WONDER: VIDEO WORLD MODEL DONE BETTER

Anonymous authors
Paper under double-blind review

ABSTRACT

We present Wonder, a general-purpose video world model for real-time, camera-controllable world exploration. Given a single image or a conditioning video, Wonder constructs a playable world in which users navigate by moving the camera, discovering unseen regions and revisiting observed ones over a long horizon. This requires a system-level co-design of control, memory, and training. We introduce a camera conditioning scheme based on a dense pixel-space coordinate field whose renderings supply spatially aligned motion and orientation cues, so the model reads camera motion directly as visual evidence. For fast retrieval over a growing context, we propose a sparse full-fidelity memory mechanism that attends to a small set of relevant tokens at inference, independent of context length. We further rectify the self-forcing distillation pipeline with a timestep-wise mixture-of-students and a camera-aware adversarial regularizer, improving control adherence while retaining the teacher's generation diversity and stability. Together, these components let Wonder synthesize diverse, minute-scale videos at 16 FPS with coherent geometry, appearance, and dynamics across long rollouts. Beyond image-to-video, Wonder natively supports video-conditioned generation, re-shooting existing dynamic scenes in real time.

1 INTRODUCTION

Video generation is evolving from fixed-duration clip synthesis toward interactive visual world modeling, with applications in game creation (Tang et al., 2025), virtual production (Rahimi et al., 2025), and robotics (Xiao et al., 2025). In these settings, a model should build a persistent visual environment from an initial image or video and let users explore it through control signals (Figure 1). This demands more than visual fidelity: interaction needs accurate control following and low latency, simulation needs long-horizon coherence, and editing needs source content preserved under control. Satisfying all three at once is difficult, because controllability, memory, and real-time efficiency place conflicting demands on the architecture and the training pipeline.

Most existing methods (Wang et al., 2026b; Li et al., 2025; Team et al., 2026a) address only a subset of them. Camera-controllable video generators condition on trajectories accurately, but are typically bidirectional, fixed-length, and expensive, precluding low-latency interaction or minute-scale rollout (Bai et al., 2025a; Yu et al., 2025). Autoregressive distillation methods improve streaming efficiency, but target video continuation rather than camera following, persistent spatial memory, or consistent revisits (Huang et al., 2025b; Cui et al., 2025; Zhu et al., 2026b). Recent interactive world models combine control, memory, and real-time generation, yet each component brings a new trade-off: abstract pose or ray conditions drift after distillation (Zhu et al., 2026a; Team et al., 2025), reconstruction-based control is limited to observed views (Yu et al., 2024; Team et al., 2026b), dense historical attention makes latency grow with rollout length (Robbyant Team, 2026), and compressed or sliding-window memory loses history details (Hong et al., 2025). Existing systems therefore still struggle to deliver accurate control, faithful memory, and real-time inference simultaneously.

We present Wonder, a general-purpose video world model for real-time, camera-controllable exploration. Wonder follows the common teacher-to-student paradigm: a bidirectional video diffusion model is adapted for camera control, converted into a causal autoregressive generator, and distilled into a few-step streaming model. Unlike prior systems that optimize these stages in isolation, Wonder treats the pipeline as a coupled system and jointly redesigns three components, so that controllability, persistence, and efficiency survive the path from teacher to real-time student.

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

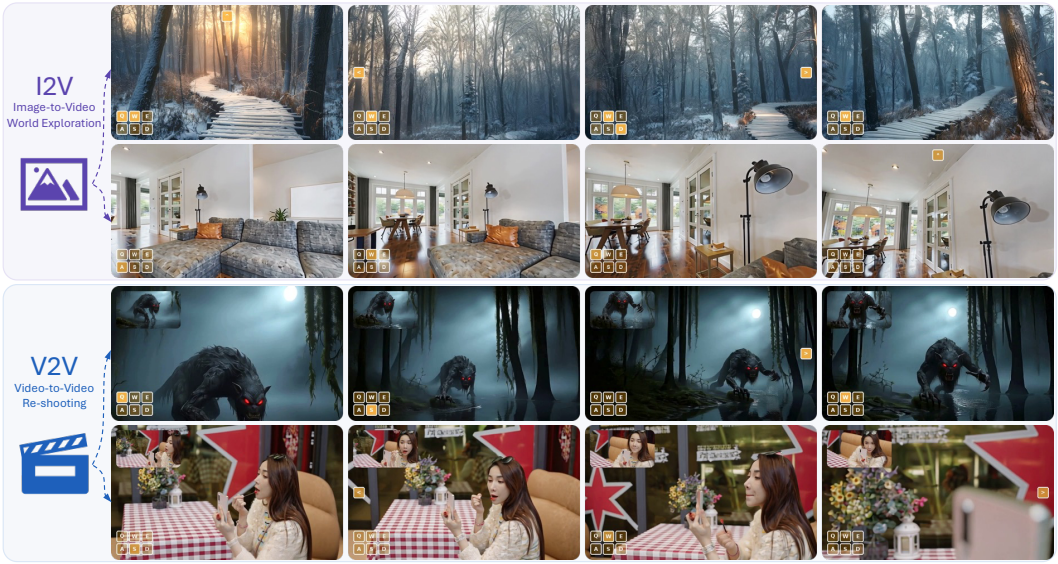

Figure 1: **Wonder enables interactive world exploration from images and videos.** From a single image, Wonder builds a navigable world in which the user moves the camera to explore both visible regions and plausible unseen areas. From a video, it builds a dynamic world that preserves the source motion while the camera trajectory is changed freely. More results can be accessed via the anonymous link: `https://canyon-tugboat-97z5b.shipped.run/`

Control. Rather than conditioning on camera geometry alone, as pose (Hong et al., 2025) and Plücker-ray (Wang et al., 2026a) do, we render a synthetic 3D scaffold and environment map along the target trajectory, turning translation, rotation, and parallax into frame-aligned cues. Because the scene is synthetic rather than reconstructed, control stays informative past the observed view, and its visual semantics survive few-step distillation better than abstract embeddings.

Memory. Wonder decouples memory storage from attention computation: it keeps the growing KV cache at full fidelity but uses pooled query-key summaries to select a constant-size set of memory chunks per rollout step, attending only to those at full resolution. A co-designed training strategy, *sparse context forcing*, activates this capability in the causal student, which learns to generate under the same retrieved-memory condition it faces at inference.

Distillation. Approximating a bidirectional many-step teacher with a causal few-step student reduces generation diversity and accumulates camera drift. We address both with a timestep-wise *mixture-of-students*, which raises effective student capacity without changing the streaming interface, and a *GAN control regularizer*, which supplies the low-frequency control supervision that distribution matching alone does not provide.

Experiments in both settings show that Wonder improves visual quality and camera-following accuracy over recent streaming world models, explores beyond observed views, preserves dynamic content when re-rendering source videos, and maintains coherent revisits over long trajectories, generating long rollouts at 16 FPS with latency that stays stable as the history grows.

2 Related Work

Modern video generators are diffusion transformers in the latent space of a 3D VAE (Yang et al., 2025; Kong et al., 2024; Wan et al., 2025); being bidirectional and fixed-length, they preclude interactive control and unbounded rollout. A second line converts them into causal generators—Diffusion Forcing (Chen et al., 2024), CausVid (Yin et al., 2025), Self-Forcing (Huang et al., 2025b), and their streaming extensions (Liu et al., 2025; Zhu et al., 2026b) providing the foundation Wonder builds on, though they optimize continuation quality rather than control fidelity or persistent memory. Camera control is usually pose-conditioned: MotionCtrl (Wang et al., 2024) injects 6-DoF extrinsics, CameraCtrl (He et al., 2025a) uses Plücker embeddings, ViewCrafter (Yu et al., 2024)

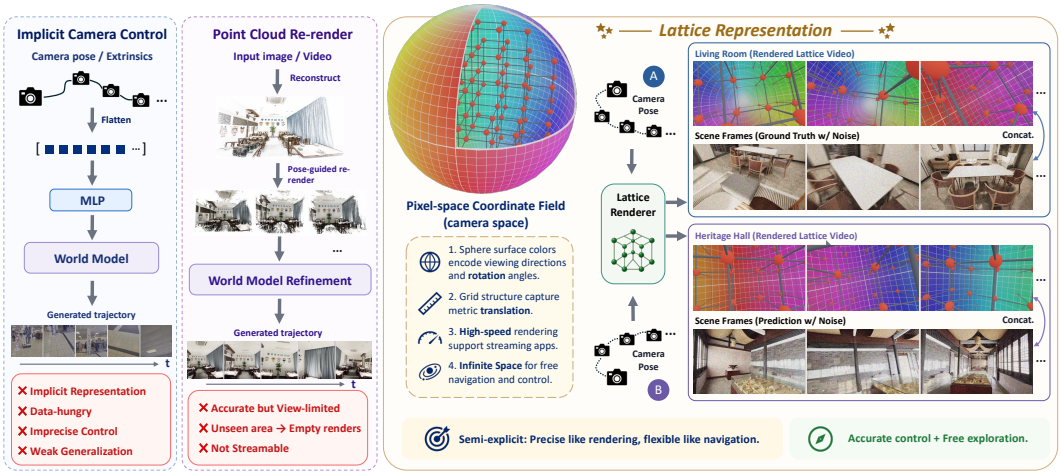

Figure 2: **Camera control with a pixel-space coordinate field.** Implicit methods encode poses with an MLP (Bai et al., 2025a) or RoPE (Wang et al., 2026a); explicit methods re-render a reconstructed point cloud (Team et al., 2026b). Ours combines their strengths: a colored spherical environment map at infinity plus a dense 3D lattice scaffold, rendered along arbitrary trajectories into pixel-aligned videos encoding viewing direction and metric translation.

