

TeleWorld: Towards Dynamic Multimodal Synthesis with a 4D World Model

TeleWorld Team

World models aim to endow AI systems with the ability to represent, generate, and interact with dynamic environments in a coherent and temporally consistent manner. While recent video generation models have demonstrated impressive visual quality, they remain limited in real-time interaction, long-horizon consistency, and persistent memory of dynamic scenes, hindering their evolution into practical world models. In this report, we present TeleWorld, a real-time multimodal 4D world modeling framework that unifies video generation, dynamic scene reconstruction, and long-term world memory within a closed-loop system. TeleWorld introduces a novel generation-reconstruction-guidance paradigm, where generated video streams are continuously reconstructed into a dynamic 4D spatio-temporal representation, which in turn guides subsequent generation to maintain spatial, temporal, and physical consistency. To support long-horizon generation with low latency, we employ an autoregressive diffusion-based video model enhanced with Macro-from-Micro Planning (MMPL)—a hierarchical planning method that reduces error accumulation from frame-level to segment-level—alongside efficient Distribution Matching Distillation (DMD), enabling real-time synthesis under practical computational budgets. Our approach achieves seamless integration of dynamic object modeling and static scene representation within a unified 4D framework, advancing world models toward practical, interactive, and computationally accessible systems. Extensive experiments demonstrate that TeleWorld achieves strong performance in both static and dynamic world understanding, long-term consistency, and real-time generation efficiency, positioning it as a practical step toward interactive, memory-enabled world models for multimodal generation and embodied intelligence.

Corresponding Author: Xuelong Li(xuelong_li@ieee.org)



1 Introduction

The pursuit of artificial intelligence systems capable of understanding, simulating, and interacting with the physical world has driven significant progress in world modeling research [Ding et al. \(2025\)](#); [Zhu et al. \(2025\)](#). Recent advances have demonstrated the promise of world models, highlighting that explicit reconstruction of the world and real-time generation are complementary and mutually reinforcing capabilities. At their core, world models aim to endow AI systems with human-like perception and interaction abilities—enabling machines to not only observe and represent dynamic environments but also to predict, generate, and meaningfully engage with them in real time. [Guo et al. \(2025b\)](#); [Xiao et al. \(2025\)](#); [Guo et al. \(2025a\)](#); [Zuo et al. \(2025\)](#); [Won et al. \(2025\)](#); [Wang et al. \(2025c\)](#); [Hu et al. \(2025\)](#); [Jin et al. \(2025\)](#); [Chen et al. \(2025a\)](#)

Let's start by discussing what a world model is. The definition of a world model varies across research communities [Ding et al. \(2025\)](#); [Zhu et al. \(2025\)](#), reflecting the multifaceted nature of this emerging field. Different researchers have varying interpretations, but broadly speaking, world models encompass several interconnected research directions, including but not limited to video generation [Wang et al. \(2025c\)](#), 3D reconstruction [Zuo et al. \(2025\)](#), embodied AI [Guo et al. \(2025b\)](#), and autonomous driving [Chen et al. \(2025a\)](#); [Jin et al. \(2025\)](#). In a general sense, any model that can naturally represent the world and interact with it may be considered a world model. However, the video generation direction has become a more popular research area within the field of world models, thanks to its higher quality output, stronger downstream multi-task capabilities (i.e., results from video generation can also be applied in areas such as embodied AI and autonomous driving), and its greater accessibility and interactivity for users.

However, video generation models themselves have several fundamental shortcomings that hinder their evolution into more practical world models [Yin et al. \(2025\)](#); [Huang et al. \(2025c\)](#); [Xiang et al. \(2025\)](#).

First, due to the structural limitations of multi-step denoising pipelines in video diffusion models, video generation is heavily restricted in meeting the real-time generation and interaction requirements of a world model. Second, long-term video generation still faces significant challenges with temporal consistency over extended durations. Extended world exploration and interaction often suffer from error accumulation and quality degradation. Third, a world model needs to retain a certain memory of the generated world, which is inherently four-dimensional—spanning the three dimensions of space and the dynamic dimension of time, just as human perception of the world. Existing world models and video generation approaches often only capture memory from past video sequences or three-dimensional representations, while achieving four-dimensional memory remains a significant difficulty in the video generation path toward world models. Finally, high-quality video generation models are typically computationally expensive, making it difficult to train and deploy them in a fast, efficient, and sustainable manner with real-time capability. The hardware demands for world models following the video generation approach remain prohibitively high for many researchers.

Here we summarize the key issues as: (1) **Modeling Dynamic 4D Scenes**: Current world models, which primarily possess 3D modeling, struggle to effectively model and memorize dynamic environments with full spatio-temporal coherence. (2) **Ensuring Long-term Consistency**: Maintaining both high fidelity and temporal consistency over extended generation periods remains difficult, often leading to issues like color shift and quality degradation. (3) **Balancing Real-time Efficiency with Quality**: Achieving real-time generation and efficient training is a primary challenge, as it requires reconciling high model quality with manageable computational cost.

In this report we propose TeleWorld, a practical real-time 4D world model that addresses these fundamental challenges through a unified framework integrating generation, reconstruction, and guidance. Firstly, we propose a "Generation-Reconstruction-Guidance" closed-loop paradigm that records the dynamic scene during video generation using a 4D spatio-temporal field. This reconstruction process runs synchronously with generation, continuously updating the world representation as new content is synthesized. The rendering results of this 4D field are then used as guidance to steer subsequent generation, ensuring spatial consistency, temporal coherence, and physical plausibility. This reconstruction-based approach achieves long-term dynamic memory through persistent, coherent understanding of the generated world. During the generation stage, we employ an autoregressive diffusion video generation model equipped with planning capabilities. Drawing insights from recent advances in planning-based generation, our Macro-from-Micro Planning (MMPL) framework [Xiang et al. \(2025\)](#) operates hierarchically: micro-planning predicts key anchor frames within short video segments to establish local temporal coherence, while macro-planning chains these segments autoregressively to achieve global consistency across long horizons. This approach reduces error accumulation from the frame level to the segment level, enabling stable, high-quality generation over extended durations. Our video generation architecture enables faster video synthesis while allowing better integration of information from the 4D scene during the planning process.

To further accelerate video synthesis, we adopt Distribution Matching Distillation (DMD) [Yin et al. \(2024\)](#) on top of TeleWorld. While DMD is critical for real-time video generation, applying it to an autoregressive model with more than 10B parameters is highly non-trivial, as it simultaneously introduces a large KV cache and requires working with three 10B-plus models—the generator, teacher, and critic. Even with Fully Sharded Data Parallelism (FSDP [Zhao et al. \(2023\)](#)), this combined memory footprint exceeds the capacity of 64 NVIDIA H100 GPUs ([Millon \(2025\)](#)). To address this challenge, we propose a novel training system for large-scale Distribution Matching Distillation. Specifically, we assign the generator, teacher, and critic to disjoint sets of GPUs and orchestrate their execution using Ray ([Moritz et al. \(2017\)](#)). In addition, we employ context parallelism to shard the generator's KV cache across devices, substantially reducing per-GPU memory consumption. Furthermore, we carefully design a pipeline execution schedule that minimizes GPU idle time (i.e., pipeline bubbles) and improves overall training efficiency. With these optimization techniques, we successfully train DMD for Teleworld-18B using only 32 H100 GPUs. Together, these system-level optimizations enable real-time video generation with modest training overhead under practical computational budgets.

Through these innovations, TeleWorld achieves seamless integration of dynamic object modeling and static scene representation within a coherent 4D framework, advancing world models toward practical, interactive, and computationally accessible systems suitable for multimodal generation and embodied intelligence applications.

