{u-c} r_{i,u}, \qquad \widehat{A}_{i,c} = \mathrm{sg}\left[\frac{G_{i,c} - \mu_c}{\sigma_c + \epsilon}\right], \tag{12}$$

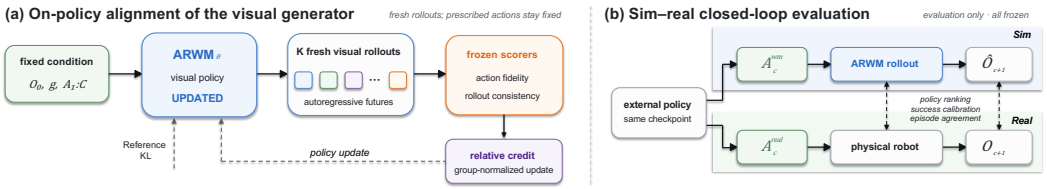


Figure 7: **Visual alignment versus frozen-policy evaluation.** Alignment samples visual futures under fixed prescribed controls and updates only the visual generator and its conditioning adapters. Evaluation runs the same frozen external policy independently in the learned world and on the robot from matched initial conditions.

where μ_c and σ_c are the mean and standard deviation across the K rollouts at clip c. This baseline requires no learned critic. For sampled token $z_{i,n}$ with causal context $\xi_{i,n}$, the clipped policy-gradient objective is

$$\rho_{i,n}(\theta) = \frac{\pi_\theta^V(z_{i,n} \mid \xi_{i,n})}{\pi_{\theta_{\text{old}}}^V(z_{i,n} \mid \xi_{i,n})}, \qquad \mathcal{L}_{\text{PG}} = -\mathbb{E}_{\text{bal}}\left[\min\left(\rho_{i,n}(\theta)\widehat{A}_{i,c}, \text{clip}(\rho_{i,n}(\theta), 1-\varepsilon, 1+\varepsilon)\widehat{A}_{i,c}\right)\right]. \tag{13}$$

$\mathbb{E}_{\text{bal}}$ uses the same clip–scale balancing principle as the supervised objective. All next-scale factors participate with balanced token normalization. Rewards and advantages are detached; gradients pass through the sampled visual-token log-probabilities.

The complete loss in Eq. (6) adds KL regularization to a frozen reference generator and autoregressive replay of real trajectories. Only the visual generator and its conditioning adapters update. The tokenizer, scorers, reference generator, and any action-producing module remain fixed. Fresh rollout batches replace old ones after a bounded number of policy updates.

D.3 Streaming and Closed-Loop Inference

Action-conditioned rollout. Given an initial synchronized observation, instruction, and control chunks, the model constructs the identity anchor once and predicts clips sequentially. Each step encodes the current controls, samples synchronized views in next-scale order, commits the completed bundle, and advances the bounded memory for the next step. Independent action-conditioned futures can be generated by forking the same committed state with different controls or sampling seeds.

Frozen-policy evaluation. The external policy consumes the latest generated multi-view observation and task instruction and returns a control chunk. An action bridge resamples this chunk to the training control rate and end-effector representation. The world model then generates synchronized head and wrist views, commits their states, and returns the new observation to the policy. After initialization, no ground-truth observation or recorded action is injected: controls come from the frozen policy. Both the policy and world model remain fixed during evaluation. This protocol tests predicted outcomes rather than optimizing the controller or visual generator.

E Generation Benchmark and Evaluation Protocols

E.1 Benchmark Construction

The embodied-generation benchmark is drawn from the heterogeneous data mixture and covers objects, scenes, viewpoints, embodiments, action verbs, and recovery behaviors. It contains 200 ID and 100 OOD tasks. The OOD conditions target linguistic and compositional changes: novel object–verb pairings, paraphrased instructions, unseen scene configurations, and composed tasks. Synchronized controls accompany 80% of evaluation samples; the other 20% use text and an initial image.

The aggregate comparison in Tab. 1 reports scores for InfinityStar, Wan2.1-1.3B, Wan2.2-14B, CogVideoX-5B, Cosmos3-Nano, and SWAN. Visual, motion, temporal/semantic, and physi-

cal/conditional metrics follow the stated WorldArena protocol (Shang et al., 2026). A missing Action Following entry is an unreported result, not a zero score. Aggregate scores alone do not isolate individual architecture or training contributions, and do not establish an OOD-only improvement.