conditions on point-cloud renderings, and ReCamMaster (Bai et al., 2025a) casts control as video-to-video re-shooting (Wang et al., 2026a; Huang et al., 2025c). Building on these, Genie 3 (Ball et al., 2025) marks the closed-source frontier, while an open-source line spanning Matrix-Game and its streaming successors (Zhang et al., 2025; He et al., 2025b; Wang et al., 2026b), Hunyuan-GameCraft (Li et al., 2025; Tang et al., 2025), Yume (Mao et al., 2025), RELIC (Hong et al., 2025), LingBot-World (Robbyant Team, 2026), and DreamX-World (Team et al., 2026a) explores streaming, long-horizon memory, and richer action interfaces, with Inspatio-World (Team et al., 2026b) closest to our setting in supporting both I2V and V2V via point-cloud re-rendering. These systems still exhibit camera drift, memory loss, and quality degradation over long rollouts; Wonder instead co-designs control, memory, and distillation as one system. Appendix C gives a full discussion.

3 Wonder World Model

Wonder generates explorable videos following a user-streamed camera trajectory, in two modes within one model: in *image-to-video* (I2V) mode a single frame initializes an explorable world, and in *video-to-video* (V2V) mode a source video anchors the appearance and motion of an existing event while the target trajectory specifies a new viewpoint of it. Both modes share one input layout, in which an optional source segment is supplied as clean latents, a binary mask separates source from target tokens, noise is applied only to generated frames, and camera-condition latents are zero-padded over the source. Following recent work (Hong et al., 2025; Robbyant Team, 2026), we distill a few-step autoregressive student from a bidirectional teacher, both conditioned on camera control, and redesign the three components it couples: the control representation (Sec. 3.1), the memory mechanism (Sec. 3.2), and the distillation objective (Sec. 3.3). Appendix B details the input layout, backbone, training schedule, runtime, and the overall pipeline (Figure 9).

3.1 Pixel-Space Coordinate Field for Camera Control

Camera-control designs fall into two families (Figure 2). *Implicit* methods encode motion as low-dimensional vectors derived from extrinsics (Robbyant Team, 2026; Bai et al., 2025a; Hong et al., 2025), injected token-wise so that every spatial token in a latent frame shares one condition; such global codes give coarse cues that are hard to associate with local, high-frequency pixel changes, and ray-based variants (Wang et al., 2026a; Team et al., 2026a) remain ambiguous in scale, origin, and depth. Distillation amplifies both weaknesses, since objectives such as DMD (Yin et al., 2024b) supply no direct supervision for camera following. *Explicit* methods (Team et al., 2026b; Yu et al., 2024) instead re-render a point cloud reconstructed from the input. This anchors geometry strongly,

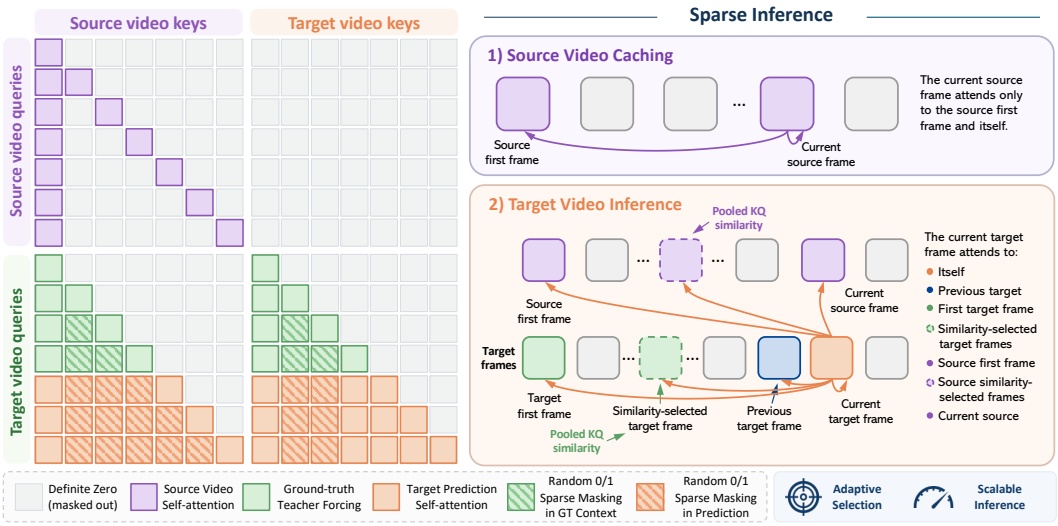

Figure 3: **Sparse context forcing (left) and sparse inference (right).** During ODE initialization, source tokens attend only to the first source frame and themselves, target tokens combine teacher-forced clean and diffusion-forced predicted context, and non-local edges are randomly dropped. At inference, each chunk attends to a constant-size set chosen by pooled query-key similarity.

but depends on external pose and depth models, is view-limited: once the camera leaves the observed region, the re-rendering goes empty and control fails, and needs frame buffering plus extra modules, which is impractical for streaming.

We therefore expose camera motion as *visual evidence*, constructing a synthetic camera space and rendering it along the target trajectory into conditioning frames spatially aligned with the video to be generated. A dense 3D scaffold in camera space resolves depth ambiguity: its projections deform under translation and induce parallax, letting the model infer metric translation and align rendered structure with the conditioning input, while a colored environment map at infinity varies with orientation and disambiguates rotation. Together they turn camera motion into dense, frame-aligned signals a video diffusion model reads directly, and because the scene is synthetic rather than reconstructed, the signal never goes blank when the camera explores unseen regions. A lightweight OpenGL renderer synthesizes these frames at 150 FPS; they are VAE-encoded and concatenated channel-wise with the noisy target latents before the DiT.

3.2 Sparse Full-Fidelity Memory

The camera-conditioned teacher is accurate but not interactive: bidirectional, densely attending, and many-step. We distill it into a few-step autoregressive student following self-forcing (Huang et al., 2025b): the student is initialized from the teacher, trained with ODE initialization to regress pre-computed teacher ODE trajectories at four denoising timesteps, and then optimized with DMD (Yin et al., 2024b) on its own rollouts.

Sparse retrieval over a full-fidelity cache. During rollout the student writes every generated chunk into the KV cache, so generation speed degrades roughly linearly with context (Robbyant Team, 2026); yet consistent revisits require persistent memory, so the cache cannot simply be truncated by a sliding window (Zhu et al., 2026a; Li et al., 2025). Latent compression of distant history (Hong et al., 2025) costs visual quality, and retrieval from field-of-view overlap (Yu et al., 2025; Team et al., 2026a) needs reliable camera geometry and fails when a location is revisited from the opposite direction. Since video diffusion models already learn rich cross-frame correspondences (Xu et al., 2025), we keep the whole KV cache at full fidelity and retrieve chunks by content-aware sparse attention (Figure 3): each chunk stores dense keys and values for attention, while a pooled key summary drives retrieval. Let $\mathcal{I}_c$ index the tokens of chunk c, with keys and values $\mathbf{K}_{\mathcal{I}_c}, \mathbf{V}_{\mathcal{I}_c}$ and pooled key $\bar{\mathbf{K}}_c$, and let $\bar{\mathbf{Q}}_t$ summarize the current query chunk. The active memory set combines the initial chunk, the r most recent, and the top-k middle-history chunks by similarity:

$$\mathcal{A}_t = \{0\} \cup \mathcal{N}_r(t) \cup \mathrm{TopK}_k \left\{ \mathrm{sim}\big(\bar{\mathbf{Q}}_t, \bar{\mathbf{K}}_c\big) \,\middle|\, c \in \mathcal{M}(t) \right\}, \qquad \mathbf{K}_t^{\mathrm{ctx}} = [\mathbf{K}_{\mathcal{I}_c}]_{c \in \mathcal{A}_t}, \tag{1}$$