The contributions of our method can be summarized as follows:

- We propose a real-time “generation–reconstruction–guidance” closed-loop framework that reconstructs long-term memory from the world model into dynamic point clouds at real-time speed while maintaining rapid world updates and temporal consistency.
- We introduce a dynamic four-dimensional world model that not only provides memory and generation capabilities in three-dimensional space but also enables the memorization and generation of moving objects within the scene, achieving true spatio-temporal coherence.
- We propose a novel training system that unlocks distillation training of large-scale autoregressive diffusion models, allowing efficient training on accessible hardware configurations while enabling real-time generation capabilities without compromising model quality.
- TeleWorld represents a comprehensive approach to world modeling that bridges video generation, 3D reconstruction, and persistent memory within a single unified system, positioning it as a practical foundation for interactive AI systems and embodied intelligence applications.

2 Related Works

2.1 World Models

The question of what constitutes a world model is a topic of frequent discussion among researchers today. The mainstream discussion centers on the idea that a world model is, in essence, an environment that can be navigated and interacted with. Consequently, much of the research has focused on how to construct such an environment. With the rise of generative models in recent years, world models have gradually branched into two main categories: 3D-based world models and video-based world models. The former first constructs a three-dimensional world and then renders it for the user, while the latter builds the world through video generation.

3D-based World Models A notable example in this field is Wonderworld [Yu et al. \(2025\)](#), which exhibits the ability to produce interactive 3D environments from just one 2D image. This highlights the possibility of building navigable virtual worlds with very limited initial data. The methodology prioritizes maintaining spatial coherence, accurate geometric interpretation, and low-latency feedback for user movement and actions.

Progress in this area has since broadened these functionalities. For instance, Matrix-3D [Yang et al. \(2025a\)](#) accomplishes extensive, all-directional 3D world creation that users can explore, utilizing panoramic 3D reconstruction techniques. Meanwhile, HunyuanWorld 1.0 [Team et al. \(2025\)](#) delivers fully immersive 360° environments by employing semantically structured 3D mesh models, ensuring smooth integration with standard computer graphics workflows. In parallel, World Labs has entered the commercial space with its first product, Marble. This multimodal world model can create high-fidelity, persistent 3D worlds from a single image, video clip, or text prompt. The company differentiates itself by focusing on generating persistent, downloadable 3D environments.

Video-based world models For instance, Cosmos [NVIDIA et al. \(2025\)](#) has achieved breakthrough performance in realistic simulations for robotics and autonomous systems. Meanwhile, Genie 3 [Ball et al. \(2025\)](#) has introduced real-time interaction capabilities, allowing users to generate and navigate controllable 3D worlds with high consistency. In contrast, models such as Hunyuan-Voyager [Huang et al. \(2025b\)](#), which outputs 3D point clouds via RGB-D video, Hunyuan-GameCraft2 [Tang et al. \(2025\)](#) designed for game videos with hybrid historical conditioning, and Adobe’s RELIC [Hong et al. \(2025\)](#), which employs a compact KV cache for long-term memory prioritize explicit 3D consistency and spatial reconstruction. The video-based approach offers distinct advantages for dynamic, user-centric applications: it delivers higher perceptual quality in motion and temporal coherence, supports more intuitive and responsive interaction due to its frame-by-generation nature, and enables rapid “cold-start” scene expansion—effectively allowing seamless “dream-outward” extension from minimal initial inputs.

However, current video-based world models are primarily limited to handling static 3D environments and often struggle to effectively model dynamic objects within these worlds.

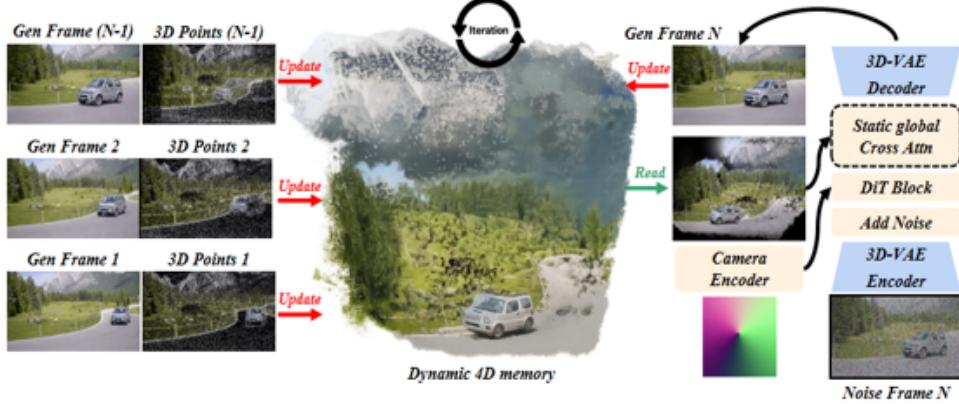


Figure 1 Structure of TeleWorld. We propose a dynamic "Generation-Reconstruction-Guidance" closed-loop framework for 4D spatio-temporal modeling. The model first generates an initial set of videos based on the user's pre-defined instructions. It then enters a loop where, in each iteration, it processes the user's real-time input instructions, reconstructs the video output from the previous round, and renders it according to the input camera poses. The rendered results serve as guidance to direct the current round of video generation and motion synthesis, and this process repeats iteratively.

2.2 Real-time Video Generation

Recent advances in long-video generation have largely been driven by autoregressive diffusion models Teng *et al.* (2025). Techniques such as Causvid Yin *et al.* (2025) and Self-Forcing Huang *et al.* (2025c) have been introduced to improve training stability and temporal coherence by conditioning each new frame on previously generated content. While these methods can produce extended sequences, they remain susceptible to error propagation over long horizons, where small inconsistencies in early frames gradually amplify and degrade visual quality. Moreover, maintaining long-range temporal consistency remains a fundamental challenge—models often "forget" earlier scene geometry or object identities, leading to incoherent narratives or visual artifacts in longer generations.

In parallel, real-time video generation has aimed to deliver low-latency, interactive synthesis. Yet scaling such systems to high-quality, high-resolution output—especially with large-parameter models such as 10B-plus architectures—presents significant difficulties. The real-time distillation of such models is particularly demanding, as it requires compressing both spatial and temporal knowledge without sacrificing fidelity, while also managing severe computational and memory constraints during deployment. Although Millon (2025) overcomes this challenge on a 14B model with dynamic KV cache management, the fundamental problem is not solved without sharding the KV cache.

These challenges—error accumulation, long-term memory decay, and the difficulty of distilling high-quality large models for real-time use—motivate the design of TeleWorld. Instead of relying solely on implicit neural representations or recurrent latent states, TeleWorld introduces an explicit 4D spatiotemporal field that continuously records and reconstructs the evolving world. This explicit representation preserves geometric and appearance information across time, effectively mitigating common failure modes such as forgetting and inconsistency, while enabling efficient, high-fidelity long-video generation and real-time inference. Moreover, thanks to the use of context parallelism to shard the generator's KV cache across devices, TeleWorld-18B can be trained for long-video distillation using only 32 H100 GPUs.

