LONGLIVE-PLUG: ONCE-FOR-ALL DISTILLATION FOR VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video diffusion models are increasingly developed into specialized models for diverse downstream tasks, and this development often includes a distillation stage, for example to accelerate sampling or to improve long-video generation. This stage is typically repeated for every specialized model. We introduce LongLive-Plug, a once-for-all distillation framework that learns reusable capabilities as LoRAs on a base model for training-free, plug-and-play deployment to compatible downstream models. These capabilities include single-pass classifier-free guidance, few-step sampling, and long-context error correction for autoregressive generation. The adapters remain reusable even when downstream models add conditioning branches, expand output channels. Despite training at a fixed guidance scale, our dedicated CFG LoRA provides text guidance control through its inference weight. Combining it with a few-step LoRA simultaneously preserves few-step generation and CFG controllability on downstream tasks. We verify training-free deployment on 54 downstream models across three backbone families and eight task categories, including world modeling, robotics, editing, and multimodal generation. The approach may support additional compatible models. Each capability can thus be distilled once per backbone family and reused without per-target retraining.

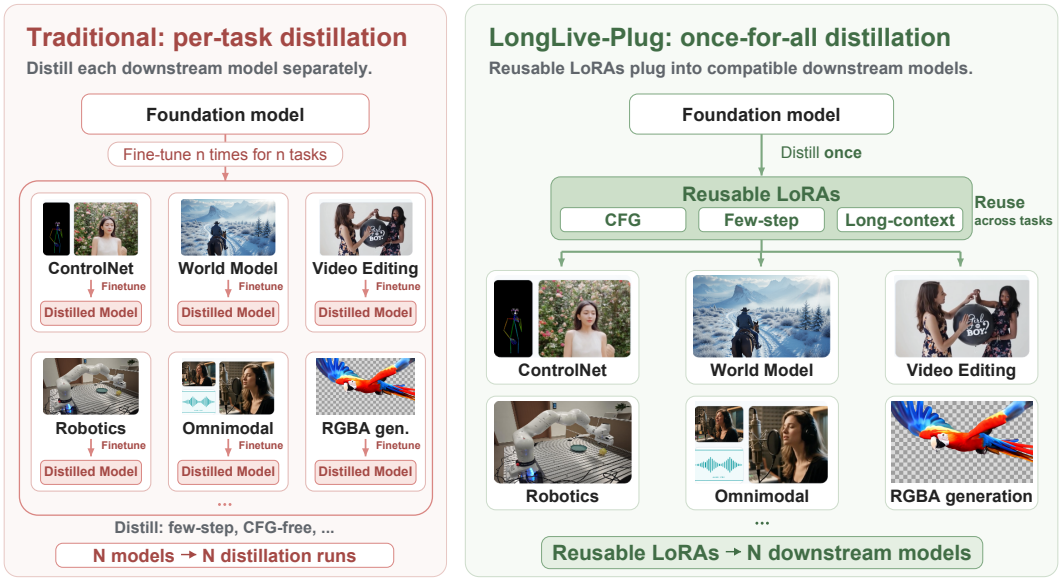

Figure 1: **Once-for-all distillation with plug-and-play deployment.** Conventional pipelines separately distill every downstream model (left). LongLive-Plug distills CFG, few-step, and long-context capabilities into reusable LoRAs on each base model, then transfers them to compatible downstream models without target-specific training (right).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Large-scale video diffusion transformers form reusable foundations for video generation (Yang et al., 2025c; Kong et al., 2024; Wan Team, 2025), while downstream applications increasingly depend on specialized models. They can be adapted to diverse tasks, including controllable generation and personalization (Wang et al., 2023a; Wei et al., 2024), video editing (Qi et al., 2023), and action-conditioned world simulation for robotics and physical AI (Rigter et al., 2024; NVIDIA et al., 2025b). Developing such a specialized model often includes a distillation stage, for example to accelerate sampling or to improve long-video generation, and this stage is typically repeated for every new model, each requiring data preparation, teacher supervision, and optimization. This repeated cost motivates a once-for-all workflow: **distill reusable capabilities once on a base model and deploy them across compatible downstream models.**

We introduce LongLive-Plug, a once-for-all distillation framework built on reusable *functional LoRAs* (Hu et al., 2022). Each adapter is learned on a base model and attached to compatible downstream models while retaining their task-specific weights (Fig. 1). We consider three capabilities: *classifier-free guidance (CFG) distillation* combines two-pass guidance into one model evaluation (Meng et al., 2023; Perez Jensen & Sadat, 2025) while retaining continuous control over guidance strength; *few-step distillation* enables four-step sampling (Yin et al., 2024a); and *long-context distillation* improves long-video quality through error correction in models that support causal autoregressive (AR) inference. Once trained on the base model, these adapters support *training-free, plug-and-play deployment* to compatible downstream models, even when those models add conditioning branches, or expand output channels.

Making these functional LoRAs transferable raises two further questions. The first concerns guidance. Existing few-step distillation methods, such as CausVid and Self Forcing (Yin et al., 2025; Huang et al., 2025), distill CFG and few-step generation jointly, which fixes guidance at the scale used during training. Downstream tasks, however, often favor different guidance strengths, so a single fixed scale cannot serve all of them. We therefore apply decoupled distillation, which distills CFG and few-step generation into separate LoRAs. The inference weight of the CFG LoRA then acts as a guidance dial: scaling it adjusts guidance strength in a near-linear manner, as we observe empirically and as prior work on scaling fine-tuning updates suggests (Wortsman et al., 2022; Ilharco et al., 2023). Adjusting this weight while keeping the few-step LoRA fixed tailors guidance to each downstream task. By contrast, rescaling a jointly distilled LoRA also perturbs few-step generation and can cause collapse.

The second question is how the training of a functional LoRA affects its transfer. We report two empirical findings. Adapter rank matters: a small adapter may fit the base teacher well yet transfer poorly, so source fit alone does not determine the capacity needed for transfer. The distillation data also matters: broad T2V prompts on the base model expose the adapter to diverse generation behavior, and broader prompt coverage improves transfer. Together, these findings indicate how to learn acceleration that remains useful beyond the checkpoint on which it was distilled.

The functional-LoRA formulation also extends to *error correction* for long contexts. We learn this capability through *long-context distillation*. Using *Streaming Long Tuning* (Yang et al., 2026a), we optimize a LoRA on a causal AR version of the base model using distribution matching distillation (DMD). The model extends its own generated history, while a teacher supervises each newly generated short clip. This exposes the adapter to errors accumulated during extended rollouts and teaches it to sustain long-video quality. The resulting functional LoRA makes long-context error correction reusable across compatible downstream models that support causal AR inference.

We verify training-free deployment on 54 downstream models from three backbone families: Wan2.1-14B, Wan2.2-TI2V-5B, and MiniMax-H3. They span eight task categories, including world modeling, robotics, controllable generation, editing, and multimodal generation. The verified coverage is listed in Tab. 5 in Appendix C; our approach may support additional compatible models. Quantitative comparisons on SCOPE and Wan2.2-Fun-5B-Control show improvements over naive four-step sampling and performance competitive with per-target distillation, without additional downstream training. Further experiments show that independent CFG control helps transfer to tasks with different guidance preferences, and that larger adapter ranks and broader distillation data can improve transfer quality. Separately, transferring the long-context LoRA to the ReWorld (Chen et al., 2026d) and Matrix-Game 3.0 (Wang et al., 2026d) world models improves video quality dur-

ing long autoregressive rollouts. These results demonstrate reusable acceleration and long-context error correction across downstream models.

2 Related Work

2.1 Specialized Video Generation

CogVideoX, Wan, HunyuanVideo, LTX-Video, and Cosmos support diverse video specializations (Yang et al., 2025c; Wan Team, 2025; Kong et al., 2024; HaCohen et al., 2024; NVIDIA et al., 2025b). Examples include DOVE for super-resolution (Chen et al., 2025b), VACE for generation and editing (Jiang et al., 2025), Matrix-Game 3.0 for action-conditioned world modeling (Wang et al., 2026d), Kiwi-Edit for guided editing (Lin et al., 2026b), HunyuanVideo-Avatar for audio-driven animation (Chen et al., 2025a), and Cosmos-Transfer1 for multimodal control (NVIDIA et al., 2025a).

Acceleration often requires distilling each specialized checkpoint: FlashMotion and StreamAvatar target trajectory control and avatar interaction (Li et al., 2026b; Sun et al., 2025), while LiveEdit and FlashVSR target streaming editing and super-resolution (Wang et al., 2026a; Zhuang et al., 2026). D2DF uses one-step consistency distillation for object removal (Chen et al., 2026e), DreamDojo uses few-step causal distillation for robot world modeling (Gao et al., 2026), and BiWM applies DMD after camera-control fine-tuning (Rui et al., 2026). LongLive-Plug instead distills acceleration once per base model for training-free reuse across compatible descendants.

2.2 Video Generation Distillation

Progressive distillation shortens sampling trajectories (Salimans & Ho, 2022), and distribution matching enables one-step generation (Yin et al., 2024b). VideoLCM uses consistency distillation (Wang et al., 2023b), T2V-Turbo adds reward feedback (Li et al., 2024), and DOLLAR combines score and consistency objectives (Ding et al., 2025). LCM-LoRA packages consistency distillation for reuse across Stable Diffusion fine-tunes (Luo et al., 2023). Guidance distillation merges the two CFG branches into one pass (Meng et al., 2023), while adapter guidance distillation reduces trainable parameters and examines transfer to image-model derivatives (Perez Jensen & Sadat, 2025). CausVid converts bidirectional teachers into autoregressive generators with KV caching (Yin et al., 2025); Self Forcing uses autoregressive training rollouts to reduce the train–test mismatch (Huang et al., 2025). We isolate distilled capabilities in reusable LoRAs for compatible video specializations: CFG and few-step distillation accelerate inference, while long-context distillation corrects errors in models that already support AR inference. Plug-and-Play Diffusion Distillation transfers a non-LoRA guide network across fine-tuned image models (Hsiao et al., 2024). CASA transfers downstream LoRAs to few-step video models (Wang et al., 2026b), whereas we transfer distilled capability LoRAs to a broader range of downstream tasks and model variants.

3 Method

3.1 Preliminaries

For each backbone family, LongLive-Plug freezes a base model F_{θ_0} and distills each capability once into LoRA parameters ϕ, then reuses them across compatible downstream models F_{θ_τ}:

$$F_{\theta_0} \xrightarrow{\text{distill once}} \phi, \qquad F_{\theta_\tau} \xrightarrow[\text{no target training}]{\oplus\phi} F_{\theta_\tau \oplus \phi}. \tag{1}$$

Here, $\oplus$ adds updates to corresponding layers while retaining the target's task-specific weights, without downstream training or re-distillation. We consider three *functional LoRAs*: CFG distillation replaces two-pass guidance with one conditional evaluation; few-step distillation uses DMD2 (Yin et al., 2024a) to enable four-step sampling; and long-context distillation corrects accumulated errors in models that already support causal autoregressive (AR) inference.

3.2 CFG Distillation and Composition

CFG distillation. For noisy latent z_t at timestep t and condition c, let $v_c = F_{\theta_0}(z_t, t, c)$ and $v_\varnothing = F_{\theta_0}(z_t, t, \varnothing)$ denote the conditional and unconditional predictions of the base model. At

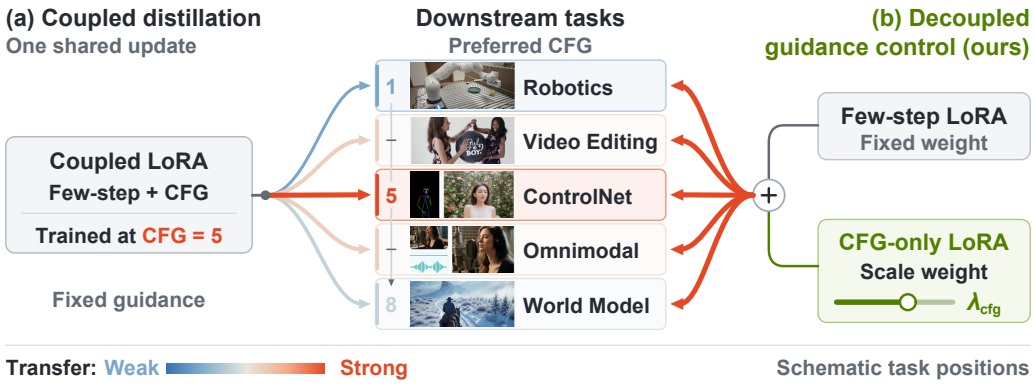

Figure 2: **Decoupled guidance control through an additional CFG-only LoRA. (a)** Scaling a single distilled LoRA changes the entire update, including its learned guidance and few-step behavior. **(b)** Adding a separately weighted CFG-only LoRA lets us adjust guidance for each target while keeping the few-step adapter weight fixed. Arrow colors and widths schematically illustrate guidance compatibility, with the warmest coupled-transfer arrow at CFG 5. CFG-only LoRA scaling provides an adjustable guidance control.

guidance scale w, the teacher predicts (Ho & Salimans, 2022)

$$v_{\text{cfg}}^{(w)} = v_{\varnothing} + w(v_c - v_{\varnothing}). \tag{2}$$

At fixed teacher scale w_{train}, we train only the CFG LoRA ϕ_{cfg} to minimize the expected squared error between its single-pass conditional output and the teacher's guided prediction, treated as a fixed target. The backbone, sampling schedule, and attention pattern remain unchanged.