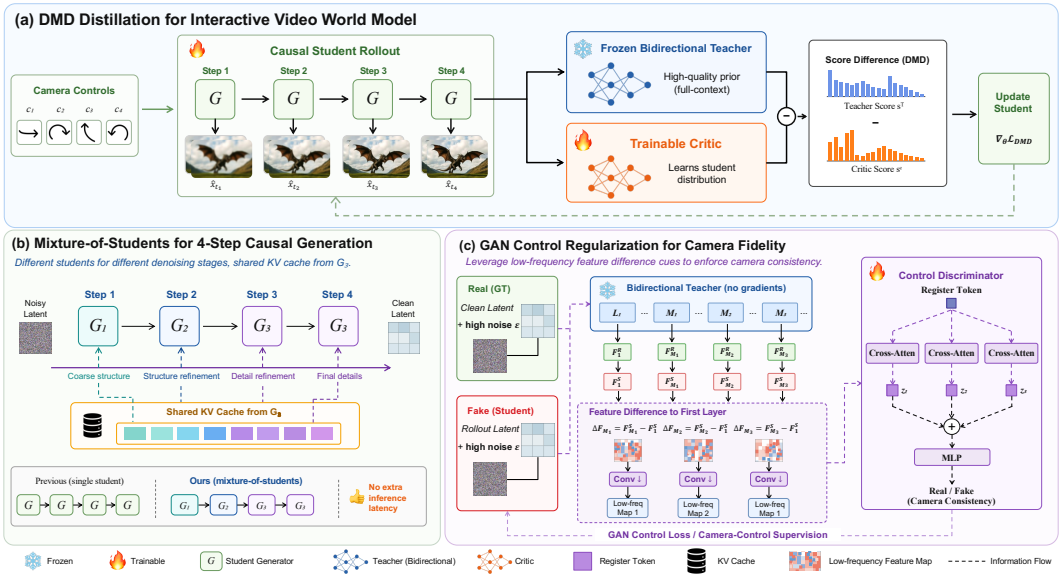

Figure 4: **Distillation with mixture-of-students and GAN control regularization.** (a) The few-step student rolls out chunks under camera control; the rollout is noised and scored by a frozen teacher and trainable critic, whose score difference gives the DMD gradient. (b) Generators specialize by denoising stage, sharing one streaming KV cache. (c) The control discriminator compares low-frequency teacher feature deltas of real and student latents under one camera condition.

with $\mathbf{V}_t^{\text{ctx}}$ formed analogously, $\mathcal{N}_r(t)$ the r most recent chunks, and $\mathcal{M}(t)$ the middle history. The initial chunk acts as an attention sink, and the recent chunks preserve motion continuity; we use $r = 2$. Retrieval runs on pooled summaries, so its overhead is negligible, while selected chunks are attended at full resolution, preserving faithful visual memory without compression. Under a fixed top-k budget, the active context stays constant regardless of rollout length, so long-horizon generation runs at stable latency.

Sparse context forcing. Applying sparse attention directly during distribution-level distillation is unstable, because the student has never learned to generate under incomplete history, so we warm it up with the same memory constraint it will meet at inference. Our ODE initialization combines teacher forcing (Gao et al., 2024; Hu et al., 2024) and diffusion forcing (Chen et al., 2024) under a causal layout: teacher-forced context supplies clean target frames simulating low-noise history, while diffusion-forced context exposes the student to previous noisy predictions, reducing the train–test mismatch (Hong et al., 2025). For V2V, the source video is cached as a reference stream whose frames attend only to themselves and the first source frame. On top of this, we randomly drop eligible historical context edges, always keeping the required ones: self-attention, first-frame anchors, and the most recent context, to maintain identity and local motion continuity, while non-local edges are resampled per layer and forward pass. With $a(i, j)$ the base causal relation between query chunk i and key chunk j, $\mathcal{R}$ the required relations and $\mathcal{O}$ the optional ones, the layer-ℓ mask is

$$M_{ij}^{(\ell)} = \mathbf{1}[a(i,j) \in \mathcal{R}] + b_{ij}^{(\ell)}\,\mathbf{1}[a(i,j) \in \mathcal{O}]\,, \qquad b_{ij}^{(\ell)} \sim \text{Bernoulli}(1 - p_\ell), \tag{2}$$

where p_ℓ controls sparsity. We use a stronger drop rate at longer horizons, pushing the model to rely on a small set of informative memories rather than dense access to the full history.

3.3 Distilling a Real-Time Student

Mixture of students. After ODE initialization, the student recovers coarse structure but not fine detail, so we optimize it with DMD (Yin et al., 2024b): rollouts are noised and passed to the teacher and critic, whose score difference updates the generator (Figure 4a). Self-forcing reduces exposure bias by training on the student's own rollouts, but the underlying problem stays hard: a causal, few-step student must approximate a bidirectional, many-step teacher, and the resulting capacity bottleneck causes mode shrinkage and degraded detail over long rollouts. We therefore adopt a timestep-wise mixture-of-students, analogous to the denoising-stage MoE in Wan2.2 (Wan et al.,

2025): instead of one generator for all steps, our 4-step sampler uses three,

$$\hat{\mathbf{x}}_0 = G_3^{(4)} \circ G_3^{(3)} \circ G_2^{(2)} \circ G_1^{(1)}(\mathbf{x}_{\tau_1}), \tag{3}$$

allocating capacity by role: G_1 forms coarse structure, G_2 refines it, G_3 handles detail and final reconstruction. The generators share the autoregressive interface and reuse G_3's KV cache rather than separate caches, so capacity rises without changing the step count or adding latency.

GAN control regularization. DMD2 (Yin et al., 2024a) attaches a GAN objective to the critic and is used in recent world-model distillation pipelines (Robbyant Team, 2026) to sharpen high-frequency detail, but it is not designed for camera-controlled autoregressive generation. We observe a distinct failure mode during DMD training: the student gradually loses camera fidelity, generated motion accelerates, deviates from the target trajectory, and collapses into blur (an error that compounds in causal rollout). DMD score residuals emphasize high-frequency artifacts but give weak gradients for the low-frequency layout changes camera motion induces, and once the rollout leaves the teacher's training distribution it can no longer be guided. A downsampled ℓ_2 loss against ground truth (Hong et al., 2025) stabilizes training but shrinks the distribution and blurs new regions.

We instead target low-frequency camera consistency adversarially (Figure 4c). We perturb the ground-truth latent $\mathbf{z}_0$ and the student rollout latent $\hat{\mathbf{z}}_0$ at a high noise level into $\mathbf{z}_\tau^r$ and $\mathbf{z}_\tau^f$, and forward both through the *frozen* camera-conditioned teacher under the same condition $\mathbf{c}$, giving a stable camera-aware feature space, empirically more effective than features from the trainable critic, whose representation keeps shifting. Rather than discriminating the final latent, the discriminator operates on teacher feature dynamics: since the camera condition enters at the input, the first-layer feature is a direct camera-aware reference, and comparing intermediate features against it measures whether a latent still matches the prescribed trajectory. For $x \in \{r, f\}$ we extract the first-layer and three middle-layer features $(\mathbf{F}_1^x, \mathbf{F}_{M_1}^x, \mathbf{F}_{M_2}^x, \mathbf{F}_{M_3}^x) = T_{\text{feat}}(\mathbf{z}_\tau^x, \mathbf{c})$, form low-frequency difference maps $\Delta\mathbf{F}_{M_i}^x = \psi_i(\eta_i(\mathbf{F}_{M_i}^x) - \eta_1(\mathbf{F}_1^x))$ with linear extractors η and a strided convolutional downsampler ψ_i that suppresses texture detail, and aggregate them with three learnable register tokens $\mathbf{q}$ into a control logit $d_{\text{ctrl}}^x = \text{MLP}([\sum_i \text{CrossAttn}_i(\mathbf{q}, \Delta\mathbf{F}_{M_i}^x), e_\tau])$, with e_τ the timestep embedding. Using relativistic softplus losses $\mathcal{L}_{\text{ctrl}}^D = \mathbb{E}[\text{softplus}(d_{\text{ctrl}}^f - d_{\text{ctrl}}^r)]$ and $\mathcal{L}_{\text{ctrl}}^G = \mathbb{E}[\text{softplus}(d_{\text{ctrl}}^r - d_{\text{ctrl}}^f)]$, The generator is optimized with $\mathcal{L}_{\text{DMD}} + \lambda_{\text{ctrl}}\mathcal{L}_{\text{ctrl}}^G$. This supplies the low-frequency supervision that DMD2's high-frequency objective lacks, stabilizing distillation and avoiding the mode shrinkage of regression-based alternatives.

4 Experiments

We evaluate Wonder in two settings: *image-to-video* (I2V), where the model receives one image and must hallucinate unseen regions while following user-specified camera controls, and *video-to-video* (V2V), where it receives a video and must preserve subject dynamics while enabling free camera navigation. Long-horizon behaviour is best judged in motion, so the supplementary material provides rendered rollouts and side-by-side video comparisons for every qualitative claim below.