3 Methods

3.1 "Generation-Reconstruction-Guidance" Loop

We introduce a dynamic "Generation-Reconstruction-Guidance" closed-loop framework for unified 4D spatial-temporal modeling. This framework constructs a real-time, native 4D world representation that updates continuously with each newly generated video segments, ensuring perfect synchronization with the evolving

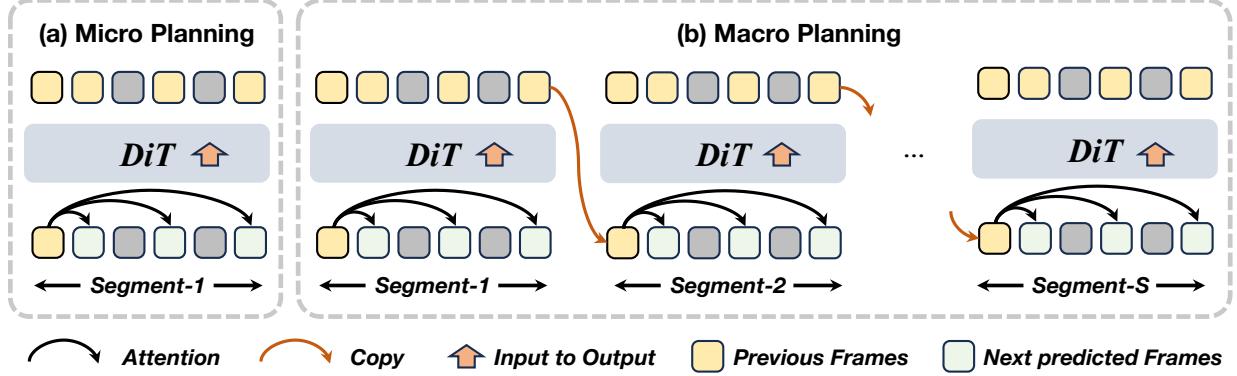


Figure 2 Our macro-from-micro planning framework is organized into two levels: (1) Micro Planning, where a sequence of frames is generated within each local segment to constrain error propagation; and (2) Macro Planning, which links segments through an autoregressive chain—each step’s output frames guide the prediction of the next, ensuring long-range temporal consistency. As shown in the figure, the three predicted frames marked in green correspond to the initial pre-planning frames, $\mathcal{P}_{\mathcal{M}_s} = \{x_s^{t_a}, x_s^{t_b}, x_s^{t_c}\}$, which serve as keyframes to maintain long-term memory and stability throughout the video sequence.

visual content. A core innovation is the seamless alignment between dynamic object modeling and static scene modeling, enabling their unified integration within a coherent spatial structure. In this loop, reconstruction refers to the process of recovering a consistent 4D scene representation from generated frames, while guidance denotes the use of both the reconstructed 4D scene and the user’s keyboard commands to direct the next round of video generation. The generation and reconstruction steps proceed in real time, with only minimal latency between guidance and generation. This cyclic process continuously updates a 4D spatial-temporal memory of the constructed dynamic scene, allowing effective motion and interaction to be driven interactively via keyboard control.

3.2 Long-memory Auto-regressive Video Generation

3.2.1 Micro and Macro Planning

Motivated by the analysis in the MMPL Xiang *et al.* (2025), we observe that autoregressive models accumulate errors proportionally to the number of propagation steps, whereas non-autoregressive models decouple errors from the step count through joint optimization. To exploit the complementary strengths of both paradigms, we introduce *Macro-from-Micro Planning* (MMPL) into our TeleWorld, a unified planning method comprising two key components: *Micro-Planning* and *Macro-Planning*.

Micro Planning. Micro Planning \mathcal{M}_s builds a short-term narrative for the s -th segment by predicting a sparse set of key frames $\mathcal{P}_{\mathcal{M}_s} = \{x_s^{t_a}, x_s^{t_b}, x_s^{t_c}\}$ from the initial frame x_s^1 . These *pre-planning frames* serve as stable anchors for subsequent synthesis, with timestamps set as $t_a = 2$ (early neighbor), $t_b = N/2$ (midpoint), and $t_c = N$ (segment end). The process is formulated as:

$$p(\mathcal{P}_{\mathcal{M}_s} | x_s^1) = p(x_s^{t_a}, x_s^{t_b}, x_s^{t_c} | x_s^1). \quad (1)$$

All frames are jointly optimized conditioned only on x_s^1 , which mutually constrains their residual errors and eliminates cumulative drift—unlike sequential autoregressive generation. This design ensures within-segment coherence and provides a drift-resistant foundation for later content population.

Macro Planning. While Micro Planning provides a segment-level temporal storyline, it remains limited in capturing global dependencies across the entire videos of the world scene. To achieve long-range coherence, we extend it into **Macro Planning**, denoted \mathcal{M}^+ . This constructs a global storyline by chaining overlapping Micro Plans sequentially across segments: the terminal pre-planning frame of one segment initializes the next, forming a segment-level autoregressive chain along the video timeline.

Formally, given a full video of length T partitioned into S segments, let x_s^1 be the initial frame of the s -th segment. The set of planning frames produced by Macro Planning is denoted $\mathcal{P}_{\mathcal{M}^+}$. The process is defined

as:

$$p(\mathcal{P}_{\mathcal{M}^+} \mid x_1^1) = \prod_{s=1}^S p(\mathcal{P}_{\mathcal{M}_s} \mid x_s^1), \quad x_{s+1}^1 := x_s^{t_c}, \quad \mathcal{P}_{\mathcal{M}^+} := \bigcup_{s=1}^S \mathcal{P}_{\mathcal{M}_s}. \quad (2)$$

Here, \mathcal{M}_s is the Micro Planning for segment s . By linking segments hierarchically, Macro Planning converts frame-by-frame autoregressive dependencies into a sparse sequence of segment-level planning steps. This ensures consistent global narrative flow, mitigates temporal drift, and reduces error accumulation from the scale of T frames to only S segments, where $S \ll T$.

This hierarchical linking enables the world model to retain long-term memory across segments. Subsequently, we anchor these memories through an online 4D reconstruction of the cross-segment anchor frames, embedding all keyframes within a coherent spatio-temporal field. This further clarifies and stabilizes the inter-segment and intra-segment memory, ensuring its precision and consistency.

However, when chaining Micro Plannings autoregressively, directly using the tail latent tokens of one segment as the prefix for the next often introduces boundary flickering and color shifts due to distribution mismatch between initial and temporally-compressed latent frames.

To stabilize inter-segment transitions, we adopt a drift-resilient re-encoding and decoding strategy. Specifically, we reconstruct a short video clip from the concatenated initial and terminal planning tokens of the current segment. To ensure temporal continuity during decoding, the terminal tokens are duplicated and inserted to form a contiguous latent sequence. The re-encoded latents of the second copy then serve as the initial condition for the next segment. For implementation details, we refer readers to our previous work MMPL [Xiang et al. \(2025\)](#).

3.2.2 MMPL-based Content Populating

Following Sec. 3.2.1, the Micro Plan \mathcal{M}_s divides each video segment into two sub-segments—e.g., $[x_s^{t_a}, x_s^{t_b}]$ and $[x_s^{t_b}, x_s^{t_c}]$ —bounded by consecutive planning frames. To synthesize the full segment by filling the remaining frames under the guidance of these planning anchors, we introduce MMPL-based Content Populating.

Micro Planning provides three types of key frames: *early* ($x_s^{t_a}$), *midpoint* ($x_s^{t_b}$), and *terminal* ($x_s^{t_c}$). Motivated by earlier frame-conditioned generation approaches, we perform content population in two sequential stages:

1. Populate the first sub-segment using the initial frame and the early planning frame as the start, and the midpoint planning frame as the end.
2. Extend the sequence by taking all frames up to the midpoint as the new start and the terminal frame as the end, thereby generating the remaining content.

The process can be formally expressed as:

$$p(\mathcal{C}_s \mid \mathcal{P}_{\mathcal{M}_s}) = p(x_s^{t_a+1:t_b-1} \mid x_s^{1:t_a}, x_s^{t_b}) \cdot p(x_s^{t_b+1:t_c-1} \mid x_s^{1:t_b}, x_s^{t_c}), \quad (3)$$

Here, \mathcal{C}_s denotes the content frames to be generated in segment s , while $x_s^{t_a}$, $x_s^{t_b}$, and $x_s^{t_c}$ represent its early, midpoint, and terminal planning frames, respectively. The notation $x_s^{1:t_a}$ and $x_s^{1:t_b}$ indicates that the generation of each sub-segment is conditioned on all preceding frames within the segment, in addition to its boundary planning frames. The intermediate frames $x_s^{t_a+1:t_b-1}$ and $x_s^{t_b+1:t_c-1}$ correspond to the content to be populated.

