WORLDRIDE: EFFICIENT REWARD-GUIDED DISTILLATION FOR AUTOREGRESSIVE WORLD MODELS

Anonymous authors
Paper under double-blind review

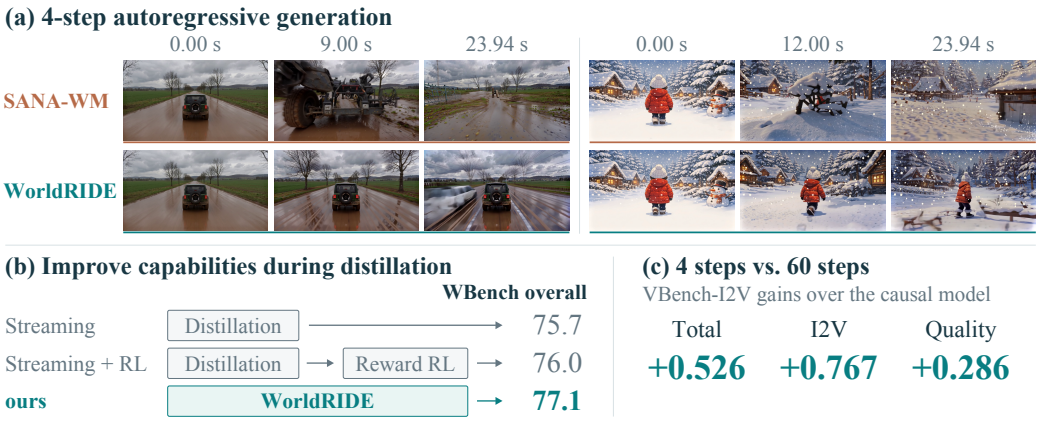

Figure 1: **WorldRIDE is based on SANA-WM and improves four-step autoregressive world models.** (a) frames from two 24-second rollouts show better performance than SANA-WM Streaming. (b) WorldRIDE achieves a WBench Overall score of 77.1 without separate reward post-training, compared with 75.7 for Streaming and 76.0 for Streaming followed by WorldCompass-style RL. (c) The four-step student outperforms the 60-step SANA-WM Chunk-Causal model on VBench-I2V Total, I2V, and Quality.

ABSTRACT

Autoregressive world models must combine low-latency generation with visual fidelity and reliable action control. Few-step distillation reduces sampling cost, but treating reward alignment as a subsequent post-training stage introduces another optimization loop. We present WorldRIDE, an efficient reward-guided distillation framework that improves a four-step autoregressive student during distillation itself, without additional post-distillation training. WorldRIDE combines three designs. Ordered reward-guided distillation gates action preferences by group-level visual quality and permits geometric refinement only among action-safe, action-matched candidates. Multi-anchor rollout reuses a single autoregressive trajectory to construct local candidate groups at multiple temporal anchors, amortizing prefix generation across training targets. Locally self-normalized distillation converts within-group rewards into detached, unit-mean weights on the distribution-matching update, retaining unweighted distillation when the quality gate fails. On WBench, WorldRIDE achieves an Overall score of 77.1. On VBench, our four-step model outperforms a 60-step causal model on all three aggregate scores, improving Total, I2V, and Quality by 0.526, 0.767, and 0.286 points, respectively. These results show that reward-guided distillation can improve world-model quality while reducing denoising steps, without a separate stage to recover performance after distillation.

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.

1 INTRODUCTION

Interactive video world models continually generate future observations conditioned on visual history and user controls, providing a generative approach to interactive virtual environments. Models such as LingBot-World (Robbyant Team et al., 2026) and Matrix-Game 2.0 (He et al., 2025) support long-horizon video generation under action controls, including camera movement and keyboard–mouse inputs. Unlike generating a complete video in a single pass, interactive generation requires timely responses to new controls while keeping subsequent observations consistent with the existing scene. Each generated clip is both the current output and the history for the next prediction, making visual quality, action following, and world-state consistency jointly important. Reducing generation cost while retaining these capabilities is a central challenge for streaming autoregressive world models.

Few-step distillation provides a path to efficient streaming generation. Distribution matching distillation (DMD) transfers the generative capabilities of a multi-step diffusion model to a few-step student through the difference between teacher and student distribution scores (Yin et al., 2024). CausVid extends this approach to autoregressive video generation through asymmetric distillation from a bidirectional teacher to a causal student (Yin et al., 2025). Self Forcing trains on student-generated histories to reduce the mismatch between training and inference conditions (Huang et al., 2025), while Causal Forcing and Causal Forcing++ improve initialization for causal generation (Zhu et al., 2026b; Zhao et al., 2026). Together, these methods establish a practical route from multi-step video generators to efficient autoregressive students.

Recent streaming world models transfer this video-generation training recipe to interactive settings. DreamX-World combines causal distillation with long self-rollouts and local training windows (DreamX Team et al., 2026), while AlayaWorld adopts standard DMD with small temporal chunks (AlayaWorld Team et al., 2026a). WorldPlay and BiWM retain the distribution-matching backbone while adapting context alignment and distribution regularization, respectively (Sun et al., 2026; Rui et al., 2026). Their distillation objectives remain centered on matching generative distributions. For an interactive world model, compression must also preserve the ability to execute new controls and maintain a coherent 3D scene under student-generated histories. This calls for distillation supervision that directly evaluates interactive capabilities alongside visual quality.

Reward feedback provides a direct signal for capability optimization. WorldCompass uses clip-level reinforcement learning to improve action following and visual quality in world models (Wang et al., 2026b). Astrolabe optimizes already-distilled autoregressive video models with online reinforcement learning (Zhang et al., 2026), and DreamX-World uses post-distillation reinforcement learning to recover visual quality and camera-control performance (DreamX Team et al., 2026). Separating compression from capability alignment introduces another cycle of rollout collection and parameter updates. Reward Forcing (Lu et al., 2026) and Reward-Aware Trajectory Shaping (Li et al., 2026) instead incorporate reward feedback into video distillation through update reweighting and trajectory supervision. For streaming world models, this direction calls for a unified design of task rewards and rollout collection: the rewards must capture the dependencies among visual quality, action execution, and geometric consistency, while the rollouts must supply comparable training signals efficiently.

Our central insight is that world-model rewards have a natural dependency structure. Visual quality establishes whether a clip is usable for assessing control, and geometric refinement becomes meaningful when the requested action has been executed. A nearly static scene, for example, can score well on geometric consistency while ignoring a camera-motion command. Treating these objectives as independent rewards obscures this distinction. We therefore order reward preferences by their conditions of validity and compare candidates under a shared prefix and control condition. Sharing the prefix makes each comparison local to the same world state; reusing one autoregressive trajectory across multiple such comparisons makes this supervision efficient to collect.

We introduce **WorldRIDE**, an efficient reward-guided distillation framework for streaming autoregressive world models. WorldRIDE integrates visual quality, action following, and 3D consistency into the distillation process through ordered local reward comparisons. A shared autoregressive trajectory supplies multiple training targets, and locally normalized reward weights guide their distribution-matching updates. Built on SANA-WM (Zhu et al., 2026a), WorldRIDE produces a

four-step student whose capabilities are optimized during distillation itself, without a separate reward post-training stage.

Specifically, **ordered reward-guided distillation** expresses the conditional dependencies among visual quality, action following, and 3D consistency through cascaded gates. Action preferences are enabled only when every candidate in a group passes the quality gate. Geometry rewards are further enabled when action performance is sufficient and matched across candidates, and geometric evidence is valid. This grounds geometric refinement in local comparisons that satisfy the action requirements, reducing opportunities to obtain high geometry scores by ignoring the action. When the quality gate fails, the entire group retains the unweighted local DMD update.

To improve the sampling efficiency of these local comparisons, we introduce **multi-anchor rollout**. Multiple temporal anchors are selected along a student-generated autoregressive spine, and local candidates are sampled with different noise from each anchor's shared prefix. Each branch generates only its target clip, allowing multiple training targets to reuse the spine's prefix computation. Our **locally self-normalized strategy** then converts within-anchor rewards into detached exponential weights with unit group mean and applies them directly to the DMD update. Every anchor retains the same total candidate weight, preventing differences in absolute reward baselines from changing the total weight assigned to different groups while preserving within-group preferences.

On WBench (Ying et al., 2026) and VBench (Huang et al., 2026), WorldRIDE combines few-step generation with improved performance. Our model achieves an Overall score of **77.1** on WBench. On VBench, the four-step student **outperforms the 60-step causal model on all three aggregate scores**, improving Total, I2V, and Quality by 0.526, 0.767, and 0.286 points, respectively. These gains require no additional post-distillation training, showing that sampling compression and improved generation can be achieved within the same distillation process.

Our contributions are threefold:

- **WorldRIDE distillation paradigm.** We propose WorldRIDE, a new distillation paradigm for interactive streaming world models. Through ordered reward-guided distillation, it organizes visual quality, action following, and 3D consistency into a conditionally activated reward hierarchy, directly guiding world-model capability optimization during distillation.
- **Multi-anchor rollout.** We construct local candidate groups at multiple temporal locations along one shared autoregressive spine, reusing prefix generation to improve rollout utilization.
- **Locally self-normalized distillation.** We turn relative within-group rewards into unit-mean distillation weights that directly modulate the DMD update while preserving the total weight assigned to each group.