4.1 Experimental Setup

Benchmarks. The I2V benchmark contains 1,000 open-source images spanning real photographs, AI-generated images, cartoons, artistic images, and gaming scenes. A VLM produces a scene caption and navigation keys for each image, which is paired with five trajectories of varying translation scale and rotation speed. The V2V benchmark contains 500 videos of diverse scenes, all with dynamic subjects such as humans, animals, and vehicles, so that dynamic preservation can be assessed alongside camera control. Each video is paired with six trajectories in three groups: *retreat*, moving backward to reveal the full scene; *follow*, tracking the moving subject; and *free*, exploring freely.

Metrics. Following prior work (Robbyant Team, 2026; Zhu et al., 2026a) we report five VBench metrics (Huang et al., 2024): imaging quality, aesthetic quality, dynamic degree, motion smoothness, and temporal flickering. For camera following we estimate relative poses with two estimators, Depth Anything 3 (Lin et al., 2025) and ViPE (Huang et al., 2025a), and average their errors. Since models produce trajectories at different scales and in different coordinate frames, each reconstructed trajectory is aligned to the canonical target with Umeyama alignment (Umeyama, 2002) before computing translational and rotational Relative Pose Error (RPE).

Table 1: **Comparison on visual quality and action accuracy** in the image-to-video (I2V) and video-to-video (V2V) settings. †Avg. is the mean of the five VBench scores. Baseline references and metric description are given in Sec. 4.1.

Model	Visual quality ↑						RPE ↓	
	Avg.†	Imaging	Aesth.	Dyn.	Motion	Flicker	Trans.	Rot.
Image-to-Video (I2V)								
HY-WorldPlay 1.5	0.7900	0.6825	**0.6915**	0.6120	**0.9893**	**0.9745**	0.0195	0.1711
RELIC	0.8225	0.5983	0.6334	0.9243	0.9889	0.9675	0.0208	0.1426
LingBot-World-Fast	0.8261	0.6553	0.6360	0.8991	0.9769	0.9633	0.0174	0.2922
SANA-WM-Streaming	0.8415	0.6480	0.6099	**1.0000**	0.9889	0.9608	0.0240	0.1155
DreamX-World	0.8385	0.6891	0.6655	0.8927	0.9847	0.9604	0.0244	0.2547
Wonder (ours)	**0.8564**	**0.7113**	0.6416	0.9874	0.9794	0.9622	**0.0132**	**0.0784**
Video-to-Video (V2V)								
Inspatio-World	0.8374	0.6756	0.5815	**1.0000**	0.9794	0.9506	0.0436	0.2470
Wonder (ours)	**0.8527**	**0.6981**	**0.6196**	**1.0000**	**0.9855**	**0.9602**	**0.0187**	**0.1119**

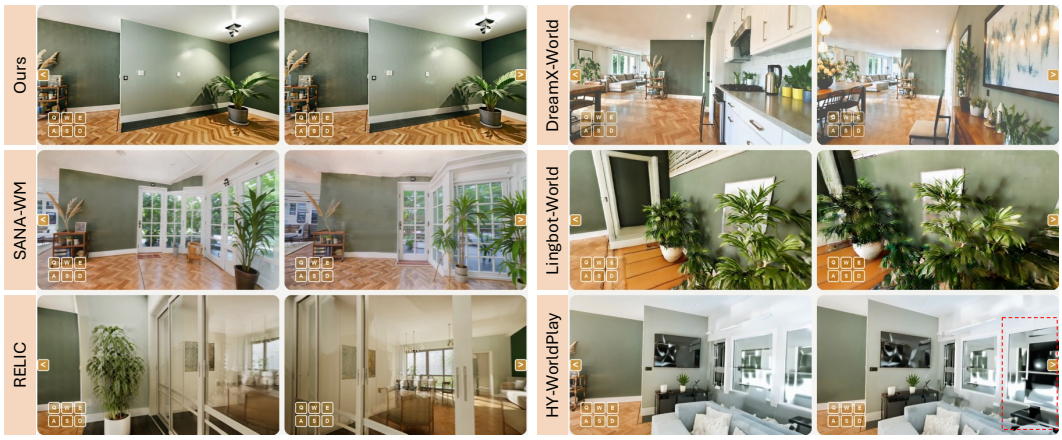

Figure 5: **Qualitative comparison on image-to-video world modeling.** Memory is probed with a revisit trajectory: the camera pans right to observe a target region, pans left away, then pans right again. Comparing the two frames captured at the same pose reveals whether scene content is preserved. Most prior models fail to retrieve long-term memory, and RELIC (Hong et al., 2025) and LingBot-World (Robbyant Team, 2026) additionally degrade in quality under long rollouts.

Baselines. For I2V we compare against five recent interactive world models available at submission time: HY-WorldPlay 1.5 (Team et al., 2025), RELIC (Hong et al., 2025), LingBot-World-Fast (Robbyant Team, 2026), SANA-WM-Streaming (Zhu et al., 2026a), and DreamX-World (Team et al., 2026a). For V2V we omit short-horizon bidirectional re-shooting models (Bai et al., 2025a; Huang et al., 2025c; Wang et al., 2026a), which target neither long-video generation nor streaming, leaving Inspatio-World (Team et al., 2026b) as the only open-source long-horizon V2V baseline. We use the officially released model or recommended inference setting for every baseline.

4.2 Main Results

Image-to-video. The upper block of Table 1 reports the I2V comparison. Wonder attains the best overall visual quality at an average score of 0.8564, driven mainly by imaging quality, where it reaches 0.7113. Some baselines score higher on individual temporal metrics such as motion smoothness or flickering, but do so at the cost of controllability, e.g., slow, smooth output is easy to keep temporally stable. On action accuracy Wonder attains the lowest translational and rotational RPE of all methods, reducing translational error from 0.0174 to 0.0132 and rotational error from 0.1155 to 0.0784 against the strongest baseline on each metric, showing that the coordinate field and the redesigned distillation improve camera following without sacrificing visual quality. Figure 5 probes long-horizon memory along a revisit trajectory: most prior methods fail to reproduce previously



Figure 6: **Qualitative comparison on video-to-video world modeling.** Because of errors in the point clouds reconstructed by external pose estimators, and the limited exploration space of a point-cloud representation, Inspatio-World (Team et al., 2026b) often fails to produce plausible camera motion or to preserve dynamic objects under large viewpoint changes, whereas Wonder keeps both the subject dynamics and the prescribed trajectory.

Table 2: **I2V ablation.** Each block changes one component under a fixed training budget. Throughput is measured at the 60 s rollout mark; *Revisit PSNR* compares matched views across a revisit.

	Variant	Visual quality ↑		Throughput (FPS) ↑	RPE ↓	
		Avg.	Revisit PSNR (dB)		Trans.	Rot.
Control	Implicit pose embedding	0.8481	15.6	–	0.0201	0.1726
	Point-cloud re-rendering	0.8427	18.8	–	0.0187	0.1524
Memory	Sliding-window cache	0.8450	16.3	16	0.0141	0.0842
	Dense full-history attention	0.8515	**22.4**	8.7	0.0133	0.0789
	w/o sparse context forcing	0.8473	19.1	16	0.0146	0.0908
Distill.	w/o mixture-of-students	0.8442	20.5	–	0.0145	0.0881
	w/o GAN control reg.	0.8266	13.2	–	0.0214	0.1905
	Wonder (full)	**0.8564**	21.9	16	**0.0132**	**0.0784**

observed regions on return, and others degrade visibly as the rollout lengthens, whereas Wonder preserves structure and appearance. Appendix D gives further results across visual domains.

Video-to-video. The lower block of Table 1 reports the V2V comparison. Wonder improves the average visual-quality score from 0.8374 to 0.8527, with consistent gains in imaging quality, aesthetic quality, motion smoothness, and temporal flickering, indicating better preservation of the input video's dynamics under new viewpoints. The gap in camera following is larger still: translational RPE drops from 0.0436 to 0.0187 and rotational RPE from 0.2470 to 0.1119. This reflects the advantage of a synthetic coordinate field over point-cloud re-rendering, whose signal vanishes exactly when the camera moves beyond the reconstructed view coverage. Unlike Wonder, Inspatio-World is not a streaming method, since its re-rendering pipeline requires the entire input video before generation begins. Figure 6 shows the corresponding qualitative behavior: the point-cloud baseline distorts dynamic subjects and loses the prescribed trajectory under large viewpoint change, whereas Wonder keeps both. Further V2V rollouts are given in Appendix D.

4.3 Ablation Study

We ablate the three contributions of Sec. 3 on the I2V benchmark, isolating control, memory, and distillation without the confound of source-video conditioning. Every variant changes exactly one component under the same schedule, step budget, and data mixture. The full model attains the best average visual quality and the lowest translational and rotational RPE at 16 FPS. Table 2 reports the quantitative effects, and Figure 7 visualizes the corresponding failure modes.