The CFG LoRA weight as a guidance dial. Varying the inference weight λ_{cfg} adjusts the learned guidance despite fixed-scale training. Under an approximately linear response,

$$\begin{aligned} F_{\theta_0 \oplus \lambda_{\text{cfg}}\phi_{\text{cfg}}} &\approx v_c + \lambda_{\text{cfg}} \left(v_{\text{cfg}}^{(w_{\text{train}})} - v_c \right) \\ &\approx v_{\text{cfg}}^{(\widetilde{w})}, \qquad \widetilde{w} = 1 + \lambda_{\text{cfg}}(w_{\text{train}} - 1). \end{aligned} \tag{3}$$

Weights 0 and 1 recover the conditional model and full distilled adapter, respectively; larger weights extrapolate. This correspondence is approximate: nonlinear responses and downstream specialization can change the effective guidance. We therefore validate control empirically in Sec. 4.3.

Decoupled guidance control. Scaling a coupled few-step LoRA also changes its learned few-step correction. To accommodate downstream guidance preferences, we add a separately trained CFG-only LoRA and adjust λ_{cfg} while fixing the few-step weight λ_{step} and sampling schedule (Fig. 2).

At inference, we add the two weighted LoRA updates to each downstream layer:

$$\widetilde{W}_{\ell}^{(\tau)} = W_{\ell}^{(\tau)} + \lambda_{\text{step}} \Delta W_{\ell,\text{step}} + \lambda_{\text{cfg}} \Delta W_{\ell,\text{cfg}}. \tag{4}$$

Here, $W_{\ell}^{(\tau)}$ is the original target-layer weight; $\Delta W_{\ell,\text{step}}$ and $\Delta W_{\ell,\text{cfg}}$ are the separately trained few-step and CFG updates. Merging both updates before sampling preserves target-specific modules without joint retraining or extra model evaluations.

3.3 Transfer-Oriented Design

Adapter rank. LoRA rank controls the distilled update's capacity. A low-rank adapter may fit the base teacher yet transfer poorly after downstream specialization. We assess capacity using both source fit and downstream transfer. Sec. 4.4 compares ranks under matched training and deployment protocols, selecting checkpoints by base-model validation.

Table 1: **Transfer results on SCOPE.** Metric names, grouping, and directions follow Table 1 of SCOPE. Native inference uses 30 steps; all accelerated methods use four steps. Best values are bolded and second-best values are underlined.

	Visual quality			Motion quality		Consistency	
Method	JEPA ↑	FVD ↓	LPIPS ↓	Flow ↑	Smooth. ↑	Photo. ↓	Depth ↓
Default 30 step	**0.868**	**382.9**	<u>0.651</u>	**22.11**	**2.418**	9.182	1.287
Naive 4 step	0.732	805.5	**0.628**	11.77	2.093	8.468	1.242
SCOPE-specific distillation	0.782	502.1	0.678	16.54	<u>2.415</u>	<u>5.695</u>	**1.203**
LongLive-Plug	<u>0.792</u>	<u>478.7</u>	0.669	<u>16.87</u>	2.399	**4.246**	<u>1.230</u>

Distillation data. We distill CFG and few-step adapters on the base model using broad T2V prompts covering diverse subjects, scenes, motions, and styles. This exposes the adapters to varied generation behavior without target training data. Varying prompt coverage under a fixed teacher isolates prompt diversity; changing both the teacher and its task data changes the distillation source jointly. Sec. 4.4 tests how prompt diversity affects transfer.

3.4 Long-Context Distillation

To correct errors accumulated during AR rollouts, we train ϕ_{long} on a frozen causal AR base model using *Streaming Long Tuning* (Yang et al., 2026a). The student generates each short clip from its cached history, while a pretrained teacher provides distribution matching distillation (DMD) supervision. Detaching the preceding history keeps gradients local to the current clip as training rollouts grow longer. The resulting LoRA transfers to compatible models without target-specific training and improves their long-context generation quality. It can be applied to models with existing causal attention.

4 Experiments

4.1 Comparison of Acceleration Strategies

Evaluation tasks. We use Wan2.2-TI2V-5B (Wan Team, 2025) as the foundation model for the main comparison and evaluate two downstream tasks: world modeling with SCOPE (Tong et al., 2026) and ControlNet-based generation with VideoX-Fun's Wan2.2-Fun-5B-Control (Alibaba PAI, 2025c). We evaluate 1,378 CrossFPS clips with SCOPE's original input and output settings and all 600 depth-conditioned PAI-Bench-C cases (Zhou et al., 2025) following its evaluation protocol. Metrics include FVD (Unterthiner et al., 2018), LPIPS (Zhang et al., 2018), SSIM (Wang et al., 2004), and DOVER (Wu et al., 2023).

Evaluation candidates. We compare four candidates: native multi-step inference, naive four-step sampling, per-target distillation, and LongLive-Plug. Native inference uses 30 steps for SCOPE and 40 for ControlNet; directly reducing it to four steps is faster but substantially degrades generation quality. Per-target DMD2 distillation (Yin et al., 2024a) recovers good four-step quality, but adds substantial per-task data collection, training, and tuning costs. Our CFG-only and few-step LoRAs are trained on the base model with the broad T2V data in Sec. 3.3; the few-step adapter uses DMD2 with CFG-guided teacher supervision. Combining these adapters enables plug-and-play transfer with **zero downstream training cost**, requiring no target data or fine-tuning. For the quantitative comparisons, LongLive-Plug uses $(\lambda_{\text{step}}, \lambda_{\text{cfg}}) = (1, 3)$ on SCOPE (Tab. 1) and $(1, 1)$ on ControlNet (Tab. 2).

Qualitative and quantitative results. LongLive-Plug reduces SCOPE FVD from 805.5 to 478.7, comparable to SCOPE-specific distillation (502.1; Tab. 1). On ControlNet, it improves all six metrics over naive four-step sampling, including depth si-RMSE (2.135 to 1.641) and DOVER (8.90 to 10.11), with metric-dependent trade-offs relative to task-specific distillation (Tab. 2). Fig. 4 shows clearer scene boundaries and finer details with preserved control fidelity, demonstrating effective four-step generation without downstream training.

Table 2: **Transfer results on Wan2.2-Fun-5B-Control.** All local runs use the same 600 depth-conditioned PAI-Bench-C cases and 720P preprocessing. SSIM, F1, si-RMSE, and mIoU measure blurred-RGB, edge, depth, and mask similarity; DOVER measures technical quality, and LPIPS measures diversity over 3,600 videos. All local quantitative runs use four steps with CFG disabled. Official scores are from the PAI-Bench-C leaderboard (Zhou et al., 2025), with undisclosed settings. Best values across the reported rows are bolded and second-best values are underlined.

	Control fidelity				Visual quality	
Method	SSIM ↑	F1 ↑	si-RMSE ↓	mIoU ↑	DOVER ↑	LPIPS ↑
Official reported (reference)	0.556	**0.106**	1.819	**0.615**	9.32	**0.481**
Naive 4 step	<u>0.560</u>	0.094	2.135	0.582	8.90	0.264
ControlNet-specific distillation	0.544	0.099	**1.515**	0.595	**10.25**	<u>0.461</u>
LongLive-Plug	**0.566**	<u>0.100</u>	<u>1.641</u>	<u>0.612</u>	<u>10.11</u>	0.426

4.2 Transfer across Backbones and Tasks

Coverage beyond the main benchmarks. We verify training-free deployment across three backbone families: Wan2.1-14B, Wan2.2-TI2V-5B (Wan Team, 2025), and MiniMax-H3 (MiniMax, 2026). Tab. 5 in Appendix C lists the 54 verified downstream models: 24 for each Wan backbone and six for H3. They span eight task categories: world modeling, robotics, structure-conditioned generation, camera and trajectory control, video editing and restoration, subject and avatar generation, audio and RGBA generation, and domain, style, and quality adaptation. Each family reuses adapters distilled on its own base model without downstream training, extending capability reuse to full fine-tunes, task LoRAs, and models with additional conditioning modules. We compare native inference with four-step, CFG-free inference after attaching the base-distilled LoRA. For models with 20–50-step native schedules, this reduces denoising steps by 5–12.5×. The approach may support additional compatible downstream models beyond this verified set.

Cumulative distillation cost. Fig. 3 tracks cumulative training cost when adding tasks to Wan2.2-TI2V-5B. Both strategies share a one-time base distillation cost of approximately 80 H100 GPU-hours: 700 iterations on 32 GPUs for about 2.5 hours. Task-specific distillation then adds 83.9, 150.0, 86.8, and 56.1 H100 GPU-hours for depth-conditioned generation, world modeling, pose-conditioned generation, and robotics simulation, respectively: 376.8 additional GPU-hours and about 456.8 in total. LongLive-Plug reuses the base-distilled adapters at a fixed cost of about 80 GPU-hours, without task-specific data collection. The depth-specific adapter requires 5,000 paired prompts and dynamic depth videos; our transfer requires no downstream training data.

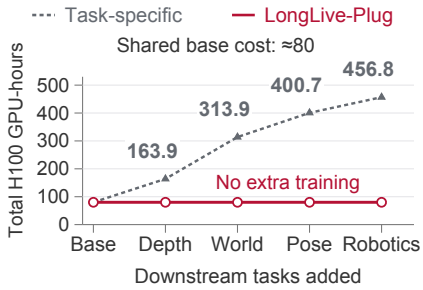

Figure 3: **Cumulative distillation cost.** Both curves include the shared base cost of approximately 80 GPU-hours.

4.3 CFG Controllability and Composition

Guidance control with a CFG-only LoRA. On Wan2.2-TI2V-5B, we vary only the CFG LoRA weight after distillation at $w_{\text{train}} = 5$, retaining the native 50-step FlowUniPC (Zhao et al., 2023) schedule. Raising λ_{cfg} from 1 to 2 or 3 strengthens the milk splash (Fig. 5) at runtime CFG 1, preserving guidance control with one conditional evaluation per step. See Appendix D (Fig. 14) for more cases.

Independent guidance after transfer. On SCOPE, a four-step coupled CFG-plus-step LoRA responds weakly to changes between the Snow Village and Crystal Maze prompts. Adding a separately weighted CFG LoRA strengthens the requested snow and crystal attributes as λ_{cfg} increases from 1 to 3 and 5, while the few-step weight stays at 1 (Fig. 6A). Matched frames preserve recognizable geometry and the foreground weapon as text control changes at a fixed few-step weight.

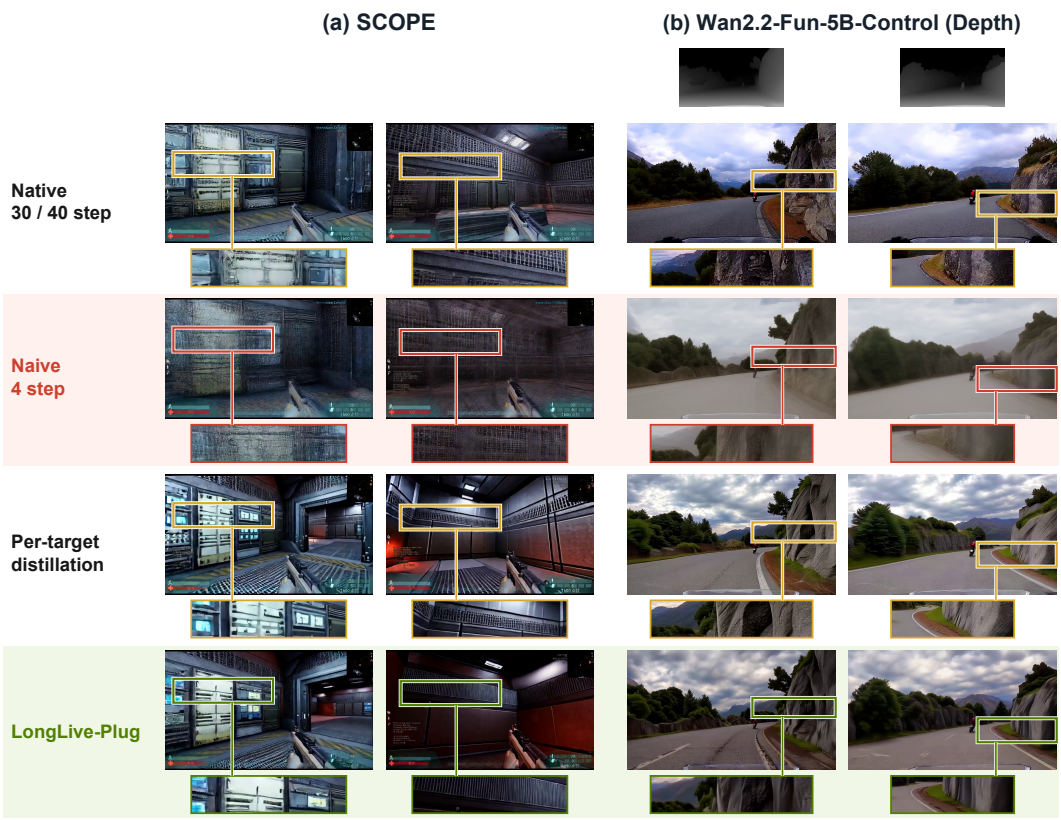

Figure 4: **Matched transfer comparisons on SCOPE and Wan2.2-Fun-5B-Control (Depth).** Two matched frames per task compare native (30/40 steps), naive four-step, per-target distilled, and LongLive-Plug outputs. Naive four-step sampling produces blurry, low-quality videos, whereas LongLive-Plug maintains high visual quality at four steps. Depth thumbnails condition ControlNet; boxes and strips show matched regions across methods. LongLive-Plug uses an additional base-distilled adapter variant. Appendix B gives full four-frame comparisons and alignment details.