2 Related Work

2.1 Autoregressive Video Distillation

Autoregressive video distillation combines few-step sampling with causal temporal generation. Distribution matching distillation (DMD) learns fast students through score-based distribution matching (Yin et al., 2024). CausVid uses asymmetric distillation to transfer a bidirectional teacher's capabilities to a causal video student (Yin et al., 2025), while Self Forcing trains under student-generated histories to align training with autoregressive inference (Huang et al., 2025). Causal Forcing and Causal Forcing++ further develop initialization for causal generation (Zhu et al., 2026b; Zhao et al., 2026).

Reward feedback provides an additional signal for optimizing fast video generators. Astrolabe uses online reinforcement learning to refine already-distilled autoregressive students (Zhang et al., 2026), whereas Reward-Aware Trajectory Shaping incorporates reward feedback into trajectory-alignment supervision for few-step generation (Li et al., 2026). Similar to Reward Forcing (Lu et al., 2026), WorldRIDE integrates rewards into distillation by reweighting DMD updates with detached reward weights.

2.2 Interactive Streaming World Models

Interactive streaming world models generate evolving environments under continuous user control. LingBot-World and Matrix-Game 2.0 support long-horizon video generation driven by camera motion and keyboard–mouse inputs (Robbyant Team et al., 2026; He et al., 2025). Memory mechanisms in WorldPlay and AlayaWorld support environment consistency across extended interactions (Sun et al., 2026; AlayaWorld Team et al., 2026a), while SANA-WM uses a hybrid linear diffusion Transformer for efficient minute-scale world modeling (Zhu et al., 2026a).

Streaming generation commonly builds on video-model distillation, including memory-conditioned DMD in WorldPlay (Sun et al., 2026) and distribution matching with supervised regression and forward-KL anchors in BiWM (Rui et al., 2026). To optimize interactive capabilities, WorldCompass employs clip-level reinforcement learning for action following and visual quality (Wang et al., 2026b), while DreamX-World applies post-distillation reinforcement learning to recover visual quality and camera control (DreamX Team et al., 2026). WorldRIDE brings visual quality, action following, and 3D consistency directly into distillation through an ordered reward hierarchy, coupling sampling compression with interactive capability optimization without an additional reward post-training stage.

3 Method

Our goal is to distill a pretrained camera-controlled world model into a four-step autoregressive student G_θ for streaming interaction. Given an initial frame I_0, a static scene caption c, and camera controls $a_{1:N}$, the student generates each video clip v_n according to $p_\theta(v_n \mid v_{<n}, I_0, c, a_{1:n})$, where $v_{<n}$ denotes previously generated clips. We optimize visual quality, action following, and 3D consistency during distillation, without a separate post-distillation training stage. Figure 2 illustrates WorldRIDE, comprising ordered reward guidance (Section 3.2), multi-anchor rollout (Section 3.3), and locally self-normalized distillation (Section 3.4).

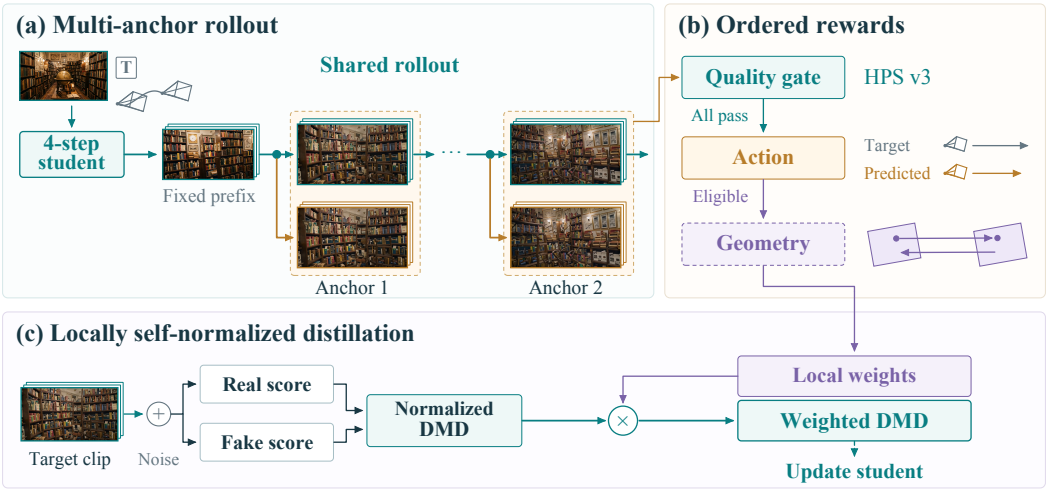

Figure 2: **WorldRIDE overview.** (a) Local candidates branch from fixed prefixes along a shared autoregressive rollout. (b) Group-level quality gating enables action rewards. Shared DA3 predictions provide action feedback and geometric refinement for action-safe, action-matched groups with valid evidence. (c) Detached, unit-mean local weights modulate normalized DMD updates. Quality failures use unweighted DMD; ineligible geometry retains action guidance. Video thumbnails illustrate the data flow.

3.1 Preliminary

SANA-WM (Zhu et al., 2026a) is a camera-controlled world model that combines high-compression video latents, a hybrid diffusion Transformer with frame-wise Gated DeltaNet (GDN) and softmax

attention, and dual-branch camera conditioning. Its Chunk-Causal variant carries the forward GDN state across chunks while restricting the backward scan to the current chunk, providing within-chunk context without access to future chunks. We use the multi-step Chunk-Causal model for ODE initialization, following Causal Forcing (Zhu et al., 2026b), and the frozen bidirectional model as the real-score teacher for DMD. Subsequent distillation follows the student-generated-prefix training of Self Forcing (Huang et al., 2025), with four denoising steps per generated clip.

3.2 Ordered Reward Guidance

Visual quality, action following, and geometry form an ordered dependency: degraded clips provide unreliable control estimates, while nearly static outputs can score well on geometry without executing the requested motion. At temporal anchor k, we compare $M \geq 2$ candidates indexed by j, sharing a generated prefix, control conditions, and sampling exit step u_k (Section 3.3).

Frozen HPSv3 (Ma et al., 2025) defines the group quality gate $g_Q^{\mathrm{grp}}(k)$, which passes only when every candidate meets both lower-quantile and minimum frame-score thresholds. For passing groups, one frozen Depth Anything 3 (DA3) (Lin et al., 2026) forward pass provides camera, depth, and confidence estimates. The action reward $r_A(k, j)$ combines action-label correctness with a motion-sufficiency score, or a stillness score for stationary commands. The geometry reward $r_G(k, j)$ measures bidirectional depth-reprojection consistency. Its group eligibility flag $v_G^{\mathrm{grp}}(k)$ requires correct actions, sufficient and similar action scores, and valid geometric evidence for every candidate. We combine them as

$$A_{k,j}^{\mathrm{ord}} = g_Q^{\mathrm{grp}}(k)\, r_A(k, j)\, (1 + \lambda_G v_G^{\mathrm{grp}}(k) r_G(k, j)), \qquad \lambda_G \geq 0. \tag{1}$$

A failed quality gate sets the group's reward signals to zero, yielding unit weights and unweighted local DMD (Section 3.4). Ineligible geometry leaves action guidance intact. This cascade makes geometry a refinement of action-valid comparisons rather than an independent objective that can favor static outputs. Appendix A details the reward calculations, normalization, and gating conditions through Figure A1.

3.3 Multi-anchor Rollout

WorldRIDE reuses a gradient-free autoregressive spine of N clips to form local candidate groups at K temporal anchors. At anchor k, the M candidates share prefix p_k, text and control conditions c_k, cached state, exit step u_k, and sampling schedule, differing only in generation noise. The spine continuation supplies the first candidate; the remaining $M - 1$ candidates branch from the same prefix and terminate after the target clip. Later anchors follow the original spine, so local rewards do not alter other anchors' prefix distributions. This yields KM training targets from $N + K(M - 1)$ clip-sampling operations; reward evaluation and optimization incur additional computation.

We evaluate rewards on decoded, detached candidates $\bar{x}_{k,j}$. To optimize the same samples, we save each target's exit-step input and context, then recompute that step with gradients to obtain $x_{k,j}^{\theta}$. Student parameters remain fixed throughout candidate generation and recomputation, which matches the original target within numerical tolerance. Gradients pass through the target exit step only. Algorithm 1 in Appendix B details cache reuse and update ordering.

3.4 Locally Self-Normalized Distillation

We convert ordered rewards into detached weights within each local group:

$$\widetilde{w}_{k,j} = \mathrm{sg}\left[\frac{M \exp(A_{k,j}^{\mathrm{ord}}/\beta_{u_k})}{\sum_{\ell=1}^{M} \exp(A_{k,\ell}^{\mathrm{ord}}/\beta_{u_k})}\right], \qquad \beta_{u_k} > 0. \tag{2}$$

Here, sg denotes stop-gradient and $\beta_{u_k} > 0$ controls preference strength. Unit group mean preserves each anchor's total weight while redistributing it among candidates. The weights are invariant to a common reward shift; equal rewards, including a failed quality gate, yield unweighted local DMD.

