

Wonder: Video World Model Done Better

Jiacong Xu^{*,1,2}, Hanwen Jiang¹, Zhixin Shu¹, Kalyan Sunkavalli¹,
Vishal M. Patel², Yiqun Mei^{*,†,1}

¹Adobe Research, ²Johns Hopkins University

*Equal Contribution, †Project Lead.

We present **Wonder**, a general-purpose video world model for real-time, camera-controllable world exploration. Given an image or a conditional video, **Wonder** constructs a playable world where users can navigate interactively by moving the camera, discovering unseen regions, and revisiting previously observed areas in real time and over a long-term horizon. Achieving this capability requires a system-level co-design of control method, memory mechanism, and training strategy. We introduce a novel camera conditioning with a dense coordinate field whose renderings provide spatially aligned motion and orientation cues, allowing the model to interpret camera motion directly as visual evidence. To support fast and precise memory retrieval over a growing generation context, we propose an efficient sparse attention-based memory mechanism, enabling the model to selectively attend to a small set of relevant context tokens at inference time, regardless of actual context length. We further develop several techniques to rectify the self-forcing-style distillation pipeline, improving the student model’s ability to respect control signals, as well as maintaining diverse generation modes and long-term memory from the teacher. Together, these components enable **Wonder** to synthesize diverse, minute-scale videos at 16 FPS while preserving coherent geometry, appearance, and dynamics across long rollouts. Beyond image-to-video generation, **Wonder** naturally supports video-conditioned generation, allowing existing dynamic scenes to be re-shot in real time.

Date: June 17th, 2026

Project Page: <https://wonder-world-model.github.io/>



1 Introduction

Video generation is rapidly evolving from fixed-duration clip synthesis toward interactive visual world modeling, with potential 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 construct a persistent visual environment from an initial condition, such as an image, a video, or a text prompt, and allow users or agents to explore it through control signals. This shift imposes requirements beyond visual fidelity alone. Interactive applications require accurate control following and low-latency response; simulation settings require long-horizon coherence; editing workflows require the model to preserve source content under control. Achieving these properties simultaneously remains challenging, because controllability, memory, and real-time efficiency often place conflicting demands on the model architecture and training pipeline.

Most existing methods (Wang et al., 2026b; Li et al., 2025; Team et al., 2026a) address only a subset of these requirements. Camera-controllable video generators provide accurate trajectory conditioning, but they are typically bidirectional, fixed-length, and computationally expensive, making them unsuitable for low-latency interaction or minute-scale rollout (Bai et al., 2025a; Yu et al., 2025). Autoregressive video diffusion and distillation methods improve streaming efficiency by converting bidirectional models into causal few-step generators, but they often focus on video continuation and do not explicitly address precise camera following, persistent spatial memory, or consistent revisits (Huang et al., 2025c; Cui et al., 2025; Zhu et al., 2026b). Recent interactive world-model systems attempt to combine control, memory, and real-time generation, yet each component introduces new trade-offs: abstract pose or ray conditions can 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 increases latency with rollout length (Robbyant Team, 2026), and compressed or sliding-window memory loses visual details (Hong et al., 2025). As a result, existing systems still struggle to achieve accurate control, faithful long-term memory, and stable real-time inference simultaneously.

In this report, we present **Wonder**, a general-purpose video world model for real-time, camera-controllable world

exploration. **Wonder** follows the common teacher-to-student paradigm: adapting a bidirectional video diffusion model for camera control, converting it into a causal autoregressive generator, and distilling it into a few-step streaming model. Unlike prior systems that optimize these stages largely in isolation, **Wonder** treats the pipeline as a coupled system. We jointly redesign the control representation, memory mechanism, and distillation strategy to preserve controllability, persistence, and efficiency from the teacher through the real-time student. As a result, **Wonder** supports both image- and video-conditioned exploration with accurate camera following, coherent long-horizon memory, open-ended generation, and stable real-time latency. We specify our designs as follows:

1. **Control.** We introduce a novel camera-control representation that exposes camera motion as rendered visual evidence rather than only as camera geometry. While pose (Hong et al., 2025) or Plücker-ray (Wang et al., 2026a) provide per-image or pixel ray geometry, they still require the diffusion model to infer how those coordinates should translate into image motion. We instead render a synthetic 3D scaffold and environment map along the target camera trajectory, directly turning translation, rotation, and parallax into frame-aligned visual cues. Because the rendered scene is synthetic rather than reconstructed from the input, the control remains informative even beyond the observed view, enabling accurate and streamable camera control for open-ended image- and video-conditioned exploration. Moreover, because the rendered cues have explicit visual semantics, they are better preserved through few-step autoregressive distillation than abstract embeddings, reducing control drift in the real-time student.
2. **Memory.** We develop a sparse full-fidelity memory mechanism for efficient long-horizon generation. Interactive world models require recent context for local motion continuity and distant context for persistent scene memory, but attending densely (Robbyant Team, 2026) to all historical KV states makes per-step latency grow with rollout length. Existing systems reduce this cost by compressing history (Hong et al., 2025), reconstructing context (Huang et al., 2025b), using sliding windows (He et al., 2025b), or relying on geometry-based retrieval (Team et al., 2026a), which trade off memory fidelity, revisit consistency, or retrieval robustness. Instead, **Wonder** decouples memory storage from attention computation: it retains the growing historical KV cache in full fidelity, but uses lightweight pooled query-key summaries to select a constant-size set of relevant memory chunks at each rollout step. Only the selected chunks are attended with full-resolution keys and values, preserving faithful visual history for revisits while keeping active attention cost and per-step latency stable as the generated sequence grows. Furthermore, we introduce a co-designed training strategy, *i.e.* sparse context forcing, that stably activates this sparse-memory capability in the student model, where the causal student can smoothly learn to generate content under the same retrieved-memory condition it will face at inference.
3. **Distillation.** We redesign few-step autoregressive distillation around the goal of preserving teacher capabilities in the real-time student. The causal student must approximate a bidirectional many-step teacher with far fewer denoising steps, which leads to both reduced generation diversity and accumulated camera drift. We address these two losses with a timestep-wise mixture-of-students, which increases effective student capacity without changing the streaming interface, and a camera-aware adversarial regularizer, which provides low-frequency control supervision during distillation. Together, they allow the few-step student to retain visual quality, diversity, and camera-following accuracy during long autoregressive rollouts.

Experiments in both image- and video-conditioned settings show that **Wonder** improves visual quality and camera-following accuracy over recent streaming world models. The model supports exploration beyond observed views, preserves dynamic content when re-rendering source videos, and maintains coherent revisits over long trajectories. It generates minute-scale rollouts at 16 FPS with stable latency as the history grows.

2 Related Work

Video Diffusion Models. Modern text- and image-to-video generators are commonly 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), can synthesize multi-second, high-fidelity video clips and have become the de facto pretrained foundations for downstream video generation tasks. However, these models are bidirectional and denoise a fixed-length chunk in a single shot, making them poorly suited for interactive control and unbounded rollout.

Autoregressive Video Generation. A second line of work transforms bidirectional video diffusion models into causal, autoregressive generators. Diffusion Forcing (Chen et al., 2024) unifies next-token prediction with diffusion by assigning

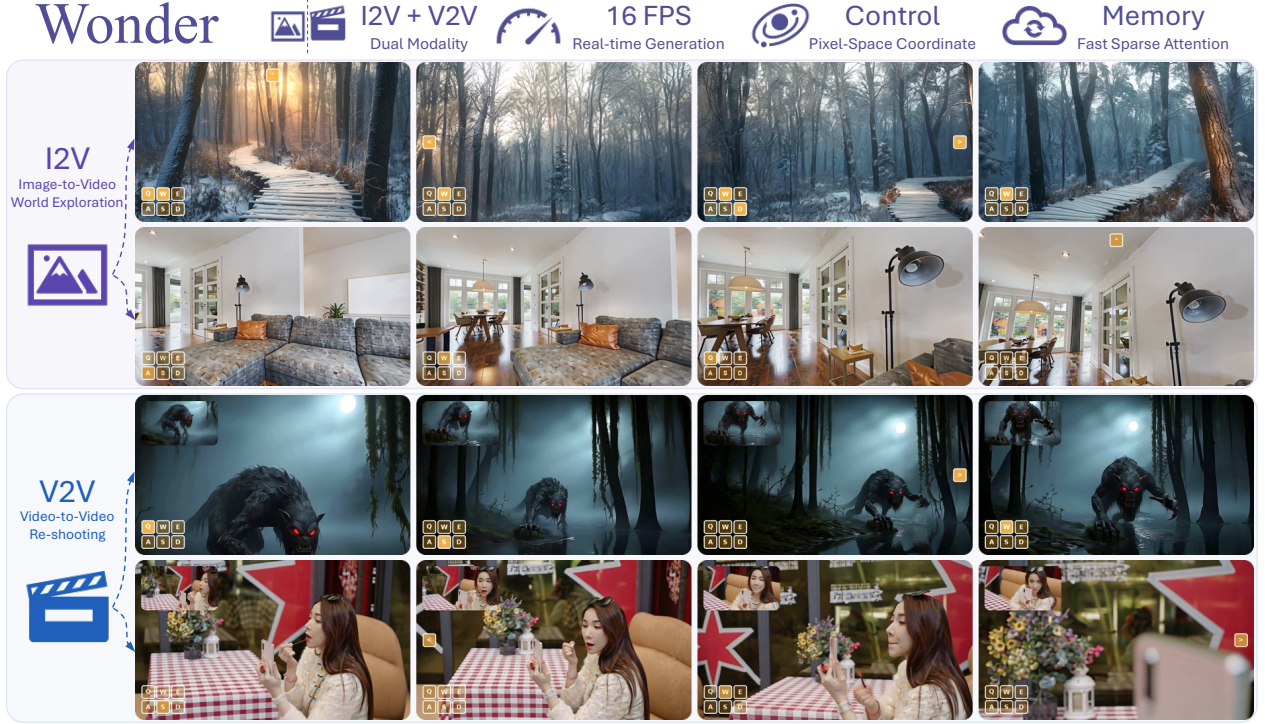


Figure 1 **Wonder** enables interactive world exploration from images and videos. Given a single image, **Wonder** turns it into a navigable world where users can move the camera to explore both visible regions and plausible unseen areas beyond the input view. Given a video, **Wonder** constructs a dynamic world that preserves the source motion while allowing users to freely change the camera trajectory, enabling camera-controllable 4D video exploration.