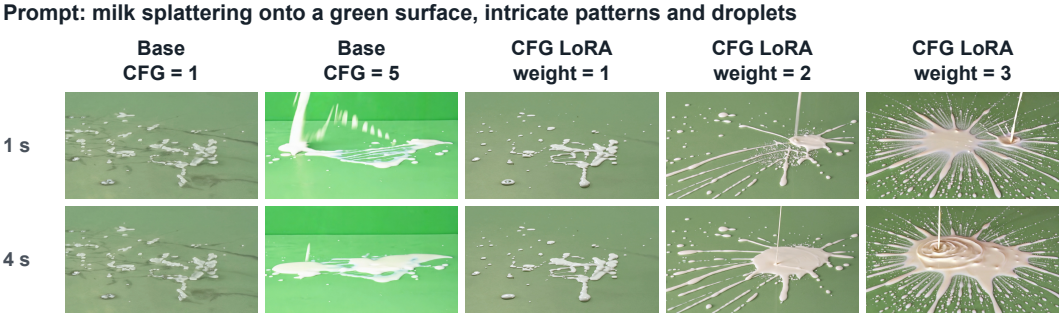

Figure 5: **Guidance control after CFG-only distillation.** The milk-splatter prompt compares native CFG references with CFG LoRA weights 1, 2, and 3. All variants use 50 sampling steps; every LoRA variant uses the distilled CFG setting and requires only one conditional forward pass per step. Rows show matched frames at 1 and 4 seconds. Higher LoRA weights follow the trend of stronger CFG, producing a more pronounced splash without exactly matching native CFG scales.

Failure of global LoRA scaling. Globally scaling the coupled LoRA fails to provide effective guidance control on SCOPE. With the checkpoint, prompt, input image, action sequence, seed, and sampler fixed, increasing the global weight darkens and distorts the scene, with severe collapse at weight 5 (Fig. 6B). Both SCOPE experiments use four steps with distilled CFG and different coupled checkpoints. See Appendix D for more cases and experimental settings.



Figure 6: **Independent CFG control on SCOPE.** (A) The coupled LoRA alone at weight 1 shows little response to the prompts. Adding a separately weighted CFG LoRA strengthens the boxed prompt attributes while the coupled LoRA weight remains at 1. Colored lines link each box to its corresponding prompt text. (B) Scaling the entire coupled LoRA instead degrades generation. Frames are matched across weights within each row. All runs use four-step sampling with distilled CFG; (A) and (B) use different coupled checkpoints.

4.4 Transfer Design Ablation

Adapter rank. Following Sec. 3.3, we vary rank while fixing the teacher, prompts, target layers, optimization budget, and checkpoint-selection rule. Each adapter transfers to SCOPE without target training and is evaluated by FVD on the full CrossFPS test set. Across three rank doublings from 16 to 128, transfer improves monotonically by 21% (Fig. 7a).

Distillation data. We vary prompt diversity with the teacher, rank, target layers, optimization budget, and number of training lines fixed. Prompt concentration is the mean pairwise cosine similarity in centred UMT5 (Chung et al., 2023) embedding space: lower values indicate broader coverage, while higher values indicate prompts clustered in one region. The number of distinct prompts covaries with concentration within the fixed line budget. Transfer degrades monotonically as diversity falls, with FVD rising by 12% across the sweep (Fig. 7b). At the same training budget, broader prompt coverage thus better supports transfer to models unseen during distillation.

4.5 Long-Context Distillation and Transfer

We evaluate the transfer of a long-context LoRA distilled on an AR Wan base model to two AR world models, ReWorld and Matrix-Game 3.0, without downstream training.

Transfer to ReWorld. On ReWorld (Chen et al., 2026d), trained on approximately 8 s windows, we compare 24-step native inference with four-step +Long over 16–64 s, using matched prompts and camera trajectories. At 64 s, +Long raises the seven-dimension mean from 73.51 to 75.77, improving visual quality and temporal scores with lower background consistency (Tab. 3, upper group). Its

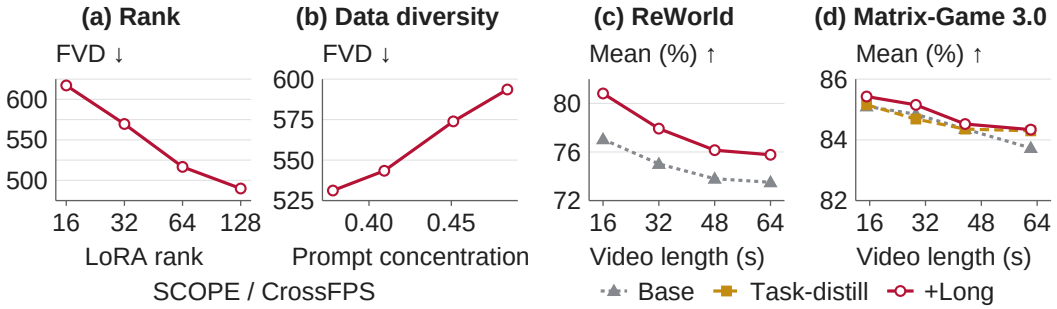

Figure 7: **Rank, data diversity, and long-context transfer.** (a–b) Higher rank and more diverse prompts (lower concentration) improve FVD after transfer. (c–d) Mean of seven VBench dimensions after transfer to ReWorld and Matrix-Game 3.0, respectively. Long-context comparisons are within each model. Our transferred long-context LoRA improves quality during long AR rollouts.

Table 3: **Long-context transfer to ReWorld and Matrix-Game 3.0.** Seven video-intrinsic VBench (Huang et al., 2024) dimensions, expressed as percentages at each model's longest tested rollout. Best quality scores are bolded within each model group.

Model	Imaging quality ↑	Aesthetic quality ↑	Subject consist. ↑	Background consist. ↑	Temporal flicker. ↑	Dynamic degree	Motion smooth. ↑	Total score ↑
ReWorld								
ReWorld-base	33.83	37.18	63.48	**88.87**	95.29	100.00	95.89	73.51
ReWorld +Long	**45.41**	**44.53**	**66.01**	80.81	**95.75**	100.00	**97.89**	**75.77**
Matrix-Game 3.0								
Matrix-Game-base	70.76	50.51	82.72	91.10	93.69	100.00	97.34	83.73
Matrix-Game-distill	**74.59**	51.91	**83.35**	91.39	91.93	100.00	96.90	84.30
Matrix-Game +Long	70.43	**54.25**	82.16	**91.51**	**94.12**	100.00	**97.90**	**84.34**

mean exceeds the base model at all tested lengths, up to $8\times$ the training duration (Fig. 7c). See Appendix E for qualitative examples.

Transfer to Matrix-Game 3.0. We transfer the same +Long adapter to Matrix-Game 3.0 (Wang et al., 2026d) and compare 50-step native inference, official three-step task-specific distillation, and four-step +Long under matched inputs and camera actions. At 62.18 s, their seven-dimension means are 83.73, 84.30, and 84.34, respectively (Tab. 3, lower group). Across tested lengths, +Long is competitive with task-specific distillation without Matrix-Game training, with metric-dependent trade-offs (Fig. 7d). Qualitative examples appear in Appendix E.2.

5 Discussion and Limitations

Once-for-all reuse requires compatible descendants of each base model. Long-context transfer requires existing causal AR inference, since LoRA updates alone do not change attention masks. Transfer quality involves task-dependent trade-offs. CFG LoRA weights provide approximate guidance control and may require adjustment after transfer.

6 Conclusion

We presented LongLive-Plug, which distills CFG, few-step sampling, and long-context error correction into reusable LoRAs once per backbone family. These adapters enable training-free transfer to compatible downstream models with adjustable guidance. Experiments demonstrate effective acceleration and improved long-video quality across tasks, reducing the need for per-target distillation.

AI USE STATEMENT

We used large language models to improve the clarity and readability of the manuscript, and AI agents to assist with experimental workflows. The authors are responsible for verifying all AI-assisted work and take full responsibility for the methods, results, and final content of this paper.

ETHICS STATEMENT

This work focuses on improving the efficiency and reuse of video generation models. We do not anticipate ethical concerns specific to our distillation framework beyond those associated with the underlying generative models, including potential misuse for misleading content and inherited biases. We encourage responsible use in accordance with the licenses and usage policies of the underlying models and datasets.

REPRODUCIBILITY STATEMENT

We will publicly release all code and artifacts developed for this work, including trained LoRA checkpoints, training and evaluation configurations, and scripts needed to reproduce our experiments. See Appendix A for implementation details.

REFERENCES

Alibaba PAI. Wan2.1-Fun-14B-Control. Official model release, 2025a. URL https://huggingface.co/alibaba-pai/Wan2.1-Fun-14B-Control. Accessed September 24, 2026.

Alibaba PAI. Wan2.1-Fun-V1.1-14B-Control-Camera. Official model release, 2025b. URL https://huggingface.co/alibaba-pai/Wan2.1-Fun-V1.1-14B-Control-Camera. Accessed September 24, 2026.

Alibaba PAI. Wan2.2-Fun-5B-Control. Hugging Face model release, 2025c. URL https://huggingface.co/alibaba-pai/Wan2.2-Fun-5B-Control.

Alibaba PAI. Wan2.2-Fun-5B-Control-Camera. Official model release, 2025d. URL https://huggingface.co/alibaba-pai/Wan2.2-Fun-5B-Control-Camera. Accessed September 24, 2026.

Alibaba PAI. Wan2.2-Fun-5B-InP. Official model release, 2025e. URL https://huggingface.co/alibaba-pai/Wan2.2-Fun-5B-InP. Accessed September 24, 2026.

Alibaba PAI. MiniMax-H3-Fun-Controlnet-Union. Official model release, 2026. URL https://huggingface.co/alibaba-pai/MiniMax-H3-Fun-Controlnet-Union. Accessed September 24, 2026.

AMD. Micro-World-I2W. Official model release, 2025. URL https://huggingface.co/amd/Micro-World-I2W. Accessed September 24, 2026.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A Unified Latent Action World Model. *arXiv preprint arXiv:2512.13030*, 2025. URL https://arxiv.org/abs/2512.13030.

BWM Team. BWM: A Low-Cost High-Fidelity World Simulator for Robot Learning. *arXiv preprint arXiv:2607.29302*, 2026. URL https://arxiv.org/abs/2607.29302.

Danze Chen, Zeqing Wang, Ziyue Lin, Xingyi Yang, and Yeying Jin. H3-World: Turning Language Understanding into World Control. *arXiv preprint arXiv:2609.01560*, 2026a. URL https://arxiv.org/abs/2609.01560.

Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. HunyuanVideo-Avatar: High-fidelity audio-driven human animation for multiple characters. *arXiv preprint arXiv:2505.20156*, 2025a. URL https://arxiv.org/abs/2505.20156.

Yiwen Chen, Guosheng Lin, and Chi Zhang. Code World Model: Coding Agent as World Brain. *arXiv preprint arXiv:2608.25927*, 2026b. URL https://arxiv.org/abs/2608.25927.

Yuzhi Chen, Ronghan Chen, Dongjie Huo, Yandan Yang, Dekang Qi, Haoyun Liu, Tong Lin, Shuang Zeng, Junjin Xiao, Xinyuan Chang, Feng Xiong, Xing Wei, Zhiheng Ma, and Mu Xu. ABot-PhysWorld: Interactive World Foundation Model for Robotic Manipulation with Physics Alignment. *arXiv preprint arXiv:2603.23376*, 2026c. URL https://arxiv.org/abs/2603.23376.

Zheng Chen, Zichen Zou, Kewei Zhang, Xiongfei Su, Xin Yuan, Yong Guo, and Yulun Zhang. DOVE: Efficient one-step diffusion model for real-world video super-resolution. In *Advances in Neural Information Processing Systems*, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7b04ec5f2b89d7f601382c422dfe07af-Abstract-Conference.html.

Zhifei Chen, Luozhou Wang, Guibao Shen, Dongyu Yan, Shuai Yang, Tianshuo Xu, Yihua Du, Wei Wang, Tianyi Gui, Lianghua Huang, and Yingcong Chen. ReWorld: An interactive world model with long-horizon memory. *arXiv preprint arXiv:2608.23565*, 2026d. URL https://arxiv.org/abs/2608.23565.

Zizhao Chen, Ping Wei, Guang Dai, Jingdong Wang, and Mengmeng Wang. From draft to draft-free: One-step video object removal via privileged distillation and fast planting. *arXiv preprint arXiv:2607.14976*, 2026e. URL https://arxiv.org/abs/2607.14976.

Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, Bin Xia, Dingdong Wang, Hongwei Yi, Xihui Liu, Hengshuang Zhao, Yu Liu, Yingya Zhang, and Yujiu Yang. WanMove: Motion-controllable Video Generation via Latent Trajectory Guidance. *arXiv preprint arXiv:2512.08765*, 2025. URL https://arxiv.org/abs/2512.08765.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. UniMax: Fairer and more Effective Language Sampling for Large-Scale Multilingual Pretraining. *arXiv preprint arXiv:2304.09151*, 2023. URL https://arxiv.org/abs/2304.09151.

Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. FantasyWorld: Geometry-Consistent World Modeling via Unified Video and 3D Prediction. *arXiv preprint arXiv:2509.21657*, 2025. URL https://arxiv.org/abs/2509.21657.

Yufan Deng, Yuanyang Yin, Xun Guo, Yizhi Wang, Jacob Zhiyuan Fang, Shenghai Yuan, Yiding Yang, Angtian Wang, Bo Liu, Haibin Huang, and Chongyang Ma. MAGREF: Masked Guidance for Any-Reference Video Generation with Subject Disentanglement. *arXiv preprint arXiv:2505.23742*, 2025. URL https://arxiv.org/abs/2505.23742.

DiffSynth-Studio. MiniMax-H3-LoRA-LineartAnime: Anime Video Line Art Colorization. Official model release, 2026. URL https://huggingface.co/DiffSynth-Studio/MiniMax-H3-LoRA-LineartAnime. Accessed September 24, 2026.

Zihan Ding, Chi Jin, Difan Liu, Haitian Zheng, Krishna Kumar Singh, Qiang Zhang, Yan Kang, Zhe Lin, and Yuchen Liu. DOLLAR: Few-step video generation via distillation and latent reward optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17961–17971, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Ding_DOLLAR_Few-Step_Video_Generation_via_Distillation_and_Latent_Reward_Optimization_ICCV_2025_paper.html.

Haotian Dong, Wenjing Wang, Chen Li, Jing Lyu, and Di Lin. Video Generation with Stable Transparency via Shiftable RGB-A Distribution Learner. *arXiv preprint arXiv:2509.24979*, 2025. URL https://arxiv.org/abs/2509.24979.

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

Hongyang Du, Junjie Ye, Xiaoyan Cong, Runhao Li, Jingcheng Ni, Aman Agarwal, Zeqi Zhou, Zekun Li, Randall Balestriero, and Yue Wang. VideoGPA: Distilling Geometry Priors for 3D-Consistent Video Generation. *arXiv preprint arXiv:2601.23286*, 2026. URL https://arxiv.org/abs/2601.23286.

Jianxiong Gao, Zhaoxi Chen, Xian Liu, Junhao Zhuang, Chengming Xu, Jianfeng Feng, Yu Qiao, Yanwei Fu, Chenyang Si, and Ziwei Liu. LongVie 2: Multimodal Controllable Ultra-Long Video World Model. *arXiv preprint arXiv:2512.13604*, 2025. URL https://arxiv.org/abs/2512.13604.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, et al. DreamDojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026. URL https://arxiv.org/abs/2602.06949.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. URL https://arxiv.org/abs/2501.00103.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Yi-Ting Hsiao, Siavash Khodadadeh, Kevin Duarte, Wei-An Lin, Hui Qu, Mingi Kwon, and Ratheesh Kalarot. Plug-and-Play Diffusion Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13743–13752, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Hsiao_Plug-and-Play_Diffusion_Distillation_CVPR_2024_paper.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Junchao Huang, Guian Fang, Shengju Qian, Xianghao Kong, Zhuoran Zhao, Wei Huang, Yihua Du, Zixin Zhang, Justin Cui, Yuchao Gu, Yukang Chen, Xinting Hu, Tianyu He, Shaoshuai Shi, Zhuotao Tian, Xin Wang, Mike Zheng Shou, and Li Jiang. SolarWM: Open Data and Scalable Training for Long-Horizon Video World Models. *arXiv preprint arXiv:2609.02886*, 2026. URL https://arxiv.org/abs/2609.02886.

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

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://github.com/Vchitect/VBench.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023.

Index Team. Index-AniSora V2.0. Official model release, 2025. URL https://huggingface.co/IndexTeam/Index-anisora. Wan2.1-14B-based V2.0 release. Accessed September 24, 2026.

Longbin Ji, Guan Wang, Xuan Wei, Chenye Yang, Xiangrui Liu, Zhenyu Zhang, Shuohuan Wang, Yu Sun, and Jingzhou He. Native Audio-Visual Alignment for Generation. *arXiv preprint arXiv:2605.30073*, 2026. URL https://arxiv.org/abs/2605.30073.

Yudong Jiang, Baohan Xu, Siqian Yang, Mingyu Yin, Jing Liu, Chao Xu, Siqi Wang, Yidi Wu, Bingwen Zhu, Xinwen Zhang, Xingyu Zheng, Jixuan Xu, Yue Zhang, Jinlong Hou, and Huyang Sun. AniSora: Exploring the Frontiers of Animation Video Generation in the Sora Era. *arXiv preprint arXiv:2412.10255*, 2024. URL https://arxiv.org/abs/2412.10255.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Jiang_VACE_All-in-One_Video_Creation_and_Editing_ICCV_2025_paper.html.

Denis Karachev. Dilated Controlnet for Wan2.1. Official source repository, 2025. URL https://github.com/TheDenk/wan2.1-dilated-controlnet. Accessed September 24, 2026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. URL https://arxiv.org/abs/2412.03603.

Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let Them Talk: Audio-Driven Multi-Person Conversational Video Generation. *arXiv preprint arXiv:2505.22647*, 2025. URL https://arxiv.org/abs/2505.22647.

Guangyuan Li, Siming Zheng, Hao Zhang, Jinwei Chen, Junsheng Luan, Binkai Ou, Lei Zhao, Bo Li, and Peng-Tao Jiang. MagicTryOn: Harnessing Diffusion Transformer for Garment-Preserving Video Virtual Try-on. *arXiv preprint arXiv:2505.21325*, 2025. URL https://arxiv.org/abs/2505.21325.

Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2V-Turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. In *Advances in Neural Information Processing Systems*, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/8a57aa8e8b57e64a42e95f7dceb0adb9-Abstract-Conference.html.

Maomao Li, Zhen Li, Kaipeng Zhang, Guosheng Yin, Zhifeng Li, and Dong Xu. OmniCustom: Sync Audio-Video Customization Via Joint Audio-Video Generation Model. *arXiv preprint arXiv:2602.12304*, 2026a. URL https://arxiv.org/abs/2602.12304.

Quanhao Li, Zhen Xing, Rui Wang, Haidong Cao, Qi Dai, Daoguo Dong, and Zuxuan Wu. FlashMotion: Few-step controllable video generation with trajectory guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8986–8996, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_FlashMotion_Few-Step_Controllable_Video_Generation_with_Trajectory_Guidance_CVPR_2026_paper.html.

Sizhe Lester Li, Evan Kim, Xingjian Bai, Tong Zhao, Tao Pang, Max Simchowitz, and Vincent Sitzmann. Turning Video Models into Generalist Robot Policies. *arXiv preprint arXiv:2605.27817*, 2026c. URL https://arxiv.org/abs/2605.27817.

Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Yuancheng Xu, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, Paul Debevec, and Ning Yu. Vista4D: Video Reshooting with 4D Point Clouds. *arXiv preprint arXiv:2604.21915*, 2026a. URL https://arxiv.org/abs/2604.21915.

Yiqi Lin, Guoqiang Liang, Ziyun Zeng, Zechen Bai, Yanzhe Chen, and Mike Zheng Shou. Kiwi-Edit: Versatile video editing via instruction and reference guidance. *arXiv preprint arXiv:2603.02175*, 2026b. URL https://arxiv.org/abs/2603.02175.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin Backbone Cross-Modal Fusion for Audio-Video Generation. *arXiv preprint arXiv:2510.01284*, 2025. URL https://arxiv.org/abs/2510.01284.

Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. LCM-LoRA: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023. URL https://arxiv.org/abs/2311.05556.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14297–14306, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Meng_On_Distillation_of_Guided_Diffusion_Models_CVPR_2023_paper.html.

MiniMax. MiniMax-H3: Official model card. Hugging Face model repository, 2026. URL https://huggingface.co/MiniMaxAI/MiniMax-H3.

Muyao Niu, Mingdeng Cao, Yifan Zhan, Qingtian Zhu, Weihang Ran, Yanhong Zeng, Xiao Sun, Zhihang Zhong, and Yinqiang Zheng. AniCrafter: Customizing Realistic Human-Centric Animation via Avatar-Background Conditioning in Video Diffusion Models. *arXiv preprint arXiv:2505.20255*, 2025. URL https://arxiv.org/abs/2505.20255.

NVIDIA, Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, et al. Cosmos-Transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025a. URL https://arxiv.org/abs/2503.14492.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025b. URL https://arxiv.org/abs/2501.03575.

Cristian Perez Jensen and Seyedmorteza Sadat. Efficient distillation of classifier-free guidance using adapters. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=uMz8FiW01.

Markus Pobitzer, Chang Liu, Chenyi Zhuang, Teng Long, Bin Ren, and Nicu Sebe. Loomis Painter: Reconstructing the Painting Process. *arXiv preprint arXiv:2511.17344*, 2025. URL https://arxiv.org/abs/2511.17344.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. FateZero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15932–15942, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/QI_FateZero_Fusing_Attentions_for_Zero-shot_Text-based_Video_Editing_ICCV_2023_paper.html.

Marc Rigter, Tarun Gupta, Agrin Hilmkil, and Chao Ma. AVID: Adapting video diffusion models to world models. *arXiv preprint arXiv:2410.12822*, 2024. URL https://arxiv.org/abs/2410.12822.

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

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Joachim Sallström. Aether Punch: Face Impact LoRA for Wan 2.2 5B. Official model release, 2025. URL https://huggingface.co/joachimsallstrom/Aether_Punch_Wan22_5b_i2v_LoRA. Accessed September 24, 2026.

Yiren Song, Cheng Liu, Yuxin Jiang, and Mike Zheng Shou. StreamingEffect: Real-Time Human-Centric Video Effect Generation. *arXiv preprint arXiv:2605.17019*, 2026. URL https://arxiv.org/abs/2605.17019.

Zhiyao Sun, Ziqiao Peng, Yifeng Ma, Yi Chen, Zhengguang Zhou, Zixiang Zhou, Guozhen Zhang, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Yong-Jin Liu. StreamAvatar: Streaming diffusion models for real-time interactive human avatars. *arXiv preprint arXiv:2512.22065*, 2025. URL https://arxiv.org/abs/2512.22065.

Zizhao Tong, Yeying Jin, Hongfeng Lai, Zeqing Wang, Zhaohu Xing, Kexu Cheng, Haoran Xu, Zhao Pu, Shangwen Zhu, Ruili Feng, Jian Zhao, Yan Zhang, Hao Tang, and Ling Shao. SCOPE: Simulating cross-game operations in playable environments for FPS world models. *arXiv preprint arXiv:2605.23345*, 2026. URL https://arxiv.org/abs/2605.23345.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards Accurate Generative Models of Video: A New Metric & Challenges. *arXiv preprint arXiv:1812.01717*, 2018. URL https://arxiv.org/abs/1812.01717.

Viggle Research. Viggle-Animate: Character Replacement in Video from a Single Repainted Frame. Official model release, 2026. URL https://huggingface.co/Viggle/Viggle-Animate. Accessed September 24, 2026.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. ATI: Any Trajectory Instruction for Controllable Video Generation. *arXiv preprint arXiv:2505.22944*, 2025a. URL https://arxiv.org/abs/2505.22944.

Boyang Wang, Xuweiyi Chen, Matheus Gadelha, and Zezhou Cheng. Frame In-N-Out: Unbounded Controllable Image-to-Video Generation. *arXiv preprint arXiv:2505.21491*, 2025b. URL https://arxiv.org/abs/2505.21491.

Mengchao Wang, Qiang Wang, Fan Jiang, Yaqi Fan, Yunpeng Zhang, Yonggang Qi, Kun Zhao, and Mu Xu. FantasyTalking: Realistic Talking Portrait Generation via Coherent Motion Synthesis. *arXiv preprint arXiv:2504.04842*, 2025c. URL https://arxiv.org/abs/2504.04842.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. VideoComposer: Compositional video synthesis with motion controllability. In *Advances in Neural Information Processing Systems*, 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/180f6184a3458fa19c28c5483bc61877-Abstract-Conference.html.

Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. VideoLCM: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023b. URL https://arxiv.org/abs/2312.09109.

Xinyu Wang, Chongbo Zhao, Fangneng Zhan, and Yue Ma. LiveEdit: Towards real-time diffusion-based streaming video editing. *arXiv preprint arXiv:2606.26740*, 2026a. URL https://arxiv.org/abs/2606.26740.

Yuchen Wang, Wenliang Zhong, Lichen Bai, Zikai Zhou, Shitong Shao, Bojun Cheng, Shuo Chen, Shuo Yang, and Zeke Xie. Exploring Data-Free LoRA Transferability for Video Diffusion Models. *arXiv preprint arXiv:2605.01929*, 2026b. URL https://arxiv.org/abs/2605.01929.

Zeqing Wang, Danze Chen, Zhaohu Xing, Zizhao Tong, Yinhan Zhang, Xingyi Yang, and Yeying Jin. ReactiveGWM: Steering NPC in Reactive Game World Models. *arXiv preprint arXiv:2605.15256*, 2026c. URL https://arxiv.org/abs/2605.15256.