Following Reward Forcing (Lu et al., 2026), we apply these weights to DMD updates. Let s_{real} be the frozen teacher score and s_{fake} estimate the student distribution. Their difference on the same

noisy candidate and context defines the detached, normalized direction $d_{k,j}$, restricted to target-clip positions (Appendix B). The diffusion time t is sampled independently of exit step u_k. We apply each reward weight once, after DMD normalization, through the pseudo-target and loss

$$y_{k,j} = \mathrm{sg}\left[x^{\theta}_{k,j} - \widetilde{w}_{k,j} d_{k,j}\right], \tag{3}$$

$$\mathcal{L}_G = \frac{1}{KM} \sum_{k=1}^{K} \sum_{j=1}^{M} \frac{1}{2} \left\| x^{\theta}_{k,j} - y_{k,j} \right\|_2^2 . \tag{4}$$

The resulting gradient scales each candidate's DMD update by $\widetilde{w}_{k,j}$, without differentiating through rewards or score networks. The fake-score network learns from unweighted student samples. We accumulate candidate gradients before updating the student; Algorithm 1 gives the complete generator update. Inference retains only the four-step student and its streaming cache.

4 Experiments

We evaluate four-step WorldRIDE on WBench and VBench-I2V and ablate its reward components.

4.1 Experimental Setup

WorldRIDE uses the SANA-WM backbone and teachers in Section 3.1. Training uses camera-annotated videos from a scene-balanced subset, GameVerse (AlayaWorld Team et al., 2026b), and DL3DV (Ling et al., 2024). Following the initialization setup of CausVid (Yin et al., 2025), we first perform ODE initialization, followed by 3,000 steps of reward-guided DMD training. We compare the resulting four-step student with SANA-WM Streaming (4 steps) and Chunk-Causal (60 steps). WBench uses 158 navigation cases; VBench-I2V uses 1,118 entries with five seeds each. The SANA-WM and WorldRIDE models use the pretrained refiner for WBench and disable it for VBench-I2V. Training details and evaluation protocols are in Appendix F.

4.2 Performance Comparison

Interactive world modeling. Table 1 reports WorldRIDE alongside interactive world models on WBench. Our four-step student achieves an Overall score of 77.1, with Quality 79.5, Setting 77.1, Interaction 82.6, Consistency 82.7, and Physical plausibility 63.6. Appendix E reports all 19 metrics.

Table 1: **WBench comparison on 158 navigation cases.** Overall corresponds to the leaderboard's Average. Higher is better for every column.

Model	Overall↑	Quality↑	Setting↑	Interact.↑	Consist.↑	Physical↑
SANA-WM Chunk-Causal	79.7	80.3	80.7	84.5	86.1	66.8
SANA-WM Streaming	75.7	79.1	77.1	81.9	79.0	61.5
DreamX-World (5B AR)†	75.0	77.5	80.8	78.6	74.9	63.3
Lyra 2.0†	76.4	77.1	73.2	85.6	79.3	66.7
LingBot-World (Fast)†	77.4	79.4	77.9	79.2	84.9	65.7
MatrixGame3.0†	71.3	75.5	63.6	83.6	74.5	59.3
WorldRIDE (ours)	77.1	79.5	77.1	82.6	82.7	63.6

1st 2nd 3rd Top three values per column, including the 60-step teacher; ties share a color.
†: official leaderboard (WBench, 2026). SANA-WM and WorldRIDE use the refiner.

WBench visual comparisons. Figure 3 shows three 24-second rollouts featuring a monitor lizard, a ghost dog, and a rowboat. In the lizard scene, Streaming produces deformed limbs and disconnected body parts; it also duplicates the ghost dog in later frames. WorldRIDE and the 60-step Chunk-Causal reference retain a more coherent lizard and a single ghost dog. The full-frame comparisons show subject appearance and surrounding scene structure at matched times.



Figure 3: **WBench visual comparisons.** Three selected 24-second rollouts from SANA-WM Chunk-Causal (60 steps), SANA-WM Streaming (4 steps), and WorldRIDE (4 steps). Each model row shows four complete frames at matched timestamps across methods. Yellow boxes highlight the lizard and ghost-dog comparison regions. All models use model-specific refiners (Appendix D).

Image-to-video generation. Table 2 evaluates the student outputs without refinement. WorldRIDE attains the highest Total and Quality scores, 88.219 and 79.383, respectively. Its four-step student exceeds the 60-step causal model on all three aggregate scores, improving Total, I2V, and Quality by 0.526, 0.767, and 0.286 points. Relative to SANA-WM Streaming, Total improves by 0.667 points and Quality by 2.008 points, while Streaming retains the highest I2V score.

Table 2: **VBench-I2V results without the refiner.** Each model generates 5,590 videos. Scores are percentages (higher is better); ΔTotal is relative to SANA-WM Streaming, computed before rounding. Bold indicates columnwise best results excluding the gray 60-step teacher reference.

Model	Steps	Quality	I2V	Total	ΔTotal
SANA-WM Chunk-Causal	60	79.097	96.288	87.692	+0.141
SANA-WM Streaming	4	77.375	**97.728**	87.552	+0.000
WorldRIDE (ours)	4	**79.383**	97.055	**88.219**	**+0.667**

4.3 Qualitative Comparisons

Camera response to mixed actions. Figure 4 pairs generated frames with reconstructed motion and input camera controls. The library combines forward motion with turns, and the vineyard follows a four-direction translation sequence. Compared with SANA-WM Streaming, WorldRIDE's recovered trajectories follow the intended direction changes and relative path shape more closely in these selected examples. The 60-step Chunk-Causal model provides an additional reference.

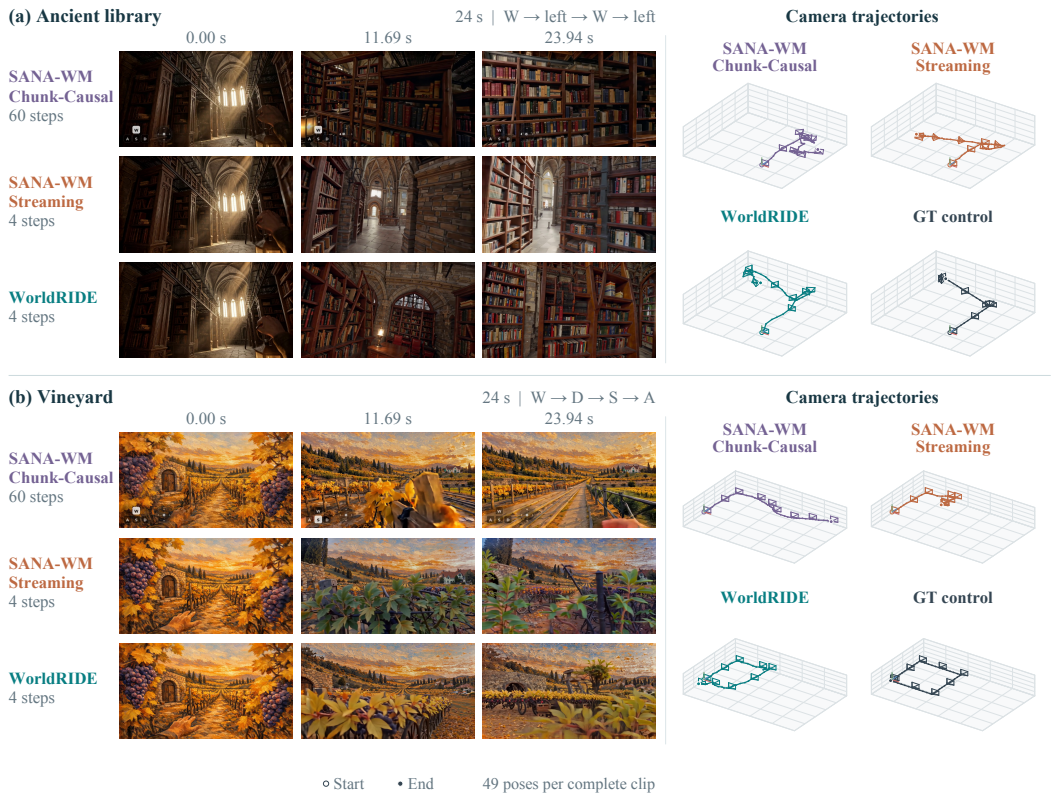

Figure 4: **Generated frames and camera motion under mixed controls.** Left: three matched frames per model from each complete clip. Right: image-only Pi3X trajectories reconstructed from 49 samples, compared with the supplied GT control. Each trajectory uses its first-camera coordinates and one scale factor, with shared projection, bounds, and glyph size within each scene (Appendix D). Open and filled circles mark start and end; nine glyphs show orientation. Models use their native refiners.

4.4 Reward Ablation

Table 3 compares the step-3,000 checkpoints with and without reward components. Full WorldRIDE scores 77.1 Overall, versus 76.7 without Geometry and 76.9 without Action and Geometry. Perspective Consistency is 67.0, compared with 62.6 and 62.7, respectively. Navigation and Geometric Consistency do not improve in this comparison; the observed gains concentrate in viewpoint consistency.

Table 3: **WBench reward-component ablation.** All variants use step-3,000 checkpoints, four-step denoising, seed 42, and the refiner. Selected metrics are shown; higher is better. Bold marks the best value in each column.

Variant	Overall	Quality	Navigation	Perspective Consistency	Geometric Consistency
WorldRIDE (full)	**77.1**	79.5	82.6	**67.0**	87.4
w/o Geometry	76.7	**79.8**	82.7	62.6	**88.3**
w/o Action + Geometry	76.9	79.6	**82.9**	62.7	87.8

Training-stage efficiency. In our comparison, the WorldCompass-style baseline applies 3,000 RL updates to SANA-WM Streaming after its 3,000-step Self Forcing distillation. WorldRIDE combines distillation and reward guidance within 3,000 DMD training steps, avoiding a separate reward post-training stage. Appendix G reports this comparison, refiner ablations, and 60-second generation results. Additional qualitative comparisons appear in Appendix C.

