CLEAN FORCING: DRIFT-RESISTANT AUTOREGRESSIVE VIDEO DIFFUSION WITH A FROZEN BASE

Anonymous authors
Paper under double-blind review



Figure 1: **Clean Forcing reduces drift during AR rollout.** The first 20 s of two 50 s generations with the same prompt and seed are shown. The adapted AR base (top) becomes oversaturated and loses the glass, while the same model with the merged Clean Forcing corrector (bottom) preserves the scene and pouring motion.

ABSTRACT

Autoregressive (AR) video diffusion models generate streaming video of arbitrary length by conditioning each new chunk of frames on those already generated. However, they suffer from exposure bias, where a model trained on clean context must continue from its own imperfect outputs, and errors compound until quality collapses within seconds (drift). Existing methods either require cluster-scale base retraining, apply training-free inference corrections with inconsistent results, or train frozen-base correctors for sampling and cache errors. We instead find that drift is largely deterministic and therefore learnable, as holding the noisy state fixed while varying only the history reveals a counterfactual velocity gap that is $\sim 95\%$ systematic across noise seeds. Based on this observation, we introduce *Clean Forcing*, which trains a 5.9M-parameter LoRA corrector on the *frozen* base model to regress this gap, with a closed-loop stage that adds DAgger-style aggregation and a drift-contraction objective to expose the corrector to its own rollouts. Because the clean histories can be the base model's own single-shot generations, training needs *no real videos*, and the merged corrector adds no inference cost. In 50 s text-to-video generation, Clean Forcing reduces Δ-drift from $+11.3$ to $+1.65$ without real video and to $\mathbf{-0.16}$ with 40 real clips, improving on the best published result of $+3.57$ with over $\mathbf{150}\times$ less data. Extensive experiments demonstrate that Clean Forcing outperforms prior methods in aesthetic quality and data efficiency with reduced drift for long-horizon video generation.

1 INTRODUCTION

Video diffusion models (Wan Team, 2025; Polyak et al., 2024) generate high-quality clips, making them promising for live content creation, simulation, and interactive world modeling. However, most use bidirectional attention (Yin et al., 2025) and jointly denoise a clip whose length is fixed before generation begins. Playback therefore cannot begin until the clip has been generated. This inherently limits their applicability to real-time streaming and long-horizon video generation.

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.

Autoregressive (AR) video diffusion models address some of these limitations by generating frames in chunks, with each chunk conditioned on those generated before it (Chen et al., 2024; Yin et al., 2025; Huang et al., 2025; Sand.ai, 2025; Skywork AI, 2025). This allows rollouts of flexible length and lets playback begin before generation is complete. However, models trained to continue from *clean* context must condition on their *own imperfect outputs* at inference. Errors are therefore fed back into subsequent chunks and accumulate over time, often causing the video to drift toward oversaturated, textureless frames within seconds (drift; Fig. 1, top). This train–inference mismatch, known as *exposure bias* (Yin et al., 2025; Huang et al., 2025; Po et al., 2025), remains a central challenge for AR video generation.

Existing drift-reduction methods fall into three categories. *Base retraining* incorporates the model's own rollouts into training (Huang et al., 2025; Po et al., 2025), which is effective but expensive: the strongest published result uses 136K clips and 16×H100 GPUs (Po et al., 2025). *Training-free methods* keep the weights fixed and modify inference by noising the context (Chen et al., 2024; Skywork AI, 2025), guiding generation against a cleaner history (Song et al., 2025), correcting the sampling trajectory (Xiang et al., 2026), or managing the cache (Meng et al., 2026). These methods either treat drift as uniform noise or re-anchor generation to a prediction that multi-step models do not produce, and none that we evaluate reduces drift at 50 s (§4). *Frozen-base correctors* learn lightweight corrections, but existing methods target sampling noise (Yu et al., 2025) or cache-approximation error (Peng et al., 2025), rather than error induced by generated history. Thus, deployed checkpoints still lack an efficient learned correction that directly targets exposure-bias drift.

In this paper, we introduce **Clean Forcing**, an efficient drift-reduction method for frozen AR video diffusion models. We first characterize history-induced drift by holding a noisy state z_t fixed and comparing the model's denoising velocities under drifted and clean histories. This counterfactual isolates the error induced by the generated history, and our analysis shows that $\sim 95\%$ of the resulting velocity gap is consistent across noise seeds, indicating that drift is largely systematic and learnable. Based on this insight, we train a rank-16 LoRA corrector r_ϕ (5.9M parameters, 0.4% of the base) to predict this velocity gap while keeping the base model frozen. To expose the corrector to states induced by its own predictions, we introduce a closed-loop stage that regenerates training pairs with the corrector active (Ross et al., 2011) and penalizes the downstream drift caused by each committed prediction. The base model's own single-shot generations serve as clean references, so training requires no real videos. On a lightly adapted causal Wan2.1-1.3B (Wan Team, 2025), Clean Forcing reduces Δ-drift from $+11.3$ to -0.16 over 50 s on 128 MovieGen prompts (Polyak et al., 2024). Without real videos, it achieves $+1.65$, outperforming the best published result of $+3.57$ (Po et al., 2025) while using over $150\times$ less data. The corrector trains in one GPU-day and merges into the base model without adding inference overhead.

Our contributions are summarized as follows:

- We identify that the velocity error induced by a drifted history is largely systematic across noise seeds, which makes the drift correction learnable.
- We introduce Clean Forcing, a 5.9M-parameter LoRA corrector trained on a frozen base against clean-history predictions at the same noisy state, with a closed-loop stage for DAgger aggregation and drift contraction, and a self-teaching procedure that requires no real videos.
- Clean Forcing eliminates measurable drift with far less data than base retraining and no inference overhead. It achieves the lowest drift and highest aesthetic score among the compared methods and ranks second in subject and background consistency, imaging quality, and MUSIQ. It also transfers to an independent base, scales to a $10\times$ larger base, and reduces drift in two other AR world models.

2 Related Work

Video Generation. Early video generation used GANs (Vondrick et al., 2016; Tulyakov et al., 2018) or autoregressed over discrete visual tokens (Yan et al., 2021). GANs offered fast sampling but were difficult to train, while token-based models faced growing generation costs as video sequences lengthened. Diffusion models (Ho et al., 2022b; Blattmann et al., 2023) improved training stability and visual quality, enabling large-scale systems such as (Brooks et al., 2024; Yang et al., 2024; Kong et al., 2024; Wan Team, 2025; Polyak et al., 2024). Most of these models jointly denoise a predetermined clip with bidirectional spatiotemporal attention (Peebles & Xie, 2023; Ho et al.,

2022a; Bar-Tal et al., 2024). Joint denoising promotes temporal coherence, but requires a fixed clip length and delays playback until the full clip is generated, limiting real-time streaming and long-horizon generation.

Autoregressive Video Diffusion. AR diffusion generates video in chunks, enabling streaming and flexible-length rollouts. However, training on clean histories but conditioning on imperfect outputs at inference creates exposure bias, causing errors to compound into drift. To address this mismatch, Diffusion Forcing trains on frames with different noise levels (Chen et al., 2024), a formulation that DFoT extends to transformers with guidance over the generated history (Song et al., 2025). Self Forcing takes a different route by training on the model's own rollouts (Huang et al., 2025). For interactive generation, CausVid distills a bidirectional teacher into a few-step causal student (Yin et al., 2025), an approach further developed by Causal Forcing (Zhu et al., 2026). MAGI-1 adds noise to clean context during training to reduce exposure bias (Sand.ai, 2025), while SkyReels-V2 uses Diffusion Forcing and noised, overlapping context to improve long-video consistency (Skywork AI, 2025). These methods mitigate error accumulation without isolating the prediction error caused by the generated history. In contrast, Clean Forcing uses the frozen model's own generations to learn the velocity gap between clean and drifted histories at the same noisy state, reducing drift while preserving streaming, flexible-length rollouts, and inference speed.

Mitigating Exposure Bias. In imitation learning, errors caused by visiting self-generated states are addressed by collecting training examples from the learner's own rollouts (Ross et al., 2011). Video generation methods take a similar approach by retraining the base on its outputs, using distributional targets (Huang et al., 2025), reversed rollouts (Po et al., 2025), and large-scale rollout training (Liu et al., 2025; Cui et al., 2025; Guo et al., 2025; Shin et al., 2026; Yang et al., 2025). This improves long-horizon stability, but requires a full model update and 10^4–10^5+ clips. Concurrent work post-trains few-step generators on matched-state targets from real videos (Liu et al., 2026). Training-free methods instead noise the context (Chen et al., 2024; Skywork AI, 2025), guide generation with a cleaner history (Song et al., 2025), adjust the sampling path or activations (Xiang et al., 2026; Tian et al., 2026), or manage the cache (Meng et al., 2026; Yi et al., 2025). They require no training, but do not reliably reduce drift (§4). Frozen-base correctors offer a lightweight alternative, but AutoRefiner targets sampling noise (Yu et al., 2025), while ERTACache addresses the separate problem of cache-approximation error (Peng et al., 2025). Unlike prior methods, Clean Forcing directly learns the history-induced velocity gap while keeping the base frozen. With at most 340 clips and no real video, it matches or surpasses full retraining in drift reduction and adds no inference cost after merging.

3 METHOD

3.1 PRELIMINARIES: BLOCK-CAUSAL VIDEO DIFFUSION

An autoregressive (AR) video model generates a video as K latent chunks $x = (x^{(1)}, \ldots, x^{(K)})$, each containing f frames, conditioned on a text prompt c:

$$p_\theta(x \mid c) = \prod_{k=1}^{K} p_\theta\Big(x^{(k)} \mid h^{(k)}, c\Big), \qquad h^{(k)} = x^{(<k)}, \tag{1}$$

where θ denotes the pretrained model parameters and $h^{(k)}$ is the history of committed chunks. Block-causal attention restricts each chunk to itself and its history, whose representations are stored in a KV cache. Each conditional is modeled by flow matching (Lipman et al., 2023). Given a clean latent chunk x_0, Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, and noise level $\sigma_t \in [0, 1]$ at solver step t, the noisy state is $z_t = (1 - \sigma_t)x_0 + \sigma_t\epsilon$. The network $v_\theta(z_t, h, t)$ predicts the path velocity and is trained with $\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{x_0,\epsilon,t}\|v_\theta(z_t, h, t) - (\epsilon - x_0)\|_2^2$. At inference, each chunk is initialized from noise and denoised by integrating $\frac{\mathrm{d}z}{\mathrm{d}\sigma} = v_\theta$ from $\sigma = 1$ to 0 over T solver steps, then committed to the history.

Exposure bias. The model is trained on clean histories but conditioned on its own imperfect outputs at inference, leading to different conditional distributions under the model's generated rollout history h_{gen} and the reference history h_{clean}: $p_\theta(x^{(k)} \mid h_{\mathrm{gen}}, c) \neq p_\theta(x^{(k)} \mid h_{\mathrm{clean}}, c)$ (Ranzato et al., 2016; Schmidt, 2019; Ning et al., 2023; Cui et al., 2025). Because each generated chunk becomes

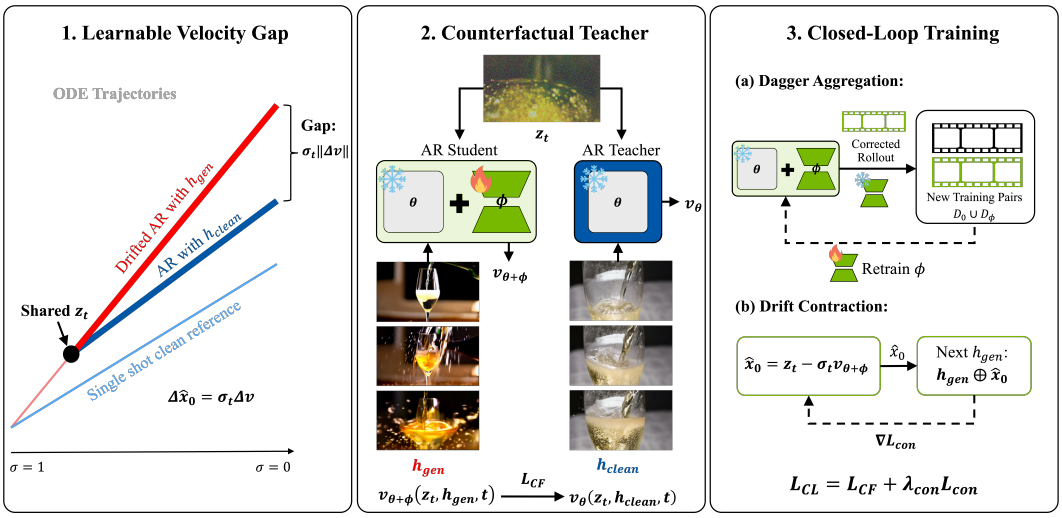

Figure 2: **Overview of Clean Forcing.** **(1)** At the shared noisy state z_t, $\Delta v := v_\theta(z_t, h_{\text{clean}}, t) - v_\theta(z_t, h_{\text{gen}}, t)$ is the history-induced velocity gap. The corresponding one-step clean-latent predictions differ by $\sigma_t\|\Delta v\|$, contributing to drift over successive chunks. **(2)** The LoRA student conditions on h_{gen} and matches the frozen teacher's prediction under h_{clean} ($\mathcal{L}_{\text{CF}}$). **(3a)** We generate new training pairs D_ϕ with the corrector active and continue training on $D_0 \cup D_\phi$, where D_0 contains pairs from uncorrected rollouts. **(3b)** $\mathcal{L}_{\text{con}}$ backpropagates the next chunk's error through the undetached prediction $\hat{x}_0$.

part of the history, errors compound across K chunks into $O(K^2\varepsilon)$ cumulative error, manifesting as progressive quality drift over longer rollouts (Ross et al., 2011). Our goal is to align these conditional distributions during inference, such that $p_\theta(x^{(k)} \mid h_{\text{gen}}, c) \approx p_\theta(x^{(k)} \mid h_{\text{clean}}, c)$.

3.2 Clean Forcing

Clean Forcing converts our measurement into a learned corrector in three steps (Fig. 2). First, we show that the velocity error induced by a drifted history is largely systematic across noise seeds and can therefore be corrected by a deterministic function (§3.2.1; panel 1). Second, we train a lightweight LoRA on the frozen base to map the drifted-history velocity to the model's counterfactual clean-history prediction at the same noisy state (§3.2.2; panel 2). Third, we close the train–inference gap through DAgger-style data aggregation and a drift-contraction objective (§3.2.3; panel 3). Finally, we show that clean references need not come from real videos because the same checkpoint can generate its own drift-free supervision in bidirectional mode (§3.2.4).

At inference, the corrected velocity is

$$v_{\text{rect}} = v_\theta + \alpha\,(v_{\theta+\text{LoRA}} - v_\theta)\,, \tag{2}$$

where α controls the correction strength. Setting $\alpha = 1$ recovers the LoRA-merged forward pass and therefore adds no inference overhead. The measured systematic fraction $\alpha^*(t)$ provides a principled correction strength at each solver step. It remains nearly constant at ≈ 0.95 on our base model, so we use the full merged correction without a time-dependent gating schedule.