Zhenzhi Wang, Jian Wang, Ke Ma, Dahua Lin, and Bing Zhou. TalkVerse: Democratizing Minute-Long Audio-Driven Video Generation. *arXiv preprint arXiv:2512.14938*, 2025d. URL https://arxiv.org/abs/2512.14938.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://ieeexplore.ieee.org/document/1284395.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, Yidan Xietian, Jiangbo Pei, Liang Hu, Boyi Jiang, Hua Xue, Zidong Wang, Haofeng Sun, Wei Li, Wanli Ouyang, Xianglong He, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-Game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026d. URL https://arxiv.org/abs/2604.08995.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. DreamVideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6537–6549, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wei_DreamVideo_Composing_Your_Dream_Videos_with_Customized_Subject_and_Motion_CVPR_2024_paper.html.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7959–7971, 2022.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20144–20154, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wu_Exploring_Video_Quality_Assessment_on_User_Generated_Contents_from_Aesthetic_ICCV_2023_paper.html.

Bowen Xue, Zheng-Peng Duan, Qixin Yan, Wenjing Wang, Hao Liu, Chun-Le Guo, Chongyi Li, Chen Li, and Jing Lyu. Stand-In: A Lightweight and Plug-and-Play Identity Control for Video Generation. *arXiv preprint arXiv:2508.07901*, 2025. URL https://arxiv.org/abs/2508.07901.

Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, and Xiaoming Wei. InfiniteTalk: Audio-driven Video Generation for Sparse-Frame Video Dubbing. *arXiv preprint arXiv:2508.14033*, 2025a. URL https://arxiv.org/abs/2508.14033.

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. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=nCAODkpsPJ.

Xiangpeng Yang, Ji Xie, Yiyuan Yang, Yue Ma, Yan Huang, Min Xu, and Qiang Wu. VideoCoF: Unified Video Editing with Temporal Reasoner. *arXiv preprint arXiv:2512.07469*, 2025b. URL https://arxiv.org/abs/2512.07469.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. NeoVerse: Enhancing 4D World Model with in-the-wild Monocular Videos. *arXiv preprint arXiv:2601.00393*, 2026b. URL https://arxiv.org/abs/2601.00393.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=LQzN6TRFg9.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William

Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World Action Models are Zero-shot Policies. *arXiv preprint arXiv:2602.15922*, 2026. URL https://arxiv.org/abs/2602.15922.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, 2024a. URL https://arxiv.org/abs/2405.14867.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yin_One-step_Diffusion_with_Distribution_Matching_Distillation_CVPR_2024_paper.html.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22963–22974, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yin_From_Slow_Bidirectional_to_Fast_Autoregressive_Video_Diffusion_Models_CVPR_2025_paper.html.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do World Action Models Need Test-time Future Imagination? *arXiv preprint arXiv:2603.16666*, 2026. URL https://arxiv.org/abs/2603.16666.

Guozhen Zhang, Zixiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. UniAVGen: Unified Audio and Video Generation with Asymmetric Cross-Modal Interactions. *arXiv preprint arXiv:2511.03334*, 2025. URL https://arxiv.org/abs/2511.03334.

Qiyuan Zhang, Biao Gong, Shuai Tan, Zheng Zhang, Yujun Shen, Xing Zhu, Yuyuan Li, Kelu Yao, Chunhua Shen, and Changqing Zou. PhysRVG: Physics-Aware Unified Reinforcement Learning for Video Generative Models. *arXiv preprint arXiv:2601.11087*, 2026a. URL https://arxiv.org/abs/2601.11087.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Xinyao Zhang, Wenkai Dong, Yuxin Song, Bo Fang, Qi Zhang, Jing Wang, Fan Chen, Hui Zhang, Haocheng Feng, Yu Lu, Hang Zhou, Chun Yuan, and Jingdong Wang. SAMA: Factorized Semantic Anchoring and Motion Alignment for Instruction-Guided Video Editing. *arXiv preprint arXiv:2603.19228*, 2026b. URL https://arxiv.org/abs/2603.19228.

Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video Generation with Updatable Spatial Memory. *arXiv preprint arXiv:2512.15716*, 2025. URL https://arxiv.org/abs/2512.15716.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A Unified Predictor-Corrector Framework for Fast Sampling of Diffusion Models. *arXiv preprint arXiv:2302.04867*, 2023. URL https://arxiv.org/abs/2302.04867.

Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. PAI-Bench: A comprehensive benchmark for physical AI. *arXiv preprint arXiv:2512.01989*, 2025. URL https://arxiv.org/abs/2512.01989.

Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. FlashVSR: Towards real-time diffusion-based streaming video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhuang_FlashVSR_Towards_Real-time_Diffusion-Based_Streaming_Video_Super_Resolution_CVPR_2026_paper.html.

A Implementation Details

A.1 CFG-Only LoRA Training

Guided flow regression. The student and teacher use the same backbone and attention mode. Given a noisy state z_t and prompt c, the teacher forms the guided flow using Eq. (2). The negative-prompt branch supplies $v_\varnothing$; "unconditional" therefore denotes the configured negative condition. The student predicts the target in one conditional pass.

Training-state construction. A captioned clean video latent is noised at a sampled scheduler timestep, and teacher predictions are evaluated online. We write $z_t = \mathcal{N}_t(x, \epsilon)$, where $\mathcal{N}_t$ denotes the scheduler's forward-noising operation. For bidirectional T2V, one timestep is shared by the whole clip, with timestep indices sampled between 2% and 98% of the training schedule.

Algorithm 1 CFG-only LoRA distillation

Require: Frozen teacher F_{θ_0}, trainable adapter ϕ_{cfg}, teacher scale w_{train}, captioned latents, optimizer

1: **for** each training iteration **do**
2: Sample captioned latent (x, c), timestep t, and noise ϵ
3: $z_t \leftarrow \mathcal{N}_t(x, \epsilon)$
4: Evaluate frozen $v_c \leftarrow F_{\theta_0}(z_t, t, c)$ and $v_\varnothing \leftarrow F_{\theta_0}(z_t, t, \varnothing)$
5: $v_{\text{cfg}} \leftarrow v_\varnothing + w_{\text{train}}(v_c - v_\varnothing)$
6: $v_s \leftarrow F_{\theta_0 \oplus \phi_{\text{cfg}}}(z_t, t, c)$
7: Compute the flow-regression loss
8: Backpropagate only through v_s; clip gradients; update ϕ_{cfg}
9: **end for**
10: **return** ϕ_{cfg}; deploy with the native schedule and distilled CFG

A.2 Few-Step LoRA Training

Few-step LoRA training follows the distribution-matching objective of DMD2 (Yin et al., 2024a), with a frozen real-score teacher, a generator adapter ϕ_{step}, and a trainable fake-score adapter ψ.

Distribution-matching update. Given a student sample x_ϕ, re-noise its detached value at an independently sampled score timestep t. The timestep is warped as $u \mapsto su/[1 + (s-1)u]$ and clamped to $[0.02, 0.98]$ of the training time range. Let $\widehat{x}_\psi$ and $\widehat{x}^{(w)}_{\text{real}}$ be the fake-score and guided real-score clean predictions at this state. The code normalizes their difference within each temporal block b:

$$\begin{aligned} n_b &= \text{mean}_{f \in b, C, H, W} \, |x_\phi - \widehat{x}^{(w)}_{\text{real}}|, \\ g_b &= \text{nan_to_num}\left[(\widehat{x}_\psi - \widehat{x}^{(w)}_{\text{real}})/n_b\right], \\ \mathcal{L}_G &= \tfrac{1}{2} \,\text{mean}\, \|x_\phi - \text{sg}[x_\phi - g]\|^2 . \end{aligned} \tag{5}$$

The detached target makes the generator gradient proportional to g; gradients do not pass through either score network in this update. For the fake-score update, another detached student sample is noised and the fake-score adapter predicts the flow target $\epsilon - x_\phi$:

$$\mathcal{L}_D = \text{mean} \, \|v_\psi(\mathcal{N}_t(\text{sg}[x_\phi], \epsilon), t, c) - (\epsilon - \text{sg}[x_\phi])\|^2 . \tag{6}$$

The selected flow-loss path has no adversarial discriminator term.

Algorithm 2 Wan2.2 few-step LoRA distillation with a learned fake score

Require: Frozen real-score teacher, generator adapter ϕ_{step}, fake-score adapter ψ, four-step schedule, update ratio $R = 5$

for iteration $j = 0, \ldots, J-1$ **do**
  Sample a training prompt c
  if $j \bmod R = 0$ **then**
    Generate a student sample with the four-step schedule
    Re-noise the detached clean prediction at a random score timestep
    Evaluate conditional fake score and CFG-guided frozen real score
    Form detached g and $\mathcal{L}_G$ using Eq. (5)
    Accumulate gradients for ϕ_{step} only
  end if
  Generate a separate student sample without gradients
  Re-noise it; compute $\mathcal{L}_D$ using Eq. (6)
  Accumulate gradients for ψ only
  Clip and apply the scheduled generator update and the fake-score update
end for
return ϕ_{step}; discard the fake-score network for inference

A.3 Optimization and Deployment

Unless varied in the rank ablation, we train each CFG, few-step, and long-context LoRA at rank $r = 128$. The CFG and few-step branches are trained separately. Tab. 4 summarizes the Wan reference optimization settings.

Table 4: **Wan training rank and reference optimization settings.** All branches are trained at rank 128. Batch sizes are per process. For DMD, the generator updates once per five fake-score updates. The CFG column reports the reference CFG-only optimization settings.

Setting	CFG-only, Wan2.2	Few-step, Wan2.2	Few-step, Wan2.1
LoRA training rank r	128	128	128
Fake-score LoRA	None	Yes	Yes
Generator / critic LR	10^{-5} / —	10^{-5} / 2×10^{-6}	10^{-5} / 2×10^{-6}
AdamW (β_1, β_2)	$(0, 0.999)$	$(0, 0.999)$	$(0, 0.999)$
Weight decay	0	0	0.01
LoRA dropout	0	0	0
Per-process batch / accumulation	1 / 1	1 / 1	1 / 1
Generator EMA	Off	0.99 from step 200	Off
Standard teacher CFG w	5	4	5
Latent frames $\times C \times H \times W$	$32\times48\times22\times40$	$32\times48\times44\times80$	$21\times16\times60\times104$
Sampling steps	50	4	4

Training uses mixed precision, FSDP, and gradient checkpointing. CFG regression uses FP32 targets and residuals, with gradient clipping at norm 10. A DMD checkpoint iteration counts fake-score updates; the generator updates only at iterations divisible by five.

At deployment, we merge the CFG and few-step updates into the downstream weights as in Eq. (4), preserving target-specific modules. We fix $\lambda_{\text{step}} = 1$ and adjust λ_{cfg} for each downstream task.

B Additional Acceleration Comparisons

This section extends Sec. 4.1. We show more cases in Figs. 8 and 9.

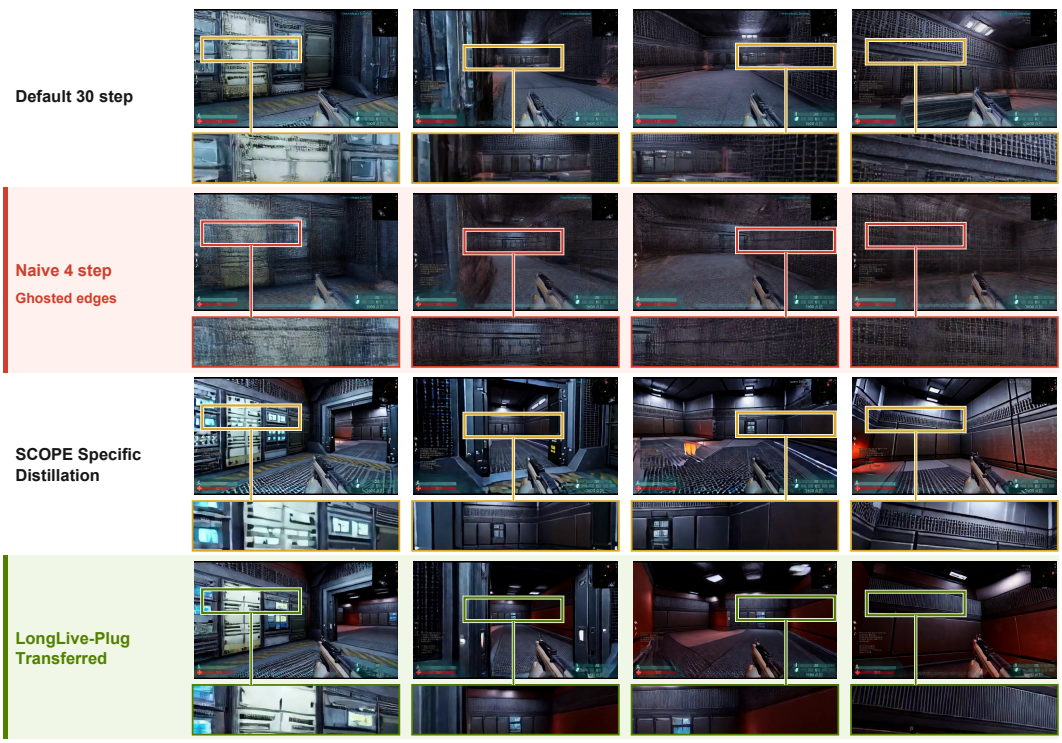

Figure 8: **Keyframe comparison on SCOPE.** Four matched keyframes from an 81-frame sequence at 20 fps. Rows compare default 30-step inference, naive four-step sampling, SCOPE-specific distillation, and transferred LoRAs, using the same case and seed. Boxes mark identical image coordinates across methods; the strips below each frame magnify these regions. Red highlights the naive four-step row's ghosted wall and door edges, while the distilled variants retain clearer boundaries and surface detail.