Figure 7: **Qualitative ablations under two revisit trajectories.** Left/right pairs use pan right–left–right and backward–forward–backward motion, with inputs inset and outputs at matched return poses. Implicit and point-cloud controls respectively miss the return pose and extrapolate poorly; sliding-window memory and sparse retrieval without context forcing lose history. Without GAN control regularization, drift degrades the scene; a single generator loses detail and structural stability.

Control. Replacing the coordinate field with an implicit pose embedding removes the frame-aligned cue and forces the distilled model to infer image motion from a global code. Average visual quality falls from 0.8564 to 0.8481, while translational/rotational RPE rises from $0.0132/0.0784$ to $0.0201/0.1726$. Point-cloud re-rendering reduces camera control error to $0.0187/0.1524$, but still trails the full model because its control signal is limited to reconstructed views.

Memory. The sliding-window cache maintains 16 FPS but discards distant history, reducing revisit PSNR from 21.9 to 16.3 dB. Dense full-history attention reaches 22.4 dB but slows generation to 8.7 FPS. Sparse retrieval therefore retains nearly the same revisit fidelity at $1.84\times$ the throughput. This requires matched training: without sparse context forcing, PSNR falls to 19.1 dB, indicating history loss due to mismatching of training and inference.

Distillation strategy. Using a single generator across all four denoising steps reduces the student's effective capacity, lowering visual quality to 0.8442. Removing GAN control regularization destabilizes the distillation process and has the largest effect: 0.8266 visual quality, 13.2 dB PSNR, and $0.0214/0.1905$ RPE. These results show that mixture-of-students preserves detail, while low-frequency control supervision prevents camera drift from compounding over long rollouts.

5 Conclusion

We presented Wonder, a real-time camera-controllable video world model that turns a single image into an explorable world and re-renders an observed video along user-specified trajectories while preserving appearance and dynamics. Wonder addresses controllability, persistence, and efficiency jointly: a dense pixel-space coordinate field supplies spatially aligned camera cues, a sparse full-fidelity memory mechanism supports coherent revisits at bounded attention cost, and an improved distillation pipeline carries visual quality and camera fidelity into a few-step causal student. Across both benchmarks, Wonder achieves stronger visual quality and camera-following accuracy than recent streaming world models, enabling minute-scale generation at 16 FPS with stable latency.

Limitations. Wonder inherits the visual priors of its pretrained backbone, and its control is metric only up to the scale implied by the synthetic scaffold. The interface exposes camera motion alone, leaving object-level interaction to future work, and although retrieval keeps per-step latency constant, the full-fidelity KV cache still grows in memory with rollout length, bounding session duration on a fixed GPU budget.

AI USE STATEMENT

We used generative AI tools as general-purpose writing and coding assistants. Specifically, large language models were used to polish the wording of the manuscript, and to assist with routine engineering tasks such as refactoring data-processing and evaluation scripts. A vision-language model is additionally used as a component of our data pipeline, where it produces scene captions, navigation keys, and quality judgments for training clips and benchmark items (Sec. 4.1, Appendix A); this is a methodological use of an existing model rather than a use of AI in the research. We did not use generative AI to originate the research ideas, to design the method, or to produce, analyze, or report experimental results. All AI-assisted text and code were reviewed and verified by the authors, and all reported numbers come from our own experiments. The authors take full responsibility for the content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies controllable video generation and does not involve human subjects, private data, or personally identifying information. Training data are drawn from publicly available video datasets and from synthetic environments we render ourselves in Unreal Engine and Blender; benchmark images are sourced from publicly available material. We inherit the biases present in the pretrained video backbone and in the public data sources, and we make no claim that the model's synthesized content is a faithful reconstruction of any real place. As with other high-quality generative video models, a real-time, camera-controllable world model could be misused to fabricate convincing footage of scenes that were never captured. We view provenance signaling, such as watermarking and content credentials, as a necessary companion to any release, and we will accompany a release with the corresponding usage restrictions.

REPRODUCIBILITY STATEMENT

The method is specified in Sec. 3, with the control representation in Sec. 3.1, the memory mechanism and its training strategy in Sec. 3.2, and the distillation objectives in Sec. 3.3. Appendix A documents the full data curation pipeline, including sources, clipping, filtering, captioning, pose estimation, and augmentation. Appendix B gives the backbone configuration, the multi-task multi-horizon training schedule, the training infrastructure, and the inference-time optimizations needed to reproduce our latency measurements. The evaluation protocol, including benchmark construction, metrics, pose estimators, and trajectory alignment, is described in Sec. 4.1. The supplementary material contains rendered rollouts and side-by-side video comparisons for every qualitative claim in the paper.

REFERENCES

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14834–14844, 2025a.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong,

Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models, 2025. URL https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Hyung Won Chung, Noah Constant, Xavier García, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *ArXiv*, abs/2304.09151, 2023. URL https://api.semanticscholar.org/CorpusID:258187051.

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.

Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing. *arXiv preprint arXiv:2411.16375*, 2024.

Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025b.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, et al. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Jinyi Hu, Shengding Hu, Yuxuan Song, Yufei Huang, Mingxuan Wang, Hao Zhou, Zhiyuan Liu, Wei-Ying Ma, and Maosong Sun. Acdit: Interpolating autoregressive conditional modeling and diffusion transformer. *arXiv preprint arXiv:2412.07720*, 2024.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception. In *NVIDIA Research Whitepapers*, 2025a.

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*, 2025b.

Zhening Huang, Hyeonho Jeong, Xuelin Chen, Yulia Gryaditskaya, Tuanfeng Y Wang, Joan Lasenby, and Chun-Hao Huang. Spacetimepilot: Generative rendering of dynamic scenes across space and time. *arXiv preprint arXiv:2512.25075*, 2025c.

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.

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

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2025.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

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.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

Fatema Rahimi, Abolghasem Sadeghi-Niaraki, and Soo-Mi Choi. Generative ai meets virtual reality: A comprehensive survey on applications, challenges, and future direction. *IEEE Access*, 2025.

Robbyant Team. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *ArXiv*, abs/2407.08608, 2024. URL https://api.semanticscholar.org/CorpusID:271098045.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Schechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. *arXiv preprint arXiv:2511.01266*, 2025.

Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, Linfeng Zhang, et al. Hunyuan-gamecraft-2: Instruction-following interactive game world model. *arXiv preprint arXiv:2511.23429*, 2025.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026a.

HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, et al. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025.

InSpatio Team, Donghui Shen, Guofeng Zhang, Haomin Liu, Haoyu Ji, Hujun Bao, Hongjia Zhai, Jialin Liu, Jing Guo, Nan Wang, et al. Inspatio-world: A real-time 4d world simulator via spatiotemporal autoregressive modeling. *arXiv preprint arXiv:2604.07209*, 2026b.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4):376–380, 2002.

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

Yiming Wang, Qihang Zhang, Shengqu Cai, Tong Wu, Jan Ackermann, Zhengfei Kuang, Yang Zheng, Frano Rajič, Siyu Tang, and Gordon Wetzstein. Bullettime: Decoupled control of time and camera pose for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18319–18330, 2026a.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026b.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Junjin Xiao, Yandan Yang, Xinyuan Chang, Ronghan Chen, Feng Xiong, Mu Xu, Wei-Shi Zheng, and Qing Zhang. World-env: Leveraging world model as a virtual environment for vla post-training. *arXiv preprint arXiv:2509.24948*, 2025.

Jiacong Xu, Yiqun Mei, Ke Zhang, and Vishal M Patel. Freevis: Training-free video stylization with inconsistent references. *arXiv preprint arXiv:2510.01686*, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo 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.

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 Computer Vision and Pattern Recognition Conference*, pp. 22963–22974, 2025.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. *arXiv preprint arXiv:2506.03141*, 2025.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. Sana-wm: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026a.

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*, 2026b.

A TRAINING DATA

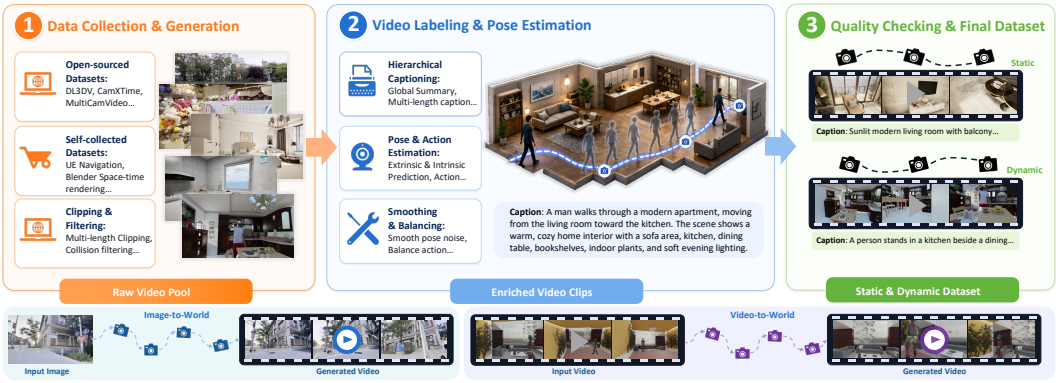

Figure 8: **Data curation pipeline.** We build a unified data engine for image-to-world (I2W) and video-to-world (V2W) training. Videos come from public datasets and self-rendered UE/Blender environments, then pass through multi-duration clipping, collision filtering, and rule-based quality screening to form a raw video pool. Each clip is enriched with hierarchical VLM captions and camera poses estimated by Depth Anything 3 (DA3), from which discrete camera actions are derived. Gaussian pose smoothing and temporal augmentations such as reverse playback and speed resampling improve trajectory quality and action coverage, and a final VLM-based quality check removes degraded clips and unnatural motion, yielding the curated datasets.