3.2.1 Drift Is a Systematic Velocity Error

For each state z_t reached by a drifted rollout, we evaluate the frozen model twice while changing only the history: $\Delta(z_t, t) = v_\theta(z_t, h_{\text{clean}}, t) - v_\theta(z_t, h_{\text{gen}}, t)$. Because both predictions use the same state, Δ isolates the counterfactual effect of history corruption. We measure the fraction of this velocity gap that is reproducible across noise seeds as

$$\alpha^*(t) = \frac{\|\mathbb{E}_s[\Delta_s]\|_2^2}{\|\mathbb{E}_s[\Delta_s]\|_2^2 + \text{Var}_s(\Delta_s)}, \tag{3}$$

where s indexes noise seeds and $\mathrm{Var}_s(\Delta_s) = \mathbb{E}_s[\|\Delta_s - \mathbb{E}_s[\Delta_s]\|_2^2]$. Across all noise levels, the measured systematic fraction $\alpha^*(t) \in [0.89, 0.99]$ on three base models (Fig. 7). Thus, approximately 95% of the history-induced velocity error is systematic, indicating that a deterministic function can learn to correct it. Because $\alpha^*(t)$ remains close to 1 throughout denoising, we apply the full correction at every step without a gating schedule, as validated in §4. The same checkpoint produces clean clips in bidirectional single-shot mode, which serve as our clean histories (§3.2.4), indicating that drift arises from autoregressive rollout rather than limited model capacity.

3.2.2 One-Step Counterfactual-Teacher Correction (r_ϕ)

We train the corrector to match the frozen model's clean-history prediction at the same noisy state:

$$\mathcal{L}_{\mathrm{CF}} = \mathbb{E}_{k,t,\epsilon}\left[\frac{\|v_{\theta+\mathrm{LoRA}}(z_t, h_{\mathrm{gen}}, t) - v_\theta(z_t, h_{\mathrm{clean}}, t)\|_2^2}{\|v_\theta(z_t, h_{\mathrm{clean}}, t) - v_\theta(z_t, h_{\mathrm{gen}}, t)\|_2^2}\right]. \tag{4}$$

Each training example selects a chunk x_0 from a cached drifted rollout and re-noises it as $z_t = (1-\sigma_t)x_0 + \sigma_t\epsilon$. The student evaluates z_t under h_{gen} with the corrector active, while the frozen teacher evaluates the same state under h_{clean}. The denominator normalizes each example by the magnitude of its original drift gap. We report $R^2 = 1 - \mathcal{L}_{\mathrm{CF}}$, the average fraction of each example's drift gap that the corrector closes. The corrector r_ϕ consists of rank-16 LoRA adapters on the self-attention projections, totaling 5.9M parameters (0.4% of the base). We initialize each LoRA output matrix B to zero, so the corrected model initially matches the frozen base exactly.

Anchoring. Anchoring z_t to a ground-truth trajectory instead can pair the generated history with a state from a different scene, producing a mismatched correction that increases drift in our experiments (§4). Under flow matching, the gap between the predicted clean latents scales linearly with the noise level, $\|\Delta\hat{x}_0\| = \sigma_t\|\Delta v\|$ (Fig. 2, panel 1).

3.2.3 Closed-Loop Training: DAgger and Drift Contraction

The one-step objective is trained on states visited by the uncorrected model but deployed on states visited by the corrected model, creating the covariate shift addressed by DAgger (Ross et al., 2011). To close this gap, we augment one-step training with DAgger aggregation and drift contraction.

DAgger aggregation. We regenerate paired training data with the current corrector active and add it to the existing dataset. Training then samples from this aggregated pool, exposing the corrector to states encountered during its own rollouts.

Drift contraction. From the corrected velocity, we estimate the clean latent as

$$\hat{x}_0 = z_t - \sigma_t v_{\theta+\mathrm{LoRA}}(z_t, h_{\mathrm{gen}}, t). \tag{5}$$

We append $\hat{x}_0$ to the student history while retaining its gradient, then evaluate the normalized velocity gap on the re-noised next chunk x_0^+:

$$\mathcal{L}_{\mathrm{con}} = \mathcal{L}_{\mathrm{CF}}\big(x_0^+;\, h_{\mathrm{gen}} \oplus \hat{x}_0,\, h_{\mathrm{clean}} \oplus \bar{x}_0\big).$$

Here $\oplus$ denotes concatenation, and $\bar{x}_0$ is the clean reference chunk corresponding to $\hat{x}_0$. Backpropagating through $\hat{x}_0$ trains the corrector to reduce the drift its committed prediction causes in the next chunk. The total closed-loop training loss is

$$\mathcal{L}_{\mathrm{CL}} = \mathcal{L}_{\mathrm{CF}} + \lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}}, \tag{6}$$

where $\mathcal{L}_{\mathrm{con}}$ denotes the drift-contraction loss and $\lambda_{\mathrm{con}} = 0.5$. Ablations in §4 show that both DAgger aggregation and drift contraction contribute to reducing long-horizon drift.

3.2.4 Clean-History Sources

The teacher requires clean histories, but these need not come from real videos. We evaluate three settings: (a) *real clips*, for which 40 clips suffice; (b) *prompts only*, where the checkpoint generates single-shot 5 s clips in bidirectional mode, which avoids autoregressive drift and provides clean self-supervision; and (c) *no clean reference*, where reference-free distribution matching fails to reduce drift (§4). Because a bidirectional checkpoint can extend an existing clip but cannot reliably generate from a prompt alone, we causally adapt it using the same synthetic clips and merge the adapter into the base (Appendix A), after which Clean Forcing corrects the remaining drift, with the two providing complementary improvements (§4).

4 EXPERIMENTS

Implementation details. Our main base is Wan2.1-T2V-1.3B (Wan Team, 2025), configured with block-causal attention over 3-latent chunks, a rolling 21-latent KV cache, and a 20-step solver at 832×480, 16 fps, CFG 6, and shift 8. The *adapted base* has the rank-64 causal adapter from §3.2.4 merged into its weights. We also evaluate Wan2.1-T2V-14B, which is $10\times$ larger (Appendix G.1), and an independently trained Causal Forcing checkpoint (Zhu et al., 2026) (Appendix G.2). The corrector places rank-16 LoRA adapters on the q, k, v, and o self-attention projections in all 30 DiT blocks, yielding 5.9M trainable parameters (0.4% of the base). We initialize the LoRA output matrices to zero, use unit scaling, and train with AdamW at learning rates of 5×10^{-4} for one-step correction and 2×10^{-4} for closed-loop training. The causal adapter uses the same projections and a learning rate of 10^{-4}. Training uses at most 340 clips (40 real and 300 synthetic); the zero-real-video variant uses only the 300 synthetic clips, versus the 55K source clips and 81K generated trajectories of Po et al. (2025). The full pipeline, including generation of teacher clips, takes approximately one day on a single NVIDIA GB300 GPU in bf16 and requires at least 48 GB of GPU memory (Appendix A).

Baselines. We compare three groups. *(A) Methods implemented on our base*: context noising with $\sigma = 0.2$ (Chen et al., 2024; Skywork AI, 2025), history guidance with $w = 1.2$ (Song et al., 2025), and pathwise test-time correction under both published schedules (Xiang et al., 2026). We run these methods on the adapted base using their published hyperparameters. Table 4 documents the implementations and three in-domain controls evaluated on the unadapted base: DMD supervision without clean references (Yu et al., 2025; Yin et al., 2024), naive fine-tuning, and architecture variants. We also evaluate the independently adapted Causal Forcing checkpoint (Zhu et al., 2026) with and without our correction under the same protocol, since its release does not specify an inference configuration (Table 8). *(B) Released checkpoints*: Self Forcing (Huang et al., 2025) and SkyReels-V2 (Skywork AI, 2025), evaluated on the same 128 prompts. *(C) Published results*: BAgger (Po et al., 2025), reported on the same calibrated scale but using a different base and prompt set, making direct comparison indicative. All groups use the Wan2.1-1.3B architecture (Appendix B).

Evaluation metrics. Our main evaluation uses 50 s generations from 128 LLM-refined MovieGen-VideoBench prompts (Polyak et al., 2024) held out from training. All methods use identical prompt and seed pairs (Table 1). We report official VBench scores (Huang et al., 2024) and Δ-drift, defined as the MUSIQ score (Ke et al., 2021) of the first 20% of frames minus that of the full video (Zhang et al., 2025; Po et al., 2025). To check for metric-specific gains, we also report ΔQualityDrift (Meng et al., 2026), quality over time (Yin et al., 2025), color shift, anchored survival (Xiang et al., 2026), and temporal LPIPS (Table 11; Appendix C). We compare dynamic degree and flicker against ground-truth clips. Error bars show the SEM across prompts. Scoring every frame changes MUSIQ by at most 1.8 and Δ-drift by at most 0.26, without changing the Δ-drift ranking. Appendix I additionally evaluates *in-domain extension* using 5 held-out clips extended by 48 latent frames with 3 paired seeds (Table 10), and *short-horizon quality* using 200 VBench prompts at 5 s (Table 12).

4.1 RESULTS

Quantitative results. Table 1 validates our claim that a frozen base can match the drift reduction of models fully retrained at cluster scale. On the shared base, Clean Forcing reduces Δ-drift from $+11.27$ to $+1.65$ without real video and to $\mathbf{-0.16}$ with only 40 real clips, indicating no measurable quality decay over 50 s. Against released, retrained models evaluated on identical prompts, the $+$real variant is lower by more than 3 SEM, while the zero-real variant ties them using over $150\times$ less data. Even the zero-real variant outperforms the best published corrective fine-tuning result on the calibrated scale, with $+1.65$ versus $+3.57$ for R3†.

Drift reduction does not trade off overall quality. Every VBench dimension improves except motion smoothness, which decreases by 2.0 points as dynamic degree rises from 29.7 to 49.2: frame-difference metrics reward stillness, so recovering motion costs smoothness (Appendix C). On our adapted base, Clean Forcing improves subject consistency and imaging quality by 15–16 points and raises prompt adherence from 58.8 to 70.3/69.9, comparable to the distilled model's 70.1 (Table 13). Its aesthetic score reaches 61.4 on this base and a table-leading 62.9 on the Causal-Forcing base. Quality also remains stable when the same videos are evaluated at 10, 30, and 50 s (Table 14).

Table 1: **Long-horizon OOD generation.** Results on 128 MovieGen prompts at 50 s, with ± denoting SEM. Δ-drift is first-20% minus full-video MUSIQ (Po et al., 2025). Across its variants, Clean Forcing achieves the lowest drift and highest aesthetics. It ranks second in subject and background consistency, imaging quality, and MUSIQ. Among shared-prompt rows, bold with light red marks the best value and underline with light green the second best. Published † values use different prompts, and dynamic degree only checks for motion collapse; neither is ranked (Appendix C).

	Official VBench ↑					Motion	Long-Horizon Quality	
Method	Subj.	Backg.	Aesth.	Imag.	Smooth	Dynamic	MUSIQ↑	Δ-drift↓
Matched base								
Adapted base	65.7	76.7	44.9	47.9	91.1	29.7	44.8	+11.27±0.81
Context noise ($\sigma = 0.2$)	67.2	77.8	45.2	51.4	85.8	32.0	49.1	+12.05±0.76
History guidance ($w = 1.2$)	64.6	76.1	42.4	47.7	94.5	0.8	42.3	+20.55±0.65
Pathwise TTC ($\{500, 250\}$)	59.7	74.0	40.2	37.4	91.0	64.1	35.5	+13.96±0.91
Pathwise TTC ($\{750, 500\}$)	61.5	75.4	42.6	39.9	91.2	71.1	38.0	+14.51±0.80
Clean Forcing (zero real videos)	81.0	86.6	60.5	63.9	89.0	48.4	66.9	+1.65±0.35
Clean Forcing (+40 real clips)	80.9	86.4	61.4	64.1	89.1	49.2	67.0	**−0.16**±0.12
Independent base								
Causal-Forcing base	64.9	75.7	45.8	56.1	90.6	33.6	52.0	+11.99±0.67
Clean Forcing (zero real videos)	81.8	87.9	**62.9**	62.0	91.1	70.3	62.1	+2.80±0.42
Released models								
Self Forcing	80.2	84.4	60.9	**67.7**	94.9	39.8	**69.3**	+1.08±0.39
SkyReels-V2-DF	**87.0**	**90.7**	61.0	62.8	**96.6**	50.8	61.5	+1.00±0.30
Published results (†)								
BAgger DF base†	80.7	87.7	53.1	60.0	98.1	81.3	—	+7.34
BAgger R3†	84.1	89.6	55.4	63.4	98.6	76.6	—	+3.57

SkyReels remains strongest in consistency and smoothness, two dimensions that benefit from large-scale base training.

All training-free interventions we evaluate increase drift relative to the base (+11.3): history guidance yields +20.6, the two pathwise TTC schedules yield +14.0 and +14.5, and context noising yields +12.1 (Appendix C). Among the methods in Table 1, only Clean Forcing achieves Δ-drift below +2 without retraining the base at cluster scale (Huang et al., 2025; Skywork AI, 2025).

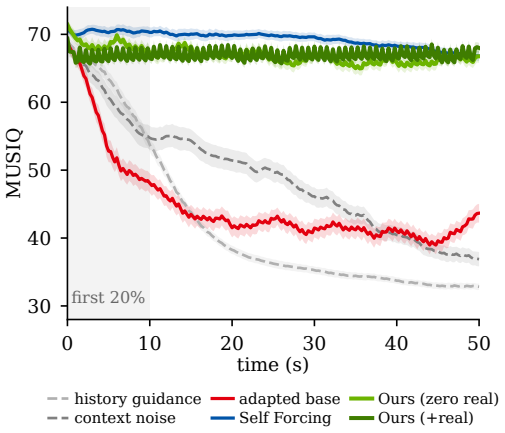

Figure 3: **Per-frame quality over 50 s.** Mean MUSIQ ± SEM across 128 prompts. Both Clean Forcing variants remain stable as the base and training-free baselines decline. Ripples mark chunk boundaries.

Clean Forcing recipe transfers without model-specific changes or hyperparameter search. On the independently adapted Causal-Forcing base, it reduces drift by 77%, from +11.99 to +2.80, recovers all consistency and aesthetic dimensions to the level of our adapted base, achieves the highest dynamic degree in the table at 70.3, and restores prompt adherence to 71.0 (Appendix G.2). At 10× scale, it removes 94% of drift on Wan2.1-14B without changing dynamic degree (Appendix G.1). With a lower correction strength, it also reduces drift in two additional autoregressive world models: HY-WorldPlay and a Cosmos-based driving model (Appendix G.3).