5 Conclusion and Limitations

We introduced WorldRIDE, a reward-guided distillation framework for interactive streaming world models that improves a four-step autoregressive student without a separate reward post-training stage. Ordered reward guidance gates action preferences by visual quality and enables geometric refinement among action-safe, action-matched candidates with valid evidence. Multi-anchor rollout amortizes prefix generation across training targets, while locally self-normalized weights integrate relative reward preferences directly into DMD updates. Built on SANA-WM, WorldRIDE achieves 77.1 WBench Overall and surpasses the 60-step causal model on all three VBench-I2V aggregate scores. Together, these results show that distillation can jointly deliver sampling compression and capability improvement.

Limitations. WorldRIDE has so far been evaluated on SANA-WM, and future work will extend validation to more world models. The current reward design also leaves room for improvement; we will further refine the reward signals and their coordination.

AI Usage Statement

We used AI to populate tables based on experimental results and to draft initial versions of selected sections. After completing the full manuscript, we used AI to refine its grammar and sentence phrasing.

REFERENCES

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Ruicong Liu, Xiaojie Xu, Xuangeng Chu, Zhen Li, Zhengyuan Lin, Zhixiang Wang, Zian Meng, and Zihui Gao. AlayaWorld: Long-horizon and playable video world generation. *arXiv preprint arXiv:2607.06291*, 2026a. URL https://arxiv.org/abs/2607.06291.

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Mingliang Zhai, Ruicong Liu, Xiaojie Xu, Xuangeng Chu, Zhen Li, Zhengyuan Lin, Zhixiang Wang, Zian Meng, and Zihui Gao. AlayaWorld: Interactive long-horizon world modeling – full technical report. *arXiv preprint arXiv:2607.18367*, 2026b. URL https://arxiv.org/abs/2607.18367.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, Geng Li, Jifan Li, Ruimin Lin, Qingfeng Shi, Bingze Song, Lei Sun, Jing Tang, Ruitian Tian, Jun Wang, Jiahong Wu, Pengfei Zhang, Shen Zhang, and Jiashu Zhu. DreamX-World 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026. URL https://arxiv.org/abs/2606.16993.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-Game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025. URL https://arxiv.org/abs/2508.13009.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 167283–167308, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3268–3285, 2026. doi: 10.1109/TPAMI.2025.3633890. URL https://doi.org/10.1109/TPAMI.2025.3633890.

Rui Li, Bingyu Li, Yuanzhi Liang, Haibin Huang, Chi Zhang, and XueLong Li. Reward-aware trajectory shaping for few-step visual generation. *arXiv preprint arXiv:2604.14910*, 2026. URL https://arxiv.org/abs/2604.14910v4.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e4cd50120b6d7e8daff1749d6bbaa889-Abstract-Conference.html.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ling_DL3DV-10K_A_Large-Scale_Scene_Dataset_for_Deep_Learning-based_3D_Vision_CVPR_2024_paper.html.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, Yujun Shen, and Min Zhang. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34385–34397,

2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lu_Reward_Forcing_Efficient_Streaming_Video_Generation_with_Rewarded_Distribution_Matching_CVPR_2026_paper.html.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ma_HPSv3_Towards_Wide-Spectrum_Human_Preference_Score_ICCV_2025_paper.html.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026. URL https://arxiv.org/abs/2601.20540.

Shaohao Rui, Xiaofeng Mao, Zhanyu Zhang, Peijia Lin, Yansong Zhu, Yibo Zhang, Haibin Wan, Zhangrui Zhao, and Weijie Ma. BiWM: Advancing open-source interactive video world models with bidirectional autoregression. *arXiv preprint arXiv:2606.10135*, 2026. URL https://arxiv.org/abs/2606.10135.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. WorldPlay: Towards long-term geometric consistency for real-time interactive world modeling. In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65111.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/11a09e0aaa74867c6b0719c639fc09f8-Abstract-Conference.html.

Zehan Wang, Tengfei Wang, Haiyu Zhang, Xuhui Zuo, Junta Wu, Haoyuan Wang, Wenqiang Sun, Zhenwei Wang, Chenjie Cao, Hengshuang Zhao, Chunchao Guo, and Zhou Zhao. WorldCompass: Reinforcement learning for long-horizon world models. In *Proceedings of the International Conference on Machine Learning*, 2026b. URL https://icml.cc/virtual/2026/poster/60526.

WBench. WBench official leaderboard, 2026. URL https://meituan-longcat.github.io/WBench/. Navigation split (158 cases). Updated September 17, 2026; accessed September 19, 2026.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.07772.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. WBench: A comprehensive multi-turn benchmark for interactive video world model evaluation. *arXiv preprint arXiv:2605.25874*, 2026. URL https://arxiv.org/abs/2605.25874.

Songchun Zhang, Zeyue Xue, Siming Fu, Jie Huang, Xianghao Kong, Y Ma, Haoyang Huang, Nan Duan, and Anyi Rao. Astrolabe: Steering forward-process reinforcement learning for distilled autoregressive video models. *arXiv preprint arXiv:2603.17051*, 2026. URL https://arxiv.org/abs/2603.17051.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026. URL https://arxiv.org/abs/2605.15141.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. SANA-WM: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026a. URL https://arxiv.org/abs/2605.15178.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. In *Proceedings of the International Conference on Machine Learning*, 2026b. URL https://icml.cc/virtual/2026/poster/65646.

A ORDERED REWARD COMPUTATION

Figure A1 expands the ordered guidance in Section 3.2. At anchor k, all M candidates share the generated prefix, caption c, controls, and exit step u_k. Panel (a) checks whether the group supports reward-based comparison; panels (b) and (c) compute action and geometry feedback from shared 3D predictions.

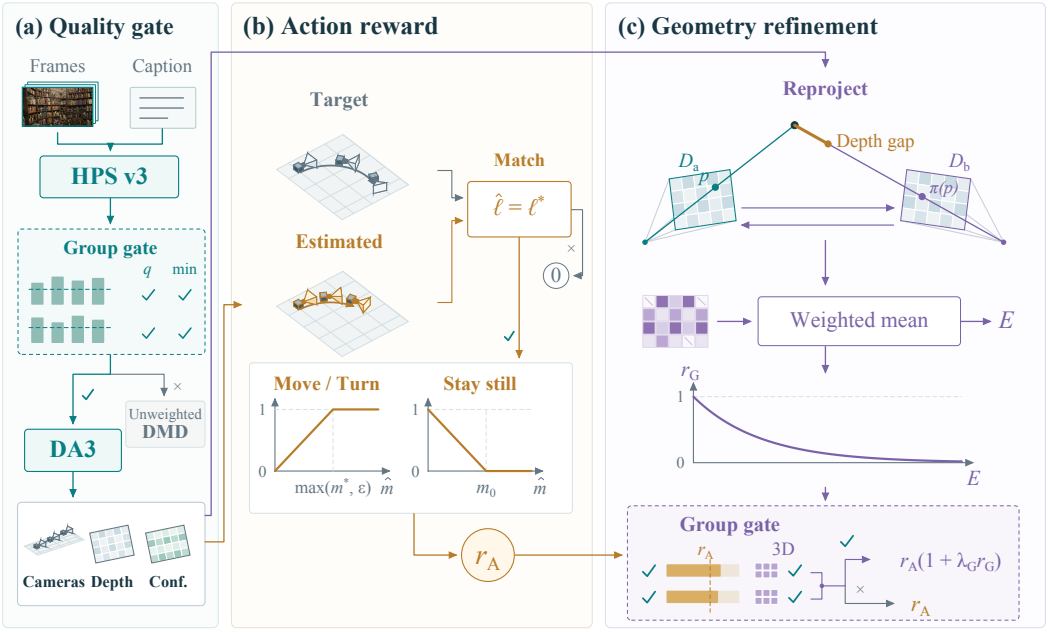

Figure A1: **Ordered reward guidance.** (a) HPSv3 quality gating and shared DA3 predictions. (b) Action-label matching with motion or stillness scoring. (c) Depth-reprojection scoring and group-level geometry eligibility. The flow follows the computations in this appendix. Camera views share scale and projection; curves follow the reward definitions, and other graphics are schematic.

A.1 VISUAL QUALITY GATING

Figure A1(a) checks quality before 3D estimation. Frozen HPSv3 (Ma et al., 2025) scores a fixed set of target frames against caption c. Let $q_{k,j}$ and $q_{k,j}^{\min}$ denote the lower quantile and minimum of candidate j's frame scores. The candidate and group gates are

$$g_Q(k,j) = \mathbf{1}[q_{k,j} \geq \tau_{Q,u_k}]\mathbf{1}[q_{k,j}^{\min} \geq \tau_{\min,u_k}],$$
$$g_Q^{\text{grp}}(k) = \prod_{j=1}^{M} g_Q(k,j). \tag{A1}$$

The lower quantile captures low-quality portions; the minimum rejects severely degraded frames. Thresholds are calibrated per exit step using the frozen ODE-initialized student and held fixed during training. Every candidate must pass, or the group retains unweighted local DMD. Passing groups send the final prefix frame and target clip to DA3, whose camera, depth, and confidence predictions are shared by action and geometry scoring.