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., 2025c) further narrows the training–inference gap by rolling out the student on its own predictions during training, while Rolling Forcing (Liu et al., 2025) and Causal Forcing (Zhu et al., 2026b) extend this paradigm to long-horizon, real-time streaming generation. Together, these methods establish the low-latency, KV-cache-enabled autoregressive foundation that makes interactive video world models practical.

Camera-controllable Video Generation. Pose-conditioned video diffusion has become a standard approach for trajectory-controlled video synthesis. MotionCtrl (Wang et al., 2024) injects 6-DoF camera extrinsics into video diffusion models, 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 works (Wang et al., 2026a; Huang et al., 2025d) further disentangle camera motion from scene dynamics and spatiotemporal correlations. These camera-conditioning designs have inspired recent advances in video world models.

Interactive Video World Models. Most closely related to our work are neural simulators that respond to user actions. Genie 3 (Ball et al., 2025) generates general-purpose, text-promptable interactive worlds, representing the current closed-source frontier. In parallel, a rapidly growing open-source line of work, including Matrix-Game and its real-time/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 prior works advance video world modeling from different perspectives, but still face persistent challenges such as camera drift, memory loss, and degraded visual quality during long-horizon rollout. In contrast, **Wonder** adopts a system-level co-design across control representation, memory mechanism, and distillation



Figure 2 Data curation pipeline. We construct a unified data engine for image-to-world (I2W) and video-to-world (V2W) training. We first collect videos from public datasets and self-rendered UE/Blender environments, then perform multi-duration clipping, collision filtering, and rule-based quality screening to build a raw video pool. Each clip is further enriched with hierarchical VLM captions and camera poses estimated by Depth Anything 3 (DA3), from which we derive discrete camera actions. To improve trajectory quality and action coverage, we apply Gaussian pose smoothing and temporal augmentations such as reverse playback and speed resampling. A final VLM-based quality check removes visually degraded clips and sequences with unnatural camera motion, yielding the curated static and dynamic datasets used for controllable world-model training.

strategy, effectively alleviating these issues and enabling more stable interactive world exploration.

3 Training Data

Training a controllable world model requires large-scale videos with diverse visual content, accurate camera motion, and sufficiently long temporal horizons. Compared with short video generation, long-horizon world modeling requires the model to enforce memory, preserving scene geometry, appearance, and dynamics over extended rollouts. Therefore, we construct a unified data curation pipeline, as shown in Fig. 2, to collect, annotate, and filter video clips.

For I2V training, we first leverage DL3DV (Ling et al., 2024) as a major source of real-world navigation videos. It contains diverse scenes and naturally captured camera trajectories, making it useful for learning realistic geometry-aware motion. However, the camera movements in such real-world videos are often relatively simple and smooth, while interactive users may perform more challenging controls such as sharp turns, lateral movements, backward motion, and compound camera actions. To better cover these hard cases, we additionally build a large-scale synthetic I2V dataset with Unreal Engine (UE). By rendering diverse indoor and outdoor environments with user-controlled-like trajectories, the synthetic data provides richer supervision for robust camera-controllable world exploration.

For V2V training, the data requirement is more challenging because the model needs paired videos that share the same underlying dynamic scene but differ in camera trajectory or temporal evolution. Although open-source datasets such as MultiCamVideo (Bai et al., 2025a) and CamXTime (Huang et al., 2025d) provide useful paired videos, their clips are typically short, often around 5 seconds, which is insufficient for long-horizon world modeling. We therefore render a large amount of long paired V2V data using Blender. The rendered pairs include standard paired trajectories, speed-varied sequences, and bullet-time videos, where camera motion and scene time are partially decoupled. These data help the model learn the distinction between spatial viewpoint changes and temporal scene dynamics, which is essential for controllable V2V generation.

After constructing the raw video pool, we apply a unified annotation and filtering pipeline. We first clip videos into multiple durations, including 5s, 10s, and 20s, to support progressive training from short to long horizons. Rule-based filters are used to remove corrupted videos, black frames, failed renderings, and trajectories with collision or clearly invalid motion. We then use a VLM (Bai et al., 2025b) to generate hierarchical captions (Robbyant Team, 2026) for each long clip, consisting of a global video-level summary and fine-grained captions for its short temporal sub-clips. For metric camera supervision, we use Depth Anything 3 (DA3) (Lin et al., 2025) to estimate camera intrinsics and extrinsics

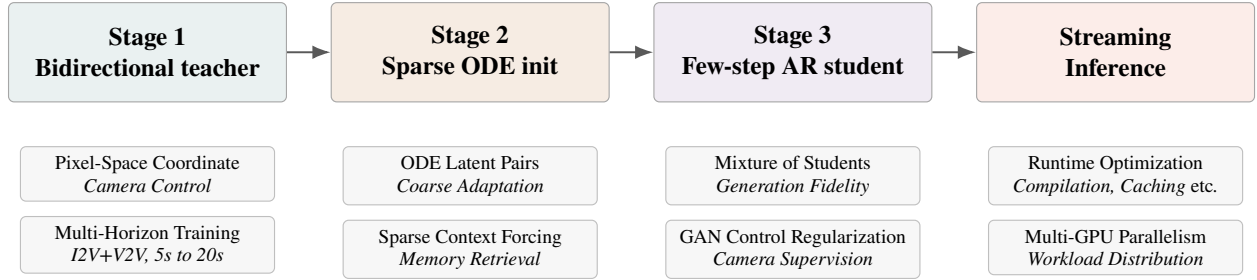


Figure 3 Training and deployment pipeline. **Wonder** first learns high-quality camera control in a bidirectional teacher, then converts the model to a sparse-context causal student, distills it into a few-step autoregressive generator, and finally deploys it with a resident streaming runtime.

for both real and synthetic videos. The recovered trajectories are converted into discrete camera actions. Since estimated poses may contain high-frequency jitter, we apply Gaussian smoothing before action discretization, which suppresses noisy pose fluctuations while preserving the overall camera path. We further use temporal augmentations such as reverse playback and speed resampling to balance the action distribution. Finally, an additional VLM-based quality checking stage removes videos with low visual quality, unnatural camera motion, or inconsistent temporal dynamics, resulting in the final curated static and dynamic datasets for training.

4 Wonder World Model

Wonder generates explorable videos that follow a user-streamed camera trajectory. It supports two inference modes. In image-to-video mode, a single input frame initializes an explorable world. In video-to-video mode, a source video anchors the appearance and motion of an existing event, while the target camera trajectory specifies a new viewpoint of that event. By unifying both modes within a single model, **Wonder** enables multimodal video world modeling.

Following the common framework of recent works (Hong et al., 2025; Robbyant Team, 2026), we adopt a two-stage pipeline that distills a few-step autoregressive video diffusion model from a bidirectional video diffusion teacher, with both models conditioned on camera-control signals. While this framework has shown promising early results, achieving high-quality, low-latency, memory-aware long-horizon rollout remains an open challenge. Existing approaches often fail to satisfy these requirements simultaneously: some (Robbyant Team, 2026; Team et al., 2025) struggle to faithfully follow camera instructions; many (Zhu et al., 2026a; He et al., 2025b; Li et al., 2025) lack an explicit memory mechanism or rely on lossy scene summaries (Hong et al., 2025); and methods that preserve memory by retaining the full historical KV cache (Robbyant Team, 2026) incur latency that grows linearly with context length. These limitations motivate us to systematically redesign the pipeline, including the control representation, memory retrieval mechanism, and autoregressive training strategy.

We highlight several key innovations in our framework, as shown in Figure 3. First, we redesign the bidirectional video diffusion teacher to condition on a dense pixel-space camera representation, enabling camera motion to be interpreted as a pixel-aligned visual signal and yielding strong action-following behavior (Sec. 4.1.1). We then distill this teacher into a few-step autoregressive video diffusion model for low-latency interactive streaming. To support long-horizon generation without increasing inference cost, we introduce a lossless yet efficient memory mechanism that selectively retrieves a compact set of faithful historical video tokens from the KV cache. This mechanism is paired with a training strategy that induces emergent memory retrieval during autoregressive rollout (Sec. 4.2.1). We further improve the self-forcing training paradigm with two techniques that mitigate quality degradation and camera drift during distillation (Sec. 4.2.2 & 4.2.3), substantially improving both training stability and controllability. Finally, we apply runtime optimizations that enable real-time inference and minute-scale generation.