Qualitative results. Figure 3 shows when drift emerges: the base and all training-free heuristics degrade within 15 s, while both Clean Forcing variants remain stable through 50 s. Figure 4 shows the corresponding frames, in which the base exhibits saturation burnout, the heuristics cause fragmentation or freezing, and Self Forcing develops a late color shift. Appendix K presents six additional prompts, Figures 8 and 10 show the same comparison at 14B and on the Causal-Forcing base, the supplementary material includes videos for all figure prompts. Compared with models retrained

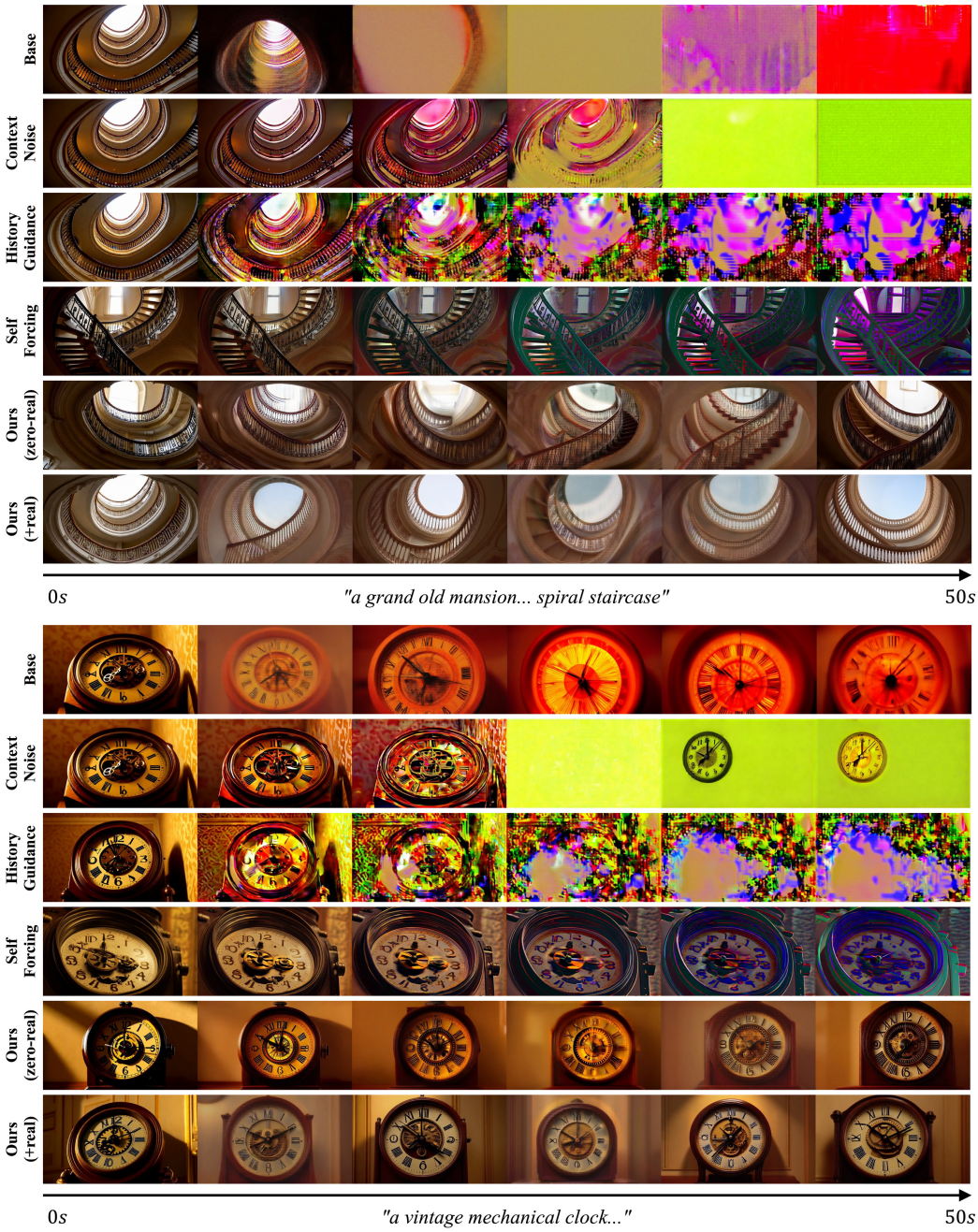

Figure 4: **Qualitative comparison over 50 s.** Frames are sampled at equal intervals for the spiral staircase (top, p098) and mechanical clock (bottom, p101) prompts. The base and training-free baselines develop saturation or structural artifacts, while Self Forcing shifts color. Both Clean Forcing variants retain the scene throughout the rollout.

at cluster scale on the same prompts, Clean Forcing achieves the lowest drift and highest aesthetic score and ranks second in subject and background consistency, imaging quality, and MUSIQ.

Efficiency analysis. At full correction strength, the corrected velocity is exactly the LoRA-merged forward pass, so r_ϕ folds into the base weights (Hu et al., 2022) without changing computation, parameter count, peak memory, or NFE ($\Delta\mathrm{NFE} = 0$). Wall-clock latency is statistically indistinguishable from the base (Table 5 in Appendix E), with a 3.5% increase and Welch $p = 0.40$.

Table 2: **Ablation studies.** We extend held-out clips by 48 latent frames (∼12 s), using paired clips and seeds across comparable rows. Bold and underlining indicate the best and second-best reported values. (Drift removed is measured relative to the unadapted base; only this metric was recorded for the DAgger-only checkpoint. Factual targets[a] collapse motion, making drift and MUSIQ uninformative (Appendix C). GT-anchored targets[b] were evaluated on a separate dataset and increased drift, so their values are not directly comparable.)

Configuration	Sat. drift↓	Drift removed↑	Late MUSIQ↑
Progressive training			
Unadapted base	0.491	—	29.6
One-step correction	<u>0.166</u>	66%	<u>36.9</u>
+ DAgger aggregation	—	<u>70%</u>	—
+ Drift contraction (**Ours**)	**0.131**	**73%**	**60.5**
Alternative targets			
No clean reference (DMD targets)	0.492	0%	29.6
Factual targets[a]	—	—	—
GT-anchored targets[b]	—	—	—

Merging with a nonconstant schedule $\alpha^*(t)$ is discussed in Appendix E. An unfused corrector adds 35% latency, while history guidance doubles NFE and pathwise TTC requires additional denoising passes. Thus, only retraining and Clean Forcing add no inference overhead, and only Clean Forcing is also cheap to train.

User study. Following Yin et al. (2025), 33 raters (29 from Prolific and 4 directly invited) each evaluated 15 randomized pairs of 10 s late-window excerpts ($t = 40$–50 s), with attention checks and at least 5 ratings per pair, yielding 858 judgments. The +real variant of Clean Forcing is preferred over the base in $\mathbf{93.0}\%$ of judgments and over Self Forcing in $\mathbf{83.9}\%$, both at $p < 10^{-16}$. Its comparison with SkyReels-V2 is not significantly different from an even split (42.7%, $p = 0.09$), although SkyReels-V2 is preferred for stability and outputs at 24 fps versus 16 fps for all other methods. Table 16 reports per-question results, and Appendix J details the protocol and catch-trial analysis.

4.2 Ablations

Table 2 validates the contribution of each main component of Clean Forcing through progressive additions and individual ablations. We observe that one-step counterfactual correction removes 66% of drift, supporting the clean-history velocity gap as a correction target. Furthermore, the closed-loop components provide further improvements. DAgger raises this to 70%, and drift contraction to 73%, while late-window MUSIQ rises from 36.9 to 60.5. With fixed training data at full scale, closed-loop training reduces Δ from $+4.77$ to $+1.65$. Additionally, the target ablations explain why the counterfactual matters: reference-free distribution matching removes no drift, factual targets collapse motion, and GT-anchored targets pair different scenes and increase drift. Further ablations cover capacity, architecture, LoRA rank, correction schedules, robustness, scene progression, and checkpoint transfer (Table 9; Appendix I; Appendix C; Appendix D; Appendix E).

5 Conclusion

We present Clean Forcing, a 5.9M-parameter LoRA corrector that learns the same-state velocity gap between drifted and clean histories, is refined on its own rollouts, and merges into the frozen base without inference overhead. Because $\sim 95\%$ of this gap is systematic, Clean Forcing matches cluster-scale retraining over 50 s using over $150\times$ less data, while its zero-real-video variant achieves Δ-drift of $+1.65$. Future work could apply the same loss during causal adaptation (Appendix F) or combine clean-history supervision with distribution matching to preserve progression. Exposure bias need not be trained away when its effects can be measured and corrected.

AI USE STATEMENT

We used generative AI tools to proofread and edit the manuscript and to help automate experiments and ablations. The authors reviewed all AI-assisted edits, independently verified the results, and take responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

All reported experiments run on a single GPU. We will release the training code for the corrector and causal adaptation, merged and unfused corrector checkpoints, the exact 128 evaluation prompts and their seeds, and implementations of every Group-A baseline. A provenance table maps each baseline to its source equations and hyperparameters (Appendix B). The release will also include the evaluation pipeline, official VBench custom-input evaluation, and all reported drift metrics and sanity checks. Appendix A specifies the hyperparameters for each training stage. The zero-real-video variant requires only prompts and approximately one GPU-day.

ETHICS STATEMENT

Improving the temporal stability of generated video carries the dual-use risks of synthetic media, including misinformation. Our method corrects drift in content produced by the base model. Training uses either the base model's own outputs in the zero-real-video variant or a small set of licensed clips; it collects no personal data. The preference study in §4.1 is the only component involving human participants. Its recruitment, protocol, and exclusion criteria are documented in Appendix J. We follow the base model's license and will release our code for research use.

REFERENCES

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia*, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, and Dominik Lorenz. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *NeurIPS*, 2024.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025.

Yuwei Guo, Ceyuan Yang, Hao He, Yang Zhao, Meng Wei, Zhenheng Yang, Weilin Huang, and Dahua Lin. End-to-end training for autoregressive video diffusion via self-resampling. *arXiv preprint arXiv:2512.15702*, 2025.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *NeurIPS*, 2022b.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *NeurIPS*, 2025. arXiv:2506.08009.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, 2021.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jia-Liang Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ICLR*, 2023.

Hongyu Liu, Chun Wang, Feng Gao, Xuanhua He, Yue Ma, Ziyu Wan, Yong Zhang, Xiaoming Wei, and Qifeng Chen. OPSD-V: On-policy self-distillation for post-training few-step autoregressive video generators. *arXiv preprint arXiv:2607.08766*, 2026.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Yu Meng, Xiangyang Luo, Letian Li, Wenyuan Jiang, Chen Gao, Xinlei Chen, Yong Li, and Xiao-Ping Zhang. Tethercache: Stabilizing autoregressive long-form video generation with gated recall and trusted alignment. *arXiv preprint arXiv:2606.13035*, 2026.

Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. *arXiv preprint arXiv:2308.15321*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Xurui Peng, Chenqian Yan, Hong Liu, Rui Ma, Fangmin Chen, Xing Wang, Zhihua Wu, Songwei Liu, and Mingbao Lin. Ertacache: Error rectification and timesteps adjustment for efficient diffusion. *arXiv preprint arXiv:2508.21091*, 2025.

Ryan Po, Eric Ryan Chan, Changan Chen, and Gordon Wetzstein. Bagger: Backwards aggregation for mitigating drift in autoregressive video diffusion models. *arXiv preprint arXiv:2512.12080*, 2025.

Adam Polyak et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *ICLR*, 2016.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011.

Sand.ai. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Florian Schmidt. Generalization in generation: A closer look at exposure bias. In *EMNLP Workshop on Neural Generation and Translation*, 2019.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls, 2026. URL https://arxiv.org/abs/2511.01266.

Skywork AI. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *ICML*, 2025. arXiv:2502.06764.

Jiahao Tian, Chenxi Song, Wei Cheng, and Chi Zhang. Free-lunch long video generation via layer-adaptive o.o.d correction. *arXiv preprint arXiv:2603.25209*, 2026.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 2018.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NeurIPS*, 2016.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xunzhi Xiang, Zixuan Duan, Guiyu Zhang, Haiyu Zhang, Zhe Gao, Junta Wu, Shaofeng Zhang, Tengfei Wang, Qi Fan, and Chunchao Guo. Pathwise test-time correction for autoregressive long video generation. *arXiv preprint arXiv:2602.05871*, 2026.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025. arXiv:2412.07772.

Zhengyang Yu, Akio Hayakawa, Masato Ishii, Qingtao Yu, Takashi Shibuya, Jing Zhang, and Yuki Mitsufuji. Autorefiner: Improving autoregressive video diffusion models via reflective refinement over the stochastic sampling path. *arXiv preprint arXiv:2512.11203*, 2025.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *NeurIPS*, 2025.

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. *arXiv preprint arXiv:2602.02214*, 2026.

APPENDICES

The following appendices supplement the main text.

– Appendix A Training Details and Hyperparameters

– Appendix B Baseline Implementation Fidelity

– Appendix C Evaluation Checks for Degenerate Outputs

– Appendix D The Drift–Progression Trade

– Appendix E Applicability, Scaling, and Transfer

– Appendix F Clean Forcing as a Training Regularizer

– Appendix G Transfer to Other Bases

  G.1 Scale: Wan2.1-14B

  G.2 Independently Trained Base: Causal Forcing

  G.3 Other Autoregressive World Models

– Appendix H Extended Ablations

– Appendix I Additional Results on Drift, Quality, and Diversity

– Appendix J User Study Details

– Appendix K Additional Qualitative Results

– Appendix L Limitations

A TRAINING DETAILS AND HYPERPARAMETERS

Algorithm 1 Clean Forcing training

Require: Frozen AR base v_θ; corrector r_ϕ (LoRA on v_θ, zero-init); weight λ_{con}
Require: Clean references h_{clean} (bidirectional single-shot); drifted rollouts cached from the adapted base
loop ▷ stage 1: one-step
  Sample chunk x_0 from a drifted rollout, timestep t, $\epsilon \sim \mathcal{N}(0, I)$
  Set $z_t \leftarrow (1 - \sigma_t)\, x_0 + \sigma_t \epsilon$
  Update r_ϕ on $\mathcal{L}_{\text{CF}}$, the velocity gap at z_t between drifted and clean history ▷ Eq. 4
end loop
Regenerate rollouts with r_ϕ active; aggregate into the pool ▷ DAgger
loop ▷ stage 2: closed-loop
  Sample from the pool; set $\hat{x}_0 \leftarrow$ corrected prediction ▷ Eq. 5
  Commit $\hat{x}_0$ to the next chunk's history, with gradient
  Update r_ϕ on $\mathcal{L}_{\text{CF}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}}$ ▷ Eq. 6
end loop
Merge r_ϕ into v_θ at α=1 ▷ Eq. 2; no inference overhead

All training stages run on one GPU in bf16, with LoRA adapters kept in fp32 and gradient checkpointing enabled. We use AdamW with PyTorch defaults (weight decay 0.01 and $\beta = (0.9, 0.999)$), linear warmup followed by a constant learning rate, gradient clipping at 1.0, batch size 1, and no EMA. The base uses block-causal attention with 3-latent-frame chunks, a rolling 21-latent KV window, and 20 solver steps. Generation uses CFG 6.0, shift 8.0, 832×480 resolution, and 16 fps, yielding clips of $21\times16\times60\times104$ latents. Data sampling uses seed 0, while rollout noise is seeded per clip. Table 3 lists the settings for each stage.

Causal adaptation. We adapt the base using the Diffusion Forcing objective. For each of the seven chunks, we independently sample $u \sim U(0, 1)$, transform it to $\sigma = 8u/(1 + 7u)$, and compute the

Table 3: **Training hyperparameters.** Training uses AdamW, bf16, batch size 1, linear warmup, and gradient clipping at 1.0.