E.2 Action-Conditioned Evaluation

Action-conditioned evaluation supplies a synchronized trajectory over the prediction horizon. *Action Following* assesses whether distinct control conditions lead to distinguishable futures from the same visual context. *Trajectory Accuracy* measures the spatial and temporal agreement of generated robot motion with its reference trajectory. The former tests sensitivity to the chosen action branch; the latter tests whether the resulting motion matches that branch. Neither metric is a task-success score.

Action-branch visualizations fork the same initial observation and change the prescribed controls (Fig. 3). This model-side intervention differs from physical rollback data, whose initial states are only approximately matched (Appendix A.2). Keeping these settings separate avoids interpreting a collection-time reset as an exact physical counterfactual.

E.3 Relation to Long-Horizon and Closed-Loop Tests

Streaming evaluation in Sec. 4.3 extends prediction up to 60 seconds and examines trajectory, state, boundary, perceptual, and cross-view consistency. Its time-resolved diagnostics are separate from the aggregate generation table. The two-stage comparison in Sec. 4.4 evaluates the effect of visual alignment.

In Sec. 4.5, frozen external policy checkpoints interact independently with the world model and the physical robot under matched initial configurations. Generated observations, rather than recorded future observations, drive the model-side policy loop (Appendix D.3). The auxiliary control experiment in Sec. 4.6 instead deploys SWAN's action expert on the physical platform. These are different tests: agreement of policy outcomes evaluates the simulator; action-expert deployment evaluates the control extension.

F Extended Generation Results and Diagnostics

F.1 Remaining Aggregate Metrics

Table 4 supplies the eight metrics omitted from Table 1; together they retain all 14 metrics for the six original models and the no-action-condition variant without repeating columns. Baselines are InfinityStar (Liu et al., 2025), Wan2.1 (Team Wan et al., 2025), Wan2.2 (Wan Team, 2025), CogVideoX (Yang et al., 2025), and Cosmos3-Nano (NVIDIA, 2026).

Model	Aesthetic quality	Dynamic degree	Flow score	Background consist.	Semantic align.	Perspectivity	Instruction following	Depth accuracy
InfinityStar	0.321	0.253	0.044	0.879	0.860	0.768	0.406	0.749
Wan2.1-1.3B	0.345	0.329	0.098	0.650	0.856	0.858	0.380	0.715
Wan2.2-14B	0.337	0.427	0.181	0.862	0.836	0.836	0.394	0.657
CogVideoX-5B	0.429	0.308	0.143	0.823	0.890	**0.900**	0.380	0.774
Cosmos3-Nano	0.439	0.425	0.208	0.897	0.903	0.874	0.410	0.816
SWAN w/o action conditioning	0.441	0.418	0.188	0.898	0.901	0.876	0.448	0.792
SWAN	**0.453**	**0.435**	**0.213**	**0.903**	**0.912**	0.885	**0.471**	**0.831**

Table 4: **Additional generation metrics.** All columns ↑; bold marks the largest reported value. The six main-table metrics and these eight columns jointly cover the complete benchmark.

Action-conditioning ablation. The lower blocks of Tables 1 and 4 compare the no-action-condition variant with the full model across all 14 metrics. The largest absolute changes are Action Following ($0.019 \rightarrow 0.290$) and Trajectory Accuracy ($0.418 \rightarrow 0.539$), while Image Quality

changes from 0.683 to 0.697. Thus, visual plausibility alone does not establish that predictions respect the evaluated controls. The comparison concerns the availability of action information, not the superiority of scale alignment over another action-injection architecture. Retraining without actions and masking them only at inference test different questions and must not be conflated.

F.2 Supplementary Visual Comparisons

Figure 8 shows 5-, 30-, and 60-second snapshots for object sorting and pouring. It is a qualitative comparison, not a controlled ablation of one architectural component.