4.1 Stage 1: Camera-Conditioned Teacher for Long Video Generation

Training a strong bidirectional teacher with precise camera following and multimodal input support is essential in our setting, as it serves both as the initialization point for the autoregressive student and as the supervision signal for distillation. We next describe the techniques used to obtain this bidirectional teacher.

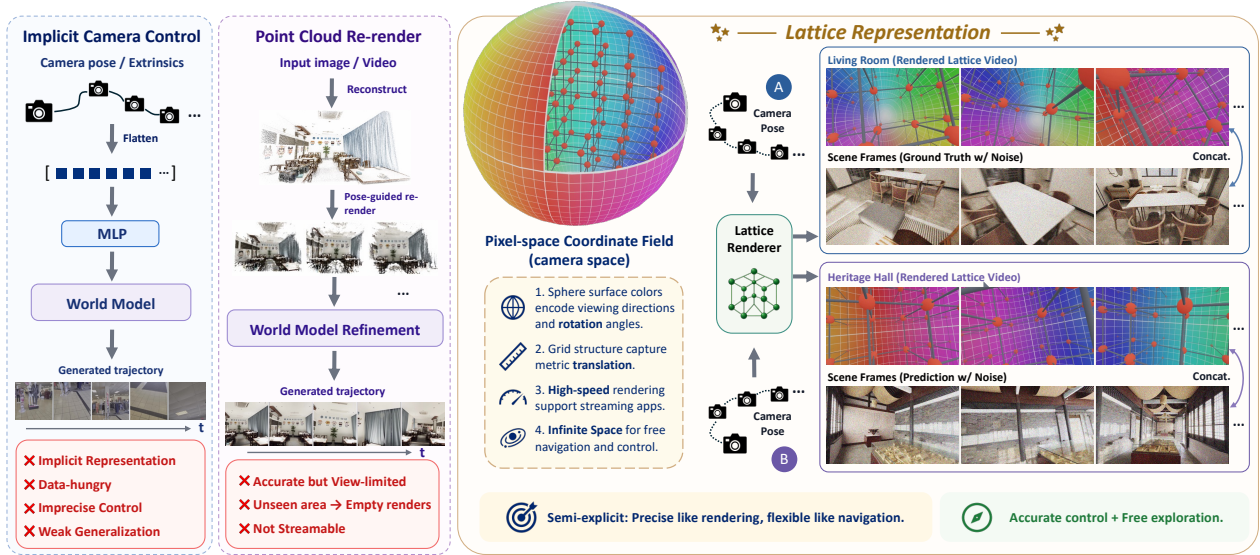


Figure 4 Camera control using Pixel-Space Coordinate Field. Prior camera-control methods are either implicit, encoding camera poses with an MLP (Bai et al., 2025a) or RoPE (Wang et al., 2026a), or explicit, re-rendering reconstructed point clouds (Team et al., 2026b). The former is data-hungry and often imprecise, while the latter is accurate but view-limited and fails when the camera explores unseen regions. Our lattice representation combines the strengths of both. We construct a pixel-space coordinate field with a colored spherical environment map at infinity and a dense 3D lattice scaffold inside it. A lightweight renderer converts arbitrary camera trajectories into pixel-aligned lattice videos that encode viewing direction and metric translation. Concatenating these renders with noisy scene inputs provides precise, streamable, and exploration-friendly camera control for world modeling.

4.1.1 Pixel-Space Coordinate Field

As shown in Figure 4, existing camera-control methods can be broadly grouped into two categories: implicit and explicit. Implicit methods typically represent camera motion as a sequence of low-dimensional vectors derived from camera extrinsics (Robbyant Team, 2026; Bai et al., 2025a; Hong et al., 2025). These vectors are projected into high-dimensional embeddings and injected token-wise into the diffusion transformer, where all spatial tokens within the same latent frame share an identical camera condition. Although compact, such global numerical representations provide only coarse motion cues and are therefore difficult for the model to associate with subtle, local, or high-frequency pixel-space changes, even with large-scale training. Ray-based representations, as well as their combinations with RoPE (Wang et al., 2026a; Team et al., 2026a), offer a more geometric alternative, but still suffer from ambiguities in scale, origin, and depth. These limitations become even more pronounced during distillation: commonly used objectives such as DMD (Yin et al., 2024b) provide no direct supervision for camera following, which often results in camera drift and weak generalization in distilled models for diverse camera movements.

In contrast, explicit methods (Team et al., 2026b; Yu et al., 2024) rely on offline pose or geometry estimation models to reconstruct a point cloud from the input image or video, which is then re-rendered from user-specified viewpoints. Such representations provide a strong geometric anchor and reduce the role of the world model to refinement or inpainting. However, they also introduce several practical limitations. First, their effectiveness heavily depends on the quality and robustness of external pose, depth, or reconstruction models. Second, the resulting control signal is inherently view-limited: when the camera moves beyond the regions observed in the input, the re-rendered point cloud often becomes empty or black, providing little useful camera guidance and causing control failures. Finally, accurate reconstruction typically requires buffering multiple frames and running additional external modules, which introduces latency and complicates both training and inference pipelines. These requirements make explicit re-rendering difficult to deploy in low-latency streaming applications.

To overcome these limitations, we introduce a pixel-space camera representation that exposes camera motion as visual evidence. Specifically, we construct a synthetic camera space and render it from the target camera trajectory, producing conditioning frames that are spatially aligned with the video to be generated. The synthetic scene consists of two complementary components. First, we place a dense 3D scaffold in the camera space, which provides an explicit spatial parameterization of the environment and alleviates the depth ambiguity. Its 2D projections deform under camera

translation and induce parallax across frames, allowing the model to infer metric camera translation and align the rendered structure with the conditional image or video. Second, we place a colored environment map at infinity, whose appearance varies with camera orientation and thus provides a clear cue for rotation. Together, the scaffold and the environment map convert translation and rotation into dense, frame-aligned visual signals that are easy for the video diffusion model to interpret. We implement this representation with a lightweight OpenGL renderer, which synthesizes conditioning frames at 150 FPS with negligible overhead. The rendered frames are encoded by the VAE and concatenated channel-wise with the target noisy latents before being processed by the DiT. Empirically, this representation substantially improves camera controllability and stabilizes the distillation process.

4.1.2 Base Architecture

We build our framework on Wan2.1-I2V-14B (Wan et al., 2025), a bidirectional video diffusion model originally developed for short-horizon image-to-video generation. Wan consists of two main components: a spatio-temporal VAE and a diffusion transformer backbone. The VAE maps input videos into compact latent sequences with a 3D causal encoder-decoder, downsampling the spatial resolution by a factor of 8 and the temporal resolution by a factor of 4. Its causal design and feature-caching mechanism provide a natural and efficient interface for streaming video generation in our system. The diffusion backbone operates on patchified video latents. After a stack of transformer blocks, the latent tokens are mapped back to video latents 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 and used to modulate the transformer blocks. This architecture provides a strong bidirectional teacher initialization, which we further adapt for camera-controllable world exploration.

To support both image- and video-conditioned generation within a unified framework, we represent the model input as an optional source segment followed by a target segment. For video-conditioned generation, the source segment corresponds to the conditioning video and is provided as clean latents. For image-conditioned generation, there is no source segment; instead, the first frame of the target video is directly provided as a clean, noise-free latent. Noise is applied only to the target frames that need to be generated, while the conditioning latents remain clean throughout both training and inference. Importantly, the binary mask is used to distinguish source and target segments. The mask value is set to 0 for the source video and 1 for the target video. Similarly, the encoded camera-condition latents are zero-padded only for the source segment, and are provided for the entire target trajectory. Formally, let $\mathbf{z}^s$ denote the optional source-video latents, and let the target latents be decomposed as $\mathbf{z}^t = [\mathbf{z}^a, \mathbf{z}^p]_T$, where $\mathbf{z}^a$ denotes the clean target anchor frame and $\mathbf{z}^p$ denotes the target frames to be generated. The model input at diffusion timestep τ is constructed as

$$\mathbf{u}_\tau = [[\mathbf{z}^s, \mathbf{z}^a, \alpha_\tau \mathbf{z}^p + \sigma_\tau \epsilon]_T, [\mathbf{0}_m^s, \mathbf{1}_m^a, \mathbf{1}_m^p]_T, [\mathbf{0}_c^s, \mathbf{z}^{c,a}, \mathbf{z}^{c,p}]_T]_C. \quad (1)$$

Here, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is Gaussian noise, and $[\cdot]_T$ and $[\cdot]_C$ indicate concatenation along the temporal and channel dimensions. For positional encoding, we mirror the RoPE coordinates of the target video onto the source video, and add separate learnable positional embeddings to help the model distinguish source and target tokens.