	Causal adaptation	Correction (one-step)	Correction (closed-loop)	Teacher clips
LoRA rank	64	16	16 (from one-step)	—
Learning rate	10^{-4}	5×10^{-4}	2×10^{-4}	—
Warmup steps	100	60	40	—
Steps / clips	6,000 (4K deployed)	800	600	300 clips
Loss / sampling	DF velocity MSE	Eq. 4	Eq. 6	30-step single shot
Length	21 latents (7 chunks)	48 (real) / 21 (synth.) latents	48 / 21 latents + 1 contraction chunk	21 latents
Data	300 synthetic clips	40 real or 300 synthetic clips	Round 0 + one DAgger pool	MovieGen prompts

velocity MSE. We merge the rank-64 LoRA weights into the base after adaptation and reserve 24 of the 300 synthetic clips for validation.

One-step corrector. At each step, we sample a rollout position $k \in [9, 45]$ and a timestep from the 20-step solver schedule. We re-noise the corresponding drifted rollout chunk as $z_t = (1-\sigma_t)x_0+\sigma_t\epsilon$ and optimize the normalized loss in Eq. 4. The clean-history teacher uses the same base network with the corrector LoRA scaled to zero. We initialize the LoRA output matrices B to zero and sample entries of A from $\mathcal{N}(0, 1/r)$, so the corrected network initially matches the base. Validation uses $\max(4, N/8)$ held-out clips.

Training pairs. Each pair aligns a drifted rollout with a clean reference by latent position $k \in [9, 45]$ ($k \in [9, 18]$ for 21-latent synthetic pairs). The adapted base generates the rollout from the reference's first three latents and caption. For chunk k, h_{gen} and h_{clean} contain the preceding nine latents from the rollout and reference, respectively, while $\bar{x}_0$ is the aligned reference chunk. References and rollouts span 48 latents for the 40 real clips and 21 latents for synthetic clips. DAgger rollouts reuse the same references, captions, and seeds with the corrector active. In both losses, the denominator uses frozen-base velocities computed without gradients and adds 10^{-8} for numerical stability.

Closed-loop corrector. Starting from the one-step checkpoint, we regenerate paired rollouts with the corrector active at scale 1, reusing the clips and noise seeds from the initial pool. We then sample uniformly from both pools. For drift contraction, we append the corrected estimate $\hat{x}_0 = z_t - \sigma_t v_{\theta+\text{LoRA}}$ to the next chunk's history without detaching its gradient and penalize that chunk's normalized velocity gap against the clean teacher. This term has weight 0.5. We sample positions $k \in [9, 42]$ ($k \in [9, 15]$ for synthetic pairs) to leave room for the subsequent chunk.

Synthetic teacher clips. We run the checkpoint in bidirectional single-shot mode with full attention, 30 solver steps, and default guidance. Using MovieGen-bench prompts disjoint from all evaluation prompts and a fixed seed for each clip, we generate 21-latent clips. These clips provide clean histories for the zero-real-video variant and data for causal adaptation.

B Baseline Implementation Fidelity

We implement the Group-A baselines on our model because the published methods do not all provide code for this backbone or specify an operation compatible with a deterministic multi-step solver. Table 4 records the source and settings for each implementation, together with its agreement with published qualitative findings. We use the same base and evaluation protocol for the first three methods and evaluate the remaining controls in domain on the unadapted base. These comparisons test the methods in our setting; they do not reproduce every experimental condition of the source papers.

The context-noising and history-guidance results agree with the published observation that neither intervention alone resolves long-horizon drift. Our implementations of the remaining controls likewise recover their reported failure modes or isolate the effect of changing the corrector architecture. We will release all Group-A implementations and their provenance settings.

Table 4: **Group-A baseline implementations and outcomes.** The first three baselines use the 128-prompt, 50 s protocol (Table 1). The remaining controls use in-domain extensions on the unadapted base (Tables 2 and 9).

Baseline	Implementation	Observed outcome
Context noise ($\sigma = 0.2$)	DF-style context noising (Chen et al., 2024; Skywork AI, 2025).	Increases drift, consistent with the finding that noising alone does not prevent it (Po et al., 2025).
History guidance ($w = 1.2$)	DFoT text and history guidance with fractional history masking; a simplified variant was also tested (Song et al., 2025).	Neither variant reduces drift, consistent with the replication in Po et al. (2025).
Pathwise TTC	Algorithm 1 of Xiang et al. (2026) with placements $\{500, 250\}$ and $\{750, 500\}$. We re-encode the reference chunk at the adjacent RoPE position.	Both schedules worsen drift. On our multistep base, the reference prediction is blurry when committed as history.
DMD-LoRA	AutoRefiner-style DMD supervision without a clean reference, adapted to our deterministic solver (Yu et al., 2025; Yin et al., 2024).	Removes 0% of drift.
Naive LoRA fine-tune	Flow matching on clean clips, following the +LoRA control in Yu et al. (2025).	Collapses motion, as reported for that control.
Full fine-tune / side network	Counterfactual loss with alternative corrector parameterizations.	Removes 23% / 25% of drift, versus 73% for rank-16 LoRA (Table 9).

The released Group-B checkpoints also differ in training resources. Self Forcing and related DMD approaches train a 1.3B causal generator using signal from a 14B bidirectional teacher alongside a 1.3B fake critic, whereas our corrected models use the frozen 1.3B base as their only teacher. This difference can affect absolute quality and should be considered when comparing checkpoints, although it does not by itself explain the observed performance gap.

C Evaluation Checks for Degenerate Outputs

A drift metric can improve even when the generated video deteriorates. Several baselines illustrate distinct ways this happens. Naive LoRA fine-tuning obtains subject-consistency and imaging scores of 66.0 and 58.5, but its output becomes nearly static: dynamic degree falls to 12.5 on the finals set and 5.2 in domain. History guidance reaches 94.5 motion smoothness at a dynamic degree of only 0.8. At the opposite extreme, collapsing unadapted rollouts register dynamic degrees of 81–100 because unstable changes count as motion. Pathwise TTC likewise reaches 64–71, roughly $2.2\times$ the base, although its outputs are blurred. That blur also lowers consecutive-frame flicker from 15.9 for the base to 9.2 for TTC by making adjacent frames more alike.

We therefore interpret dynamic degree and flicker relative to ground-truth clips, compare quality at matched levels of motion (Xiang et al., 2026), and inspect frame strips for each batch. Reading the metrics together helps distinguish sustained, coherent motion from stillness, blur, and unstable changes.

D The Drift–Progression Trade

Qualitative inspection revealed a limitation that the main drift score does not capture: corrected scenes can remain close to their opening rather than progress. We measure this tendency as *anchoring*, the DINO-feature similarity between the last fifth of a video and its opening. This measures scene progression and is distinct from the noisy-state anchoring discussed in §3. We also compare each prompt with Self Forcing to account for differences in the progression implied by the prompt.

Figure 5 shows that stronger drift reduction generally coincides with greater anchoring across our variants. One possible mechanism is that drift contraction favors predictions compatible with the clean references, many of which are five-second clips with limited scene changes. Sweeping the

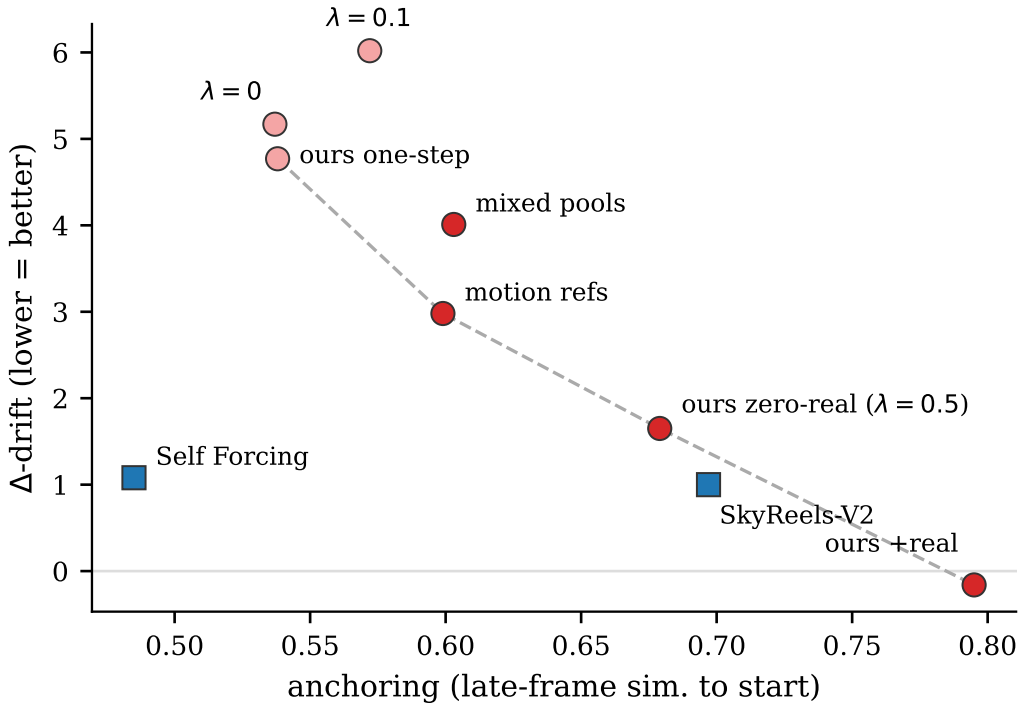

Figure 5: **Drift reduction and scene progression.** Anchoring is the DINO (Caron et al., 2021) similarity between the opening and final fifths of a video; higher values indicate less progression. Hollow circles show Clean Forcing variants, filled circles its reported zero-real and +real configurations, and squares other systems. Across Clean Forcing settings, lower Δ-drift generally coincides with higher anchoring. Self Forcing, trained on over 10^5 clips, achieves low drift with less anchoring.

contraction weight λ moves along this tradeoff, while changing the reference distribution shifts it modestly. SkyReels-V2 lies near the same frontier, with anchoring 0.70 at $\Delta = +1.00$. Self Forcing attains more progression at low drift after distribution matching on over 10^5 clips, although the comparison involves a different training recipe and scale. Anchoring alone does not account for our drift reduction: on the less anchored half of each configuration, Δ remains $+2.2$ for zero real videos and -0.1 for +real, versus $+11.3$ for the base. Across videos, the correlation between anchoring and drift is -0.05.

We also tested whether access to future frames could improve this tradeoff. A bidirectional teacher conditioned on clean past, z_t, and clean future reaches anchoring 0.64 at $\Delta \approx +0.1$, compared with approximately 0.78 on the measured frontier, but its videos visibly blur. Replacing the clean future with the model's predicted future removes much of the blur and raises anchoring to 0.70. These results suggest that future information can improve measured progression, but an unpredictable future creates an ambiguous regression target.

A separate limitation is frequent shot changes. Our zero-real videos average about 0.9 cuts (0.04 for +real), versus 0.1–0.3 for the retrained models; the uncorrected base averages 2.8 (Table 15). This behavior is partly inherited from the base: extending causal adaptation from 4K to 20K steps roughly halves cuts and flicker and increases MUSIQ by 13 points (Appendix E). We report anchoring and cut rate alongside drift because neither is captured reliably by the headline quality metrics.

E Applicability, Scaling, and Transfer

Scaling the base instead. Figure 6 evaluates uncorrected bases after 4K–20K causal adaptation steps using the same data and training recipe. Quality continues to improve through 20K, with a

Table 5: **Inference overhead.** Timings use three runs of 21 steady-state chunks on one GPU (bf16, batch 1), with matched prompt and seed. Chunk latency excludes decoding; decoded fps includes it. Merging adds no NFE or peak memory, with no significant latency difference from the base (Welch $p = 0.40$). First-frame latency is unchanged.

	Chunk latency		Compute		Throughput (fps)	
Method	s/chunk↓	vs. base	ΔNFE	Peak VRAM (GiB)	DiT↑	With decode↑
Adapted base	2.90 ± 0.35	—	—	29.49	4.10	3.88
Merged corrector	3.00 ± 0.89	$+3.5\%$	0	29.49	4.01	3.80
Unfused corrector	3.92 ± 0.49	$+35.3\%$	0	29.51	3.02	2.91

substantial gain between 14K and 20K. The corrected 4K base nevertheless reaches 66.9 MUSIQ, above the uncorrected checkpoints on this curve, with 1,400 corrector updates versus 16,000 additional adaptation steps. Retraining a corrector on the 20K base yields 62.3 MUSIQ, below the corrected 4K base. Thus, improving the base does not guarantee a better corrected rollout: the corrector must fit the errors of the particular checkpoint to which it is applied.

Increasing the adaptation chunk size from three to seven latents also improves the uncorrected base at the same training budget, giving late MUSIQ 54.2 and 0.25 cuts per video. It remains below the corrected three-latent base at approximately 67.0 late MUSIQ, and the periodic chunk artifact persists at the new cadence. Longer adaptation and larger chunks improve base quality, but neither eliminates the measured drift on its own.

Checkpoint transfer. When we apply a corrector to a sibling checkpoint, late MUSIQ remains 56.6, compared with 59.4 on its native checkpoint, but it retains only about 25% of its drift calibration. Transfer across model types is weaker: a corrector trained on a multi-step velocity model gives $R^2 = +0.03$ on a distilled four-step x_0 model, compared with $+0.39$ on its source. The training recipe transfers to other bases, but the learned weights are specific to the base whose error they correct.

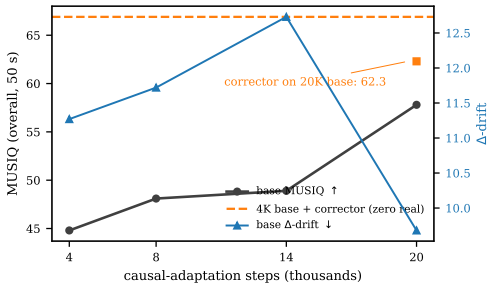

Figure 6: **Longer causal adaptation does not replace correction.** MUSIQ (black, left axis) improves from 4K to 20K steps, but Δ-drift (blue, right axis) remains substantial. The corrected 4K base reaches 66.9 MUSIQ (dashed line); a corrector retrained on the 20K base reaches 62.3 (square).

Within-family evidence. The same recipe yields a weaker corrected result on a base trained for five times as many adaptation steps, despite that base having less drift before correction. A trajectory-conditioned probe also fails to explain additional gap variance beyond the current state under cross-validation with swept regularization. Together, these results indicate that the correctable signal depends on the particular base and is not recovered simply by adding trajectory context.

Correction schedules. Eq. 2 blends model outputs, whereas merging a scaled LoRA update yields $v_{\theta+\alpha\Delta W}$. Because the network is nonlinear in its weights, these expressions generally differ for $0 < \alpha < 1$; they coincide at $\alpha \in \{0, 1\}$. All main results use $\alpha = 1$. Timestep-dependent strength $\alpha^*(t)$ requires no additional forward pass when a merged weight set is stored for each distinct strength. The reduced strengths in Appendix G.3 use this implementation.