Importantly, the factorization in Eq. 3 shows that content population within each sub-segment depends solely on its corresponding planning frames. This allows multiple sub-segments to be optimized in parallel once their internal planning frames are ready. By distributing segment-wise optimization across multiple GPUs, the proposed MMPL-based Content Populating enables concurrent execution, significantly accelerating the synthesis of long videos.

3.3 Real-time 4D Reconstruction

3.3.1 Key-frame Reconstruction

As discussed in Introduction, we propose a real-time 4D reconstruction module to further provide dynamic memory of moving objects within the scene. Considering the planning strategy in the MMPL architecture,

our reconstruction process also follows macro planning synchronously. The reconstruction task continuously progresses backward along with the macro structure, allowing the reconstruction speed to closely follow the generation process. Meanwhile, micro planning uses the rendered results of the reconstruction under corresponding manipulations as guidance.

In this way, the overhead of reconstruction is minimized, and the input to reconstruction is kept as sparse as possible to prevent the reconstruction task from failing over long sequences due to extended world generation. We term this approach key-frame reconstruction.

Specifically, only the sparse set of *pre-planning frames* $\mathcal{P}_{\mathcal{M}_s} = \{x_s^{t_a}, x_s^{t_b}, x_s^{t_c}\}$ need to conduct 4D reconstruction. These planning frames essentially serve as anchors within the video—they are generated first with minimal error and highest quality, and they determine the motion trajectory of the video. Using them for 4D reconstruction also introduces sufficiently rich records for long-video generation tasks in world models. The beginning, middle, and end of each video segment will be used to record information in the 4D spatiotemporal field. During content population, the intermediate motion is then filled in based on these recorded cues.

3.3.2 Move Object Segmentation

Inspired by 4D-VGGT Wang *et al.* (2025b), we utilize its dynamic saliency map as the dynamic masks. To aggregate temporal information, we employ an interframe sliding-window strategy across frames, defined as $\mathcal{W}(t) = \{t - n, \dots, t - 1, t + 1, \dots, t + n\}$. Within this window and across three set of layers L , including shallow, middle, and deep layers Shallow, middle, and deep correspond to different layer ranges (i, j) . w_{shallow} captures semantic saliency, w_{middle} reflects motion instability, and w_{deep} provides a spatial prior to suppress outliers. Finally, a per-frame dynamic mask is obtained by thresholding: $M_t = [\text{Dyn} > \alpha]$, followed by feature clustering for refinement. A network-level early-stage masking strategy for 4D reconstruction and stacking is also conducted in our framework. Static scene elements are merged and progressively expanded, while sparse dynamic components are separately rendered over time. However, since our input is limited to *pre-planning frames* $\mathcal{P}_{\mathcal{M}_s} = \{x_s^{t_a}, x_s^{t_b}, x_s^{t_c}\}$, the rendered dynamic content remains highly sparse. This requires predicting subsequent dynamic regions based on earlier frames within the pre-planning sequence—a challenge we address through macro-planning in video generation. From a macroscopic perspective, smooth continuous motion is decomposed into keyframe-like dynamic segments embedded within the scene.

Specifically, following 4D-VGGT Wang *et al.* (2025b), to mitigate geometric inconsistencies introduced by dynamic pixels, we also mask dynamic image tokens only in shallow and mid-level layers (layers 1~5) by suppressing their Key (K) vectors.

3.4 Guidance

3.4.1 Keyboard Control

As the widespread adoption of keyboard control in world models Mao *et al.* (2025); Tang *et al.* (2025); Hong *et al.* (2025), we also utilize the four WASD keys along with the arrow keys to simulate movement and perspective changes, as illustrated below. These inputs are correspondingly mapped to camera poses.

These signals are conditioned to guide the model’s generation. We map these controls into camera motion movements along with input frame depth scales.

$$\begin{aligned} \text{perspective changes} &= \left\{ \begin{array}{l} \rightarrow : \text{Camera turns right } (\rightarrow). \\ \leftarrow : \text{Camera turns left } (\leftarrow). \\ \uparrow : \text{Camera tilts up } (\uparrow). \\ \downarrow : \text{Camera tilts down } (\downarrow). \\ \uparrow\rightarrow : \text{Camera tilts up and turns right } (\uparrow\rightarrow). \\ \downarrow\rightarrow : \text{Camera tilts down and turns right } (\downarrow\rightarrow). \\ \downarrow\leftarrow : \text{Camera tilts down and turns left } (\downarrow\leftarrow). \\ \cdot : \text{Camera remains still } (\cdot). \end{array} \right. \\ \text{camera movement} &= \left\{ \begin{array}{l} \text{W : Camera moves forward (W).} \\ \text{A : Camera moves left (A).} \\ \text{S : Camera moves backward (S).} \\ \text{D : Camera moves right (D).} \\ \text{W+A : Camera moves forward and left (W+A).} \\ \text{W+D : Camera moves forward and right (W+D).} \\ \text{S+D : Camera moves backward and right (S+D).} \\ \text{S+A : Camera moves backward and left (S+A).} \\ \text{None : Camera stands still (None).} \end{array} \right. \end{aligned}$$

Furthermore, to enhance the continuity and coherence of video generation as much as possible, we endeavor to avoid maintaining a static camera position. Therefore, even when no keyboard input is provided by the user, the camera pose will drift forward at a very slow speed—a feature we refer to as the standby animation.

3.4.2 View-Conditioned Guidance

Subsequently, we need to encode the processed keyboard inputs for the world model network. As discussed in ReCamMaster¹, the conditioning by frame dimension is a more effective approach for integrating target camera poses into the DiT network. Following this insight, we adopt a similar structure and incorporate the following mechanism into TeleWorld’s DiT network:

To achieve better synchronization and content consistency with the keyboard guidance video, we propose to concatenate the guidance video tokens with the target video tokens along the frame dimension:

$$\begin{cases} x_s = \text{patchify}(z_s), & x_t = \text{patchify}(z_t), \\ x_i = [x_s, x_t]_{\text{frame-dim}}, \end{cases}$$

where $x_i \in \mathbb{R}^{b \times 2f \times s \times d}$ is the input of the diffusion transformer. In other words, the input token number is doubled compared to the vanilla video generation process. Moreover, no additional attention layers are needed for cross-video aggregation, as 3D self-attention inherently processes all tokens.

3.5 Distribution Matching Distillation

Our approach integrates seamlessly with existing Distribution Matching Distillation (DMD) frameworks without requiring any architectural modifications. Specifically, the MMPL video generation pipeline adjusts the attention visibility range and prediction order during both training and inference. Building on standard self-forcing pipelines, DMD can be directly applied on top of MMPL and deployed within the TeleWorld framework.

When combined with parallelized decoding, the resulting system delivers substantial inference speedups, achieving sustained throughput exceeding 32 FPS for long-horizon video generation on the TeleWorld-1.3B model and 8 FPS on the Teleworld-18B model, both evaluated on NVIDIA H100 GPUs.

Despite its importance for real-time video generation, DMD introduces significant challenges to the training infrastructure, esp. when applied to our 18B model. The training setup requires the simultaneous coordination of three diffusion models—the autoregressive generator, the critic, and the teacher—making it infeasible to host all components within a single 80-GB HBM GPU. To address this constraint, we employ Ray, Moritz *et al.* (2017), to distribute the model weights across multiple GPUs. Furthermore, leveraging the Ulysses sequence-parallel capabilities provided by TeleTron¹, we shard the generator’s KV cache across GPUs, enabling it to fit within memory limits.