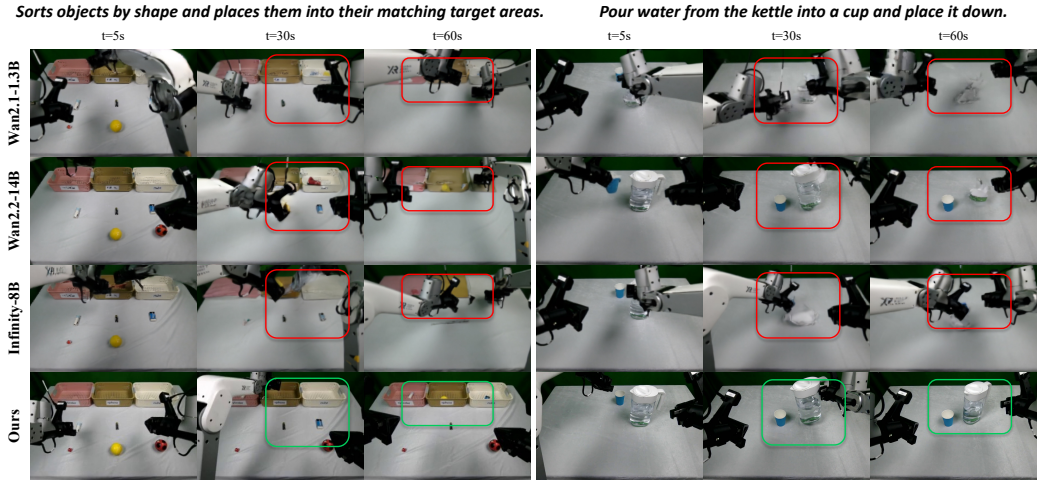


Figure 8: **Supplementary long-horizon visual comparison.** Sorting and pouring at 5, 30, and 60 seconds. Rows show Wan2.1-1.3B, Wan2.2-14B, Infinity-8B, and SWAN, as labeled in the visualization.

Figure 9 illustrates a single continuous pouring rollout. The displayed path error is an image-plane diagnostic of this example, not a guarantee across the test distribution.

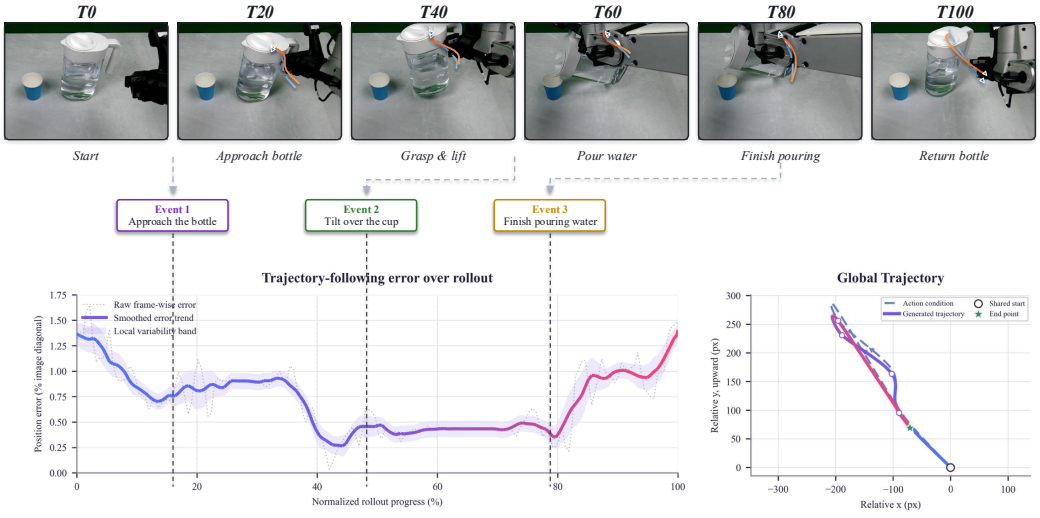


Figure 9: **Continuous action-following example.** Keyframes and tracked end-effector paths from one rollout. Its displayed error remains below 0.5% of the image diagonal during the shown pouring phase; aggregation across episodes is separate from this illustration.

F.3 Long-Horizon Metrics and Variants

Streaming starts from the same synchronized initial views and uses the same prescribed controls for all variants. No real future frame is fed back. The *recent-clip-only* variant drops distant records; *full-history KV* retains historical states at uniform precision; and *w/o dream forcing* retains the inference memory but removes self-generated-context training. This comparison does not separately quantify the effect of corrupted-history training.