When correction does not help. We applied the full recipe to Self Forcing's distilled generator, using an x_0-space loss at its four distillation timesteps. The one-step objective was learnable ($R^2 = 0.48$), but the deployed corrector reduced MUSIQ from 69.3 to 47.5 and increased Δ from $+1.08$ to $+8.3$; DAgger did not reverse this deterioration. Before training, the diagnostic showed a smaller systematic fraction, $\alpha^*(t) \in [0.65, 0.80]$, than on the bases where correction helped (0.89–0.99), and the velocity gap itself was about ten times smaller. A modest fitting error can therefore outweigh the remaining correctable signal when a short solver commits each prediction.

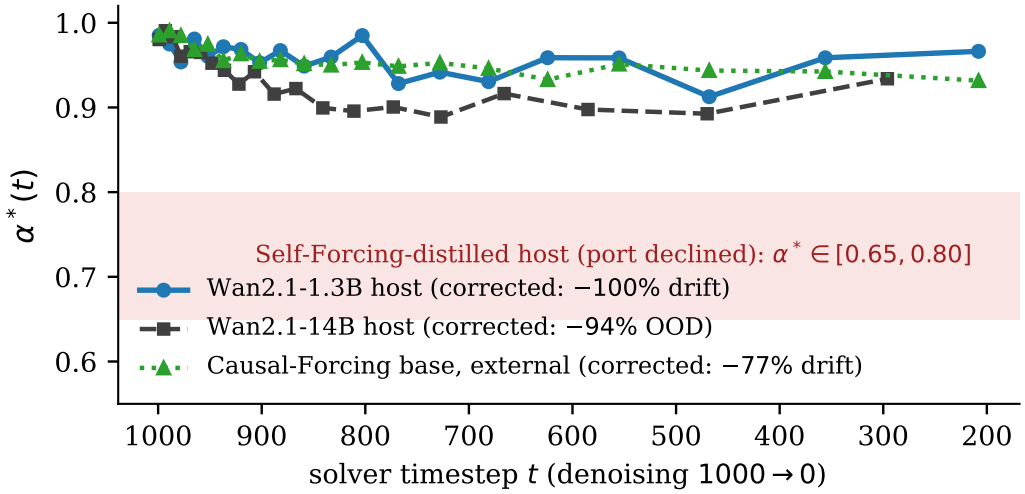

Figure 7: **Systematic velocity gap across bases.** Per-timestep $\alpha^*(t)$ (Eq. 3) is high on the three bases that benefit from correction (markers). The Self Forcing distilled generator has a lower measured range, $[0.65, 0.80]$ (shaded band), and its corrected rollout deteriorates. All measurements precede corrector training.

The product of the measured gap magnitude and its systematic fraction provides a practical indicator of potential benefit, rather than a guarantee. It identifies lightly adapted bases with a large, reproducible history-induced error as candidates for correction and suggests limited headroom on a distilled base that already has little drift. Figure 7 plots the per-timestep fractions for these bases. The observed failure of the distilled port is consistent with the possibility that large-scale distillation has already reduced much of the systematic gap, although these experiments do not isolate that mechanism.

Concurrent results on related distilled models report further gains after generator updates using real-video signal (Liu et al., 2026). This is consistent with limited room for a frozen-base corrector after distillation, though those results use a different training method.

F Clean Forcing as a Training Regularizer

The counterfactual loss could also be incorporated into causal training. If training already includes on-policy student rollouts, as in Self Forcing-style recipes, the following term would compare the student's prediction under its generated history with a slowly updated teacher's prediction under a clean history:

$$\mathcal{L} \;=\; \mathcal{L}_{\mathrm{FM}} \;+\; \lambda_{\mathrm{CF}}\,\mathbb{E}\,\frac{\left\|v_\theta(z_t, h^\theta_{\mathrm{gen}}, t) - \mathrm{sg}[v_{\bar\theta}(z_t, h_{\mathrm{clean}}, t)]\right\|^2}{\left\|v_{\bar\theta}(z_t, h_{\mathrm{clean}}, t) - v_{\bar\theta}(z_t, h^\theta_{\mathrm{gen}}, t)\right\|^2}, \tag{7}$$

Here h^θ_{gen} comes from a short rollout of the current student, $\bar\theta$ denotes an EMA teacher, and sg stops gradients through the target. The EMA could stabilize the clean-history target as the student changes. This formulation aims to reduce the conditional mismatch during base training; we have not evaluated it experimentally. It may fit naturally into training loops that already compute student rollouts and teacher predictions. Used on its own, however, it requires rollouts during base training and would incur additional cost. Our reported experiments use the corrector trained after the base is fixed; Eq. 7 is a proposed extension.

Table 6: **Wan2.1-14B in-domain extension.** Five held-out clips are extended by 48 latent frames under three paired seeds. Both correctors use the zero-real data. Parentheses give drift removed relative to the corresponding base; the closed-loop checkpoint is selected by validation performance.

	Unadapted 14B		Adapted 14B	
Method	Sat. drift↓	Late MUSIQ↑	Sat. drift↓	Late MUSIQ↑
Base	0.329	28.4	0.147	38.7
+ One-step corrector	0.109 (67%)	**72.1**	0.059 (60%)	**71.5**
+ Closed-loop corrector	**0.098** (70%)	67.3	**0.055** (63%)	69.4

Table 7: **Wan2.1-14B long-horizon generation.** Results on 32 held-out MovieGen prompts at 50 s with the zero-real closed-loop corrector merged. ± denotes SEM across prompts. Dynamic degree checks for motion collapse and is not ranked.

	Quality		Temporal		Official VBench			Motion
Method	Δ-drift↓	MUSIQ↑	Lag-2s id.↑	Cuts/video↓	Subj.↑	Aesth.↑	Imag.↑	Dynamic
Adapted base	$+15.65_{\pm 1.41}$	$46.4_{\pm 1.5}$	0.791	0.375	67.1	52.0	54.4	37.5
+ **Clean Forcing**	$\mathbf{+0.97}_{\pm 0.50}$	$\mathbf{69.0}_{\pm 1.3}$	**0.915**	**0.19**	**86.3**	**65.7**	**64.9**	37.5

G Transfer to Other Bases

G.1 Scale: Wan2.1-14B

We apply the zero-real-video recipe to Wan2.1-T2V-14B, a model with roughly ten times the parameters, using the same protocol: 300 synthetic teacher clips, rank-64 causal adaptation, one-step correction, and one round of DAgger followed by drift contraction. The complete pipeline takes about three GPU-days.

Diagnostic and adaptation. The unadapted 14B base collapses within its first two generated chunks. Its counterfactual velocity gap remains predominantly systematic, with pooled $\alpha^* = 0.934$, $\alpha^* = 0.961$ at $t \geq 900$, and a relative gap of 0.205. Following the same checkpoint selection rule as at 1.3B, we deploy step 4K from a 6K adaptation run. This checkpoint has a saturation ramp of 0.035 and late-window MUSIQ of 65.0.

Corrector training. The one-step corrector reaches validation $R^2 = 0.482$. The closed-loop stage peaks at 0.643 before falling to 0.415 at its final checkpoint, so we deploy the validation-peak snapshot. This oscillation makes checkpoint selection consequential at 14B.

Results. On the paired in-domain extension task (Table 6), correction reduces saturation drift by 67–70% on the unadapted 14B and 60–63% on the adapted base, with gains of up to 43 points in late-window MUSIQ. On 32 held-out MovieGen prompts at 50 s, the closed-loop corrector reduces Δ-drift by 94%, from $+15.65$ to $+0.97$, and raises MUSIQ from 46.4 to 69.0 (Table 7; Fig. 8). Every reported VBench dimension improves while paired dynamic degree remains 37.5, supporting drift reduction without a loss of measured motion.

G.2 Independently Trained Base: Causal Forcing

Setup. To test transfer beyond our causal adaptation, we apply the same recipe to the released chunkwise autoregressive teacher from Causal Forcing (Zhu et al., 2026). It shares the Wan2.1-T2V-1.3B architecture and three-latent chunk structure with our base; all 825 checkpoint tensors load strictly. Its independent training uses teacher forcing, so the model sees clean ground-truth history rather than its own generated context. It produces sharp early frames, with first-20% MUSIQ of about 64, but its 50 s rollouts deteriorate within 10–15 s and reach $\Delta = +11.99$ (Table 1). This separates strong short-clip quality from robustness under self-generated history.

Protocol. Because the release does not specify a complete inference configuration, we evaluate it using our fixed setup: a rolling 21-latent KV window without an attention sink, 20-step UniPC

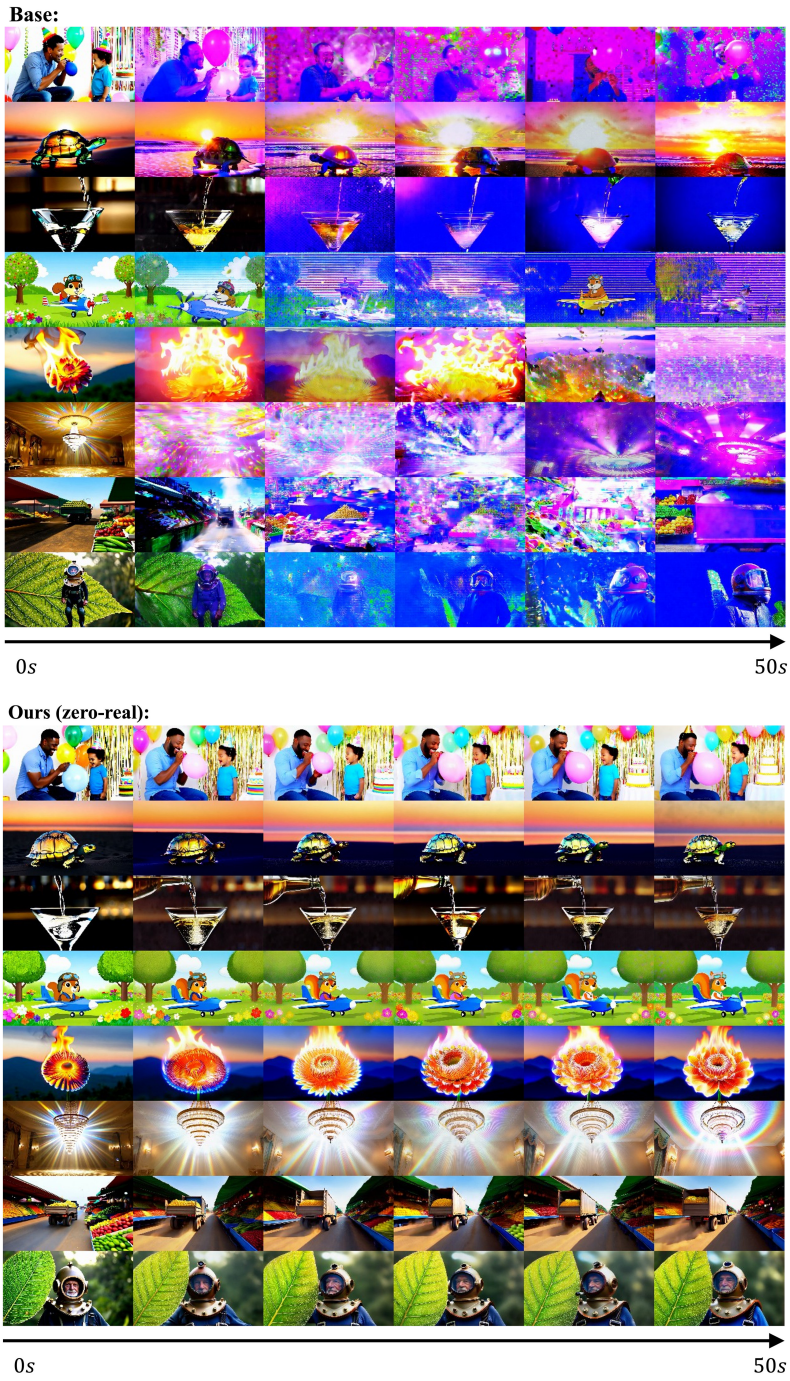

Figure 8: **Qualitative comparison on Wan2.1-14B.** Six frames sampled over 50 s for eight held-out prompts, with seeds matched between the adapted base (top) and the same base with the zero-real Clean Forcing corrector merged (bottom). The base progressively saturates and loses scene structure; the corrected model preserves the scenes.

Table 8: **Causal-Forcing inference robustness.** All configurations use the same 16 prompt–seed pairs at 50 s; ± denotes SEM across prompts. Rolling uses a 21-latent KV window. Anchoring and cuts are defined in Table 15.

	Quality		Temporal diagnostics	
Configuration	MUSIQ↑	Δ-drift↓	Anchoring	Cuts/video↓
Rolling window (our protocol)	$51.5_{\pm 2.3}$	$+11.07_{\pm 2.2}$	0.254	1.94
Full KV cache (no eviction)	$46.4_{\pm 3.9}$	$+16.34_{\pm 2.8}$	0.449	1.00
Rolling + sink (first chunk)	$54.7_{\pm 2.6}$	$+8.48_{\pm 1.3}$	0.410	2.56
Rolling + sink (half window)	$57.2_{\pm 3.6}$	$+5.29_{\pm 1.3}$	0.663	1.56
Rolling + context noise ($\sigma = 0.2$)	$47.8_{\pm 2.1}$	$+12.59_{\pm 2.2}$	0.180	2.75
+ Clean Forcing (our protocol)	$58.8_{\pm 3.6}$	$+3.94_{\pm 1.4}$	0.685	**0.50**
+ Clean Forcing + sink (half window)	$\mathbf{59.4}_{\pm 3.7}$	$\mathbf{+3.23}_{\pm 0.8}$	0.794	0.56

sampling, and seed equal to the prompt index. This differs from its 50-step sampling default. We therefore also test alternative cache configurations on the same 16 prompts (Table 8) to assess how much this choice affects the comparison.

Inference robustness. A full KV cache without eviction increases drift relative to the rolling window, from $+11.07$ to $+16.34$ on the subset. Attention sinks improve the uncorrected base, reaching $+5.29$ with a half-window sink; context noise worsens it to $+12.59$. Under the rolling-window protocol, the corrector reaches $+3.94$ on those prompts. Combining correction with the half-window sink reduces drift further to $+3.23$. Thus, the correction helps under the fixed protocol and remains useful alongside the strongest cache setting tested. These cache comparisons do not establish the best possible inference configuration for the released checkpoint.

Diagnostic and result. Before corrector training, the systematic fraction is high across timesteps, $\alpha^*(t) \in [0.93, 0.99]$ (Fig. 7), and the preliminary drift measurement is $\Delta = +10.83$. Both indicate substantial room for correction. On the full 128-prompt evaluation, one 36-hour run with a one-step stage, one DAgger round, and 600 closed-loop steps reduces drift from $+11.99$ to $+2.80$ (77%) and increases MUSIQ from 52.0 to 62.1, without a model-specific hyperparameter search or checkpoint selection.