4.1.3 Multi-task Multi-Horizon Training

We train a single student model that supports both image-to-video and video-to-video generation, while being able to roll out to long horizons. Following the progressive-horizon distillation strategy of (Hong et al., 2025), we gradually extend the training horizon instead of directly training on the longest sequences. Concretely, we first post-train the base model on 5-second clips with both I2V and V2V tasks jointly enabled. We then extend the horizon to 10 seconds and finally to 20 seconds, where the RoPE positions are extrapolated with YaRN (Peng et al., 2023) to support longer temporal contexts.

A direct multi-task extension to long horizons is computationally inefficient, since V2V samples contain both source and target videos and therefore incur substantially higher memory and compute cost. We therefore adopt a mixed-length training schedule in the long-horizon stages. For most training iterations, we keep V2V samples at the 5-second horizon while training I2V samples at the current long horizon, i.e., 10 or 20 seconds. This design is based on the observation that long-horizon camera following can be effectively learned from I2V data, while 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 strategy maximizes training throughput while enabling the model to generalize to long-horizon V2V dynamic world exploration.

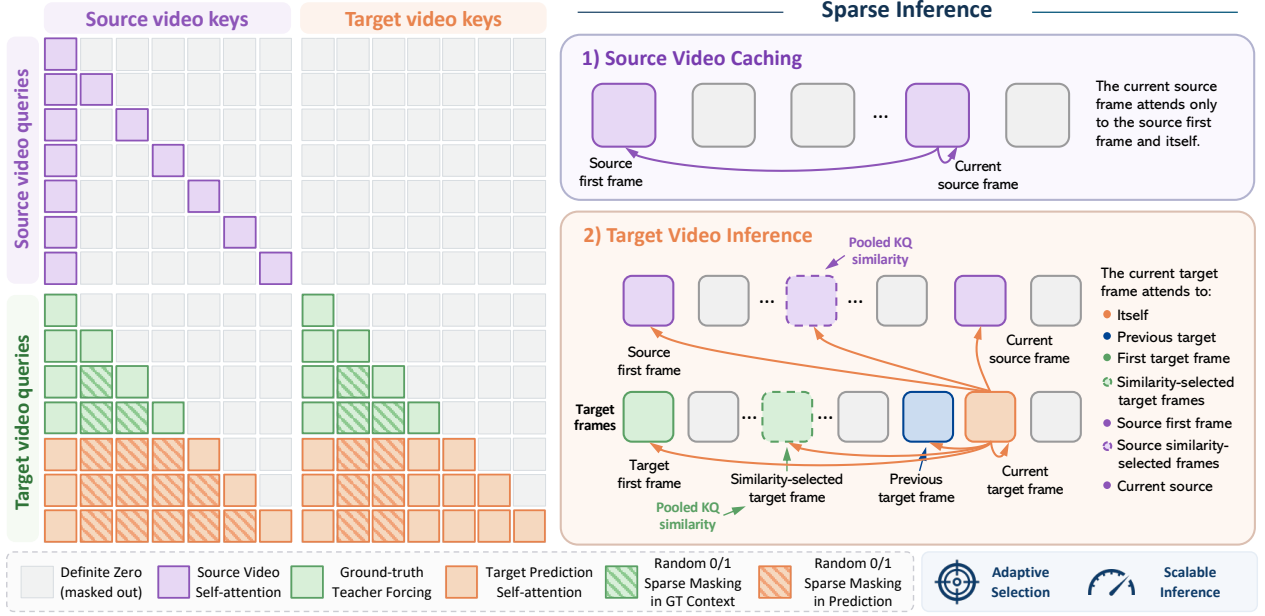


Figure 5 Sparse context forcing and sparse inference. Left: during ODE initialization, we train the causal student with a unified attention layout for I2V and V2V. Source-video tokens attend only to the first source frame and themselves, while target-video tokens combine teacher-forced clean context and diffusion-forced predicted context. To prepare the student for sparse-memory inference, we randomly drop eligible non-local context edges in both ground-truth and prediction regions, while preserving required anchors such as self-attention, first-frame anchors, and recent context. Right: at inference time, the source video is cached as a reference stream, and each target chunk attends to a constant-size set of memories, including itself, recent target chunks, first-frame anchors, current/source references, and additional source/target chunks selected by pooled query-key similarity.

4.2 Stage 2: Fast Autoregressive Student for Real-Time Streaming

The camera-conditioned teacher provides accurate control, but it is not directly interactive: it is bidirectional, relies on dense attention, and requires many denoising steps, making live control prohibitively expensive. Following recent video world model pipelines (Robbyant Team, 2026; Team et al., 2025; Hong et al., 2025), we adapt and distill this teacher into an autoregressive few-step student for low-latency streaming generation. We base our distillation framework on self-forcing (Huang et al., 2025c), a rollout-based distillation paradigm designed to reduce exposure bias in long-horizon generation. Specifically, the student is initialized from the teacher and first trained with ODE initialization, where it regresses precomputed teacher ODE trajectories at four denoising timesteps. We then apply DMD (Yin et al., 2024b) on the student’s own autoregressive rollouts, minimizing the reverse KL divergence between the diffused real-data distribution and the student’s self-induced generation distribution over sampled timesteps.

4.2.1 Sparse Context Forcing

During autoregressive rollout, the student continuously writes generated chunks into the KV cache. As the rollout grows longer, the cache size increases accordingly, causing the generation speed to decrease approximately linearly with the context length (Robbyant Team, 2026). However, world modeling requires persistent memory to support consistent revisits to previously observed regions. Therefore, the cache cannot be simply discarded with a sliding-window strategy (Zhu et al., 2026a; Li et al., 2025; Team et al., 2026a). Existing solutions are still limited. One approach compresses distant historical frames in the latent space (Hong et al., 2025), but aggressive compression often leads to noticeable degradation in visual quality and weakens long-term consistency. Another line of work retrieves relevant memory by estimating the field-of-view overlap between the current viewpoint and historical viewpoints (Yu et al., 2025; Team et al., 2026a). While more adaptive, such geometry-based heuristics are brittle in practice: they depend on reliable camera geometry and can fail in corner cases, such as revisiting the same location from opposite viewpoints.

Video diffusion models naturally learn rich cross-frame correspondences (Xu et al., 2025). Instead of compressing or discarding historical KV states, we therefore keep the entire KV cache in full fidelity and let the model retrieve relevant chunks through content-aware sparse attention. As shown in Figure 5, each historical chunk stores its dense keys and

values for attention, while only a lightweight pooled key summary is used for retrieval. At each rollout step, the current query summary is compared with the summaries of historical chunks to select a small set of active memory chunks. We always keep the initial chunk and the most recent r chunks active, where the former serves as an attention sink, and the latter preserves short-term motion continuity.

Let $\mathcal{I}_c$ denote the token index set of cache chunk c , and let $\mathbf{K}_{\mathcal{I}_c}$ and $\mathbf{V}_{\mathcal{I}_c}$ be the corresponding full-resolution keys and values. We summarize each historical chunk by a pooled key $\bar{\mathbf{K}}_c$, and summarize the current query chunk as $\bar{\mathbf{Q}}_t$. The active memory set $\mathcal{A}_t$ is then selected by combining the initial chunk, the most recent r chunks, and the top- k middle-history chunks with the highest query-key similarity:

$$\begin{aligned}\mathcal{A}_t &= \{0\} \cup \mathcal{N}_r(t) \cup \text{TopK}_k \left\{ \text{sim}(\bar{\mathbf{Q}}_t, \bar{\mathbf{K}}_c) \mid c \in \mathcal{M}(t) \right\}, \\ \mathbf{K}_t^{\text{ctx}} &= [\mathbf{K}_{\mathcal{I}_c}]_{c \in \mathcal{A}_t}, \quad \mathbf{V}_t^{\text{ctx}} = [\mathbf{V}_{\mathcal{I}_c}]_{c \in \mathcal{A}_t}.\end{aligned}\tag{2}$$

Here, $\mathcal{N}_r(t)$ denotes the most recent r chunks before the current chunk, and $\mathcal{M}(t)$ denotes the middle-history chunks excluding the initial and recent chunks. TopK_k returns the indices of the k chunks with the highest similarity scores, and $[\cdot]_{c \in \mathcal{A}_t}$ denotes concatenation over the selected chunks in their original temporal order. In practice, we set $r = 2$ to preserve local motion smoothness. Since retrieval is performed only on lightweight pooled summaries, the selection overhead is negligible. Meanwhile, the selected chunks are still attended with their full-resolution keys and values, allowing the model to preserve faithful visual memory without latent compression. Under a fixed top- k budget, the active attention context remains constant regardless of rollout length, enabling long-horizon generation with stable latency.

Directly applying sparse attention during distribution-level distillation is unstable, since the student has not yet learned to generate under incomplete historical context. We therefore introduce *sparse context forcing* during ODE initialization, so that the student is warmed up with the same memory constraint it will encounter at inference time. Our ODE initialization combines teacher forcing (Gao et al., 2024; Hu et al., 2024) and diffusion forcing (Chen et al., 2024) under a causal attention layout. Teacher-forced context provides clean target frames to simulate low-noise historical context, while diffusion-forced context exposes the student to previous noisy target frames, reducing the train–test mismatch during autoregressive rollout (Hong et al., 2025). For video-to-video training, the source video is additionally cached as a reference stream, where each source frame attends only to itself and the first source frame. For image-to-video training, the layout naturally degenerates to the target-only case.