To mitigate GPU underutilization caused by model parallelism, we design a novel **pipelined training schedule** that overlaps the computation of the generator, critic, and teacher models, thereby minimizing GPU idle time (i.e., pipeline bubbles). The execution schedules for the generator and critic steps are illustrated in Figure 3. For the generator step, enabling the degree of overlap shown in the figure requires carefully matching the combined execution time of the generator forward and backward stages to that of the critic/teacher stage through explicit resource allocation. In practice, we find that a generator:critic:teacher GPU ratio of 4:1:1 achieves near-perfect overlap. In addition, to simplify DMD optimization and ensure predictable stage durations, we fix the number of denoising steps in the generator pipeline during training rather than randomly sampling them. We note that two copies of the KV cache must be maintained to support correct backpropagation; however, this overhead is manageable since the KV cache is already sharded across devices using context parallelism. As a result, our pipelined system achieves an approximately 50% end-to-end training speedup compared to a non-pipelined baseline.

Taken together, efficient KV-cache sharding, model parallelism, and pipelined execution position our training system to scale naturally to future auto-regressive diffusion models with substantially larger parameter counts.

3.6 Streaming and Scheduled Generation with Online Video Super-resolution

3.6.1 Scheduled Generation

Although content populating across different segments can be parallelized (Sec. 3.2.2), a key limitation remains: parallel execution cannot begin until planning frames for all segments are fully generated, leading

¹<https://github.com/Tele-AI/TeleTron>

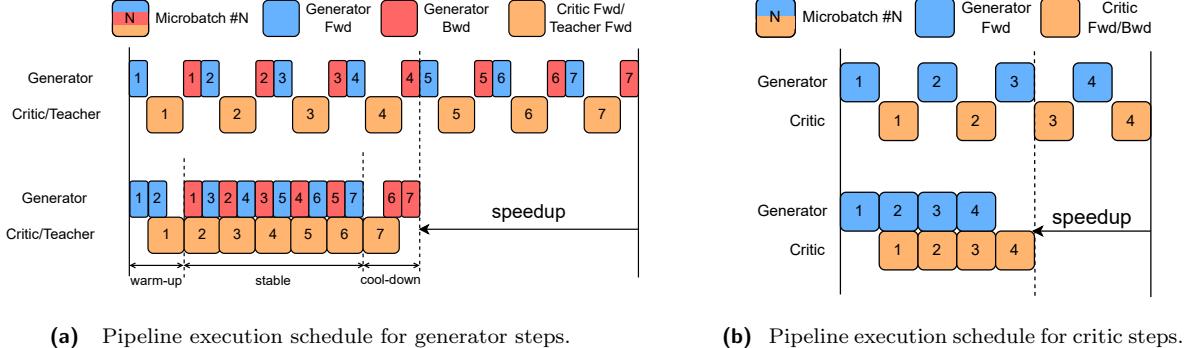


Figure 3 Pipeline execution schedules for Distribution-Matching Distillation. (a) Generator-step pipeline with 7 micro-batches. Cell length denotes execution time. The critic and teacher works in parallel, so their cells are merged together for simplicity, and their cell length denotes the maximum of their execution time. The upper half of the figure is the non-pipelining baseline, which introduces a lot of GPU bubbles (i.e. GPU idle time). The lower half is our proposed pipeline schedule. In the stable phase, the generator backward stage of micro-batch i and the generator forward stage of micro-batch $i+2$ are executed concurrently with the critic/teacher forward stage of micro-batch $i+1$. The execution time of all stages are carefully balanced by allocating appropriate numbers of GPUs to each component, enabling near-perfect overlap. This method minimizes GPU bubbles and achieves efficient parallelization of generator, teacher, and critic workloads in the proposed system. (b) Critic-step pipeline with 4 micro-batches. Since the generator parameters remain frozen during the critic update, the pipeline follows a simpler producer-consumer execution pattern.

to an unavoidable prefix delay that reduces overall throughput.

To address this, we introduce an *adaptive workload scheduling* strategy that dynamically orders the execution of Micro Planning, Macro Planning, and Content Populating to maximize parallelism. Since Macro Planning forms an autoregressive chain of segment-level Micro Plannings, the planning frames are generated sequentially across segments. This allows the Content Populating of an earlier segment to start as soon as its own planning frames are ready, without waiting for subsequent segments.

For illustration, with $t_a = 2$, $t_b = 6$, and $t_c = 10$, the planning frame $x_s^{t_c}$ from the current segment immediately serves as the initial frame x_{s+1}^1 for the next segment. Thus, the next segment can begin its Micro Planning while the current one is still populating its intermediate frames (e.g., $x_s^{t_a+1:t_b-1}$). This staged independence naturally enables segment-parallel generation, as formally expressed in Eq. (4):

$$\begin{aligned} \text{Segment } s: \quad & x_s^{t_a+1:t_b-1} \sim p_\theta(x | x_s^1, x_s^{t_a}, x_s^{t_b}), \\ \text{Segment } s+1: \quad & \{x_{s+1}^{t_a}, x_{s+1}^{t_b}, x_{s+1}^{t_c}\} \sim p_\theta(x | x_{s+1}^1), \quad x_{s+1}^1 \in \{x_s^{t_b}, x_s^{t_c}\}. \end{aligned} \quad (4)$$

Here, the initial frame x_{s+1}^1 of the next segment can be selected either as $x_s^{t_b}$ or $x_s^{t_c}$. In order to keep the real-time practical generation, we choose the maximum throughput prediction as follows:

To minimize latency as much as possible, we use the **Minimum Memory Peak Prediction** strategy. When $x_s^{t_b}$ is used as x_{s+1}^1 , intermediate frames $x^{t_b+1} : x^{t_c-1}$ are skipped, bypassing the region with the deepest temporal context and highest generation latency. This mode minimizes peak memory usage and reduces per-segment latency but introduces frame reuse between segments, slightly reducing overall throughput. As illustrated in Fig. 4, f_4^0 and f_6^1 are in fact generated synchronously. This means that any immediate user input manipulation is only rendered after three latent chunks, resulting in a feedback latency of approximately one second. Consequently, the world output currently being observed corresponds to the pre-buffered changes captured one second prior to the user's input.

3.6.2 Streamed VAE

To achieve real-time video generation for live streaming, we designed a streaming-capable VAE based on the principles of StreamDiffusionV2 Feng *et al.* (2025). The core challenge in a live setting is to minimize the “time to first frame” and ensure continuous, low-latency output, which is fundamentally different from batch-based video generation that processes long sequences offline. Our Stream-VAE is a low-latency Video-VAE

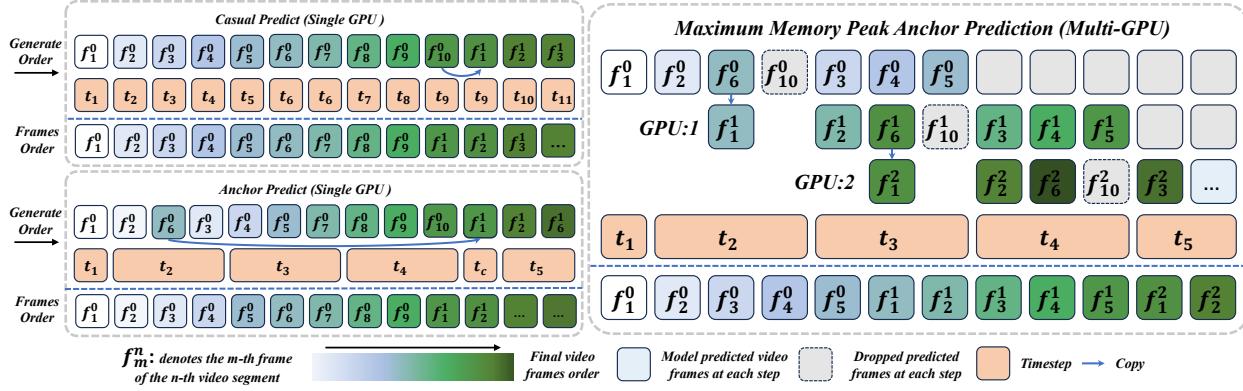


Figure 4 Multi-GPU parallel inference via adaptive workload scheduling. Given the initial frame f_1^0 , segment 0 first generates its planning frames f_2^0 , f_6^0 , and f_{10}^0 . These planning frames then guide the content population of the intermediate frames f_3^0 , f_4^0 , and f_5^0 . While segment 0 is still populating these frames, segment 1 can immediately start its Micro Planning by taking f_{10}^0 as the initial frame f_1^1 and generating its own planning frames f_2^1 , f_6^1 , and f_{10}^1 . This staged execution enables overlapping planning and populating across segments, maximizing multi-GPU parallelism. Here, each t_i denotes an inference step in the diffusion sampling process.