The corrected checkpoint also improves subject consistency from 64.9 to 81.8, background consistency from 75.7 to 87.9, aesthetics from 45.8 to 62.9, dynamic degree from 33.6 to 70.3, and prompt adherence from 57.5 to 71.0 (Tables 1 and 13). Increased dynamic degree rules out simply freezing the output. The quality curve remains comparatively stable over 50 s (Fig. 9), while lagged identity improves and the cut rate falls by roughly a factor of three (Table 15). Compared with our primary corrected base, this external base retains more residual drift ($+2.80$ versus $+1.65$) and lower absolute MUSIQ (62.1 versus 66.9), so successful transfer does not imply identical final quality.

Artifacts across bases. The corrected Causal-Forcing rollouts exhibit the same luminance pulse at the three-latent chunk cadence as our primary system (Appendix L). Periodogram power at that cadence exceeds broadband power in both the uncorrected and corrected runs, with ratios of 309 and 693, respectively; neither has a corresponding peak at the 21-latent window cadence. This association with chunk boundaries across two bases suggests an artifact of chunked generation, although correction can make it more visible by reducing background degradation. Anchoring is likewise similar to the primary corrected base (last-fifth DINO similarity 0.680 versus 0.679; Table 15), while motion energy remains stable from the first fifth to the last in both arms.

On a 16-prompt subset, regenerate-and-blend decoding with two overlap steps adds one-third to decode NFE and lowers the cadence-band power ratio from 538 to 99 and autocorrelation from 0.69 to 0.53. Motion energy and anchoring remain approximately unchanged. The overlap experiment offers a partial mitigation for the pulse, with additional decode cost.

G.3 Other Autoregressive World Models

We apply the zero-real-video recipe to two additional autoregressive world models: HY-WorldPlay (5B), which supports interactive camera control, and a 2B driving model built on Cosmos-Predict2.5.

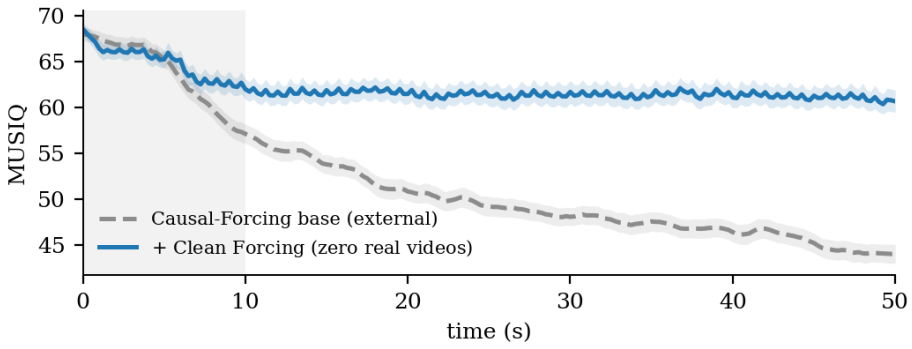

Figure 9: **Quality over 50 s on the Causal Forcing base.** Mean per-frame MUSIQ over 128 prompts, smoothed over five frames; bands show SEM. The shaded region marks the first 20% of each video. Both models start at similar quality, but the base declines throughout the rollout while the corrected model stabilizes after 10 s.

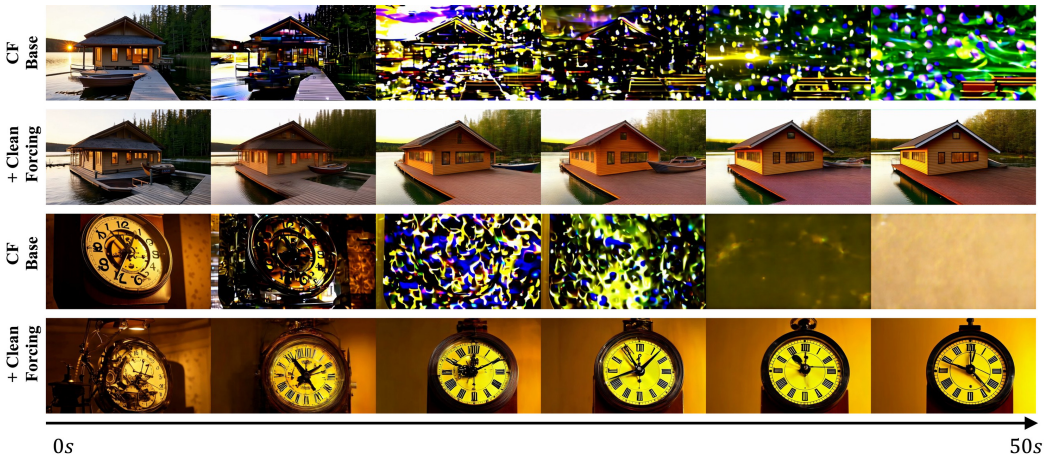

Figure 10: **Qualitative transfer to the Causal Forcing base.** Six frames sampled over 50 s for a lakeside cabin (top pair, p040) and a mechanical clock (bottom pair, p101). Each pair uses the same prompt and seed, with the base above and + Clean Forcing below. The base loses scene structure during rollout, while the corrected model retains recognizable content through 50 s.

Unlike our Wan bases, whose systematic fraction $\alpha^*(t)$ remains within 0.89–0.99, both models show substantial variation across timesteps (0.81 $\rightarrow$ 0.53 for HY-WorldPlay). We therefore use correction strengths of $0.5\alpha^*(t)$ and $0.25\alpha^*(t)$, respectively. Static-camera HY-WorldPlay rollouts show no drift and are left uncorrected. On held-out rollouts, correction reduces Δ-drift by 48% from +1.70 on HY-WorldPlay while preserving dynamic degree (19.3 $\rightarrow$ 20.3). On the driving model, it reduces Δ-drift by 60% from +2.44 and raises late-window MUSIQ by 5%, with 17% lower motion. Representative rollouts appear in the supplementary material.

H Extended Ablations

Capacity and placement. Table 9 changes the corrector architecture while holding the objective, data, and evaluation protocol fixed. Full fine-tuning and a side network each reduce drift by roughly one quarter, compared with 73% for rank-16 LoRA. Full fine-tuning uses 283M trainable parameters and reaches validation $R^2 = 0.763$, yet gives only a 23% drift reduction in rollout. Its validation fit is lower than the rank-16 LoRA's 0.854 in the rank sweep (Table 9); the difference in experimental settings should be considered when comparing those values. The 1.9M-parameter side network also

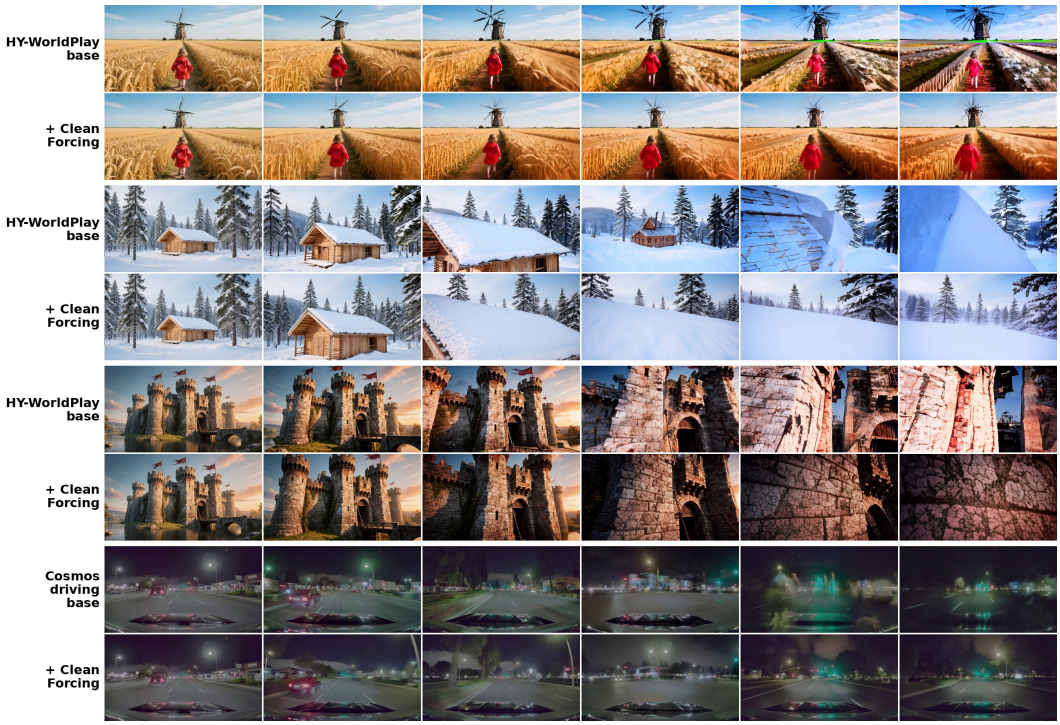

Figure 11: **Correction on other autoregressive world models.** Six evenly spaced frames from three HY-WorldPlay rollouts ($\sim$24 s) and one Cosmos-based driving rollout ($\sim$80 s). Each base and corrected pair shares its prompt, conditioning, and seed. We use strengths of $0.5\alpha^*(t)$ for HY-WorldPlay and $0.25\alpha^*(t)$ for the driving model. The corrected rollouts better retain color, detail, and scene structure as the base rollouts degrade.

Table 9: **Corrector capacity, placement, and rank.** Paired in-domain evaluation on 48-latent-frame extensions ($\sim$12 s). The upper panel compares corrector parameterizations on the unadapted base; the lower panel varies LoRA rank on the adapted base. Drift removal is relative to each panel's base. Bold marks the best scores and the two MUSIQ losses relative to rank 16; underlining marks second-best scores.

Configuration	Params	Sat. drift↓	Drift removed↑	Late MUSIQ↑
Rank-16 LoRA in backbone (**Ours**)	5.9M	**0.131**	**73%**	**60.5**
w/o low rank (full-rank fine-tune)	283M	0.379	23%	—
w/o in-backbone placement (side network)	1.9M	<u>0.367</u>	<u>25%</u>	—
Unadapted base	0	0.491	—	29.6

Rank	Params	Val. R^2	Sat. drift↓	Late MUSIQ↑	ΔMUSIQ vs. r16
8	3.0M	0.829	$0.048_{\pm 0.027}$ (89%)	<u>$55.3_{\pm 7.2}$</u>	$-1.0_{\pm 0.8}$ (n.s.)
16 (**Ours**)	5.9M	**0.854**	<u>$0.047_{\pm 0.033}$</u> (89%)	**$56.3_{\pm 8.5}$**	—
32	11.8M	<u>0.843</u>	$0.050_{\pm 0.012}$ (88%)	$49.1_{\pm 4.7}$	$-7.2_{\pm 2.7}$
64	23.6M	<u>0.843</u>	**$0.031_{\pm 0.015}$** (93%)	$51.7_{\pm 6.7}$	$-4.7_{\pm 1.4}$
Adapted base	0	—	$0.426_{\pm 0.085}$	$38.3_{\pm 11.7}$	—

underperforms despite having fewer parameters than LoRA, reducing drift by 25%. A LoRA modifies the attention projections within the velocity model, whereas the side network adds a residual after those representations are formed. The rank sweep in Table 9 further tests how capacity affects the balance between one-step fit and rollout quality.

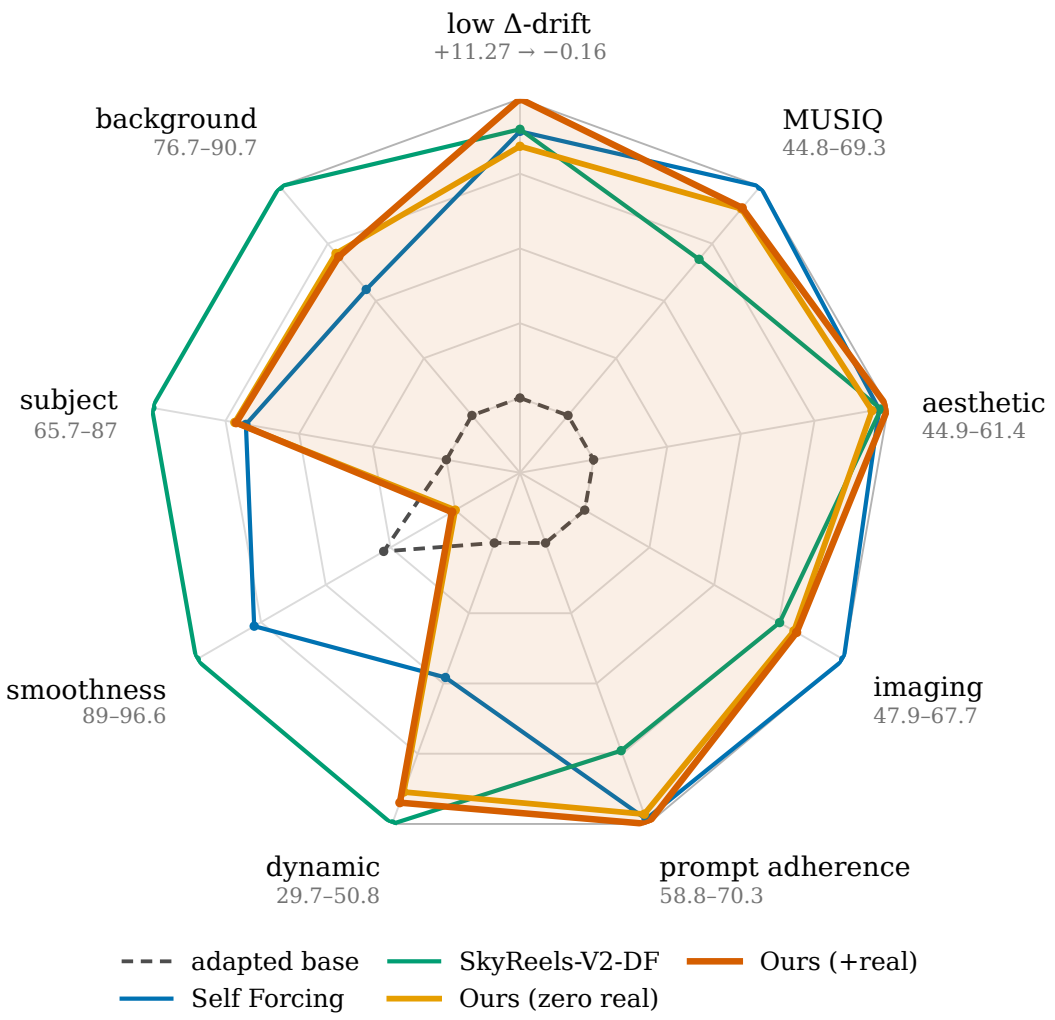

Figure 12: **VBench and drift comparison.** Scores from Tables 1 and 13 are normalized by axis; the drift axis is reversed so that outward is better. Clean Forcing improves the adapted base across most measures and is competitive with the retrained models, though it remains lower on smoothness and consistency.

LoRA rank. Table 9 varies only the rank in a matched one-step experiment on the adapted base. Validation R^2 is similar from rank 8 to 64 (0.829–0.854), indicating that one-step fit largely saturates at low rank. Late-window rollout quality instead peaks at rank 16 in this sweep. Ranks 32 and 64 score 7.2 and 4.7 MUSIQ lower than rank 16 on paired clips, without a corresponding improvement in drift. Rank 8 differs from rank 16 by only 1.0 ± 0.8 MUSIQ. These results motivate rank 16 as the deployed setting while showing that more capacity does not necessarily improve long-horizon output.