To make the student robust to sparse memory retrieval, we randomly drop eligible historical context edges during ODE initialization. Required edges, such as self-attention, the first-frame anchors, and the most recent context frames, are always preserved to maintain identity and local motion continuity. Non-local historical edges are treated as optional memory and are randomly kept on each layer and each forward pass. Formally, let $a(i, j)$ denote the base causal relation between query chunk i and key/value chunk j . We define $\mathcal{R}$ as the set of required relations and $\mathcal{O}$ as the set of optional historical-context relations. The attention mask at layer ℓ is sampled as

$$M_{ij}^{(\ell)} = \mathbf{1}[a(i, j) \in \mathcal{R}] + b_{ij}^{(\ell)} \mathbf{1}[a(i, j) \in \mathcal{O}], \quad b_{ij}^{(\ell)} \sim \text{Bernoulli}(1 - p_\ell).\tag{3}$$

Here, p_ℓ controls the sparsity level of optional memory edges. We use a stronger drop rate for longer target horizons, encouraging the model to rely on a smaller set of informative memories instead of dense access to the full history. This training strategy aligns the ODE-initialized student with our sparse inference pattern: the model learns to preserve quality and camera consistency even when only a limited subset of historical chunks is available.

4.2.2 Mixture of Students

After ODE initialization, the student can already recover coarse scene structure and layout, but still struggles to synthesize fine-grained local details. We therefore further optimize the student with DMD (Yin et al., 2024b) to better align its rollout distribution with the data distribution. As illustrated in Figure 6, the few-step causal student autoregressively generates full video rollouts conditioned on previously generated clean chunks. These rollouts are then perturbed with noise and fed into the bidirectional teacher and critic for score estimation. The resulting score difference provides the gradient signal for updating the student generator.

Self-forcing (Huang et al., 2025c) reduces exposure bias by training the student on its own autoregressive rollouts. However, the underlying distillation problem remains highly challenging: the student is required to approximate a strong bidirectional, many-step teacher with a causal architecture and only a few denoising steps. These architectural and sampling constraints significantly reduce the effective modeling capacity of a single student, making it difficult to capture

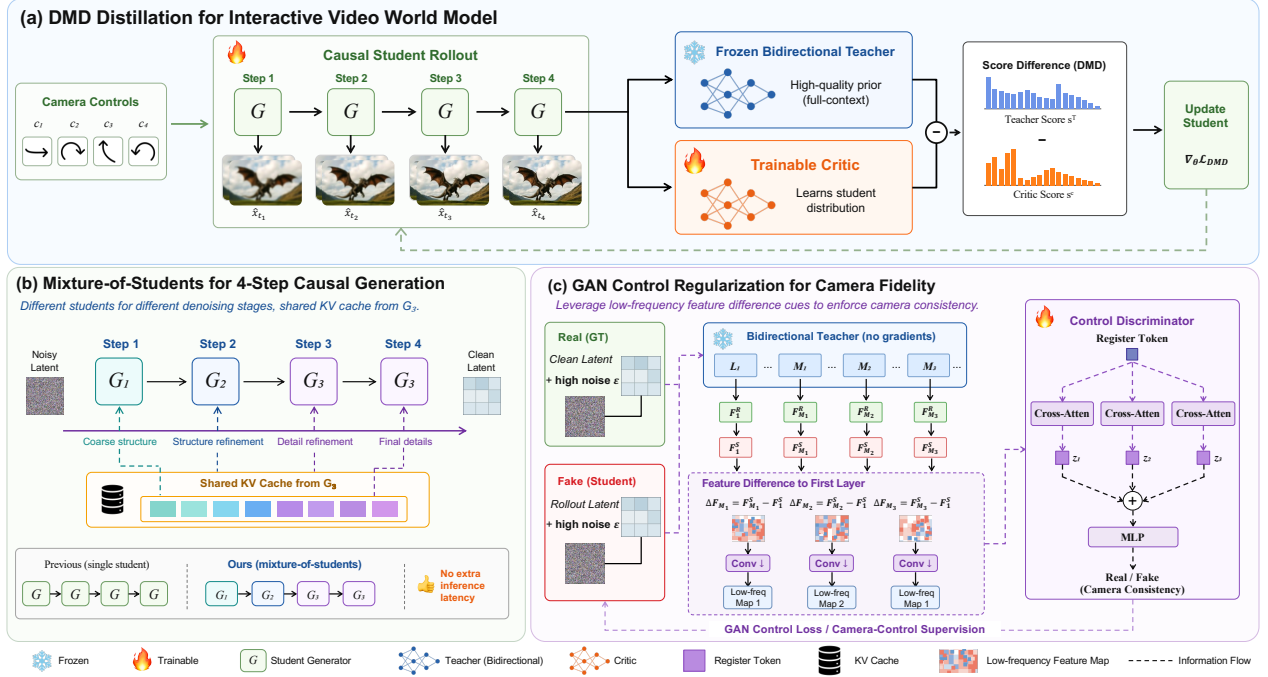


Figure 6 Distillation framework with Mixture-of-Students and GAN Control Regularization. (a) The causal few-step student autoregressively rolls out video chunks under camera controls. The generated rollout is perturbed with noise and evaluated by a frozen bidirectional teacher and a trainable critic, whose score difference provides the DMD gradient for updating the student. (b) To increase student-side capacity, we specialize different generators for different denoising stages. The 4-step sampler uses G_1 for coarse structure, G_2 for structure refinement, and G_3 for detail refinement and final reconstruction, while sharing the same streaming KV cache to avoid extra inference latency. (c) To improve camera fidelity, we introduce a control discriminator built from low-frequency teacher feature differences. Real and fake noised latents are passed through the frozen teacher with the same camera condition. Intermediate features are compared with the first-layer feature, downsampled into low-frequency maps, and aggregated by a register-token discriminator to distinguish camera-consistent real samples from student rollouts.

the diverse modes of the teacher distribution. In practice, this often leads to mode shrinkage, weakened local realism, and degraded fine details during long-horizon rollout. This motivates us to introduce a mixture-of-students design, which increases the student-side capacity while preserving the same autoregressive few-step inference interface.

Motivated by the capacity bottleneck discussed above, we adopt a timestep-wise mixture-of-students design, analogous to the denoising-stage MoE in Wan2.2 (Wan et al., 2025). Instead of using a single generator for all few-step updates, our 4-step sampler uses three student generators: G_1 for the first step, G_2 for the second step, and G_3 for the last two steps:

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

where the superscript denotes the denoising step. This design allocates capacity according to the role of each denoising stage: G_1 focuses on coarse structure formation, G_2 performs structure refinement, and G_3 is responsible for detail refinement and final reconstruction.

To preserve streaming efficiency, we do not maintain separate KV caches for different students. Instead, all generators share the autoregressive interface and reuse the KV cache produced by G_3 . Therefore, the mixture-of-students increases the effective student-side capacity without changing the number of denoising steps or introducing additional inference latency. Empirically, this design improves distribution matching to the teacher, leading to more diverse outputs and better fine-grained visual quality during long-horizon rollout.

4.2.3 GAN Control Regularization

Built upon DMD (Yin et al., 2024b), DMD2 (Yin et al., 2024a) attaches an additional GAN objective to the critic and has been adopted in recent world-model distillation pipelines (Robbyant Team, 2026) to improve high-frequency details and local realism. However, these objectives are not specifically designed for camera-controlled autoregressive video

generation. In our setting, we observe a distinct failure mode during DMD training: the student gradually loses camera fidelity, where the generated camera motion becomes faster, deviates from the target trajectory, and eventually leads to severe blur or scene collapse. This issue is amplified in causal rollout, since camera-control errors accumulate over time.

We attribute this camera drifting problem to insufficient control supervision in the standard DMD objective. The score residuals produced by DMD mainly emphasize high-frequency local artifacts, such as edges and textures, but provide weak gradients for the low-frequency layout changes induced by camera motion. Once the student rollout drifts far from the teacher’s training distribution, e.g., when the camera has already flown away or the scene is heavily degraded, the frozen bidirectional teacher can no longer provide reliable score guidance. Prior work mitigates this instability by adding an additional downsampled ℓ_2 loss between self-rollouts and ground-truth videos (Hong et al., 2025). While effective for stabilization, such direct regression tends to shrink the distribution and often produces blurry newly explored regions.

To address this issue, we introduce *GAN control regularization*, an adversarial objective that focuses on low-frequency camera-consistency cues rather than high-frequency texture realism. As shown in Figure 6(c), we construct real and fake noised latents by perturbing the ground-truth latent and the student rollout latent with a high noise level:

$$\mathbf{z}_\tau^r = \alpha_\tau \mathbf{z}_0 + \sigma_\tau \epsilon, \quad \mathbf{z}_\tau^f = \alpha_\tau \hat{\mathbf{z}}_0 + \sigma_\tau \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (5)$$

where $\mathbf{z}_0$ denotes the clean ground-truth latent, $\hat{\mathbf{z}}_0$ denotes the student rollout latent, and τ is sampled from a high-noise range. Both latents are forwarded through the frozen camera-conditioned bidirectional teacher under the same camera condition $\mathbf{c}$. Since the teacher is fixed, its intermediate representations provide a stable camera-aware feature space for comparing real and fake trajectories. We empirically find this design more effective than extracting features from the trainable critic, whose representation keeps changing during distillation training.