Trajectory error tracks generated end effectors and compares them with the reference path using normalized dynamic time warping. Boundary error uses robot- and object-centric motion tracks around clip commits. Persistent-state consistency evaluates robot pose, manipulated objects, contact relations, and task-relevant object state over successive windows. These diagnostics should not be equated with the aggregate Trajectory Accuracy score. The full-history reference is strongest in the plotted dynamics metrics. A bounded allocation alone does not establish lower measured peak device memory or latency at equal quality.

F.4 Perceptual Quality over Time

PSNR uses temporally aligned generated–reference pairs. FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018) are computed in successive three-second windows on the common evaluation set; plotted curves use a three-window moving average, whereas aggregate comparisons use the underlying window scores. The curves favor SWAN over the plotted baselines, but windowed visual fidelity does not establish correct contact dynamics or calibrated policy outcomes.

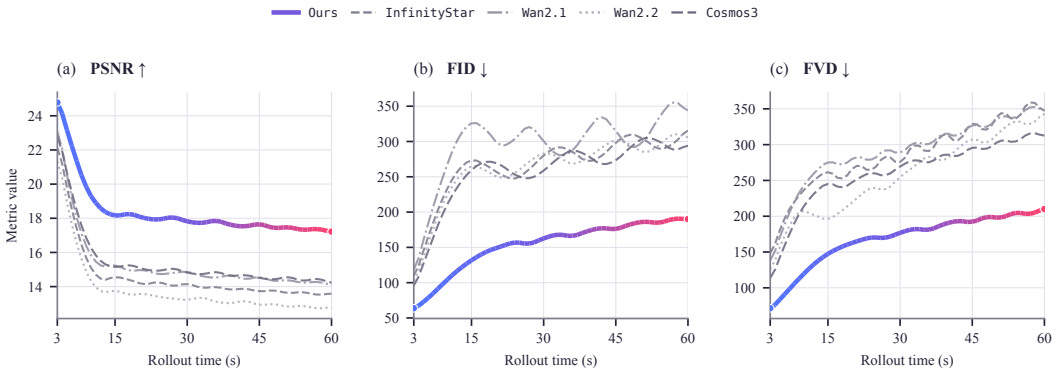


Figure 10: **Windowed perceptual diagnostics.** PSNR (↑), FID (↓), and FVD (↓) over 60-second rollouts. Curves are smoothed for display; raw-window scores determine aggregate comparisons.

F.5 Cross-View Diagnostics

Synchronized triplets consist of a head camera and two wrist cameras. A frozen linear probe evaluates recoverable 3D structure, while calibrated point matches measure cross-view correspondence error in pixels. Each plotted point aggregates valid synchronized windows. The favorable location of SWAN is descriptive evidence under this probe protocol, not a proof that every generated camera view represents a single consistent 3D state.

F.6 Scope of the Visual-Alignment Comparison

Table 2 compares complete supervised and aligned checkpoints, not individual reward or regularization terms. Time-resolved state diagnostics (Fig. 4) test memory variants and should not be treated as an isolated alignment ablation. No independent reward-term, KL-only, or replay-only effect is inferred from these comparisons. Action-information ablation is reported separately in Appendix F.1, not inferred from the supervised/aligned checkpoint pair.

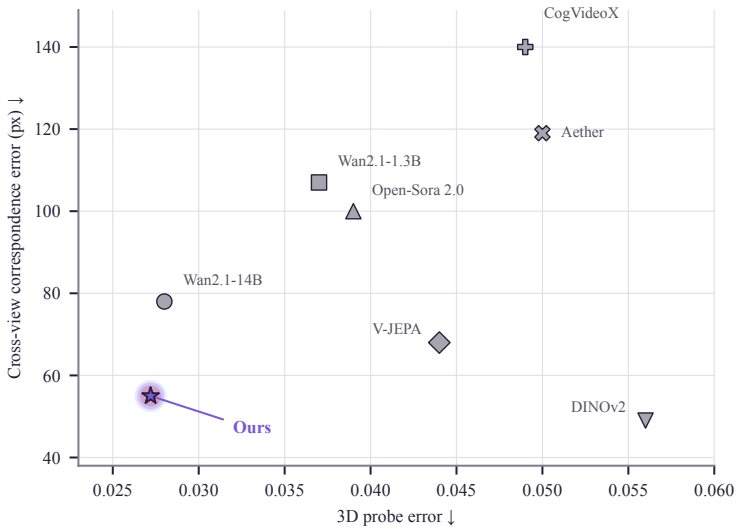


Figure 11: **Cross-view representation diagnostic.** Probe error and correspondence error (both ↓). The displayed model set differs from the generation benchmark. This is a representation-level diagnostic, not a matched comparison of video-generation systems.