I Additional Results on Drift, Quality, and Diversity

Isolating the corrector. The paired 2×2 experiment separates correction from causal adaptation (Table 10). Correction reduces saturation drift by 66–73% on the unadapted base and 80–85% on the adapted base. Adaptation enables prompt-only generation and raises uncorrected late MUSIQ from 29.6 to 38.5 in this setting; correction then reduces the remaining drift. On the unadapted base, the closed-loop stage also raises late MUSIQ from 36.9 after one-step correction to 60.5. Alternative metrics preserve the principal drift ranking (Table 11). The two corrected variants include the lowest

Table 10: **In-domain extension.** Five held-out clips are extended by 48 latent frames ($\sim$12 s) under three paired seeds. Values are means $\pm$ standard deviations across the three seed means. Bold marks the best mean in each metric.

Base	Correction	Sat. drift↓	Late MUSIQ↑
Unadapted	None	0.491 ± 0.045	29.6 ± 1.5
	One-step r_ϕ	0.166 ± 0.083	36.9 ± 2.1
	Closed-loop r_ϕ	0.131 ± 0.024	60.5 ± 1.2
Adapted	None	0.415 ± 0.017	38.5 ± 0.8
	One-step r_ϕ	$\mathbf{0.061} \pm 0.013$	59.4 ± 0.2
	Closed-loop r_ϕ	0.082 ± 0.006	$\mathbf{63.3} \pm 0.5$

TetherCache ΔQualityDrift ($+0.36$ versus $+4.6$ for Self Forcing), with strong color stability and anchored content survival. Their temporal LPIPS remains relatively poor among outputs that retain motion, consistent with the visible chunk seam.

Zero real videos. Single-shot bidirectional generations from the base provide the clean references needed to train the causal adapter and corrector using prompts alone (Table 1). This variant reaches MUSIQ 66.9 and $\Delta = +1.65$, compared with 67.0 and -0.16 using 40 real clips. Its drift score is also lower than the published $+3.57^\dagger$ result, which uses a different base and prompt set and is therefore an indicative comparison. Reference-free distribution matching removes no drift in our ablation. The small training set is consistent with a highly systematic velocity gap (§3.2.1), and drift changes little across the tested 16, 40, and 300 clip regimes.

Short-horizon quality. The corrector is trained on long rollouts, so we also measure its effect before substantial drift develops. On 200 VBench prompts at 5 s, the merged corrector improves five of six reported dimensions, including imaging quality ($+4.0$), subject consistency ($+2.3$), and dynamic degree ($+12.5$). Motion smoothness falls by 1.7 points as motion increases (Table 12).

Inter-seed diversity. Correction preserves similar semantic diversity across seeds. DINO inter-seed dissimilarity, measured over three seed pairs, eight extension frames, and five clips, is 0.31–0.38 for the corrected configurations versus 0.33 for the base, with no consistent trend across ranks or closed-loop training. Pixel-level LPIPS falls by 14–18% for the deployed correctors, but seed-dependent saturation artifacts in the base can inflate this pixel-space measure. The DINO comparison therefore provides a complementary check on semantic diversity.

Prompt adherence. Table 13 adds the VBench temporal-flickering and overall-consistency dimensions available on custom prompts. Overall consistency uses the official ViCLIP video–text similarity against each generation prompt, truncated to 32 tokens. Eight other dimensions require labeled prompts from the official suite, human action relies on a Kinetics label parsed from the filename, and temporal style duplicates overall consistency in this setting. The uncorrected base loses 11.5 points in overall consistency as it drifts, while the TTC variants score 47–50.5. Correction restores prompt adherence to approximately 70, close to the distilled models, with the zero-real variant within 0.4 points of +real. Temporal flickering is less informative here: the nearly static history-guidance output scores best among 16 fps rows, as discussed in Appendix C.

Horizon sweep. We truncate the same 128 videos at 10, 30, and 50 s and score each segment (Table 14); the 50 s results match Table 1. The base already reaches $\Delta = +10.7$ and MUSIQ 56.1 at 10 s, then loses another 11.3 MUSIQ by 50 s. The +real variant remains stable across horizons, while the zero-real variant shows a smaller, gradual decline. The base's Δ peaks at 30 s and then decreases despite continued loss of absolute quality. As the collapsed tail occupies more of the video, a first-versus-overall score can saturate, so Δ should be reported alongside an absolute quality measure.

Table 15 reports anchoring, two-second lagged identity, cut rate, and intent-conditioned progression excess for all 128 prompts at 50 s. Scoring every frame rather than using the standard stride changes MUSIQ by at most 1.8 and Δ by at most 0.26, without changing the Δ-drift ranking; only the two Clean Forcing variants swap MUSIQ order, by at most 0.5. The resulting MUSIQ/Δ pairs are $65.7/+1.60$ for zero real videos, $65.2/-0.10$ for +real, $68.5/+1.34$ for Self Forcing, and $61.5/+0.98$ for SkyReels. The release will include per-dimension VBench results, per-prompt

Table 11: **Drift diagnostics on 50 s videos.** Each configuration uses 32 videos from the evaluation set. ΔQD denotes ΔQualityDrift. Bold marks the best values among the four drift measures. Temporal LPIPS and flicker can reward low motion, as illustrated by the nearly static history-guidance run* (Appendix C).

	Drift measures				Frame differences	
Method	ΔQD↓	Color L1↓	Color corr.↑	Survival↑	t-LPIPS↓	Flicker↓
Adapted base	+22.9	1.43	0.16	0.44	0.23	15.9
Context noise	+31.4	1.44	0.18	0.45	0.19	11.1
History guidance*	+36.5	1.25	0.15	0.37	0.12	8.2
Self Forcing	+4.6	1.33	0.06	0.64	0.15	9.5
Clean Forcing (zero real)	+3.5	0.95	0.48	**0.70**	0.33	22.8
Clean Forcing (+real)	**+0.36**	**0.85**	**0.57**	0.69	0.59	32.8

Table 12: **Short-horizon quality.** Official VBench scores on 200 prompts at 5 s. Merged Clean Forcing improves four of five quality dimensions and increases dynamic degree. Bold marks the higher quality score; dynamic degree is not ranked.

	Official VBench ↑					Motion
Method	Subj.	Backg.	Aesth.	Imag.	Smooth	Dynamic
Adapted base	91.6	93.1	57.0	64.9	**97.2**	21.5
+ Clean Forcing (merged)	**93.9**	**94.5**	**59.0**	**68.9**	95.5	34.0

Table 13: **Prompt adherence and flickering.** Official VBench custom-input scores on 128 matched 50 s videos, reported on the calibrated scale with ± SEM. Prompt adherence denotes overall consistency. Flickering is a motion-sensitive diagnostic, and SkyReels' 24 fps score is not directly comparable with the 16 fps rows (Appendix C). Bold and underlining mark the two highest mean prompt-adherence scores.

Method	Prompt adherence↑	Flickering
Adapted base	58.81 ± 1.32	90.25 ± 0.72
+ Context noise ($\sigma = 0.2$)	61.68 ± 1.33	86.38 ± 1.20
+ History guidance ($w = 1.2$)	62.15 ± 1.08	93.94 ± 0.18
+ Pathwise TTC ($\{500, 250\}$)	47.00 ± 1.52	89.94 ± 0.64
+ Pathwise TTC ($\{750, 500\}$)	50.53 ± 1.55	89.86 ± 0.68
+ **Clean Forcing (zero real videos)**	69.85 ± 1.23	87.29 ± 0.61
+ **Clean Forcing (+40 real clips)**	$\underline{70.28} \pm 1.11$	86.75 ± 0.36
Causal-Forcing base	57.46 ± 1.28	88.42 ± 1.10
+ **Clean Forcing (zero real videos)**	$\mathbf{70.97} \pm 1.11$	89.59 ± 0.72
Self Forcing	70.12 ± 1.19	93.37 ± 0.68
SkyReels-V2-DF	67.31 ± 1.29	93.97 ± 0.56

MUSIQ curves, color-shift and survival trajectories, and temporal LPIPS for each configuration in Table 1.

J User Study Details

Protocol. We use the two-alternative forced-choice format of Yin et al. (2025): raters view two videos side by side with the generation prompt and select the better one. The survey runs on Qualtrics. We choose 26 finals prompts at fixed indices $0, 5, \ldots, 125$ and compare the +real variant with the adapted base, Self Forcing, and SkyReels-V2. Each comparison is rendered with A/B labels, and left–right placement is randomized using a fixed seed. Because our videos last 50 s, raters view excerpts from $t = 40$–50 s, where late drift is visible, rather than the full approximately 10 s videos used by Yin et al. (2025).

Table 14: **Horizon sweep.** For each method, the same 128 videos are rescored at 10, 30, and 50 s (± SEM). The base's Δ-drift shrinks after 30 s even as MUSIQ falls, so drift should be read alongside absolute quality. Bold marks the best mean at each horizon for each metric.

Method	Metric	10 s	30 s	50 s
Adapted base	Δ-drift↓	$+10.74_{\pm 0.69}$	$+13.17_{\pm 0.89}$	$+11.27_{\pm 0.81}$
	MUSIQ↑	$56.1_{\pm 1.2}$	$47.4_{\pm 1.0}$	$44.8_{\pm 0.95}$
+ **Clean Forcing (zero real)**	Δ-drift↓	$+1.25_{\pm 0.29}$	$+1.30_{\pm 0.39}$	$+1.65_{\pm 0.35}$
	MUSIQ↑	$68.6_{\pm 0.8}$	$67.5_{\pm 0.8}$	$66.9_{\pm 0.80}$
+ **Clean Forcing (+40 real)**	Δ-drift↓	$+0.39_{\pm 0.26}$	$\mathbf{-0.01}_{\pm 0.16}$	$\mathbf{-0.16}_{\pm 0.12}$
	MUSIQ↑	$66.8_{\pm 0.6}$	$67.0_{\pm 0.6}$	$67.0_{\pm 0.61}$
Causal-Forcing base	Δ-drift↓	$+3.60_{\pm 0.42}$	$+10.94_{\pm 0.66}$	$+11.99_{\pm 0.67}$
	MUSIQ↑	$63.9_{\pm 0.8}$	$55.7_{\pm 0.8}$	$52.0_{\pm 0.74}$
+ **Clean Forcing (zero real)**	Δ-drift↓	$+2.07_{\pm 0.31}$	$+3.64_{\pm 0.47}$	$+2.80_{\pm 0.42}$
	MUSIQ↑	$64.9_{\pm 0.9}$	$62.7_{\pm 1.0}$	$62.1_{\pm 0.99}$
Self Forcing	Δ-drift↓	$\mathbf{-0.31}_{\pm 0.25}$	$+0.29_{\pm 0.31}$	$+1.08_{\pm 0.39}$
	MUSIQ↑	$\mathbf{70.4}_{\pm 0.7}$	$\mathbf{70.1}_{\pm 0.7}$	$\mathbf{69.3}_{\pm 0.60}$
SkyReels-V2-DF	Δ-drift↓	$+0.35_{\pm 0.22}$	$+0.82_{\pm 0.27}$	$+1.00_{\pm 0.30}$
	MUSIQ↑	$62.5_{\pm 1.0}$	$61.9_{\pm 1.0}$	$61.5_{\pm 1.04}$

Table 15: **Scene progression and temporal continuity.** Results on 128 matched 50 s videos scored through one pipeline. Anchoring is DINO similarity between the opening and final fifth; paired excess subtracts Self Forcing's anchoring prompt by prompt. Lagged identity uses a 2 s offset, and cuts are detected between adjacent frames. Anchoring reflects the drift–progression trade (Appendix D).

	Opening similarity		Temporal continuity	
Method	Anchoring	Excess vs. SF	Lag-2s identity↑	Cuts/video↓
Adapted base	0.336	−0.148	0.812	2.82
+ **Clean Forcing (zero real)**	0.679	+0.194	0.830	0.89
+ **Clean Forcing (+40 real)**	0.795	+0.311	0.746	0.04
Causal-Forcing base	0.249	−0.235	0.807	1.71
+ **Clean Forcing (zero real)**	0.680	+0.196	0.850	0.55
Self Forcing	0.485	0.000	0.922	0.26
SkyReels-V2-DF	0.697	+0.213	0.927	0.12

Each rater evaluates a block of 15 pairs: 13 comparisons drawn from the three pairings and two catch trials that compare a healthy video with a collapsed one. The block takes approximately 10 minutes. For each pair, raters answer which video is better overall in visual quality and prompt alignment and which remains cleaner and more stable over time. Figure 13 shows the instruction page and one trial as raters saw them.

Raters and exclusions. The analysis includes 33 raters, 29 recruited through Prolific and four directly invited to complete an underfilled block, with at least five ratings per pair. Under the specified catch-failure rule, we excluded one submission that failed both graded catches in its block; we also removed one prelaunch test session. Item analysis identified two of six catch items as nondiscriminative: roughly half of otherwise successful raters failed them, possibly because the collapsed output appeared stylistic. The four retained catch items had failure rates of 0–10%. Table 16 reports the resulting preference rates, Wilson intervals, and exact binomial tests.