A.2 MAGNITUDE-WEIGHTED ACTION REWARD

Figure A1(b) compares the target action label ℓ_k^* and normalized magnitude m_k^* with the estimates $\hat{\ell}_{k,j}$ and $\hat{m}_{k,j}$ from frozen DA3 (Lin et al., 2026). Its input includes the final prefix frame and target clip, capturing the boundary response. Translation is normalized by local depth; translation and rotation are then expressed per unit time and made dimensionless using fixed reference scales. The

camera drawings show direction and magnitude, and the curves illustrate the magnitude term after label matching.

The action reward combines label correctness with a magnitude term:

$$r_A(k,j) = \mathbf{1}[\hat{\ell}_{k,j} = \ell_k^*] \begin{cases} \min\left(\dfrac{\hat{m}_{k,j}}{\max(m_k^*, \epsilon)}, 1\right), & \ell_k^* \neq \text{NOOP}, \\ \left[1 - \dfrac{\hat{m}_{k,j}}{m_0}\right]_+, & \ell_k^* = \text{NOOP}. \end{cases} \tag{A2}$$

Here, $[z]_+ = \max(z, 0)$, $\epsilon > 0$ stabilizes division, and $m_0 > 0$ is the stationary-motion tolerance. Motion scores saturate at the target magnitude, measuring sufficiency; stationary commands instead reward remaining still.

A.3 Conditional Geometry Refinement

Figure A1(c) reuses the same DA3 depth, cameras, and confidence to measure bidirectional depth-reprojection consistency. Projecting pixel p's 3D point from frame a into frame b gives location $\pi_{a\to b}(p)$ and depth $z_{a\to b}(p)$. With D_b the predicted depth map of frame b, the correspondence error is

$$e_{a\to b}(p) = \frac{|z_{a\to b}(p) - D_b(\pi_{a\to b}(p))|}{|z_{a\to b}(p)| + |D_b(\pi_{a\to b}(p))| + \epsilon}. \tag{A3}$$

Within valid-projection and confidence masks, we clip extreme errors at a high quantile and apply confidence weighting. Averaging both projection directions and the predefined local frame pairs gives $E_{k,j}$. The pairs span the final prefix frame to the first target frame, and the first to the last target frame. The geometry reward is

$$r_G(k,j) = \exp(-E_{k,j}/\sigma_G), \tag{A4}$$

where $\sigma_G > 0$ controls the error scale. If the valid-pixel count or coverage is insufficient, we set evidence validity $v_G(k,j) = 0$ and use zero as a placeholder score, not as evidence of poor geometry. Otherwise, $v_G(k,j) = 1$.

Geometry comparison further requires correct action labels, sufficient action scores, and a small score difference within the group:

$$\begin{aligned} g_{A\to G}(k) = \mathbf{1}[\forall j,\ \hat{\ell}_{k,j} = \ell_k^*]\, \mathbf{1}[\min_j r_A(k,j) \geq \tau_A] \\ \cdot \mathbf{1}[\max_j r_A(k,j) - \min_j r_A(k,j) \leq \epsilon_A], \\ v_G^{\text{grp}}(k) = g_{A\to G}(k) \prod_{j=1}^{M} v_G(k,j). \end{aligned} \tag{A5}$$

A.4 Calibration and Update Behavior

In our configuration, HPSv3 evaluates four target frames, and $q_{k,j}$ is their lower quartile. Each anchor contains $M = 2$ candidates, with geometry weight $\lambda_G = 0.25$. Quality and action-comparability thresholds, geometry scales, and reward temperatures are calibrated with the frozen ODE-initialized student and remain fixed during reward-guided training (Appendix F).

The gates determine which terms enter Equation (1). If any candidate fails quality gating, all $A_{k,j}^{\text{ord}}$ are zero, so Equation (2) gives unit weights. If quality passes but actions are incorrect, insufficient, or dissimilar, or any candidate lacks valid geometry, $v_G^{\text{grp}}(k) = 0$ and the signal is $r_A(k,j)$. Only eligible groups use the geometric multiplier $1 + \lambda_G r_G(k,j)$. The rewards determine detached, unit-mean local weights through exponentiation and within-group normalization before weighting DMD updates; no gradients pass through HPSv3 or DA3.

B TRAINING ALGORITHM

Algorithm 1 expands Sections 3.3 and 3.4 into one generator update after ODE initialization. Reward definitions and calibration follow Appendix A. Sampling, decoding, rewards, and score evaluation use stop-gradient; only the saved target exit step is recomputed with gradients.

Algorithm 1 WorldRIDE: one generator update

Require: Student G_θ, fake score s_{fake}, frozen teacher, HPSv3, and DA3
Require: Conditions $(I_0, c, a_{1:N})$; K anchors, $M \geq 2$ candidates; fixed reward calibration
Hold student parameters fixed at $\theta_0 = \theta$; clear generator gradients
Sample a gradient-free spine of N clips; retain prefix caches and exit-step inputs
Select K temporal anchors along the spine
for each anchor $k = 1, \ldots, K$ **do**
    Fix prefix p_k, conditions c_k, cache snapshot, exit step u_k, and schedule
    Use the spine target as $\bar{x}_{k,1}$; retain its exit-step input and context
    for $j = 2, \ldots, M$ **do**
        Sample $\bar{x}_{k,j}$ with independent noise from the same prefix cache
        Save its exit-step input and context; terminate this branch at the target clip
    end for
    Decode all $\bar{x}_{k,j}$ to $\bar{v}_{k,j}$; initialize $A^{\text{ord}}_{k,j} = 0$ for all j
    Score target frames with HPSv3; compute $g^{\text{grp}}_Q(k)$ using (A1)
    if $g^{\text{grp}}_Q(k) = 1$ **then**
        Run DA3 on each final prefix frame and target clip
        Reuse camera, depth, and confidence predictions for both rewards
        Compute $r_A(k, j)$, $r_G(k, j)$, and validity $v_G(k, j)$ for all candidates
        Compute group geometry eligibility $v^{\text{grp}}_G(k)$ using (A5)
        Set $A^{\text{ord}}_{k,j} = r_A(k, j)[1 + \lambda_G v^{\text{grp}}_G(k) r_G(k, j)]$
    end if
    Compute detached $\widetilde{w}_{k,j}$ by (2) with FP32 log-sum-exp
    for each candidate $j = 1, \ldots, M$ **do**
        Sample diffusion time t independently of u_k; noise candidate and context
        Evaluate real and fake scores on the same noisy input
        Form detached normalized direction $d_{k,j}$ using (A8)
        Recompute saved exit step with gradients to obtain $x^\theta_{k,j}$
        Set $y_{k,j} = \text{sg}[x^\theta_{k,j} - \widetilde{w}_{k,j} d_{k,j}]$
        Accumulate $\nabla_\theta \left[\|x^\theta_{k,j} - y_{k,j}\|_2^2 / (2KM)\right]$
    end for
end for
Update θ once with the accumulated generator gradients

The fake-score network is trained separately on student samples without reward weighting or resampling. In our configuration it updates every training iteration, while Algorithm 1 runs every five iterations (Appendix F). Teacher and reward-model parameters remain frozen.

B.1 Shared Prefixes and Target Recomputation

For fixed prefix p_k, conditions c_k, and exit step u_k, the detached candidates are

$$\bar{x}_{k,j} = \mathrm{sg}[G_\theta(\xi_{k,j}; p_k, c_k, u_k)], \qquad j = 1, \ldots, M. \tag{A6}$$

where $\xi_{k,j}$ contains generation noise and sg denotes stop-gradient. The sampling schedule and student parameters are identical within a group. Prefix caches retain the backbone's attention sink and historical state. Branches read the same cache snapshot without overwriting it, and later anchors always use the pre-generated spine rather than a reward-selected branch.

For one target clip per candidate, the spine and local branches require

$$N + K(M-1) \tag{A7}$$

clip-sampling operations to supply KM targets. This counts trajectory generation; reward evaluation, score-network computation, and differentiable target recomputation contribute separately. Replaying each saved exit-step input with the unchanged student gives $x^\theta_{k,j} \approx \bar{x}_{k,j}$ within numerical tolerance. The shared prefix and preceding denoising steps remain detached.

B.2 Weighting and Gradient Accumulation

Equation (2) is evaluated as $\widetilde{w}_{k,j} = \mathrm{sg}\{M \exp[z_{k,j} - \mathrm{LSE}_\ell(z_{k,\ell})]\}$, with $z_{k,j} = A^{\mathrm{ord}}_{k,j}/\beta_{u_k}$ and FP32 log-sum-exp. It guarantees $\sum_j \widetilde{w}_{k,j} = M$ and invariance to a common additive reward shift. This fixes total weight, not gradient norm; reward-difference scale still affects preference strength through β_{u_k}.

At diffusion time t, sampled independently of u_k, both score networks receive the same noisy candidate and context. On target-clip positions, the DMD direction is

$$d_{k,j} = \mathrm{sg}\Big[\mathrm{Norm}_{\mathrm{DMD}}\left(s_{\mathrm{fake}} - s_{\mathrm{real}}\right)_{k,j}\Big]. \tag{A8}$$

using the gradient-descent sign convention. $\mathrm{Norm}_{\mathrm{DMD}}$ retains the underlying score-parameterization conversion and normalization. Applying rewards after this operation prevents subsequent normalization from canceling their effect.