G Closed-Loop Policy Evaluation Details

G.1 Tasks and Matched Executions

Five frozen manipulation policy checkpoints, spanning validation competence, are evaluated on six held-out tasks with 20 matched initial configurations per task. This produces 30 task–checkpoint cells and 600 generated–real episode pairs. Each environment feeds its own observations to the same policy checkpoint; matching the initial state does not force subsequent actions to be identical. The generated branch uses the interface of Appendix D.3 without real-future feedback.

Task	Description	Ordered subtasks
T1	Put glasses on shelf	Grasp glasses; place upright on shelf
T2	Put fruits in basket	Grasp fruit in order; place fruit in basket in order
T3	Pour water into cup	Grasp bottle; align with cup; pour; set bottle down
T4	Put ring on rod	Grasp ring; align with rod; slide ring onto rod
T5	Put item in drawer	Open drawer; grasp item; place item; close drawer
T6	Put garment in hamper	Grasp garment; move over hamper; place garment

Table 5: **Closed-loop tasks.** Task-specific ordered subtasks for the six-task evaluation. Common-stage diagnostics are reported separately in Fig. 12.

Blinded annotators label episode outcomes and ordered-subtask completion. Unrecoverable generation artifacts count as failures rather than being filtered. Final episode outcomes and partial subtask completion are distinct labeling targets.

G.2 Calibration and Checkpoint Selection

Let $S^{\text{real}}_{q,p}$ and $S^{\text{wm}}_{q,p}$ be empirical success rates for task q and checkpoint p. Calibration aggregates 30 cells; ranking compares checkpoints within tasks, avoiding confounding by task difficulty. Selection regret is

$$R_q = \max_p S^{\text{real}}_{q,p} - S^{\text{real}}_{q,\hat{p}_q}, \qquad \hat{p}_q = \arg\max_p S^{\text{wm}}_{q,p}. \tag{14}$$

The generated outcomes choose the best checkpoint on four of six tasks; mean regret is 0.025 and the worst-task value is 0.10. We do not interpret nonzero selection regret as statistically equivalent to zero merely because there are only 20 trials per cell.

Level	Metric	Estimate	Reported 95% CI
Calibration	Success-rate MAE	0.062	$[0.06, 0.11]$
	Signed bias	$+0.028$	$[-0.01, 0.06]$
	Slope	0.84	$[0.65, 0.97]$
	Pearson correlation	0.93	–
Within-task ranking	Pairwise score	0.89	$[0.76, 0.95]$
	Mean selection regret	0.025	–
	Mean Spearman	0.88	–
Episode agreement	Balanced accuracy	0.88	$[0.83, 0.93]$
	Success recall	0.90	–
	Failure recall	0.86	–
	False-positive rate	0.136	$[0.09, 0.19]$
	MCC	0.76	–

Table 6: **Complete closed-loop summary.** Calibration, within-task ranking, and episode agreement, with reported 95% confidence intervals. Dashes denote intervals not reported. FPR is the fraction of real failures predicted as successes.

The pairwise score compares checkpoints within tasks after excluding tied real success rates, giving 59 eligible pairs. Reported intervals use stratified bootstrap. The score depends on predicted-tie handling and aggregation weights; uncertainty also depends on the paired outcomes and shared initial configurations across checkpoints.

G.3 Episode Agreement and Residual Bias

The confusion matrix contains 240 true successes, 28 missed successes, 45 false successes, and 287 true failures, totaling 600 pairs. These counts are consistent with rounded success recall 0.90, failure recall 0.86, balanced accuracy 0.88, and FPR $45/332 \simeq 0.136$. Agreement rates do not by themselves establish annotation reliability or independence across paired initial configurations.