Interpretation. The +real variant is preferred overall to the adapted base in 93.0% of judgments and to Self Forcing in 83.9%. Its 42.7% preference rate against SkyReels-V2 does not differ significantly from an even split ($p = 0.09$). On temporal stability, raters prefer our method to the drifting base (81.8%), report no clear preference against Self Forcing (46.9%), and favor SkyReels-V2 (21.0% for ours). SkyReels-V2 also runs at 24 fps, compared with 16 fps for the other evaluated videos, which may affect perceived smoothness. These judgments concern late excerpts and should not be

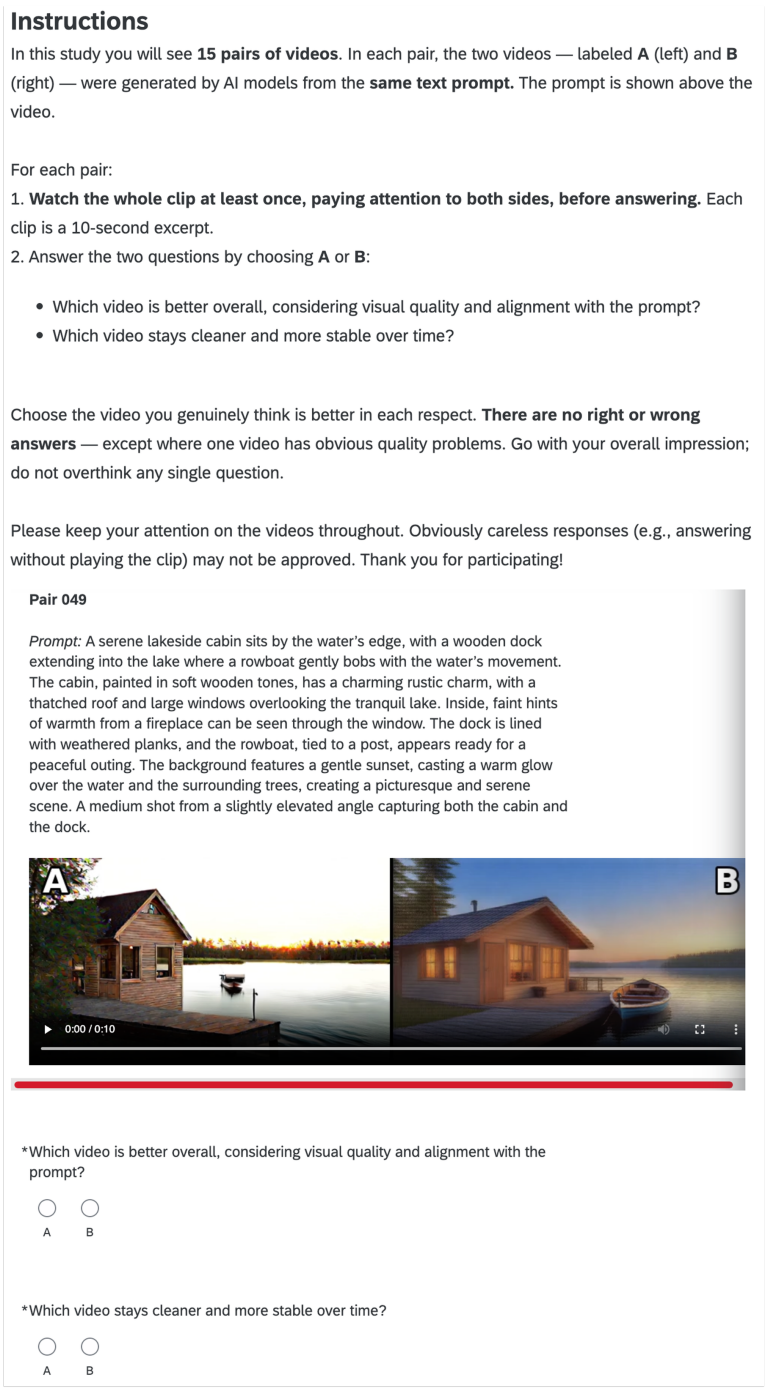

Figure 13: **User study interface.** Instructions (top) ask raters to watch each 10 s pair before judging overall quality and temporal stability. A sample trial (bottom) shows the prompt above the two videos, labeled A and B, followed by the two questions.

read as whole-video preferences. Rater instructions and the analysis script will accompany the code release.

Table 16: **User study.** Preference for Clean Forcing (+40 real clips) on 10 s excerpts from $t = 40$–50 s, with 143 judgments per opponent and question. Brackets give Wilson 95% intervals; p is the exact two-sided binomial test against 50%. Bold marks significant preference for Clean Forcing.

Opponent	Question	Clean Forcing preferred [95% CI]	Binomial p
Adapted base	Overall	**93.0**% [87.6, 96.2]	1×10^{-28}
	Temporal stability	**81.8**% [74.7, 87.3]	6×10^{-15}
Self Forcing	Overall	**83.9**% [77.0, 89.0]	5×10^{-17}
	Temporal stability	46.9% [38.9, 55.0]	n.s.
SkyReels-V2-DF	Overall	42.7% [34.8, 50.9]	0.09
	Temporal stability	21.0% [15.1, 28.4]	2×10^{-12}

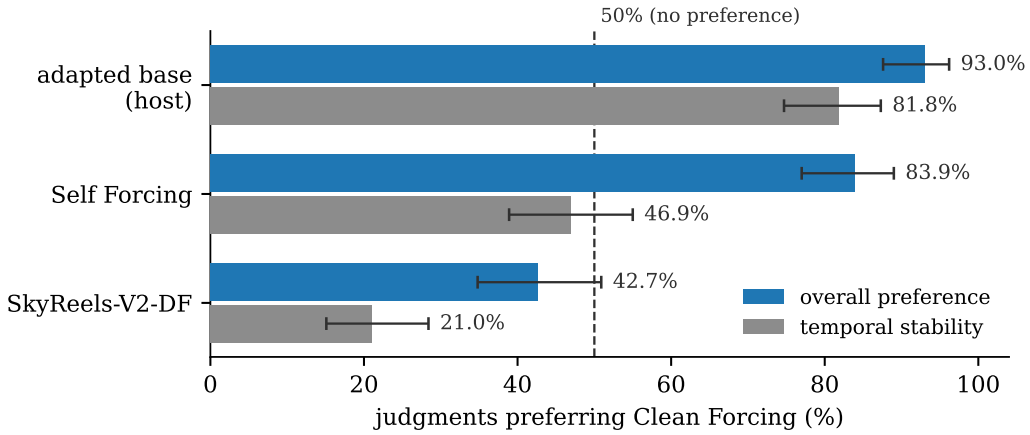

Figure 14: **User study preferences.** Percentage of judgments favoring Clean Forcing (+40 real clips) over each baseline for overall quality (blue) and temporal stability (gray). Whiskers show Wilson 95% confidence intervals; the dashed line marks 50% preference (Table 16).

K Additional Qualitative Results

Figure 4 in the main text and Figures 15–20 show evenly spaced frames across 50 s for eight prompts covering wildlife, vehicles, fluids, and architecture. Seeds are matched across methods for each prompt, and the rows follow the ordering in Figure 4. The supplementary material includes videos for every prompt shown in the figures, including the teaser (Figure 1), method comparisons, and 14B and Causal Forcing examples. It also includes side-by-side clips of the final 10 s ($t = 40$–50 s) from the base and Clean Forcing (+real) for all 128 prompts. Full-resolution videos and per-video measurements for all 128 prompts will accompany the code release.

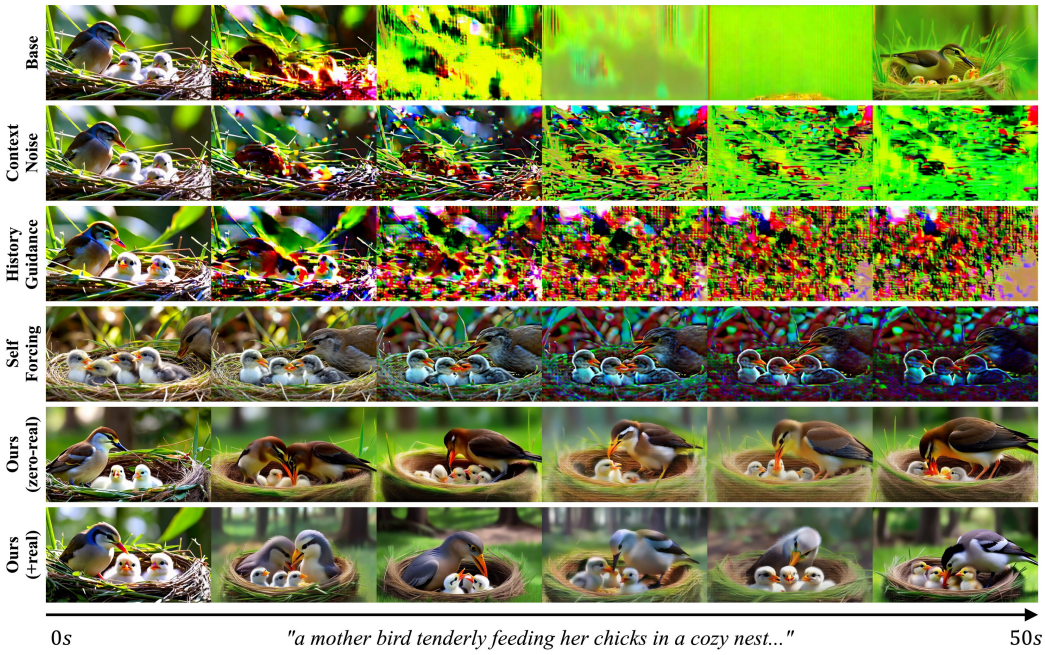

Figure 15: Frames from evaluation prompt 113 (mother bird feeding chicks).

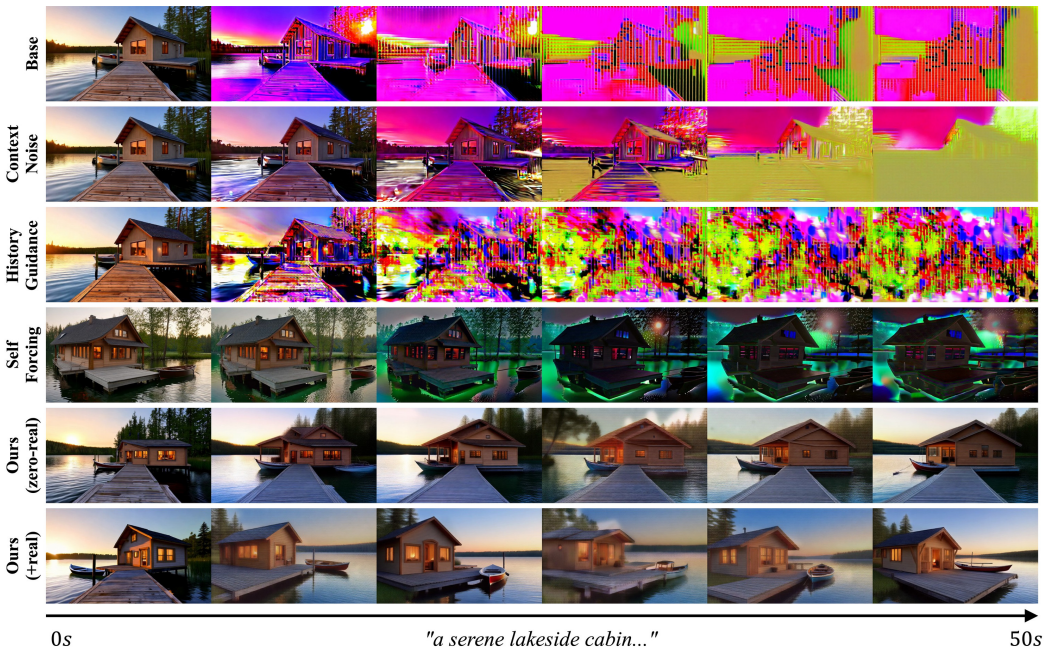

Figure 16: Frames from evaluation prompt 40 (lakeside cabin).

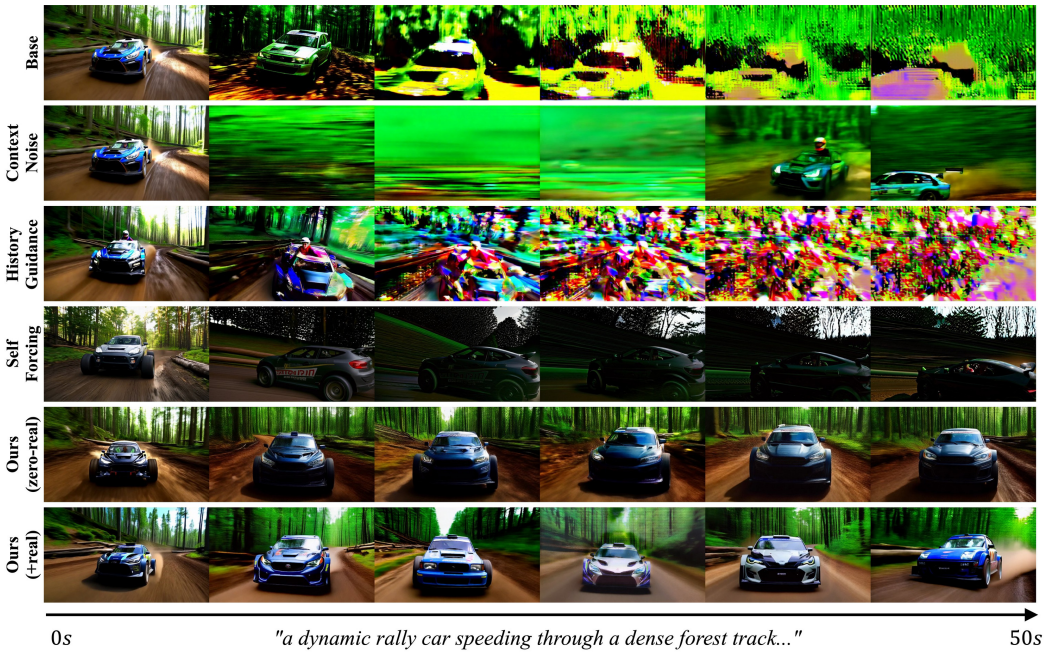

Figure 17: Frames from evaluation prompt 8 (rally car through a forest track).

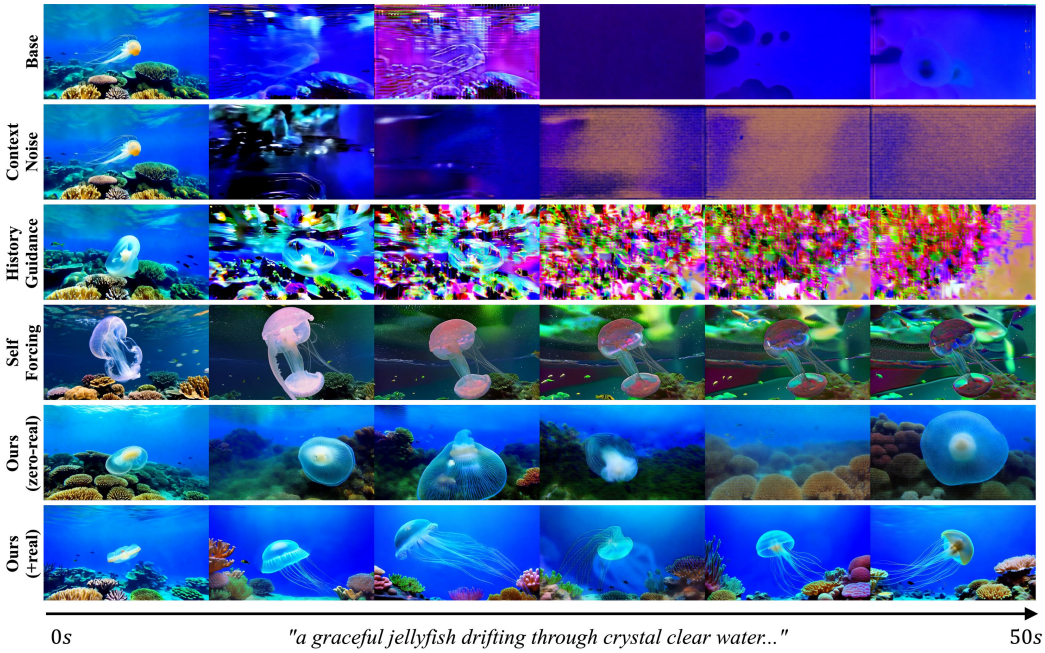

Figure 18: Frames from evaluation prompt 28 (jellyfish in clear water).



Figure 19: Frames from evaluation prompt 32 (martini pour, slow motion).



Figure 20: Frames from evaluation prompt 30 (drink poured over ice).

L LIMITATIONS

Correction substantially reduces quality drift, but three limitations remain. Stronger reference correction can suppress scene progression, a tradeoff the contraction weight adjusts but does not eliminate (§D). Each new checkpoint or domain requires a new corrector. The base model's luminance pulse at chunk boundaries also persists (Appendix G.2).