Instead of discriminating the final latent directly, our control discriminator operates on teacher feature dynamics. Since the camera condition is injected at the input, the first-layer feature provides a direct camera-aware reference. We compare intermediate teacher features against this reference to assess whether the input latent remains consistent with the prescribed camera trajectory. For $x \in \{r, f\}$, where r denotes the real ground-truth latent and f denotes the fake student rollout, we extract the first-layer and three selected middle-layer features from the frozen teacher:

$$(\mathbf{F}_1^x, \mathbf{F}_{M_1}^x, \mathbf{F}_{M_2}^x, \mathbf{F}_{M_3}^x) = T_{\text{feat}}(\mathbf{z}_\tau^x, \mathbf{c}), \quad (6)$$

where T_{feat} denotes the frozen camera-conditioned teacher feature extractor. For each middle layer, we compute a low-frequency feature-difference map with respect to the first-layer feature:

$$\Delta \mathbf{F}_{M_i}^x = \psi_i(\eta_i(\mathbf{F}_{M_i}^x) - \eta_1(\mathbf{F}_1^x)), \quad i \in \{1, 2, 3\}. \quad (7)$$

Here, η_1 and η_i are linear feature extractors for the first layer and layer M_i , and ψ_i is a strided convolutional downsampler that suppresses high-frequency texture details and encourages the discriminator to focus on low-frequency layout and camera-motion consistency. We then aggregate the resulting feature-difference maps with 3 learnable register tokens. The register tokens attend to their corresponding low-frequency maps through cross-attention, and the resulting token features are summed to produce a control-consistency logit:

$$\mathbf{h}_i^x = \text{CrossAttn}_i(\mathbf{q}, \Delta \mathbf{F}_{M_i}^x), \quad d_{\text{ctrl}}^x = \text{MLP}\left(\left[\sum_{i=1}^3 \mathbf{h}_i^x, e_\tau\right]\right), \quad x \in \{r, f\}. \quad (8)$$

Here, $\mathbf{q}$ is the learnable register token, e_τ refers to the timestep embedding, and d_{ctrl}^r and d_{ctrl}^f denote the discriminator logits for real ground-truth latents and fake student rollout latents, respectively. We train the control discriminator and the student generator with a relativistic softplus adversarial loss:

$$\mathcal{L}_{\text{ctrl}}^D = \mathbb{E}_{\tau, \epsilon} \left[\text{softplus} \left(d_{\text{ctrl}}^f - d_{\text{ctrl}}^r \right) \right], \quad \mathcal{L}_{\text{ctrl}}^G = \mathbb{E}_{\tau, \epsilon} \left[\text{softplus} \left(d_{\text{ctrl}}^r - d_{\text{ctrl}}^f \right) \right]. \quad (9)$$

The generator is optimized with $\mathcal{L}_{\text{DMD}} + \lambda_{\text{ctrl}} \mathcal{L}_{\text{ctrl}}^G$. Compared with the high-frequency adversarial objective in DMD2, our control regularization provides a complementary low-frequency supervision signal for camera fidelity. Empirically, the proposed objective stabilizes the distillation process, substantially improves camera following during autoregressive rollout, and avoids the mode shrinkage introduced by prior regression-based stabilization losses.

4.3 Inference-time Optimization

To achieve low-latency inference, we conduct code optimizations following (Hong et al., 2025). We briefly describe techniques below. Because inference latency can be caused by GPU memory traffic, kernel-launch overhead, and repeated transformer computation, we focus our optimizations on these aspects. 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 reduce launch overhead. We adopt self-attention KV caches to preserve previously computed video context, while using text cross-attention caches to eliminate repeated computation of prompt key and value projections. For memory retrieval, we maintain a separate cache that stores compressed key summaries to support efficient retrieval of the relevant historical chunks. We further employ FlashAttention-3 (Shah et al., 2024) on Hopper GPUs, fused QKV projections, BF16 RMSNorm, cached RoPE embeddings, and timestep modulations. Finally, guided by profiling results, we manually restructure some PyTorch operations to eliminate additional runtime overhead.

During inference, we employ multi-GPU parallelism to distribute both computation and memory workloads across devices. FFN layers and cross-attention modules are distributed along the sequence dimension using sequence parallelism, whereas self-attention is partitioned across attention heads using tensor parallelism. NCCL A11-to-A11 collectives convert between these two tensor layouts. This design also partitions the KV cache along the attention-head dimension, such that each GPU stores and processes only its assigned subset of heads. Auxiliary workloads, including text encoding and VAE processing, are offloaded from the primary generation GPUs, while a Tiny VAE adapted from (Shin et al., 2025) is used for efficient decoding of conditioning inputs and generated outputs.

To support minutes-long generation beyond our training horizon, we adopt a rolling/sliding-window caching scheme (Xiao et al., 2024). Specifically, the KV cache is organized as a concatenation of [sink chunks, top-k chunks, recent context chunks]. Top-k chunks are selected dynamically by our sparse attention at each prediction step. To avoid unseen positional embeddings along the frame axis, we remap the frame indices back to the training horizon based on their relative distance. Together, these designs keep both the active context size and positional range bounded, enabling stable minute-scale rollouts without increasing per-step inference latency.

4.4 Training Infrastructure and Model Details

Training a bidirectional teacher together with three 14B student generators (MoS) over a 20-second horizon is highly memory-intensive. Following prior work (Hong et al., 2025; Robbyant Team, 2026), we combine several techniques to improve training efficiency: (1) Fully Sharded Data Parallel 2 (FSDP2) (Zhao et al., 2023) to shard model parameters, gradients, and optimizer states across GPUs; (2) sequence parallelism to distribute the video-token sequence across devices; (3) tensor parallelism to partition self-attention heads and associated computations; (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 required for differentiating through autoregressive rollouts. Together, these techniques allow us to train with a global batch size of 64 on 32 NVIDIA H200 GPUs. At inference time, our mixture-of-students sampler uses three generators for four denoising steps: G_1 is used at the first step, G_2 at the second step, and G_3 at the final two steps. We use a chunk size of two latent frames to maintain low streaming latency.

5 Experiments

In this section, we evaluate **Wonder** as a unified interactive video world model for both image- and video-conditioned generation. We compare it with recent advanced world models and show that **Wonder** achieves the best overall visual quality and camera-following accuracy. Our experiments cover two settings. In the *image-to-video* setting, the model is given a single image and must infer unseen regions while following user-specified camera controls. In the *video-to-video* setting, the model is given an input video and must preserve its subject dynamics while enabling free camera navigation, thereby constructing a dynamic and controllable world. We first describe the benchmarks and evaluation metrics, followed by quantitative and qualitative results for both tasks.

5.1 Image-to-Video World Modeling

Benchmark. We construct a general image-to-video world-modeling benchmark with 1,000 diverse open-source images, covering real-world photographs, AI-generated images, cartoons, artistic images, and gaming scenes. For each image, we use a VLM to generate a scene caption and produce navigation keys that are consistent with the image content. Each