variant specifically optimized for streaming inference. Instead of encoding an entire video sequence at once, it operates on short, contiguous video chunks—typically 4 frames in our implementation. This chunk-wise processing is critical for maintaining a steady output stream.

The architecture of the Stream-VAE incorporates strategic caching of intermediate features within its 3D convolutional layers. As each new chunk of frames is fed into the model, the network reuses relevant temporal features computed from previous chunks, thereby preserving temporal coherence across chunk boundaries without the need to re-encode a long history. This design significantly reduces redundant computation and memory overhead, enabling efficient incremental encoding and decoding. By integrating this Stream-VAE into our pipeline, we ensure that the latent representations of the video are generated and can be delivered to users with minimal delay, forming the foundational stage of our real-time streaming system.

3.6.3 Video Super-resolution

For the subsequent enhancement of video quality, we incorporate a streaming super-resolution module inspired by FlashVSR. This component is responsible for upscaling the decoded latents from the Stream-VAE into high-resolution video frames in real time. A key innovation we adopt from FlashVSR is its locality-constrained sparse attention mechanism. This mechanism restricts the self-attention operations to local spatial-temporal windows, drastically reducing the computational complexity that typically plagues video super-resolution models. It effectively bridges the resolution gap often encountered between training and inference without sacrificing the quality of fine details.

Furthermore, we leverage FlashVSR’s lightweight conditional decoder, which is engineered for fast feature reconstruction. The decoder conditions its upscaling process on the features extracted from the Stream-VAE’s output, ensuring high-fidelity results while maintaining a low computational footprint. Crucially, this super-resolution module is designed to work in harmony with our Stream-VAE in a fully streaming manner. It processes short video chunks (e.g., 5 frames) that align with the VAE’s output stream, applying super-resolution incrementally as each chunk becomes available. This integrated, chunk-wise processing pipeline enables our model to achieve super-resolution decoding at approximately 17 FPS on 960×1760 resolution videos, making high-quality real-time video generation practical.

In summary, by integrating Scheduled Generation, Streamed VAE, and Video Super-resolution techniques, our system enables the TeleWorld-18B model to achieve stable 8 FPS performance and generate high-quality 960×1760 videos on a setup of four NVIDIA H100 GPUs.

4 Experiments and Discussion

4.1 Multi-modal Dataset Preparation

We introduce the data collections here. To support large-scale training and unified evaluation, we construct TeleWorld-500K, a curated dataset tailored for controllable camera and dynamics 4D annotated videos. TeleWorld-500K is built through two pipelines.

4.1.1 Curation Pipeline

(1) Data Collection. We assembled a large-scale collection of real-world video clips through a hybrid approach combining systematic web scraping and selective manual gathering. Sources included major public platforms such as YouTube, Pexels, Pixabay, Mixkit, and Bilibili, ensuring broad coverage of diverse visual content and scenarios.

(2) Automated Quality Filtering. From the initial pool, we applied a multi-stage automated filtering pipeline to eliminate low-quality content. The LAION aesthetic scorer was used to retain clips with aesthetic ratings above 6, while PaddleOCR [Liao et al. \(2022\)](#) detected and removed videos containing prominent overlaid text, watermarks, or subtitles. Additionally, extremely short, corrupted, or visually inconsistent clips were automatically discarded to maintain overall dataset integrity.

(3) Motion-Aware Selection. To ensure the dataset contains meaningful dynamics suitable for controllable camera and object modeling, we performed motion-based filtering. Using TTT3R [Chen et al. \(2025b\)](#), we estimated per-clip camera motion and excluded sequences with negligible viewpoint changes. Furthermore, to retain videos with salient foreground object motion, the vision-language model Qwen-2.5-VL-72B [Bai et al. \(2025\)](#) analyzed each clip and filtered out those without detectable moving subjects.

(4) Expert Review and Dataset Finalization. The remaining clips underwent thorough manual inspection by twenty domain experts over 690 person-hours to remove any residual low-quality or unsuitable content. This careful curation resulted in the final TeleWorld-500K dataset, comprising 500K high-quality video clips that feature diverse real-world environments, pronounced camera motion, and rich dynamic interactions, providing a robust foundation for training world models.

4.1.2 Annotation Pipeline

(1) Motion Object Segmentation. To annotate moving objects, we first employed Segment Any Motion in Videos [Huang et al. \(2025a\)](#), which takes a video as input and predicts masks for all moving foreground objects. It provides an initial mask on the first frame for each distinct object, with unique colors assigned to maintain consistent identity labeling across frames.

(2) Camera Trajectory Annotating. Using the first-frame object masks as initialization, we employed 4D-VGGT [Wang et al. \(2025b\)](#) to recover dense motion and camera annotations. 4D-VGGT is a unified camera trajectory annotating framework that jointly estimates point clouds, depth maps, camera intrinsics, and camera poses in an end-to-end manner. For each video, it reconstructs 3D trajectories of moving objects and estimates per-frame camera poses.

(3) Semantic Description Generation. To enable precise text description, we employed the large vision–language model Qwen-2.5-VL-72B [Bai et al. \(2025\)](#) to generate textual annotations that describe the appearance and motion of both moving object and camera motion, along with the overall scene context. These captions complement the 3D trajectories of moving objects, providing comprehensive semantic information aligned with scene-level dynamics.

4.2 WorldScore Benchmark

This section evaluates TeleWorld on the WorldScore [Duan et al. \(2025\)](#) benchmark, which is currently one of the most comprehensive protocols for measuring “world generation” ability. Unlike image or short-video benchmarks that primarily assess local visual quality, WorldScore evaluates whether a model can construct and maintain a consistent world across viewpoints, scene transitions, and temporal evolution. The benchmark includes both static and dynamic settings, as well as a rich set of metrics assessing controllability, consistency,

perceptual quality, and motion behavior. All results in this section are reported from the official WorldScore leaderboard to ensure comparability.

The WorldScore evaluation consists of two primary aggregate dimensions. First, WorldScore-Static measures whether the generated world remains stable and coherent while the camera moves through multiple viewpoints. This focuses on spatial fidelity, layout preservation, and cross-view semantic consistency. Second, WorldScore-Dynamic measures world evolution over time, including object motion, scene changes, and temporal stability. This dimension evaluates whether a model generates motion patterns that are coherent, semantically grounded, and structurally consistent with the underlying world. The official evaluation pipeline computes a set of sub-metrics and integrates them into the two final aggregate scores.