With the detached pseudo-target in Equation (3), the generator gradient is

$$\nabla_\theta \mathcal{L}_G = \frac{1}{KM} \sum_{k=1}^{K} \sum_{j=1}^{M} \left(J_\theta x^\theta_{k,j}\right)^\top \widetilde{w}_{k,j} d_{k,j}. \tag{A9}$$

where $J_\theta x^\theta_{k,j}$ is the output Jacobian. Each reward weight is applied exactly once. HPSv3, DA3, the score networks, and shared history receive no generator gradients. All candidates use the same student parameter snapshot until their gradients have been accumulated and the optimizer takes one step. Inference uses the four-step student and streaming cache alone.

C Additional Qualitative Comparisons

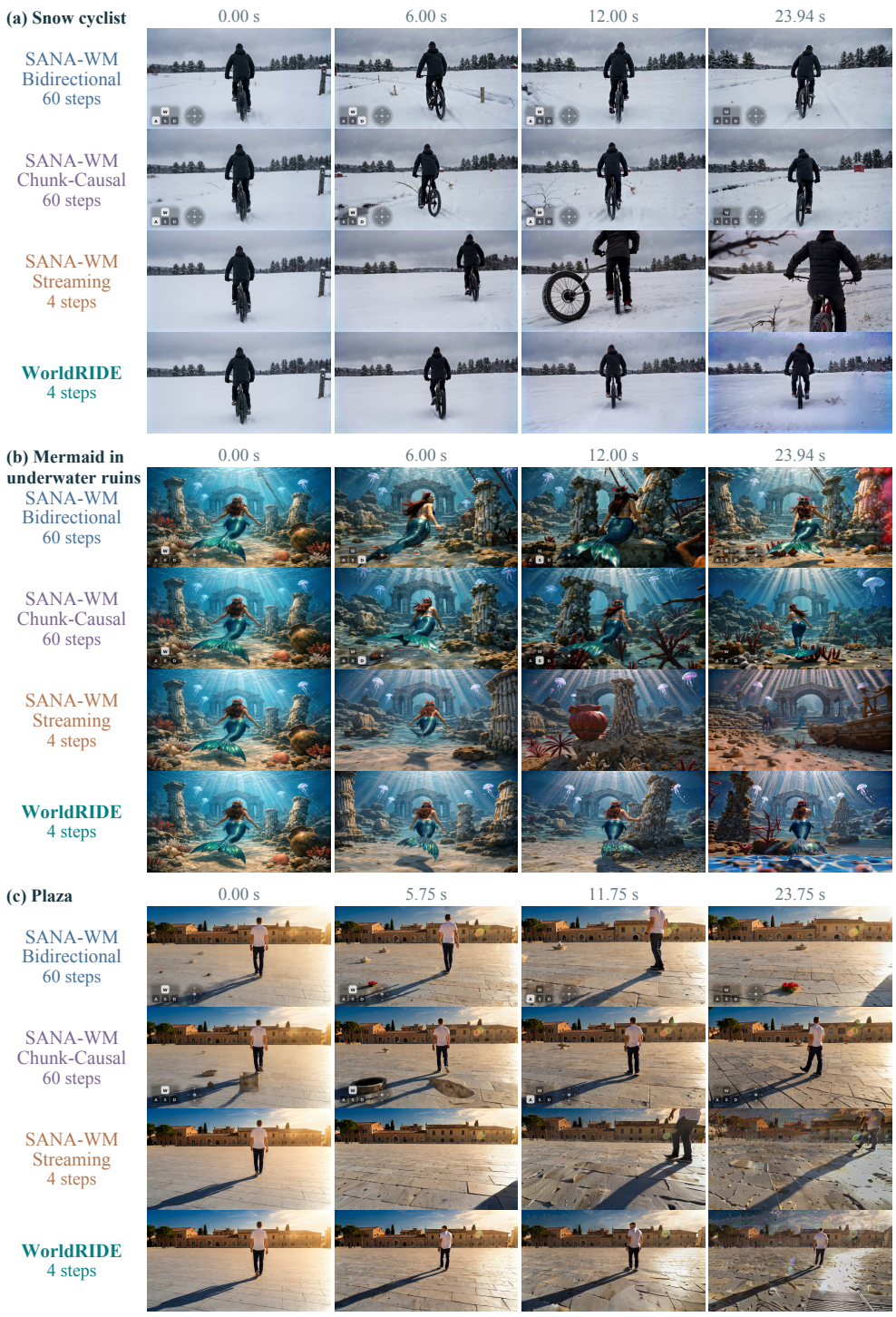

Figure A2: **Additional WBench visual comparisons.** Four-model comparisons on the snow cyclist, mermaid, and plaza. SANA-WM Bidirectional and Chunk-Causal use 60 steps; Streaming and WorldRIDE use four. Rows show four matched full frames. The first two cases use model-specific refiners; the plaza uses a shared bidirectional refiner.

VBench camera-motion examples. Figure A3 shows five-second generations from VBench-I2V camera-motion inputs depicting bubbles and the Parthenon. In the Parthenon case, Streaming replaces foreground ruins with building facades, whereas WorldRIDE retains stone steps. The three models use a common non-streaming sampling setup and a shared bidirectional refiner.

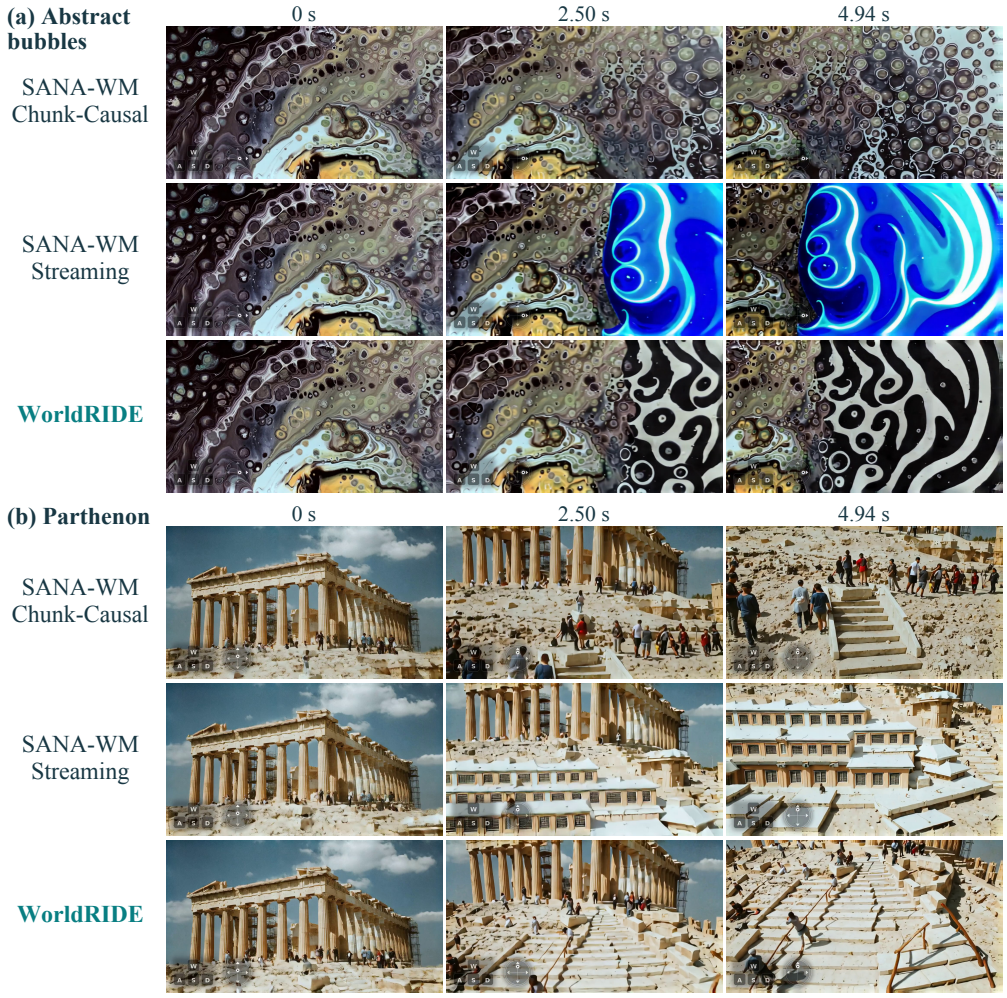

Figure A3: **VBench-I2V camera-motion examples.** Two selected five-second clips from the 60-step Chunk-Causal model and the four-step Streaming and WorldRIDE checkpoints, using VBench-I2V camera-motion inputs. Columns show decoded frames 0, 40, and 79 at 16 FPS. All three models use the same non-streaming sampling setup and bidirectional refiner.

Recoverable scene structure. Figure A4 compares Pi3X reconstructions from all three models under a shared view for each scene. WorldRIDE recovers repeated staircase treads and bordering walls, as well as recognizable museum wall surfaces and picture frames. The 60-step and streaming references show the corresponding scene structures recovered from their generated videos.