Training a controllable world model requires large-scale video with diverse visual content, accurate camera motion, and long temporal horizons. Compared with short video generation, long-horizon world modeling additionally requires the model to maintain memory, preserving scene geometry, appearance, and dynamics over extended rollouts. We therefore construct the unified curation pipeline shown in Figure 8 to collect, annotate, and filter video clips.

I2V sources. We use DL3DV (Ling et al., 2024) as the main source of real-world navigation video. It provides diverse scenes with naturally captured trajectories, useful for learning realistic geometry-aware motion. However, camera movement in such footage is typically simple and smooth, whereas interactive users perform sharp turns, lateral movement, backward motion, and compound actions. To cover these harder cases, we also build a large-scale synthetic I2V dataset with Unreal Engine, rendering diverse indoor and outdoor environments along user-control-like trajectories, which supplies richer supervision for robust camera control.

V2V sources. V2V is more demanding because the model needs paired videos that share the same underlying dynamic scene but differ in camera trajectory or temporal evolution. Open-source datasets such as MultiCamVideo (Bai et al., 2025a) and CamXTime (Huang et al., 2025c) provide useful pairs, but their clips are typically around five seconds, too short for long-horizon world modeling. We therefore render a large amount of long paired V2V data with Blender, including standard paired trajectories, speed-varied sequences, and bullet-time videos in which camera motion and scene time are partially decoupled. These pairs teach the model to distinguish spatial viewpoint change from temporal scene dynamics, which is essential for controllable V2V generation.

Annotation and filtering. Videos are first clipped into 5 s, 10 s, and 20 s durations to support progressive training from short to long horizons. Rule-based filters remove corrupted videos, black frames, failed renderings, and trajectories with collisions or clearly invalid motion. An advanced VLM (Bai et al., 2025b) then produces hierarchical captions (Robbyant Team, 2026) for each long clip, consisting of a global video-level summary plus fine-grained captions for its temporal sub-clips. For metric camera supervision we use Depth Anything 3 (Lin et al., 2025) to estimate intrinsics and extrinsics for both real and synthetic video, and convert the recovered trajectories into discrete camera actions. Because estimated poses contain high-frequency jitter, we apply Gaussian smoothing before action discretization, suppressing noisy fluctuations while preserving the overall path. Temporal augmentations such as reverse playback and speed resampling balance the action distribution. Finally, a VLM-based quality check removes clips with low visual quality, unnatural camera motion, or inconsistent temporal dynamics, producing the curated datasets used for training.

B Implementation Details

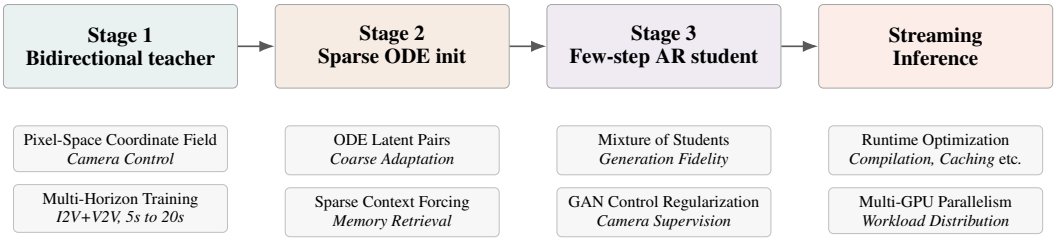

Figure 9: **Training and deployment pipeline.** Wonder first learns high-quality camera control in a bidirectional teacher, converts the model into a sparse-context causal student, distills it into a few-step autoregressive generator, and finally deploys it with a resident streaming runtime.

B.1 Base Architecture

We build on Wan2.1-I2V-14B (Wan et al., 2025), a bidirectional video diffusion model developed for short-horizon image-to-video generation. Wan comprises a spatio-temporal VAE and a diffusion transformer. The VAE maps videos into compact latent sequences with a 3D causal encoder-decoder, downsampling spatial resolution by $8\times$ and temporal resolution by $4\times$; its causal design and feature-caching mechanism give a natural interface for streaming generation. The diffusion backbone operates on patchified video latents and maps tokens back through corresponding unpatchifying layers. Text prompts are encoded by umT5 (Chung et al., 2023) and injected via cross-attention, while diffusion timesteps are embedded by a shared MLP modulating the transformer blocks. This gives a strong bidirectional teacher initialization, which we adapt for camera control.

Unified I2V/V2V input layout. We represent the input as an optional source segment followed by a target segment. Let $\mathbf{z}^s$ denote the source-video latents and let the target latents decompose as $\mathbf{z}^t = [\mathbf{z}^a, \mathbf{z}^p]_T$, where $\mathbf{z}^a$ is the clean anchor frame and $\mathbf{z}^p$ the frames to be generated. Noise is applied only to $\mathbf{z}^p$; a binary mask marks source (0) versus target (1) tokens; and camera-condition latents $\mathbf{z}^c$ are zero-padded over the source segment. The input at diffusion timestep τ is

$$\mathbf{u}_\tau = \left[[\mathbf{z}^s, \mathbf{z}^a, \alpha_\tau \mathbf{z}^p + \sigma_\tau \boldsymbol{\epsilon}]_T \, , [\mathbf{0}^s_m, \mathbf{1}^a_m, \mathbf{1}^p_m]_T \, , [\mathbf{0}^s_c, \mathbf{z}^{c,a}, \mathbf{z}^{c,p}]_T \right]_C , \tag{4}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $[\cdot]_T$, $[\cdot]_C$ denoting temporal and channel concatenation. For I2V there is no source segment and the layout degenerates to the target-only case, with the first frame supplied as a clean, noise-free latent. We mirror the RoPE coordinates of the target onto the source and add separate learnable positional embeddings so the model can distinguish the two streams.

B.2 Multi-Task Multi-Horizon Training

We train a single model supporting both I2V and V2V generation while rolling out to long horizons. Following the progressive-horizon strategy of Hong et al. (2025), we extend the training horizon gradually rather than training directly on the longest sequences. We first post-train the base model on 5-second clips with I2V and V2V jointly enabled, then extend to 10 seconds and finally to 20 seconds, extrapolating RoPE positions with YaRN (Peng et al., 2023) for the longer contexts.

A naive multi-task extension to long horizons is inefficient, because V2V samples contain both source and target videos and therefore cost substantially more memory and compute. We adopt a mixed-length schedule in the long-horizon stages: for most iterations V2V samples stay at the 5-second horizon while I2V samples are trained at the current long horizon (10 or 20 seconds). This follows from the observation that long-horizon camera following can be learned effectively from I2V data, whereas V2V mainly teaches source-target correspondence and can be learned from shorter paired clips. The remaining iterations still expose the model to full-horizon multi-task samples to preserve compatibility across tasks. This maximizes training throughput while still generalizing to long-horizon V2V dynamic world exploration.

B.3 Training Infrastructure

Training a bidirectional teacher together with three 14B student generators over a 20-second horizon is highly memory-intensive. Following prior work (Hong et al., 2025; Robbyant Team, 2026), we combine (1) Fully Sharded Data Parallel 2 (FSDP2) (Zhao et al., 2023) to shard parameters, gradients, and optimizer states across GPUs; (2) sequence parallelism to distribute the video-token sequence; (3) tensor parallelism to partition self-attention heads and associated computation; (4) gradient checkpointing; and (5) gradient accumulation. During distribution-level distillation we further adopt replayed back-propagation (Hong et al., 2025) to reduce the memory needed to differentiate through autoregressive rollouts. Together these allow training with a global batch size of 64 on 32 NVIDIA H200 GPUs. At inference our mixture-of-students sampler uses three generators over four denoising steps (G_1 at step 1, G_2 at step 2, G_3 at steps 3 and 4), with a chunk size of two latent frames to keep streaming latency low.