WorldScore reports 12 metrics. Camera Control, Object Control, and Content Alignment measure controllability. They jointly characterize how well a model follows layout constraints, preserves required entities, and responds to semantic instructions. 3D Consistency, Photometric Consistency, Style Consistency, and Subjective Quality measure structural and perceptual stability. These metrics reflect how well a model maintains consistent geometry, appearance, lighting, and aesthetics. Motion Accuracy, Motion Magnitude, and Motion Smoothness measure dynamic behavior, capturing temporal realism, motion amplitude suitability, and continuity. Together, these metrics serve as a comprehensive evaluation of static world structure and dynamic world evolution.

Model Name	WS-Static	WS-Dynamic	CamCtrl	ObjCtrl	ContAlign	3DCons	PhotoCons	StyleCons	SubjQual
TeleWorld	78.23	66.73	76.58	74.44	73.20	87.35	88.82	85.59	61.66
Voyager Huang et al. (2025b)	77.62	54.53	85.95	66.92	68.92	81.56	85.99	84.89	71.09
WonderWorld Yu et al. (2025)	72.69	50.88	92.98	51.76	71.25	86.87	85.56	70.57	49.81
LucidDreamer Chung et al. (2023)	70.40	49.28	88.93	41.18	75.00	90.37	90.20	48.10	58.99
WonderJourney Yu et al. (2023)	63.75	44.63	84.60	37.10	35.54	80.60	79.03	62.82	66.56
CogVideoX-I2V Yang et al. (2025b)	62.15	59.12	38.27	40.07	36.73	86.21	88.12	83.22	62.44
Text2Room Höllein et al. (2023)	62.10	43.47	94.01	38.93	50.79	88.71	88.36	37.23	36.69
InvisibleStitch Engstler et al. (2025)	61.12	42.78	93.20	36.51	29.53	88.51	89.19	32.37	58.50
Gen-3 Runway (2024)	60.71	57.58	29.47	62.92	50.49	68.31	87.09	62.82	63.85
Wan2.1 Wang et al. (2025a)	57.56	52.85	23.53	40.32	45.44	78.74	78.36	77.18	59.38
Hailuo HailuoAI (2024)	57.55	56.36	22.39	69.56	73.53	67.18	62.82	54.91	52.44
LTX-Video HaCohen et al. (2024)	55.44	56.54	25.06	53.41	39.73	78.41	88.92	53.50	49.08
Allegro Zhou et al. (2024)	55.31	51.97	24.84	57.47	51.48	70.50	69.89	65.60	47.41
CogVideoX-T2V Yang et al. (2025b)	54.18	48.79	40.22	51.05	68.12	68.81	64.20	42.19	44.67
EasyAnimate Xu et al. (2024)	52.85	51.65	26.72	54.50	50.76	67.29	47.35	73.05	50.31
VideoCrafter2 Chen et al. (2023)	52.57	47.49	28.92	39.07	72.46	65.14	61.85	43.79	56.74
DynamiCrafter Xing et al. (2023)	52.09	47.19	25.15	47.36	25.00	72.90	60.95	78.85	54.40
SceneScape Fridman et al. (2024)	50.73	35.51	84.99	47.44	28.64	76.54	62.88	21.85	32.75
VideoCrafter1-I2V Chen et al. (2023)	50.47	47.64	25.46	24.25	35.27	74.42	73.89	65.17	54.85
VideoCrafter1-T2V Chen et al. (2023)	47.10	43.54	21.61	50.44	60.78	64.86	51.36	38.05	42.63
T2V-Turbo Li et al. (2024)	45.65	40.20	27.80	30.68	69.14	38.72	34.84	49.65	68.74
Vchitect-2.0 Fan et al. (2025)	42.28	38.47	26.55	49.54	65.75	41.53	42.30	25.69	44.58
4D-fy Bahmani et al. (2024)	27.98	32.10	69.92	55.09	0.85	35.47	1.59	32.04	0.89

Table 1 Quantitative comparison on the WorldScore benchmark. We report the leaderboard scores for static and dynamic world generation (WorldScore-Static/Dynamic) and the corresponding controllability and consistency metrics (Camera Control, Object Control, Content Alignment, 3D/Photometric/Style Consistency, and Subjective Quality) for TeleWorld and representative baselines under the official evaluation protocol. Higher is better for all metrics.

4.2.1 Quantitative Results:

We compare TeleWorld against 23 baseline models across 3D, 4D, and video-based approaches. These baselines include 3D world generators such as Voyager [Huang et al. \(2025b\)](#), WonderWorld [Yu et al. \(2025\)](#), LucidDreamer [Chung et al. \(2023\)](#), WonderJourney [Yu et al. \(2023\)](#), Text2Room [Höllein et al. \(2023\)](#), InvisibleStitch [Engstler et al. \(2025\)](#), and SceneScape [Fridman et al. \(2024\)](#); 4D-oriented systems such as 4D-fy [Bahmani et al. \(2024\)](#); and a range of image-to-video and text-to-video systems including Gen-3 [Runway \(2024\)](#), Wan2.1 [Wang et al. \(2025a\)](#), Hailuo [HailuoAI \(2024\)](#), LTX-Video [HaCohen et al. \(2024\)](#), Allegro [Zhou et al. \(2024\)](#), CogVideoX [Yang et al. \(2025b\)](#), EasyAnimate [Xu et al. \(2024\)](#), DynamiCrafter [Xing et al. \(2023\)](#), VideoCrafter [Chen et al. \(2023\)](#), T2V-Turbo [Li et al. \(2024\)](#), and Vchitect [Fan et al. \(2025\)](#). All compared models are evaluated under the same protocol. TeleWorld is tested under the Video and I2V configuration using a single generation setup not specialized for the WorldScore benchmark.

TeleWorld achieves the strongest performance on both aggregate metrics, with a WorldScore-Static score

of 78.23 and a WorldScore-Dynamic score of 66.73. The next best models achieve 77.62 in the static setting (Voyager [Huang et al. \(2025b\)](#)) and 59.12 in the dynamic setting (CogVideoX-I2V [Yang et al. \(2025b\)](#)). TeleWorld therefore outperforms the strongest baselines by 0.61 points in static world generation and by 7.61 points in dynamic world generation. The relatively small margin in static performance indicates that TeleWorld reaches the saturation point of current static scene modeling, while the significantly larger dynamic margin suggests a distinct advantage in temporal reasoning, motion modeling, and evolving world stability. Notably, TeleWorld is the only method that simultaneously ranks first in both static and dynamic tracks, indicating that it does not favor one operational regime at the cost of the other.

In controllability, TeleWorld delivers balanced scores in Camera Control (76.58), Object Control (74.44, best among all systems), and Content Alignment (73.20). This indicates that it respects multi-modal user constraints without specializing in a single dimension. The strong Object Control score, in particular, suggests that TeleWorld maintains an implicit, persistent world state that preserves object identity and arrangement across long sequences, consistent with its closed-loop generation–reconstruction design.

TeleWorld also excels in structural and perceptual consistency, with scores of 87.35 (3D Consistency), 88.82 (Photometric Consistency), 85.59 (Style Consistency), and 61.66 (Subjective Quality). These results reflect that the generated content behaves as projections of a coherent internal 4D representation—aligned with our framework’s ability to capture and enforce global spatio-temporal structure while preserving visual fidelity.

The dynamic performance further underscores TeleWorld’s advantage. Its WorldScore-Dynamic of 66.73 decomposes into strong Motion Accuracy (53.94), moderate Motion Magnitude (31.55), and high Motion Smoothness (34.18). This profile indicates that motion is plausible, well-regulated, and free of temporal discontinuities—avoiding the under-motion or instability common in baseline systems. This stability stems from TeleWorld’s use of a learned internal state to guide temporal evolution, rather than approximating change locally.