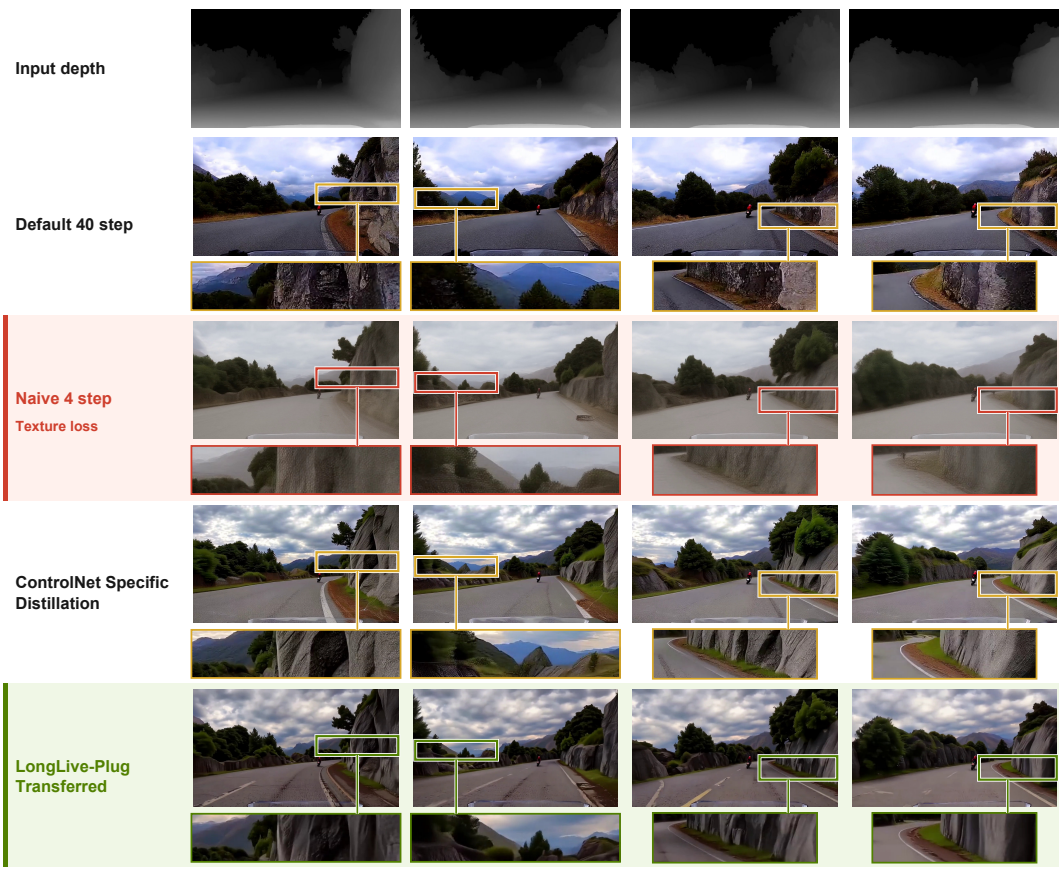

Figure 9: **Keyframe comparison on Wan2.2-Fun-5B-Control.** Four matched keyframes compare input depth and four generation methods. Boxes mark identical image coordinates across methods, magnified in the strips below each output. Red highlights the naive four-step row's smeared rock, foliage, and road details and low contrast. Columns match conditioning frame indices (depth at 30 fps, outputs at 24 fps). The transferred output uses an additional base-distilled adapter variant.

C Complete Transfer Coverage and Additional Cases

C.1 Coverage Inventory

The inventory supporting Sec. 4.2 contains 54 distinct downstream model entries: 24 descendants of Wan2.1-14B, 24 of Wan2.2-TI2V-5B, and six of MiniMax-H3.

Each family uses an adapter distilled on its own base model. MiniMax-H3 natively supports inference without CFG, so its transfer experiments omit the CFG LoRA by default. We also train a CFG-only LoRA on the MiniMax-H3 base model and find that it supports distilled CFG inference with adjustable guidance (Fig. 16).

Table 5: **Recorded transfer coverage across backbone families and tasks.** Every listed Wan-model entry is evaluated for four-step and CFG transfer. The two panels share the same backbone axis.

Backbone	World models	Robotics	ControlNet / structure	Camera / trajectory
Wan2.1-14B	FantasyWorld (Dai et al., 2025); LongVie 2 (Gao et al., 2025); Micro-World I2W (AMD, 2025)	DreamZero (Ye et al., 2026); ABot-PhysWorld (Chen et al., 2026c); VERA DROID Planner (Li et al., 2026c)	Fun Control (Alibaba PAI, 2025a); TheDenk Dilated ControlNet (Karachev, 2025)	Fun-V1.1 Control-Camera (Alibaba PAI, 2025b); Wan-Move (Chu et al., 2025); ATI (Wang et al., 2025a); NeoVerse (Yang et al., 2026b); Vista4D (Lin et al., 2026a)
Wan2.2-TI2V-5B	Matrix-Game 3.0 (Wang et al., 2026d); DreamX-World-5B AR (DreamX Team et al., 2026); SCOPE (Tong et al., 2026); ReactiveGWM (Wang et al., 2026c); Spatia (Zhao et al., 2025)	Boundless World Model (BWM Team, 2026); Fast-WAM LIBERO (Yuan et al., 2026); Motus Stage-1 VGM (Bi et al., 2025)	Fun Control (Alibaba PAI, 2025c)	Fun Control-Camera (Alibaba PAI, 2025d); FlashMotion (Li et al., 2026b); FrameINO v1.6 (Wang et al., 2025b)
MiniMax-H3	H3-World (Chen et al., 2026a); Code World Model (Chen et al., 2026b); SolarWM-H3 (Huang et al., 2026)	—	Fun ControlNet-Union (Alibaba PAI, 2026)	SolarWM-H3 (Huang et al., 2026)

Backbone	Editing / restoration	Subject / avatar	Audio / RGBA outputs	Domain / style / quality
Wan2.1-14B	VideoCoF (Yang et al., 2025b); SAMA-14B (Zhang et al., 2026b)	Stand-In (Xue et al., 2025); MAGREF (Deng et al., 2025); MagicTryOn (Li et al., 2025); InfiniteTalk (Yang et al., 2025a); MultiTalk (Kong et al., 2025); FantasyTalking (Wang et al., 2025c); AniCrafter (Niu et al., 2025)	Wan-Alpha v1/v2 (Dong et al., 2025)	Index-AniSora V2.0 (Jiang et al., 2024; Index Team, 2025)
Wan2.2-TI2V-5B	Fun InP (Alibaba PAI, 2025e); Kiwi-Edit (Lin et al., 2026b); StreamingEffect (Song et al., 2026)	TalkVerse-5B (Wang et al., 2025d)	Ovi (Low et al., 2025); OmniCustom (Li et al., 2026a); NAVA (Ji et al., 2026); UniAVGen (Zhang et al., 2025)	Loomis Painter LoRA (Pobitzer et al., 2025); Aether action/VFX LoRAs (Sallström, 2025); VideoGPA DPO LoRA (Du et al., 2026); PhysRVG (Zhang et al., 2026a)
MiniMax-H3	LineartAnime (DiffSynth-Studio, 2026)	Viggle-Animate (Viggle Research, 2026)	—	—

C.2 Additional Wan Transfer Examples

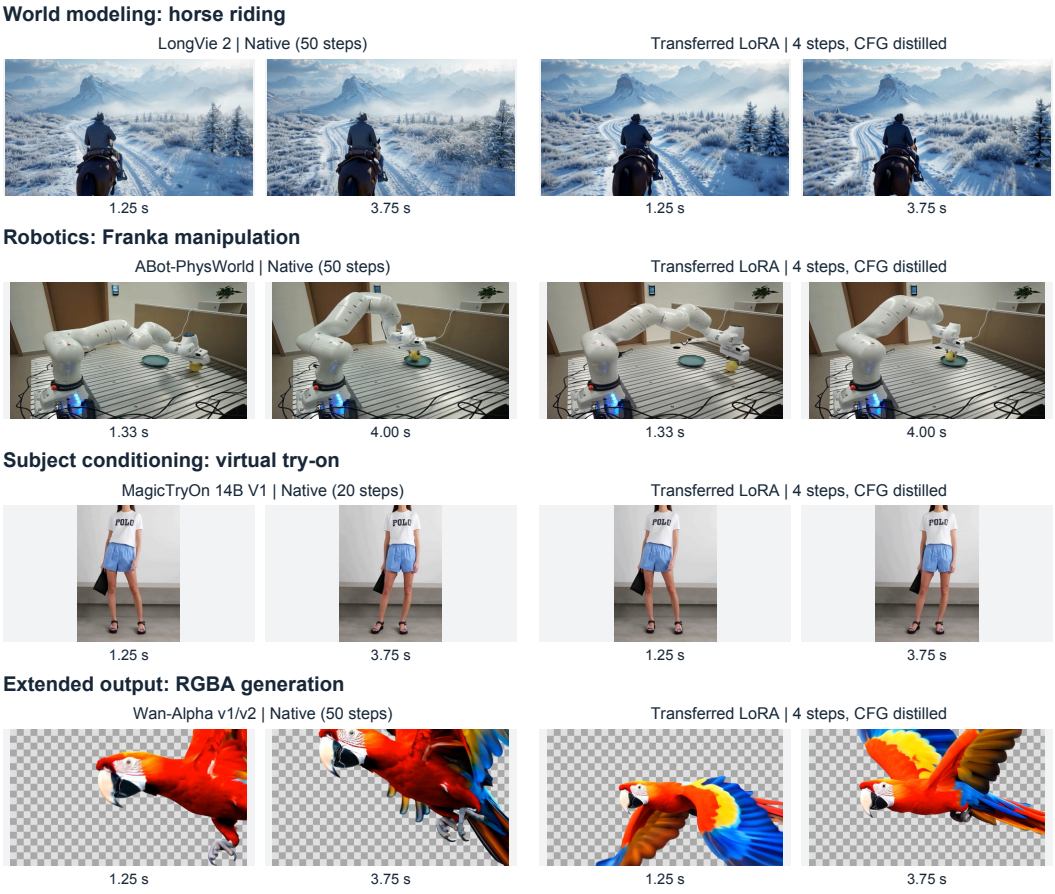

Figure 10: **Additional Wan2.1-14B transfer cases.** LongVie 2, ABot-PhysWorld, MagicTryOn, and Wan-Alpha illustrate world modeling, robotics, subject conditioning, and RGBA output. Each row compares two native frames (left) with the same two timestamps after four-step transfer (right). Inputs and task conditions are paired in the source report. The checkerboard is part of the Wan-Alpha preview, not a measurement of alpha-channel accuracy. Changes in appearance and motion remain visible after transfer.

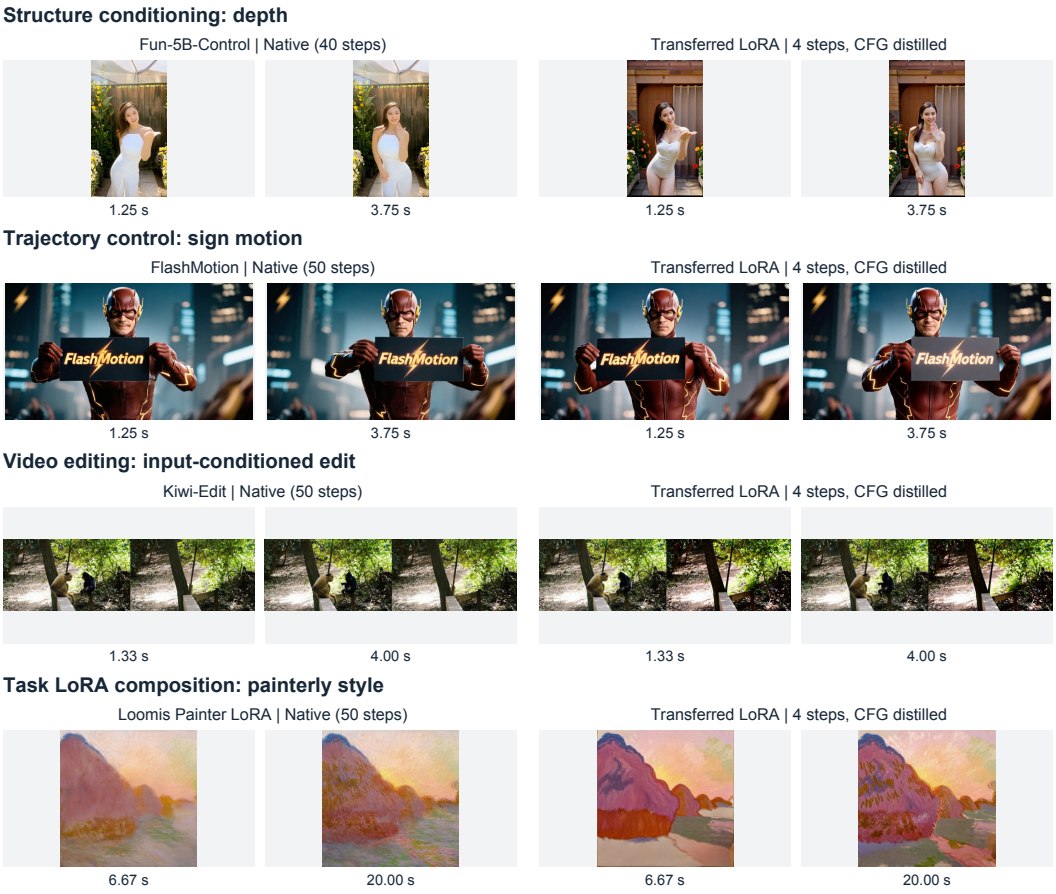

Figure 11: **Additional Wan2.2-TI2V-5B transfer cases.** Depth-conditioned Fun Control, FlashMotion, Kiwi-Edit, and Loomis Painter cover structure, trajectory, editing, and style adaptation. The native and transferred columns show identical frame indices for each paired task input. Native inference uses 50 steps for Kiwi-Edit and Loomis Painter. All transferred outputs use four steps with CFG distilled into the adapter; downstream conditioning and task adapters are retained.