B.4 Inference-Time Optimization

Inference latency stems from GPU memory traffic, kernel-launch overhead, and repeated transformer computation, so we target all three. For the repeated prediction and cache-update steps we compile the model operations once, reuse the generated GPU kernels across runs, and replay the same CUDA execution pattern to cut launch overhead. We use self-attention KV caches to preserve previously computed video context and text cross-attention caches to eliminate repeated prompt key/value projections. For memory retrieval we keep a separate cache of compressed key summaries. We further employ FlashAttention-3 (Shah et al., 2024) on Hopper GPUs, fused QKV projections, BF16 RMSNorm, cached RoPE embeddings, and cached timestep modulations, and, guided by profiling, restructure several PyTorch operations to remove residual overhead.

At deployment we use multi-GPU parallelism to distribute computation and memory. FFN layers and cross-attention modules are split along the sequence dimension with sequence parallelism, while self-attention is partitioned across attention heads with tensor parallelism; NCCL `All-to-All` collectives convert between the two layouts. This also partitions the KV cache along the head dimension, so each GPU stores and processes only its assigned heads. Auxiliary workloads such as text encoding and VAE processing are offloaded from the primary generation GPUs, and a Tiny VAE adapted from Shin et al. (2025) decodes conditioning inputs and generated outputs efficiently.

To support minute-scale generation beyond the training horizon we adopt a rolling-window caching scheme (Xiao et al., 2024): the KV cache is organized as `[sink chunks, top-`k` chunks, recent context chunks]`, where the top-k chunks are selected dynamically by sparse attention at each prediction step. To avoid unseen positional embeddings along the frame axis, frame indices are remapped back into the training horizon according to their relative distance. Together these designs bound both the active context size and the positional range, enabling stable minute-scale rollouts without growth in per-step latency.

C Extended Related Work

Video diffusion models. Modern text- and image-to-video generators are built upon diffusion transformers operating in the latent space of a 3D VAE. Open-source DiT/MMDiT-style backbones such as CogVideoX (Yang et al., 2025), HunyuanVideo (Kong et al., 2024), Wan (Wan et al., 2025), and LTX-Video (HaCohen et al., 2026) synthesize multi-second, high-fidelity clips and have become the de facto pretrained foundations for downstream video generation. They are nevertheless bidirectional and denoise a fixed-length chunk in a single shot, which makes them poorly suited to interactive control and unbounded rollout.

Autoregressive video generation. A second line of work turns bidirectional video diffusion models into causal, autoregressive generators. Diffusion Forcing (Chen et al., 2024) unifies next-token prediction with diffusion by assigning independent noise levels to individual tokens. CausVid (Yin et al., 2025) distills a bidirectional teacher into a few-step causal student through distribution matching distillation (Yin et al., 2024b). Self-Forcing (Huang et al., 2025b) narrows the train–test gap by rolling the student out on its own predictions during training, while Rolling Forcing (Liu et al., 2025) and Causal Forcing (Zhu et al., 2026b) extend the paradigm to long-horizon, real-time streaming.

Together these establish the low-latency, KV-cache-enabled autoregressive foundation that makes interactive video world models practical.

Camera-controllable video generation. Pose-conditioned video diffusion is the standard approach for trajectory-controlled synthesis. MotionCtrl (Wang et al., 2024) injects 6-DoF camera extrinsics, CameraCtrl (He et al., 2025a) introduces Plücker embeddings as a plug-in adapter, and ViewCrafter (Yu et al., 2024) conditions generation on point-cloud renderings to improve geometric consistency. ReCamMaster (Bai et al., 2025a) formulates camera control as video-to-video re-rendering along novel trajectories, while subsequent work (Wang et al., 2026a; Huang et al., 2025c) further disentangles camera motion from scene dynamics and spatiotemporal correlations. These conditioning designs directly inspired recent video world models.

Interactive video world models. Closest to our work are neural simulators that respond to user actions. Genie 3 (Ball et al., 2025) generates general-purpose, text-promptable interactive worlds and represents the current closed-source frontier. In parallel, a rapidly growing open-source line, including Matrix-Game and its real-time and streaming successors (Zhang et al., 2025; He et al., 2025b; Wang et al., 2026b), MineWorld (Guo et al., 2025), Hunyuan-GameCraft (Li et al., 2025) and its instruction-following extension Hunyuan-GameCraft-2 (Tang et al., 2025), Yume (Mao et al., 2025), and RELIC (Hong et al., 2025), has explored real-time streaming, long-horizon memory, and richer action interfaces. More recently, LingBot-World (Robbyant Team, 2026) and DreamX-World (Team et al., 2026a) achieve real-time streaming with minute-level context across photorealistic, game-style, and cartoon domains. Closest to our setting, Inspatio-World (Team et al., 2026b) provides both I2V and V2V baselines for real-time dynamic world modeling and re-rendering. These works advance video world modeling from different perspectives, but still face persistent camera drift, memory loss, and degraded visual quality during long-horizon rollout. Wonder instead adopts a system-level co-design across control representation, memory mechanism, and distillation strategy, alleviating these issues and enabling more stable interactive world exploration.

D Additional Qualitative Results

Static frames can only partially convey long-horizon behaviour. The supplementary material accompanying this submission contains an anonymized demo video and side-by-side video comparisons for both settings, including the revisit trajectories used to probe memory in Figure 5 and the dynamic-preservation cases of Figure 6, together with the additional rollouts shown below.

As Figures 10, 11, and 12 show, across diverse inputs Wonder understands and preserves the underlying scene structure while synthesizing plausible unseen regions under user-controlled camera motion. Generated views follow the prescribed instructions accurately, and newly discovered areas remain semantically consistent with the input while retaining high visual quality and diverse content. When the camera revisits previously observed regions after long-horizon exploration, Wonder preserves scene detail and appearance consistency, demonstrating the effect of the sparse memory mechanism. Benefiting from multi-task training, Wonder also exhibits dynamic world modeling in the I2V setting, producing object and subject motion consistent with the scene.

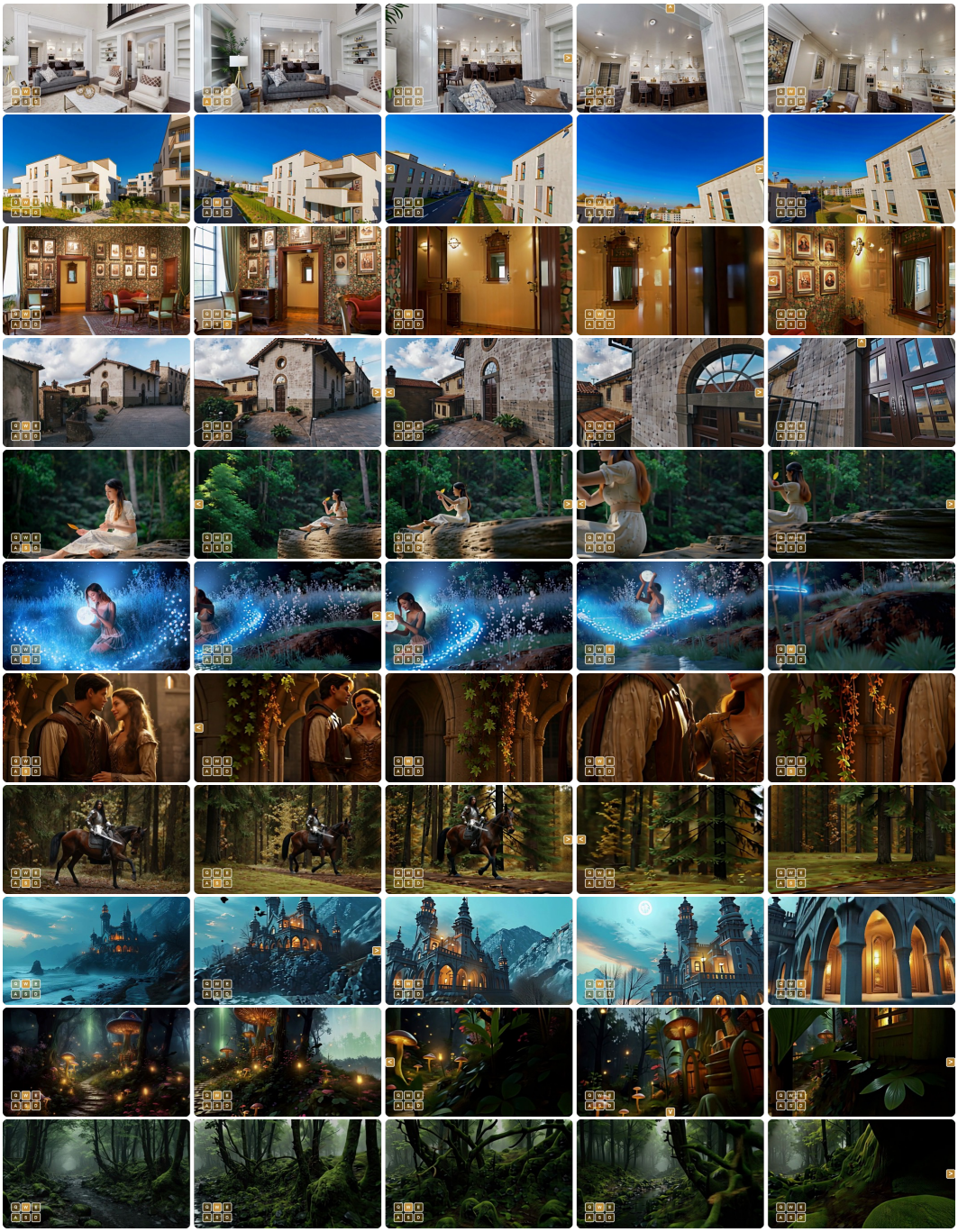

Figure 10: **Navigating into diverse images (I).** Given a single input image, Wonder follows the prescribed camera motion while synthesizing plausible content beyond the observed view. Newly discovered regions stay semantically consistent with the input, and previously observed regions are preserved when the camera returns.