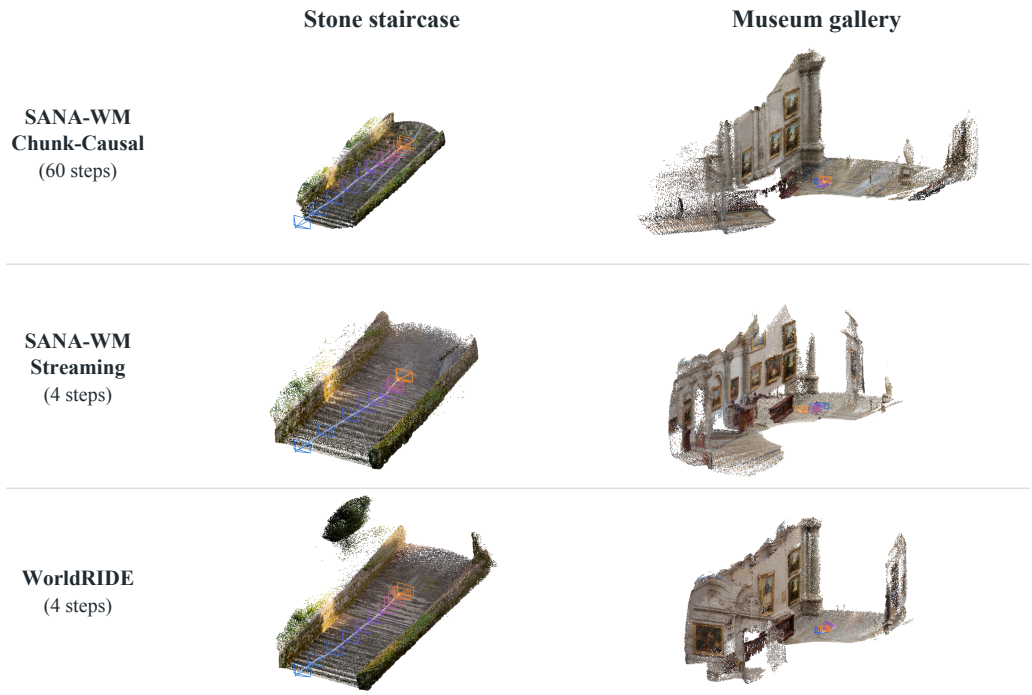

Figure A4: **Pi3X scene reconstruction comparison.** The three models are reconstructed from 24 matched frames per video. Within each scene, point clouds share the viewing direction, framing bounds, point-filtering rule, and rendering settings after independent first-camera alignment and median-distance normalization. Camera colors indicate temporal order. Chunk-Causal uses the bidirectional refiner; Streaming and WorldRIDE use the streaming refiner.

D QUALITATIVE VISUALIZATION PROTOCOL

We compare SANA-WM Chunk-Causal (60 steps), SANA-WM Streaming (4 steps), and WorldRIDE (4 steps) on WBench and VBench separately. Within each case, the 16-FPS videos use matched input images, prompts, camera controls, and seeds. Image-only Pi3X (Wang et al., 2026a) recovers camera motion and scene structure. Appendix D describes the inference and refinement settings for each comparison.

Video comparisons. All qualitative videos use 16 FPS and seed 42. SANA-WM Streaming uses the four-step schedule $[1000, 960, 889, 727, 0]$, and WorldRIDE uses $[1000, 967, 908, 764, 0]$. The Bidirectional and Chunk-Causal models use the configured 60-step sampler setting; this setting does not denote an exact count of network evaluations. Displayed columns match decoded frame indices within each case. Full video frames are decoded and resized for display, without geometric registration. Baked-in action overlays in the archived Chunk-Causal videos are retained in both displayed frames and Pi3X inputs.

WBench inference and refinement. The three cases in Figure 3 and three additional cases in Figure A2 each contain 384 decoded frames (24 seconds). The lizard, ghost dog, rowboat, snow cyclist, and mermaid cases use model-specific refiners: Chunk-Causal uses the bidirectional refiner and text encoder, whereas Streaming and WorldRIDE use the streaming refiner and text encoder. The plaza case uses a shared bidirectional refiner and text encoder. Figure A2 additionally includes the original bidirectional SANA-WM model, using its bidirectional checkpoint, 60-step sampler, and bidirectional refiner. Displayed frame indices are $[0, 144, 256, 383]$ for the lizard, $[0, 192, 335, 383]$ for the ghost dog, $[0, 96, 192, 383]$ for the rowboat, snow cyclist, and mermaid, and $[0, 92, 188, 380]$ for the plaza. The camera and point-cloud comparisons use model-specific refiners.

VBench inference and refinement. The two cases in Figure A3 use official VBench-I2V camera-motion images and prompts, with camera trajectories constructed from their motion labels: pan right for bubbles and tilt down for the Parthenon. Each video contains 80 decoded frames (5 seconds). All three models use the same non-streaming sampling setup, bidirectional refiner, and text encoder. The displayed frames are $[0, 40, 79]$. These examples come from a separate qualitative generation batch, whereas Table 2 reports the benchmark evaluation without refinement.

Camera reconstruction and normalization. Pi3X receives 49 matched image samples spanning each complete clip at 518×280 resolution, without camera or depth conditioning. The library and vineyard clips each contain 384 frames (24 seconds). The filmstrips show decoded frames 0, 187, and 383, matched across models at 16 FPS. Full generated frames retain the original action overlays where present. Saved-video frame f corresponds to input camera $f + 1$. Each trajectory is first expressed in its own initial camera coordinates, then scaled by a single positive factor so that displacement at a reference-selected frame equals one. Both displayed scenes use frame 187 (11.6875 seconds) as the normalization anchor. All 49 positions and the original orientation changes are retained; nine common samples are shown as camera glyphs. Projection, bounds, and glyph size are shared across methods within a scene. The control trajectory denotes the input to generation, and the plotted distances have normalized units.

Point-cloud display. The reconstructions in Figure A4 use 24 sampled frames per video at 518×280 resolution. Cached point clouds remove the bottom 45% of confidence scores and retain at most 420,000 points. For display, we retain the 85% of points nearest to a recovered camera, express the cloud in the first-camera coordinates, and normalize by median point distance. For each scene, the three models share one viewing direction, framing bounds pooled across their normalized clouds, and identical point-splat and camera-glyph sizes. These visualizations characterize recovered structure without a ground-truth 3D reference.

E FULL WBENCH RESULTS

Tables A1 and A2 report the seven training-strategy and refiner configurations. Each table includes Overall, the five dimension scores, and all 19 individual metrics. Metrics are averaged over their valid cases; the dimensions group 6/2/1/8/2 metrics for Quality, Setting, Interaction, Consistency, and Physical plausibility, respectively. We retain the benchmark's reported one-decimal scores and aggregation. Valid-case counts are identical across configurations.

Table A1: **Full WBench results with the refiner enabled.** All 19 metrics and six aggregate scores are reported on the 158-case navigation split. N is the valid-case count for each individual metric, shared across models. All scores are higher-is-better. Bold marks the best four-step result per row, including ties; the gray column is the 60-step reference. RL denotes Streaming followed by WorldCompass-style post-training.

Metric	N	SANA-WM Chunk-Causal	SANA-WM Streaming	Streaming + RL	WorldRIDE (ours)
Steps	–	60	4	4	4
Aggregate scores					
Overall	–	79.7	75.7	76.0	**77.1**
Quality	–	80.3	79.1	79.4	**79.5**
Setting	–	80.7	77.1	**78.3**	77.1
Interaction	–	84.5	81.9	82.0	**82.6**
Consistency	–	86.1	79.0	79.2	**82.7**
Physical plausibility	–	66.8	61.5	61.3	**63.6**
Quality					
Aesthetic quality	158	62.0	60.5	60.7	**61.3**
Imaging quality	158	67.1	63.3	63.9	**64.4**
Temporal flickering	158	91.9	93.7	**94.0**	93.9
Dynamic degree	158	96.8	**98.7**	**98.7**	96.8
Motion smoothness	158	97.4	98.1	**98.2**	98.0
HPSv3 quality	158	66.3	60.3	60.8	**62.8**
Setting					
Scene adherence	158	71.2	63.9	**65.4**	63.4
Subject adherence	74	90.1	90.4	**91.1**	90.8
Interaction					
Navigation trajectory	158	84.5	81.9	82.0	**82.6**
Consistency					
Background consistency	158	92.1	89.4	89.4	**90.2**
Spatial consistency	60	84.7	**75.1**	73.5	**75.1**
Gated spatial consistency	60	81.4	**75.1**	73.5	74.4
Perspective consistency	103	71.7	49.6	53.5	**67.0**
Segment continuity	158	99.4	91.8	90.5	**98.1**
Geometric consistency	158	90.8	86.1	87.0	**87.4**
Photometric consistency	158	79.6	80.7	81.0	**81.1**
Subject consistency	103	89.5	84.3	85.0	**87.8**
Physical plausibility					
Visual plausibility	158	57.4	54.5	54.6	**55.3**
Causal fidelity	50	76.2	68.4	68.0	**71.9**

Table A2: **Full WBench results with the refiner disabled.** All 19 metrics and six aggregate scores are reported on the 158-case navigation split. N is the valid-case count for each individual metric, shared across models. All scores are higher-is-better. Bold marks the best four-step result per row, including ties; the gray column is the 60-step reference.