C.3 MiniMax-H3 Transfer and Its Limits

The H3 comparisons show undistilled multi-step inference on the left (D), naive four-step Euler in the middle (E4), and four-step inference with LongLive-Plug on the right (S4). S4 uses fresh re-noising. The downstream conditions are fixed. E4 and S4 share the initial noise; D retains the native random-number generation path, so bitwise-identical initial noise across all three arms is not established. E4 to S4 changes both the adapter and sampler. The following examples assess the complete deployment recipe, not the isolated causal effect of adding a LoRA.



Figure 12: **H3 task transfer: action control and line-art coloring. Left:** undistilled multi-step inference. **Middle:** naive four-step sampling. **Right:** four-step inference with LongLive-Plug. H3-World shows matched frames at 1 and 4 s under a forward-action condition. LineartAnime shows frames at 1 s and the final available frame (3.71 s). S4 preserves a clearer character outline than E4 in these examples, while appearance can differ from D. These static frames do not evaluate audio quality or synchronization.

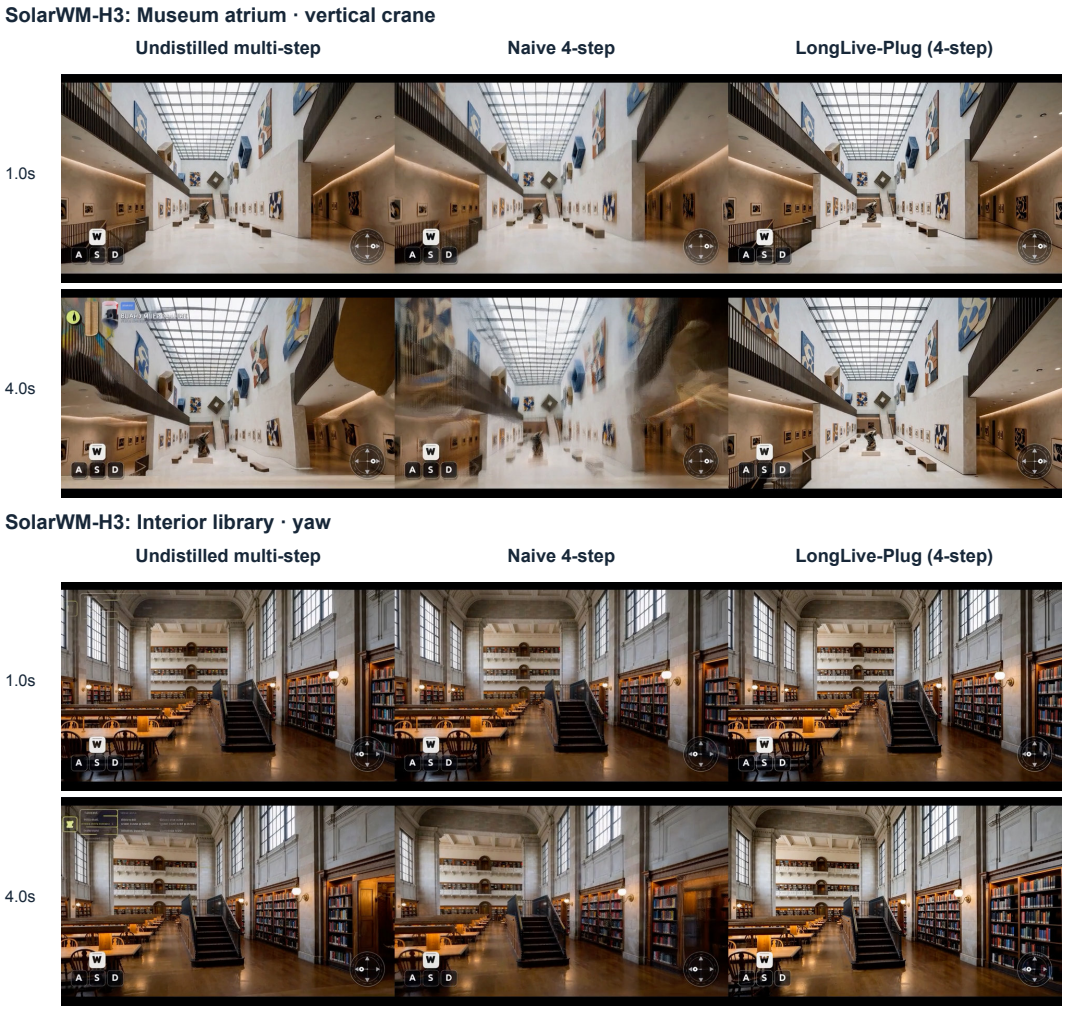

Figure 13: **H3 camera transfer includes a quality–control trade-off. Left:** undistilled multi-step inference. **Middle:** naive four-step sampling. **Right:** four-step inference with LongLive-Plug. In the museum-crane example, S4 retains clearer architectural detail and a visible upward-camera response. In the library-yaw example, S4 remains sharp but its change in framing is attenuated relative to D/E4. HUD elements are inherited from the supplied anchor images. These cases have no generated-video pose regression metric and do not establish precise trajectory adherence.

The ten H3 comparisons are single-seed cases, with no repeated-run confidence intervals. The native default is a reference operating point, not ground truth. Similarity to it cannot by itself measure action, camera, or structure-control accuracy. Runtime accounting also differs across H3 backends, so timings should only be compared within a task.

D ADDITIONAL CFG CONTROL EXPERIMENTS

This section extends Sec. 4.3 with native-schedule CFG-only control and four-step composition. Displayed LoRA weights scale adapter residuals and are not calibrated runtime CFG scales.

D.1 CFG-ONLY CONTROL WITH THE NATIVE SCHEDULE

Wan2.2-TI2V-5B. The CFG-only experiment uses the native 50-step FlowUniPC schedule, no few-step adapter, teacher scale $w_{\text{train}} = 5$, and runtime CFG 1. Fig. 14 adds a flower-opening prompt using the main-text milk-splatter comparison protocol.

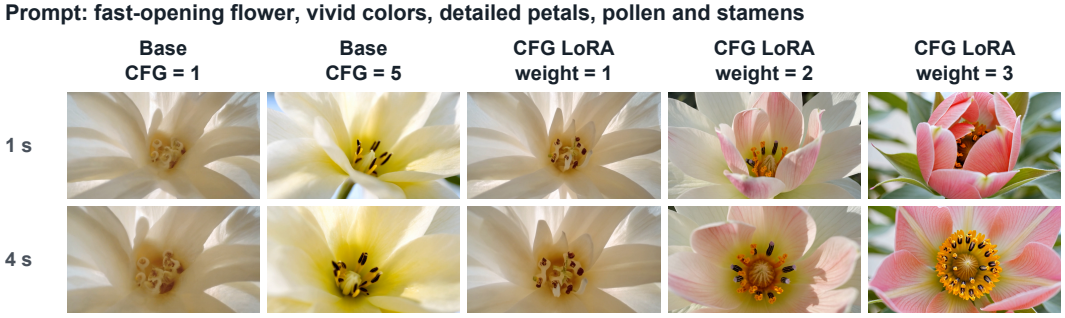

Figure 14: **Additional Wan CFG-only control example.** Native CFG references and CFG LoRA weights 1, 2, and 3 use 50 sampling steps; LoRA outputs use runtime CFG 1 with one conditional evaluation per step. Matched frames at 1 and 4 s show more pronounced flower opening as the adapter weight increases.

MiniMax-H3. Native CFG and CFG LoRA outputs are compared in Figs. 15 and 16.

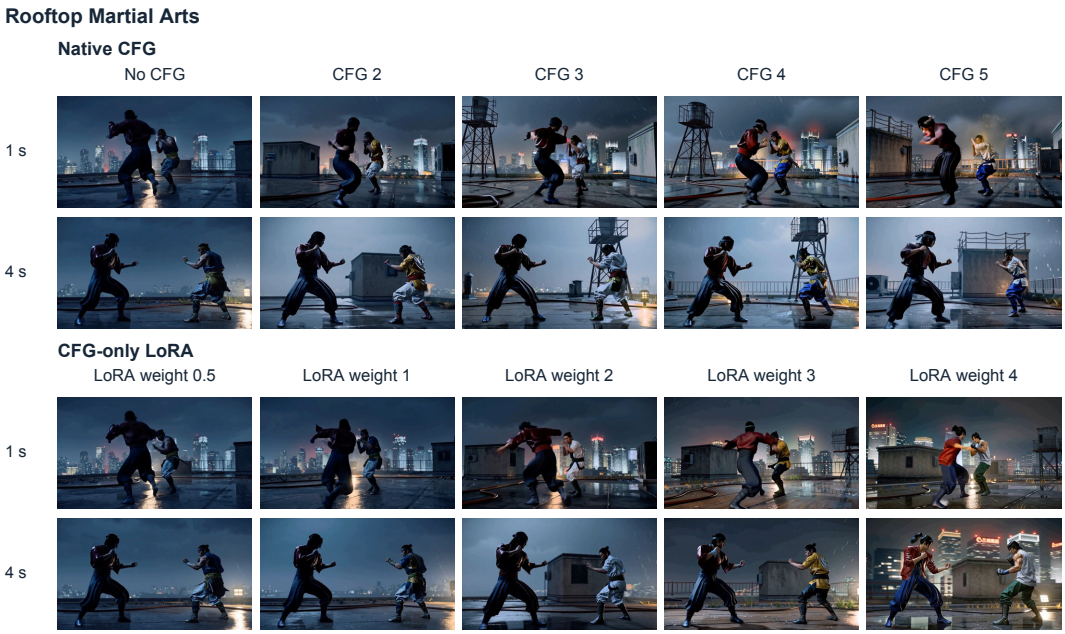

Figure 15: **Native CFG and CFG-only LoRA on Rooftop Martial Arts.** Matched frames at 1 and 4 s compare native CFG (top: no CFG and scales 2–5) with CFG LoRA (bottom: weights 0.5, 1, 2, 3, 4). The prompt, seed, and native schedule are fixed; LoRA outputs use one conditional forward pass per step. LoRA weights are not calibrated native CFG scales.

The MiniMax-H3 CFG-only report covers 34 prompts, seven native reference conditions, and six adapter weights per prompt (442 videos). The adapter is trained at CFG scale 3 and uses no external CFG. The native 50-step scheduler performs 49 denoising updates (one conditional forward per LoRA update), generating 124 frames at 24 fps and 1344×768 resolution; video/audio flow shifts are $12/3$. Figs. 15 and 16 show Rooftop Martial Arts, Night village—neutral, and Night village—subtle Van Gogh influence. Each case places native CFG above CFG LoRA with matched frames at 1 and 4 s and a fixed prompt and seed. Native columns use no CFG and scales 2–5; adapter weights are 0.5, 1, 2, 3, and 4.



Figure 16: **Native CFG and CFG-only LoRA on the two night-village cases.** Each case places native CFG above CFG LoRA at matched 1 and 4 s frames, using the sweeps in Fig. 15. The prompt, seed, and native sampling schedule are fixed within each case.

D.2 Independent CFG Control with a Few-Step Adapter

We compose the two branches as $\Delta W = \Delta W_{\text{step}} + \lambda_{\text{cfg}} \Delta W_{\text{cfg}}$. The few-step weight remains 1 while $\lambda_{\text{cfg}} \in \{0.5, 1, 2, 3, 5\}$ varies. Every output uses four denoising steps, distilled CFG, and scheduler shift 5. Within each sweep, the prompt, seed, conditioning, and action sequence are fixed. These examples show that the CFG branch remains an effective semantic control when combined with the few-step branch.