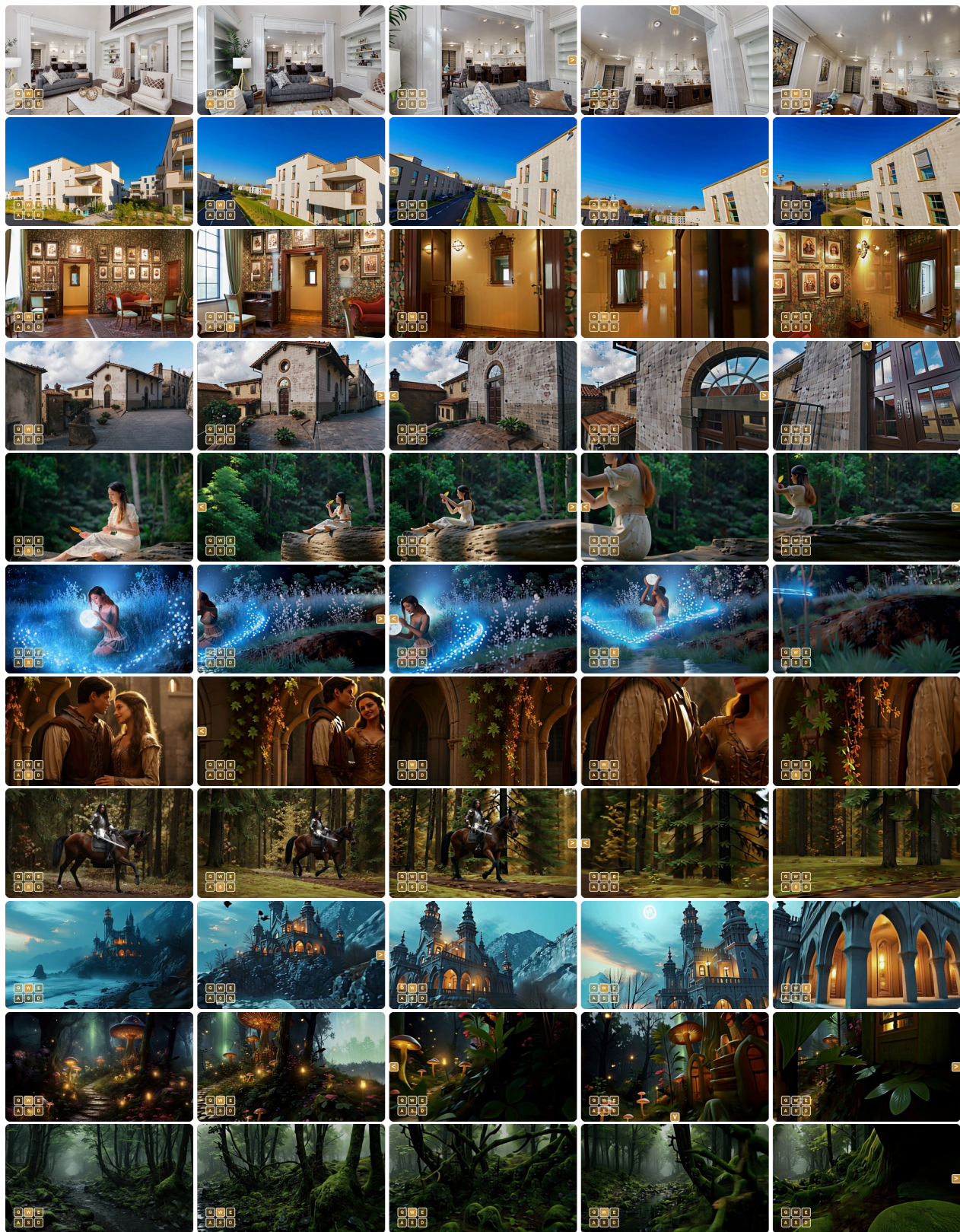


Figure 7 Navigation into diverse images.

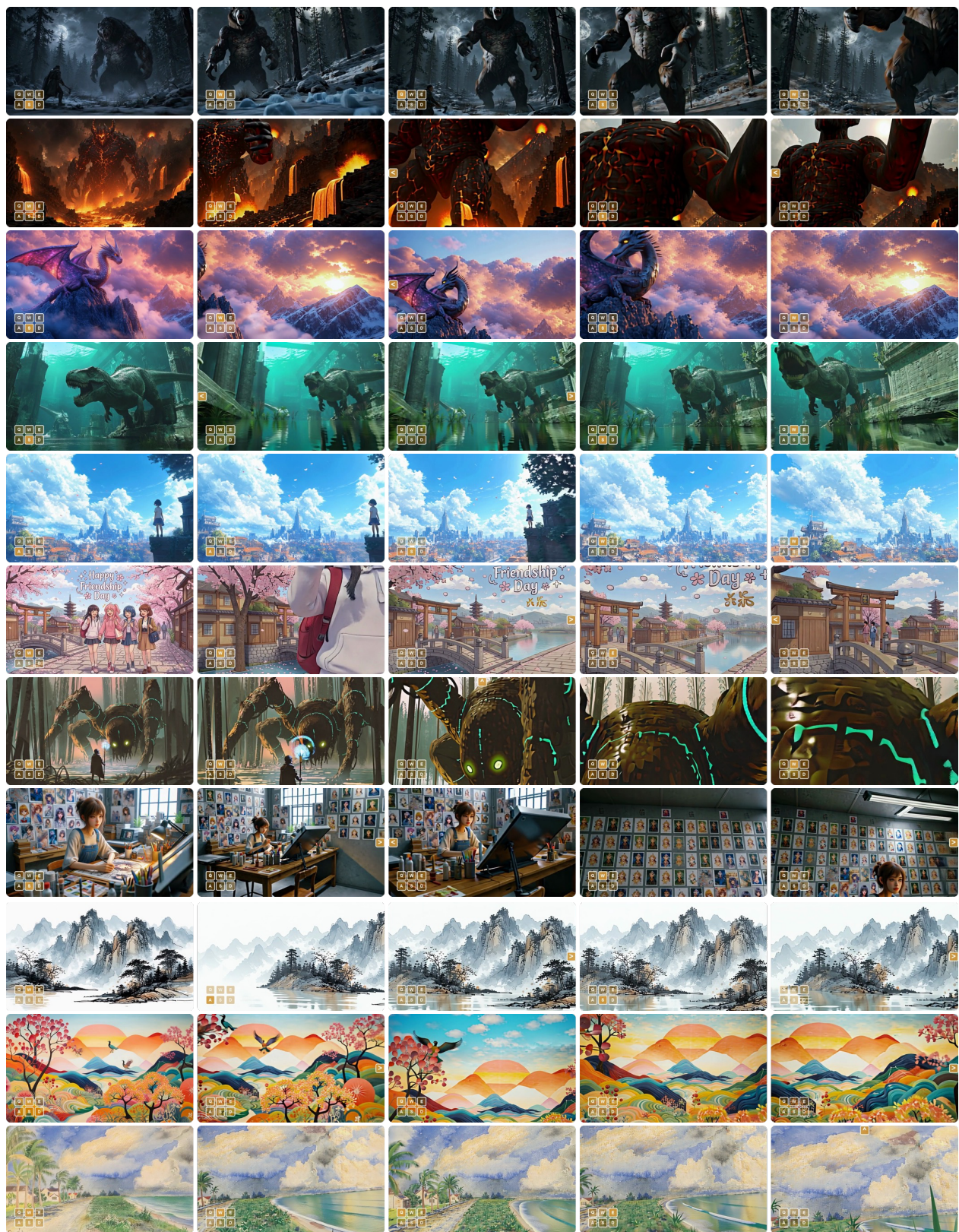


Figure 8 Navigation into diverse images.

Model	Visual quality $\uparrow$						Action accuracy (RPE $\downarrow$)	
	Avg. Score [†]	Imaging	Aesthetic	Dynamic	Motion	Flickering	Trans.	Rot.
<i>Image-to-Video (I2V)</i>								
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	0.8558	0.7113	0.6416	0.9874	0.9794	0.9622	0.0132	0.0784
<i>Video-to-Video (V2V)</i>								
Inspatio-World	0.8374	0.6756	0.5815	1.0000	0.9794	0.9506	0.0436	0.2470
Wonder	0.8527	0.6981	0.6196	1.0000	0.9855	0.9602	0.0187	0.1119

Table 1 Comparison on visual quality and action accuracy. [†] Average Score is the arithmetic mean of Imaging, Aesthetic, Dynamic, Motion, and Flickering scores. We report results under both image-to-video (I2V) and video-to-video (V2V) settings. **Wonder** achieves the best overall video synthesis quality and camera-following accuracy.

image is paired with five camera trajectories of varying translation scales and rotation speeds, allowing us to evaluate the robustness of different models under various camera-control patterns.

Metrics. Following prior work (Robbyant Team, 2026; Zhu et al., 2026a), we report five visual-quality metrics from VBench (Huang et al., 2024): imaging quality, aesthetic quality, dynamic degree, motion smoothness, and temporal flickering. To evaluate camera-following accuracy, we estimate relative camera poses from each generated video using two pose-estimation models, Depth Anything 3 (Lin et al., 2025) and ViPE (Huang et al., 2025a), and average their pose errors for a more robust assessment. Since different models may produce trajectories with different translation scales and coordinate frames, we align each reconstructed trajectory to the canonical target trajectory using Umeyama alignment (Umeyama, 2002). We then compute translational and rotational Relative Pose Error (RPE) to measure how accurately each model follows the prescribed camera motion.

Baselines. We compare **Wonder** with five recent interactive video world models available at the time of this report: 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). These models all target real-time or streaming video generation with interactive camera control, making them the most relevant baselines for our setting. For each baseline, we use the officially released model or the recommended inference setting whenever available.

Results. Table 1 reports the quantitative comparison under the image-to-video setting. **Wonder** achieves the best overall visual quality with an average score of 0.8558, outperforming all recent streaming world-model baselines. The gain mainly comes from improved imaging quality, where **Wonder** obtains the highest score of 0.7113, indicating stronger scene synthesis and fewer visual artifacts. Although some baselines achieve higher scores on individual temporal metrics such as motion smoothness or flickering, they often trade off camera controllability. In terms of action accuracy, **Wonder** achieves the lowest translational and rotational RPE among all methods. Compared with the strongest baseline, **Wonder** reduces translational error from 0.0174 to 0.0132 and rotational error from 0.1155 to 0.0784. These results demonstrate that our semi-explicit camera representation and distillation strategy substantially improve camera following while maintaining high visual quality, enabling more reliable interactive world exploration. Figure 9 presents a qualitative comparison of long-horizon memory and visual quality between **Wonder** and recent state-of-the-art models. Most prior methods fail to faithfully reconstruct previously observed regions upon revisit, while others suffer from substantial visual degradation as the rollout length increases.

We present more qualitative results of **Wonder** in Figure 7 and Figure 8. Across diverse input images, **Wonder** can understand and preserve the underlying scene structure while synthesizing plausible unseen regions under user-controlled camera motion. The generated views follow the prescribed camera instructions accurately, and newly discovered areas remain semantically consistent with the input image while exhibiting high visual quality and diverse content. Moreover, when the camera revisits previously observed regions after long-horizon exploration, **Wonder** preserves scene details and appearance consistency, demonstrating the effectiveness of our sparse memory mechanism. Benefiting from multi-task training, **Wonder** also exhibits dynamic world-modeling capability in the I2V setting, producing meaningful object and subject motion that is consistent with the scene semantics.