Metric	N	SANA-WM Chunk-Causal	SANA-WM Streaming	WorldRIDE (ours)
Steps	–	60	4	4
Aggregate scores				
Overall	–	75.0	73.6	**74.6**
Quality	–	76.0	**77.6**	77.4
Setting	–	73.5	72.7	**72.8**
Interaction	–	84.5	**84.3**	83.9
Consistency	–	84.2	78.7	**81.8**
Physical plausibility	–	57.0	54.6	**57.0**
Quality				
Aesthetic quality	158	57.9	**59.3**	59.0
Imaging quality	158	63.7	**66.8**	**66.8**
Temporal flickering	158	92.2	**90.7**	**90.7**
Dynamic degree	158	98.1	**99.4**	97.5
Motion smoothness	158	96.4	**96.6**	96.1
HPSv3 quality	158	47.8	52.7	**54.4**
Setting				
Scene adherence	158	57.5	**60.6**	56.1
Subject adherence	74	89.5	84.9	**89.6**
Interaction				
Navigation trajectory	158	84.5	**84.3**	83.9
Consistency				
Background consistency	158	90.4	88.5	**89.4**
Spatial consistency	60	81.0	**74.8**	**74.8**
Gated spatial consistency	60	79.3	**74.8**	**74.8**
Perspective consistency	103	70.1	51.2	**66.9**
Segment continuity	158	99.4	99.4	**100.0**
Geometric consistency	158	89.7	86.0	**89.0**
Photometric consistency	158	78.9	72.6	**74.2**
Subject consistency	103	84.8	82.4	**85.5**
Physical plausibility				
Visual plausibility	158	53.8	51.1	**52.2**
Causal fidelity	50	60.2	58.1	**61.8**

F EXPERIMENTAL DETAILS

Training data. We use a three-source collection of camera-annotated videos comprising a scene-balanced subset, GameVerse (AlayaWorld Team et al., 2026b), and DL3DV (Ling et al., 2024). Each example provides an initial image, a static scene caption, and a camera trajectory. Training windows contain 145 frames sampled at 16 FPS and resized to 1280×704, with camera intrinsics adjusted accordingly. The causal VAE maps each window to 19 latent frames: one conditioning frame and six target chunks of three latent frames each. Reward-guided distillation operates on student-generated continuations conditioned on cached first-frame latents. We reserve a disjoint calibration subset for reward calibration.

Implementation details. We use the SANA-WM backbone and teacher configuration described in Section 3.1. Causal ODE initialization uses 16,000 precomputed trajectories and 3,000 optimization steps with an AdamW learning rate of 2×10^{-5}. Reward-guided DMD then runs for 3,000 training iterations. The student and fake-score networks use AdamW learning rates of 10^{-5} and 2×10^{-6}, respectively, with $(\beta_1, \beta_2) = (0, 0.999)$ and zero weight decay. The fake-score network is updated every iteration and the student every five iterations. Training uses bfloat16 weights, FSDP2, and activation checkpointing. The four-step denoising schedule is $[1000, 967, 908, 764, 0]$. Streaming inference retains the backbone's attention sink and a two-chunk cache. We evaluate the student obtained at the end of reward-guided distillation without additional reward post-training.

Rollout and reward configuration. Each shared trajectory provides three anchor slots, with two action-focused slots and one geometry-eligible slot. Each anchor has $M = 2$ local candidates, and geometry uses $\lambda_G = 0.25$. HPSv3 (Ma et al., 2025) evaluates four target frames, using the lower quartile of frame scores together with the minimum-score check for group-level quality gating. Frozen DA3-GIANT-1.1 (Lin et al., 2026) supplies camera-motion and depth-reprojection evidence in a shared forward pass. Quality and action-comparability thresholds, geometry scales, and reward temperatures are calibrated with the frozen ODE-initialized student and held fixed throughout training.

Evaluation benchmarks. For interactive world modeling, we use the 158-case navigation split of WBench (Ying et al., 2026). Its 19 metrics cover visual quality, setting adherence, interaction, consistency, and physical plausibility, with an Overall score aggregated from the five dimensions.

For image-to-video generation, we follow the VBench-I2V protocol (Huang et al., 2026) with 1,118 official entries and five fixed seeds per entry, yielding 5,590 videos per model. Each video contains 81 frames at 16 FPS and 1280×704 resolution. Evaluation covers nine dimensions spanning camera motion, image conditioning, temporal consistency, dynamics, and visual quality. We use the benchmark's Total, I2V, and Quality aggregate scores.

For long-horizon generation, we use the public SANA-WM 80-scene benchmark (Zhu et al., 2026a), with one simple and one hard 60-second camera trajectory per scene, yielding 160 videos per model. Each video contains 960 source frames. The benchmark evaluates visual quality, revisit consistency, camera control, and temporal quality retention; visual and temporal metrics exclude the initial conditioning frame. We aggregate the two splits using equal weights for VBench scores, revisit-pair weights for revisit metrics, and scene weights for camera metrics.

Comparison models and inference settings. We compare against the 60-step SANA-WM Chunk-Causal model and the official four-step SANA-WM Streaming model. As reported in SANA-WM, Streaming is obtained through standard Self Forcing distillation (Zhu et al., 2026a; Huang et al., 2025). For cross-model WBench comparisons, we include DreamX-World (5B AR), Lyra 2.0, LingBot-World (Fast), and MatrixGame3.0 from the official navigation leaderboard (WBench, 2026); SANA-WM and WorldRIDE results are evaluated by us. The main WBench comparison enables SANA-WM's pretrained refiner for the SANA-WM models and WorldRIDE. Ablations compare with and without refinement and include a WorldCompass-style RL baseline (Wang et al., 2026b), initialized from SANA-WM Streaming after 3,000 Self Forcing distillation steps and trained for an additional 3,000 RL optimizer updates. VBench-I2V disables the refiner for all three backbone-matched models, whereas the 80-scene benchmark enables it. Denoising-step counts refer to the first-stage generator.

G Additional Experimental Analyses

G.1 Training Strategy and Refinement

Reward integration during distillation. Table A3 compares training strategies and refiner settings on the same SANA-WM backbone. The WorldCompass-style baseline starts from SANA-WM Streaming after 3,000 Self Forcing distillation steps and adds 3,000 RL updates. With refinement enabled, this post-training raises Streaming's Overall from 75.7 to 76.0. WorldRIDE integrates reward guidance into 3,000 DMD training steps and reaches 77.1, with Consistency improving from 79.0 to 82.7. Thus, reward-guided distillation improves the student without the baseline's additional RL stage.

Table A3: **Ablation of refinement and training strategy on WBench.** w/ and w/o denote with and without the refiner. WorldCompass RL adds 3,000 updates after Streaming's 3,000-step Self Forcing distillation. All 158 navigation cases are used; higher is better. Bold indicates columnwise best non-teacher results, including ties; gray marks 60-step teacher references.

Model	Steps	Overall	Quality	Setting	Interact.	Consist.	Physical
SANA-WM Chunk-Causal (w/o)	60	75.0	76.0	73.5	84.5	84.2	57.0
SANA-WM Chunk-Causal (w/)	60	79.7	80.3	80.7	84.5	86.1	66.8
SANA-WM Streaming (w/o)	4	73.6	77.6	72.7	**84.3**	78.7	54.6
SANA-WM Streaming (w/)	4	75.7	79.1	77.1	81.9	79.0	61.5
Streaming + WorldCompass RL (w/)	4	76.0	79.4	**78.3**	82.0	79.2	61.3
WorldRIDE (ours, w/o)	4	74.6	77.4	72.8	83.9	81.8	57.0
WorldRIDE (ours, w/)	4	**77.1**	**79.5**	77.1	82.6	**82.7**	**63.6**

Student capabilities before refinement. To isolate the first-stage generator, we examine the results without refinement in Table A3. WorldRIDE improves Overall from 73.6 to 74.6 over SANA-WM Streaming and comes within 0.4 points of the 60-step reference. Consistency improves by 3.1 points and Physical plausibility by 2.4 points, showing that these gains are present in the student output before refinement. The streaming baseline scores slightly higher in Quality and Interaction in this setting.

G.2 LONG-HORIZON GENERATION

Table A4 evaluates 60-second trajectories on the SANA-WM 80-scene benchmark. Relative to the four-step streaming baseline, WorldRIDE improves revisit PSNR from 14.258 to 15.256 dB and SSIM from 0.2666 to 0.2923, while reducing LPIPS from 0.5697 to 0.5343. Camera rotation error falls from 22.717° to 16.583°. VBench Total increases from 69.065 to 69.140, with nearly unchanged Quality (80.260 vs. 80.265). These gains concentrate on revisit fidelity and camera rotation; translation error and temporal quality drop remain higher than those of the streaming baseline.

Table A4: **SANA-WM 80-scene results with the refiner enabled.** Each model generates 160 one-minute videos. VBench reports percentages averaged over simple and hard splits; revisit uses 800 pairs (PSNR in dB). TransErr and CamMC use relative normalization; ΔIQ is first-to-last-window imaging-quality drop. Bold indicates columnwise best results excluding the gray 60-step teacher reference.

		VBench (%)			Revisit		
Model	Steps	Total↑	Quality↑	Semantic↑	PSNR↑	SSIM↑	LPIPS↓
SANA-WM Chunk-Causal	60	70.080	81.420	24.715	14.976	0.3117	0.4666
SANA-WM Streaming	4	69.065	80.260	24.295	14.258	0.2666	0.5697
WorldRIDE (ours)	4	**69.140**	**80.265**	**24.660**	**15.256**	**0.2923**	**0.5343**

		Camera			Temporal
Model	Steps	RotErr($^\circ$) ↓	TransErr↓	CamMC↓	ΔIQ ↓
SANA-WM Chunk-Causal	60	7.520	1.450	1.493	0.0166
SANA-WM Streaming	4	22.717	**2.089**	2.264	**0.0065**
WorldRIDE (ours)	4	**16.583**	2.136	**2.244**	0.0658