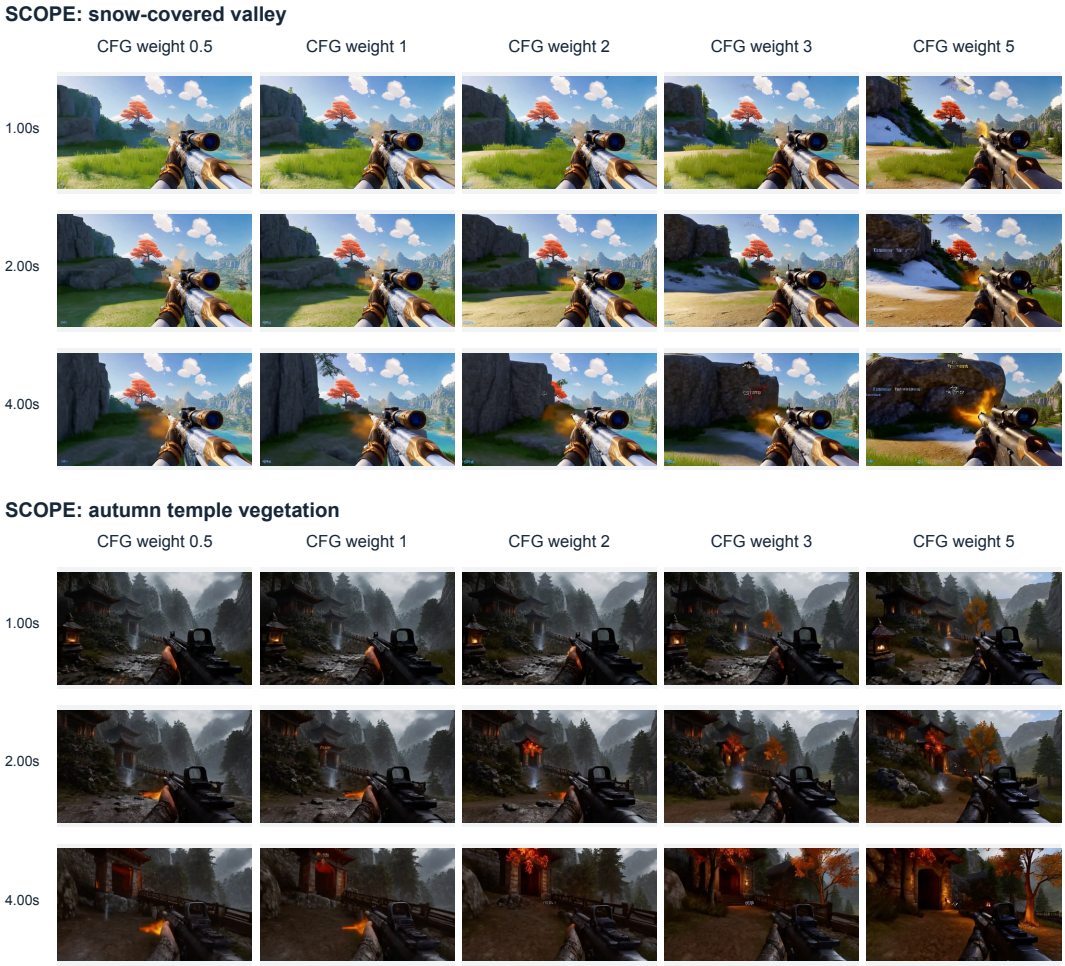

Figure 17: **CFG-branch control with four-step SCOPE generation.** The prompts request snow cover in a mountain valley and autumn vegetation around an ancient temple. Both use 81 frames at 20 fps; rows show 1, 2, and 4 s. Increasing the CFG-branch weight strengthens the snow cover and orange-red foliage, while the few-step weight stays at 1. An excessively large weight (5) degrades quality, changing geometry and introducing spurious text.

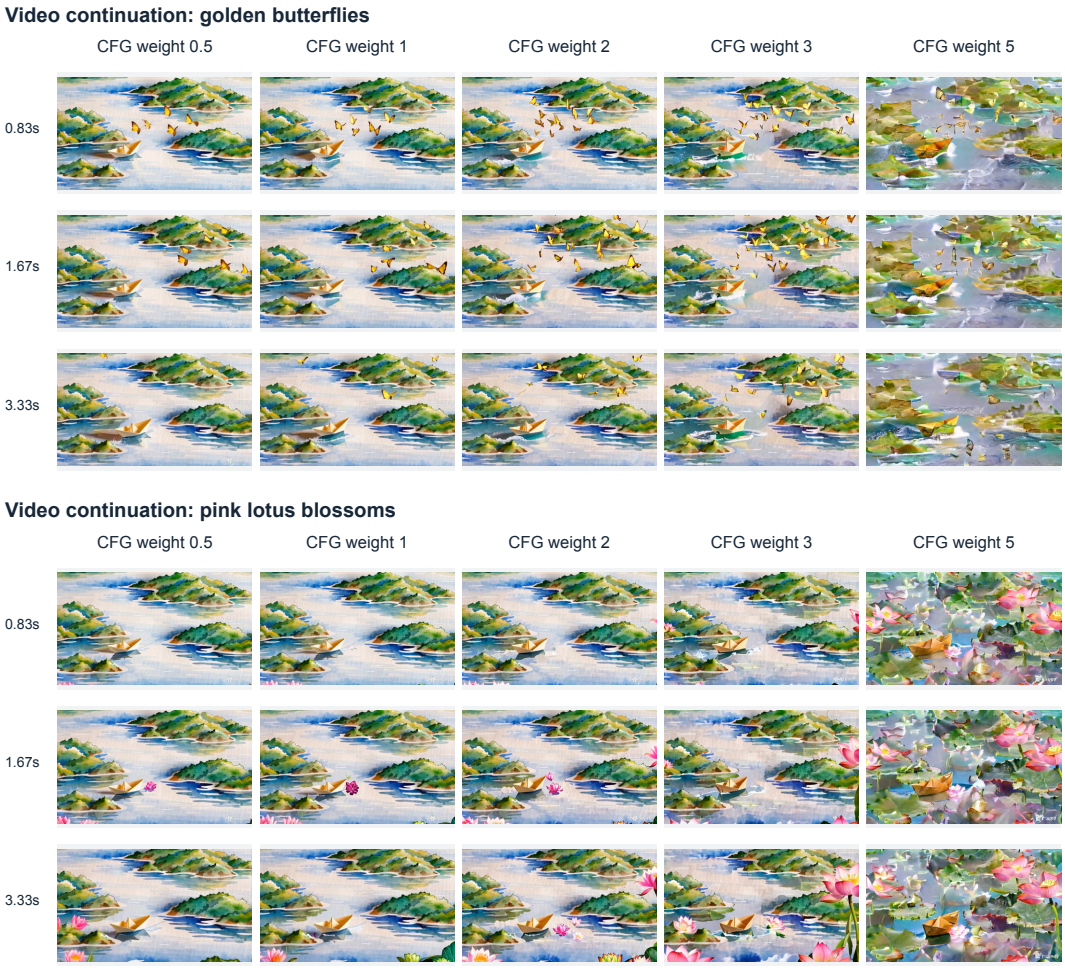

Figure 18: **CFG-branch control with four-step video continuation.** We continue the same watercolor paper-boat input with golden butterflies or pink lotus blossoms. The last 24 input frames condition 81 new frames at 24 fps; displayed times are relative to the generated continuation. Weights 2–3 produce more visible and persistent butterflies or more prominent lotus blossoms. At weight 5, dense generated content comes with fragmented scenery and stronger changes to the boat and islands. The few-step branch remains fixed at weight 1 in every column.

Figs. 17 and 18 show that increasing the CFG LoRA weight strengthens text guidance, making the requested attributes more pronounced. However, overly large weights, such as 5, degrade visual quality. The CFG LoRA weight should therefore be adjusted within a reasonable range for each task, balancing text guidance and visual quality.

E LONG-CONTEXT QUALITATIVE COMPARISONS

This section accompanies Sec. 4.5, with long rollouts on ReWorld and Matrix-Game 3.0.

E.1 REWORLD

Fig. 19 retains two 64 s cases with matched prompts, camera trajectories, and initial noise. Generated layouts can differ across methods.

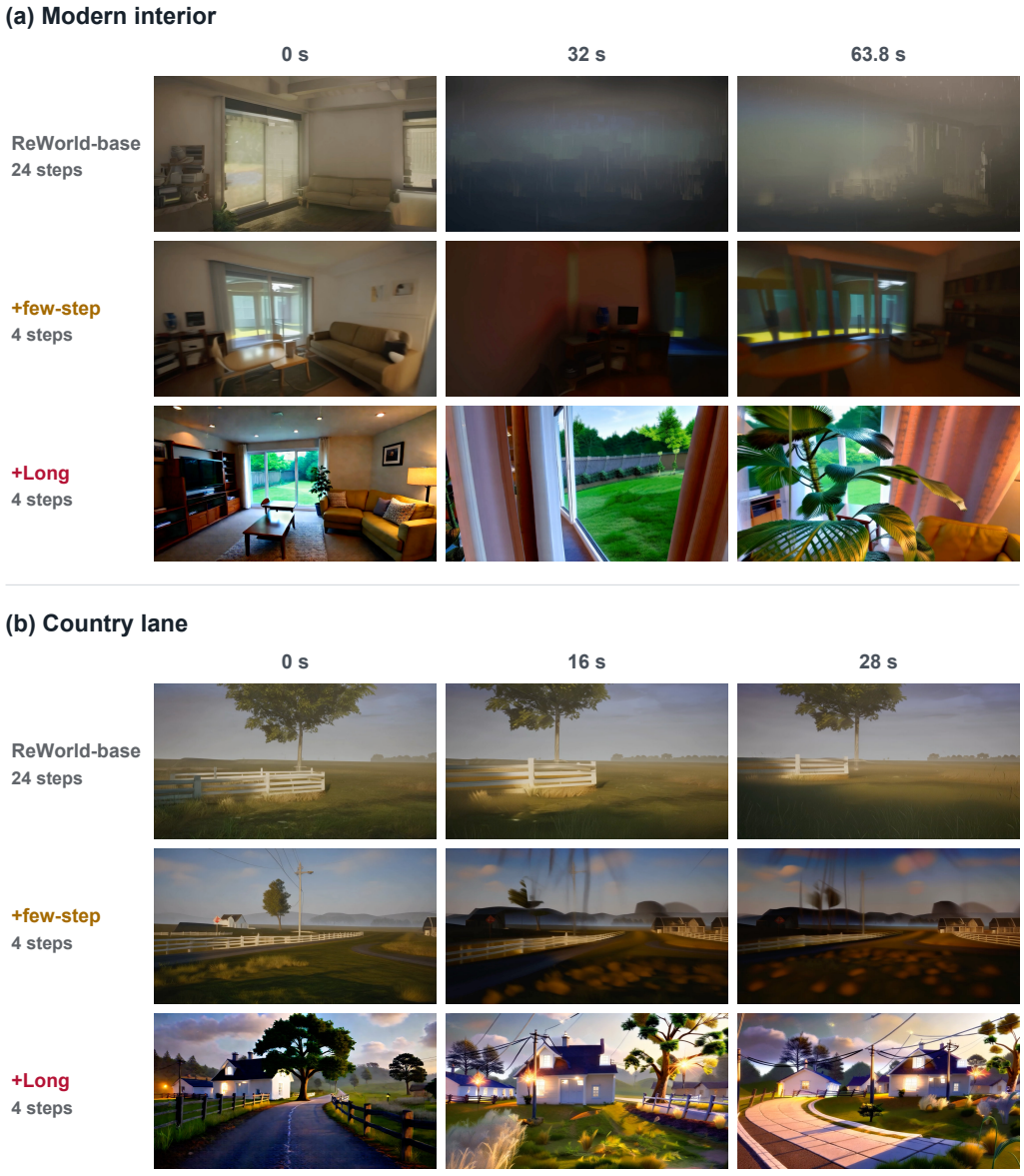

Figure 19: **Long-rollout qualitative comparisons on ReWorld.** Matched frames from a modern interior and a country lane. The +Long outputs retain more visible texture and object detail at late times.

E.2 MATRIX-GAME 3.0

We show two cases. Each rollout has 1,057 frames at 17 fps and lasts 62.18 s. All four methods share the same input image, prompt, seed, and frozen action sequence.



Figure 20: **Long-context transfer to Matrix-Game 3.0: animated city.** Columns show native frames 132, 528, 924, and 1,056 (zero-based), spanning early, middle, and late stages of the same 62.18 s rollout. The +Long output retains distinct facade edges and street objects at late times.

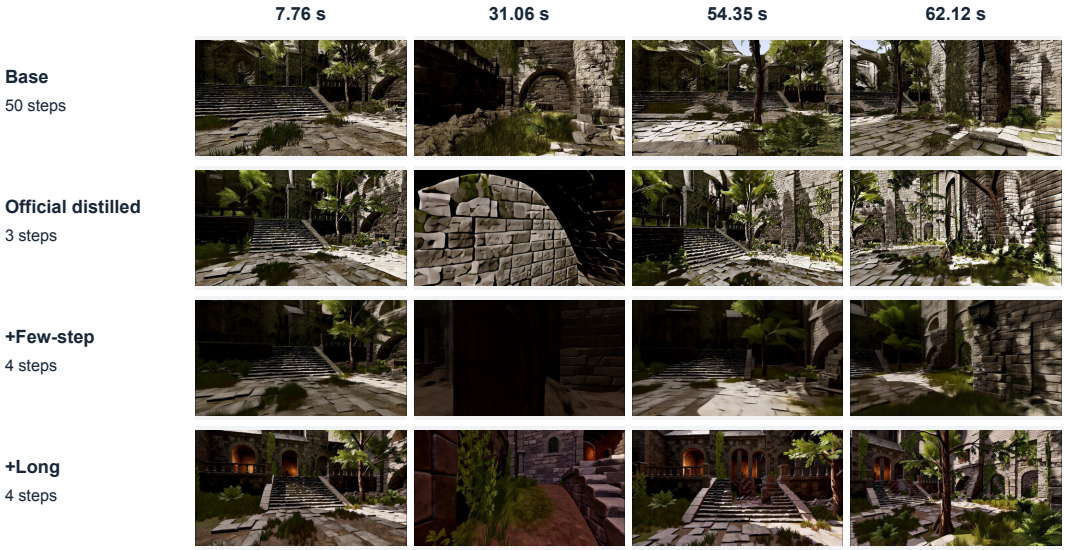

Figure 21: **Long-context transfer to Matrix-Game 3.0: overgrown temple.** The same methods and timestamps as Fig. 20 compare stone architecture and vegetation. The +Long frames preserve visible stone-block boundaries, steps, and foliage late in the rollout, while appearance and layout vary across methods.