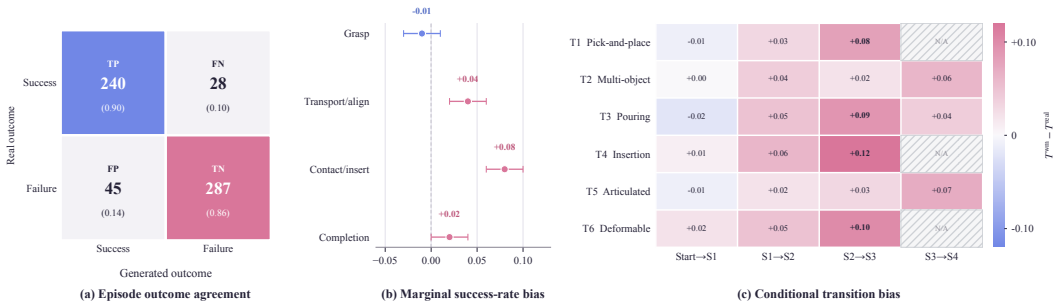


Figure 12: **Outcome and stage-level discrepancies.** Left: the 600-pair confusion matrix, with 527 matching outcomes. Middle: marginal model-minus-real stage success-rate bias. Right: conditional transition biases; unavailable task–transition combinations are hatched.

Marginal bias is -0.01 at grasp, $+0.04$ at transport/alignment, $+0.08$ at contact, and $+0.02$ at completion. The displayed conditional insertion bias reaches $+0.12$. These describe where optimism occurs, rather than proving a specific contact-dynamics cause. Marginal completion bias and whole-episode signed bias need not be identical unless their eligible samples and aggregation weights coincide.

H Auxiliary Action-Expert Details

Causal control interface. Before chunk c, the co-trained action expert reads

$$H_c^A = F_\omega^A(\mathrm{Read}_V(\mathcal{M}_c), E_s(s_c), T_g)\,, \tag{15}$$

where s_c is the current proprioceptive or plan-side state and Read_V exposes only committed visual states. Following the flow-matching action-expert formulations, noisy action slots cross-attend to this context to predict the executable control chunk A_c. The mask excludes the current visual target, current action condition, and all future records.

Executable actions and visual action conditions remain distinct. Predicted and externally supplied controls use the same grounding map $\widehat{\mathbf{X}}^A_{c,j} = \Gamma_j(\widehat{A}_c, s_c)$; the visual model never reads private expert states. Recursive imagination alternates action prediction, visual generation, and memory commit. Physical deployment updates s_c from measured proprioception.

Paired supervision. On paired teacher-forced transitions, the visual and action branches read the same pre-action memory $\mathcal{M}_c$. The demonstrated control is both the expert target and the condition for its recorded visual consequence. The joint objective is

$$\mathcal{L}_{\mathrm{cotrain}} = \mathcal{L}_{\mathrm{long}} + \lambda_A \mathcal{L}_{\mathrm{act}}. \tag{16}$$

$\mathcal{L}_{\mathrm{act}}$ follows the cited action-expert objective and is applied only to paired teacher-forced histories. Unpaired histories retain the visual loss alone; self-generated visual histories do not acquire an unverified demonstrated-action target. The visual alignment phase freezes all action-producing modules, so its reward objective does not directly optimize this controller.

H.1 Real-Robot Deployment and Results

The execution loop, seven tasks, and progress metric are given in Sec. 4.6; Fig. 6 compares external baselines, and Table 3 gives means. Best-baseline scores for headphone sorting, shape classification, and ordered button pressing are 66, 55, and 31; $\pi_{0.5}$ is stronger on ring insertion.

H.2 Real-Robot RL Ablation

Table 7 gives the seven-task scores. Unweighted means round to 67.4 and 69.1; full-model gains are 1, 2, 2, 1, 1, 2, and 3 points, averaging $12/7 \simeq 1.7$ progress points, not success-rate percentage points. Positive task-wise differences alone do not establish statistical significance.

Model	Cup	Spoon	Ring	Basket	Headphone	Shape	Button	Mean ↑
SWAN w/o RL	52	56	59	97.5	69	78	60	67.4
SWAN	**53**	**58**	**61**	**98.5**	**70**	**80**	**63**	**69.1**

Table 7: **Per-task real-robot RL ablation.** Task Progress (0–100; all columns ↑). Task names follow Sec. 4.6. Means are unweighted across seven tasks and rounded to one decimal.

The comparison concerns complete control systems. The reported means alone do not isolate visual alignment from differences in data, compute, initialization, or action-expert training.