Figure 9 Visual comparison of Image-to-Video world modeling performance. We evaluate the memory capability of each model with a revisit trajectory. Specifically, the camera first pans right to observe a target region, then pans left to move away, and finally pans right again to revisit the same region. We compare the two frames captured at the same camera position to assess whether the scene content is consistently preserved. Most prior models struggle to retrieve long-term visual memory, leading to inconsistent structures or changed appearances after revisiting. RELIC (Hong et al., 2025) and LingBot-World (Robbyant Team, 2026) further suffer from degraded visual quality under long-horizon rollout, highlighting the difficulty of maintaining both memory consistency and generation quality over extended interactive exploration.

5.2 Video-to-Video World Modeling

Benchmark & Metrics. Similar to the I2V setting, we construct a video-to-video benchmark with 500 videos covering diverse scenes. Unlike the I2V benchmark, all videos contain dynamic subjects, such as humans, animals, and vehicles, allowing us to evaluate whether a model can preserve source-video dynamics while supporting camera-controlled world exploration. We use a VLM to generate video captions and user navigation keys for each input video. Each video is paired with 6 camera trajectories, which are grouped into three types: *retreat*, where the camera moves backward to reveal the full object or scene; *follow*, where the camera tracks the moving subject and keeps it visible; and *free*, where the camera performs unconstrained exploratory navigation. We use the same visual-quality and camera-following metrics as in the I2V setting. Specifically, VBench metrics are used to evaluate generation quality, while translational and rotational RPE are computed from estimated camera trajectories to measure action-following accuracy.

Baselines. We omit short-horizon bidirectional re-shotting models (Bai et al., 2025a; Huang et al., 2025d; Wang et al., 2026a), as they are not designed for long-video generation or real-time streaming. At the time of this report, the only open-source baseline that supports long-horizon V2V world modeling is Inspatio-World (Team et al., 2026b). Inspatio-World adopts point-cloud re-rendering as its camera-control signal and uses a sliding-window strategy by discarding distant video tokens for efficient inference. However, as discussed in previous sections, point-cloud re-rendering is inherently limited by the reconstructed view coverage and is therefore less suitable for open-ended free exploration. Meanwhile, discarding distant tokens removes long-term visual memory, making it difficult to preserve consistency when revisiting previously observed regions. Notably, unlike **Wonder**, Inspatio-World is not designed for streaming inference, as its point-cloud re-rendering pipeline requires access to the entire input video.

Results. Table 1 reports the quantitative results under the video-to-video setting. Compared with Inspatio-World, **Wonder** achieves better overall visual quality, improving the average score from 0.8374 to 0.8527. The gains are consistent across most visual metrics, including imaging quality, aesthetic quality, motion smoothness, and temporal flickering. This indicates that **Wonder** can better preserve the dynamic content of the input video while synthesizing high-quality target views under user-controlled camera motion. More importantly, **Wonder** substantially improves camera-following accuracy. It reduces translational RPE from 0.0436 to 0.0187 and rotational RPE from 0.2470 to 0.1119, showing much stronger alignment with the prescribed camera trajectories. These improvements demonstrate the advantage of our semi-explicit lattice control over point-cloud re-rendering, especially when the camera moves beyond the reconstructed view coverage. The qualitative results shown in Figure 10 and Figure 11 further demonstrate that



Figure 10 Visual comparison of Video-to-Video world modeling performance. Due to errors in the point clouds reconstructed by external pose-estimation models and the limited exploration space of point-cloud representations, Inspatio-World (Team et al., 2026b) often struggles to produce plausible camera motion and to preserve dynamic objects under large viewpoint changes.

Wonder enables streaming V2V world modeling with better dynamic preservation, stronger camera following, and more stable long-horizon visual consistency. We believe **Wonder** also establishes a strong baseline for video re-shooting.

6 Conclusion

We introduced **Wonder**, a real-time camera-controllable video world model that transforms a single image into an explorable world and re-renders an observed video along user-specified camera trajectories while preserving its appearance and dynamics. **Wonder** jointly addresses controllability, persistence, and efficiency through a unified system design. A dense pixel-space coordinate field provides spatially aligned camera cues, a sparse full-fidelity memory mechanism supports coherent revisits with bounded active attention cost, and an improved autoregressive distillation pipeline preserves visual quality and camera fidelity in a few-step causal student. Across both image- and video-conditioned benchmarks, **Wonder** achieves stronger visual quality and camera-following accuracy than recent streaming world models, while preserving dynamic content under novel viewpoints and maintaining scene consistency over long rollouts. These advances enable minute-scale generation at 16 FPS with stable inference latency, unifying open-ended exploration beyond observed views and real-time re-rendering of dynamic scenes within a single model.

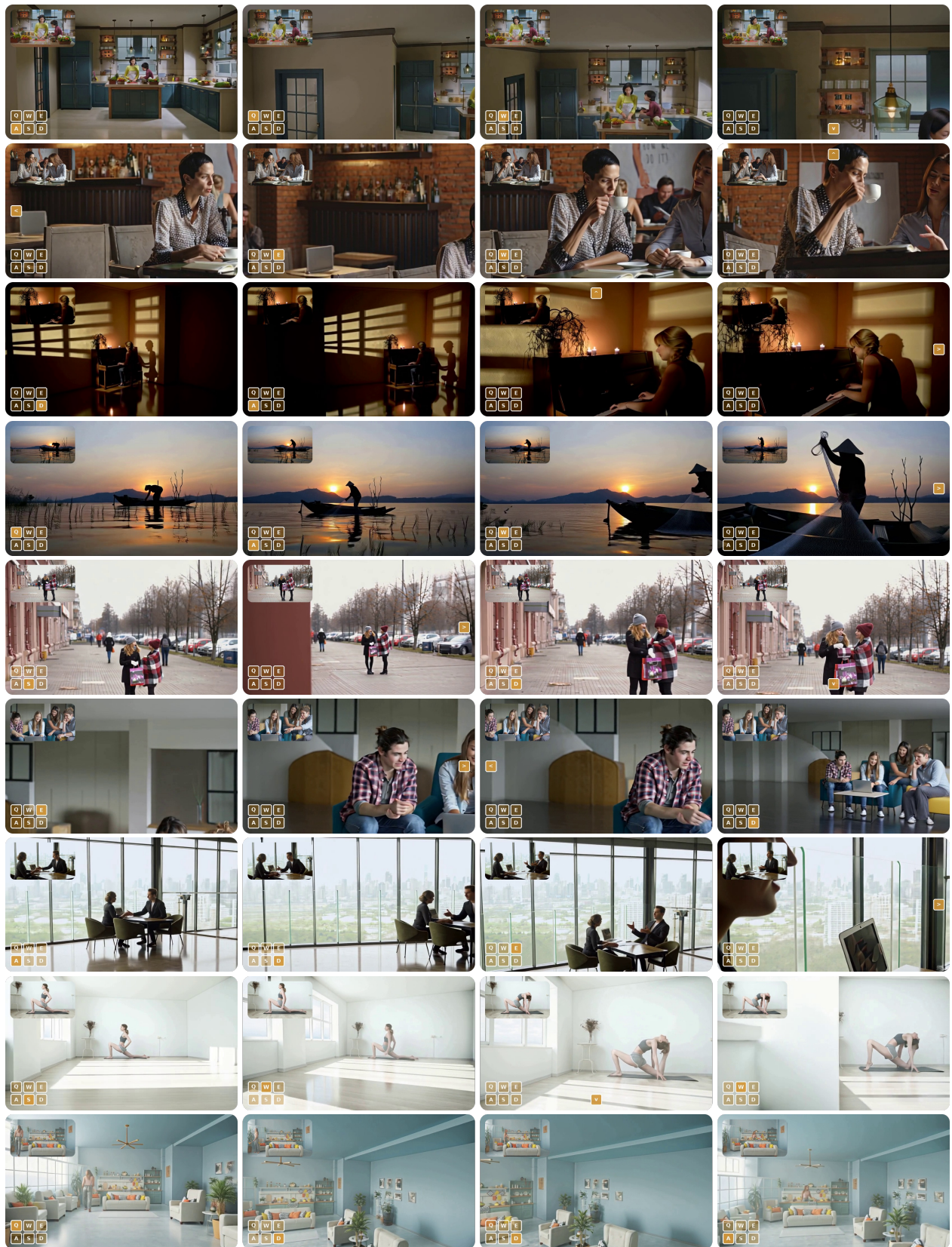


Figure 11 Navigation into videos.

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*, pages 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 Buttmore, 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 Mehning, 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 Ghahramani, Raia Hadsell, Aaron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models, 2025. <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. <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, Jiabin 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. Acddit: 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.
- Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. *arXiv preprint arXiv:2510.03198*, 2025b.
- 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*, 2025c.
- Zhenning 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*, 2025d.

- 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. D13dv-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*, pages 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. <https://api.semanticscholar.org/CorpusID:271098045>.
- Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. *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, Fei Wu, 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 Rajić, Siyu Tang, and Gordon Wetzstein. Bullethtime: Decoupled control of time and camera pose for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 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*, pages 1–11, 2024.
- Zile Wang, Zexiang Liu, Jiaying 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, pages 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, pages 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*, pages 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*, pages 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.