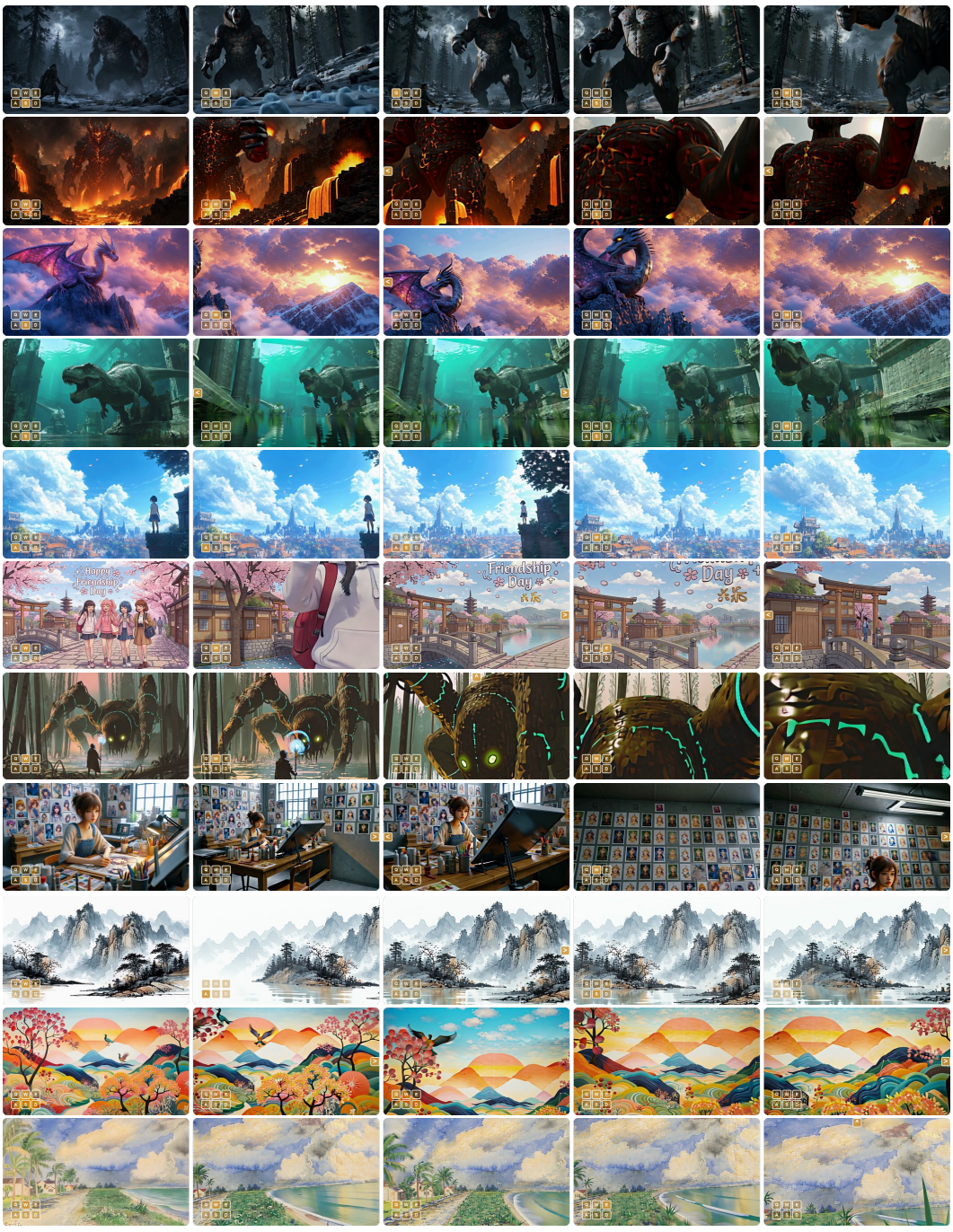

Figure 11: **Navigating into diverse images (II).** Results span photographs, AI-generated images, cartoons, and gaming scenes, showing that control and memory transfer across domains.

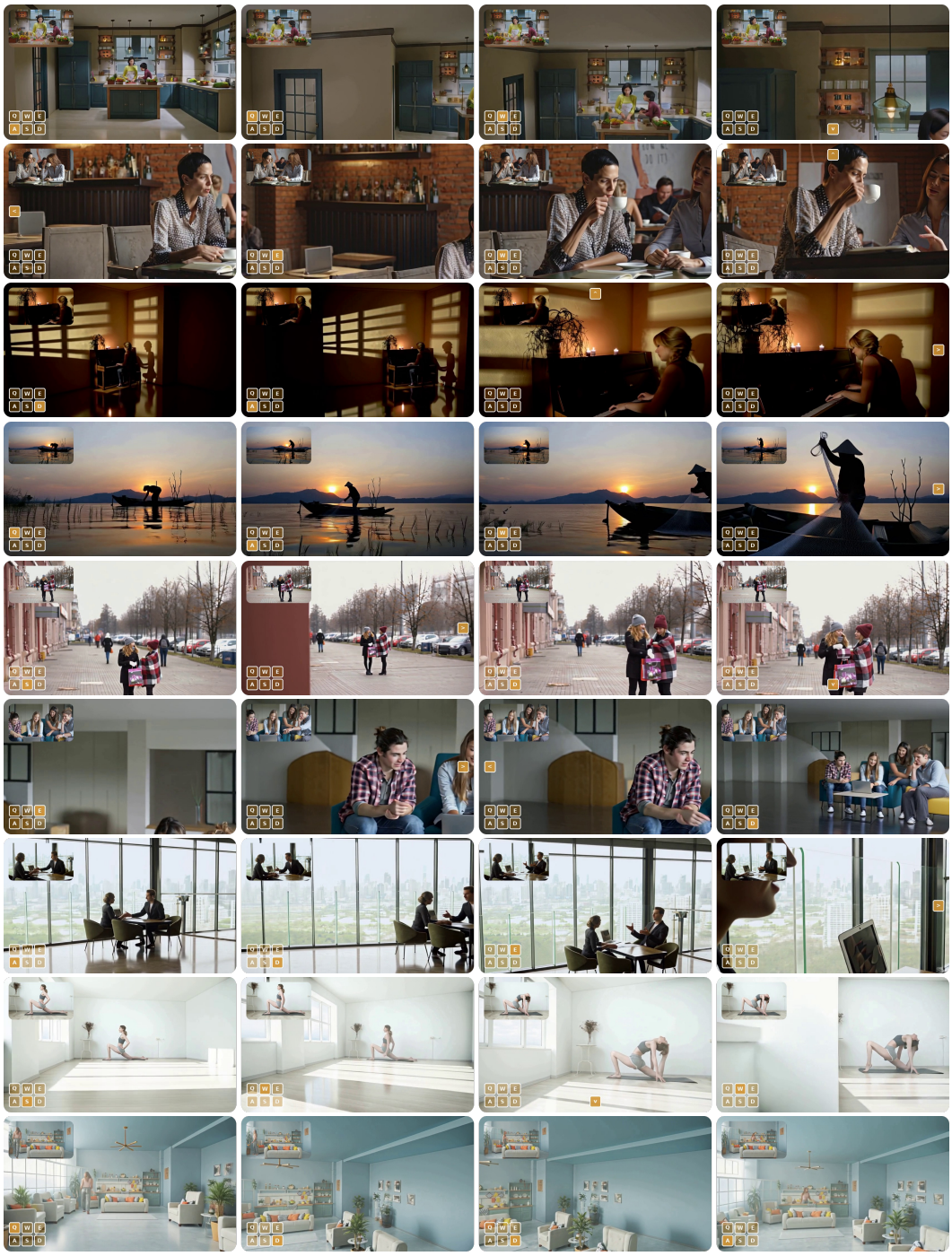

Figure 12: **Navigating into videos.** Given a source video, Wonder re-shoots the same event along a user-specified trajectory, preserving subject dynamics and appearance as the viewpoint changes.