A cross-paradigm analysis shows that TeleWorld bridges a key capability gap: it matches the structural consistency of 3D systems while retaining the conditioning flexibility of video models, and rivals the visual quality of video models while avoiding their typical failures in semantic drift and world collapse. This positions TeleWorld in a previously difficult regime—structurally grounded, flexibly conditioned, and temporally stable generation—supporting its role as a practical step toward interactive, memory-enabled world models.

In summary, the empirical evidence indicates that TeleWorld provides balanced, stable, and scalable world generation capabilities. It does not rely on extreme metric optimization or single-axis specialization. Instead, it demonstrates that a unified model can jointly optimize controllability, consistency, perceptual fidelity, and dynamic behavior. The gains observed in dynamic scores, combined with structural and semantic stability, suggest that TeleWorld is particularly suitable for long-horizon and multi-condition generative tasks. These results identify TeleWorld as a strong candidate for future research directions involving long-range video synthesis, controllable simulation, interactive environments, and world modeling tasks that require coherent spatial-temporal evolution rather than isolated visual quality.

5 Conclusion

In summary, TeleWorld is a 18B-parameter model capable of generating high-resolution video (960×1760) in real time at 8 FPS, ranking first on the WorldScores benchmark. It introduces a novel generation–reconstruction-guidance closed-loop that provides a new solution framework for 4D spatiotemporal world modeling. The model is able to produce long, spatiotemporally consistent 4D scene videos while maintaining persistent 4D memory, offering a valuable reference for subsequent research in world models.

To further speed up scene video generation, we present a scalable and efficient training system that makes Distribution Matching Distillation practical for large-scale auto-regressive video generation models. By decoupling the generator, teacher, and critic across dedicated GPU groups, sharding the generator KV cache via context parallelism, and introducing a carefully balanced pipeline execution schedule, our system overcomes the prohibitive memory and efficiency barriers of applying DMD at the 10B scale and beyond. These system-level optimizations enable DMD training of TeleWorld-18B on a limited GPU budget while sustaining high hardware utilization. In summary, our approach bridges the gap between state-of-the-art distillation techniques and large-scale video diffusion models, unlocking real-time long-horizon video synthesis

under practical computational constraints.

Contributors

Project Leaders: Haibin Huang, Chi Zhang, Xuelong Li

Core Contributors: Yabo Chen, Yuanzhi Liang, Jiepeng Wang

Contributors (Listed alphabetically): Chengcheng Zhou, Guangce Liu, Haoyuan Wang, Jialun Liu, Junfei Cheng, Junqi Liu, Junyu Zhou, Qizhen Weng, Shiwen Zhang, Tian Li, Tingxi Chen, Wei Li, Weichen Li, Xiaoyan Yang, Xin Zhang, Xuan'er Wu, Xunzhi Xiang, Yuyang Huang, Zicheng Jiang, Zixiao Gu, Zuoxin Li

References

- Bahmani, Sherwin, Skorokhodov, Ivan, Rong, Victor, Wetzstein, Gordon, Guibas, Leonidas, Wonka, Peter, Tulyakov, Sergey, Park, Jeong Joon, Tagliasacchi, Andrea, & Lindell, David B. 2024. 4D-FY: Text-to-4D Generation Using Hybrid Score Distillation Sampling. *Pages 7996–8006 of: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Bai, Shuai, Chen, Keqin, Liu, Xuejing, Wang, Jialin, Ge, Wenbin, Song, Sibo, Dang, Kai, Wang, Peng, Wang, Shijie, Tang, Jun, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Ball, Philip J., Bauer, Jakob, Belletti, Frank, Brownfield, Bethanie, Ephrat, Ariel, Fruchter, Shlomi, Gupta, Agrim, Holsheimer, Kristian, Holynski, Aleksander, Hron, Jiri, Kaplanis, Christos, Limont, Marjorie, McGill, Matt, Oliveira, Yanko, Parker-Holder, Jack, Perbet, Frank, Scully, Guy, Shar, Jeremy, Spencer, Stephen, Tov, Omer, Villegas, Ruben, Wang, Emma, Yung, Jessica, Baetu, Cip, Berbel, Jordi, Bridson, David, Bruce, Jake, Buttimore, Gavin, Chakera, Sarah, Chandra, Bilva, Collins, Paul, Cullum, Alex, Damoc, Bogdan, Dasagi, Vibha, Gazeau, Maxime, Gbadamosi, Charles, Han, Woohyun, Hirst, Ed, Kachra, Ashyana, Kerley, Lucie, Kjems, Kristian, Knoepfel, Eva, Koriakin, Vika, Lo, Jessica, Lu, Cong, Mehring, Zeb, Moufarek, Alex, Nandwani, Henna, Oliveira, Valeria, Pardo, Fabio, Park, Jane, Pierson, Andrew, Poole, Ben, Ran, Helen, Salimans, Tim, Sanchez, Manuel, Saprykin, Igor, Shen, Amy, Sidhwani, Sailesh, Smith, Duncan, Stanton, Joe, Tomlinson, Hamish, Vijaykumar, Dimple, Wang, Luyu, Wingfield, Piers, Wong, Nat, Xu, Keyang, Yew, Christopher, Young, Nick, Zubov, Vadim, Eck, Douglas, Erhan, Dumitru, Kavukcuoglu, Koray, Hassabis, Demis, Gharamani, Zoubin, Hadsell, Raia, van den Oord, Aäron, Mosseri, Inbar, Bolton, Adrian, Singh, Satinder, & Rocktäschel, Tim. 2025. Genie 3: A New Frontier for World Models.
- Chen, Anthony, Zheng, Wenzhao, Wang, Yida, Zhang, Xueyang, Zhan, Kun, Jia, Peng, Keutzer, Kurt, & Zhang, Shanghang. 2025a. *GeoDrive: 3D Geometry-Informed Driving World Model with Precise Action Control*.
- Chen, Haoxin, Xia, Menghan, He, Yingqing, Zhang, Yong, Cun, Xiaodong, Yang, Shaoshu, Xing, Jinbo, Liu, Yaofang, Chen, Qifeng, Wang, Xintao, Weng, Chao, & Shan, Ying. 2023. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation. *arXiv preprint*.
- Chen, Xingyu, Chen, Yue, Xiu, Yuliang, Geiger, Andreas, & Chen, Anpei. 2025b. TTT3R: 3D Reconstruction as Test-Time Training. *arXiv preprint arXiv:2509.26645*.
- Chung, Jaeyoung, Lee, Suyoung, Nam, Hyeongjin, Lee, Jaerin, & Lee, Kyoung Mu. 2023. LucidDreamer: Domain-free Generation of 3D Gaussian Splatting Scenes. *arXiv preprint arXiv:2311.13384*.
- Ding, Jingtao, Zhang, Yunke, Shang, Yu, Feng, Jie, Zhang, Yuheng, Zong, Zefang, Yuan, Yuan, Su, Hongyuan, Li, Nian, Piao, Jinghua, Deng, Yucheng, Sukienik, Nicholas, Gao, Chen, Xu, Fengli, & Li, Yong. 2025. *Understanding World or Predicting Future? A Comprehensive Survey of World Models*.

- Duan, Haoyi, Yu, Hong-Xing, Chen, Sirui, Fei-Fei, Li, & Wu, Jiajun. 2025. *WorldScore: A Unified Evaluation Benchmark for World Generation*.
- Engstler, Paul, Vedaldi, Andrea, Laina, Iro, & Rupprecht, Christian. 2025. Invisible Stitch: Generating Smooth 3D Scenes with Depth Inpainting. *Pages 457–468 of: 2025 International Conference on 3D Vision (3DV)*.
- Fan, Weichen, Si, Chenyang, Song, Junhao, Yang, Zhenyu, He, Yinan, Zhuo, Long, Huang, Ziqi, Dong, Ziyue, He, Jingwen, Pan, Dongwei, *et al.* 2025. VChitect-2.0: Parallel Transformer for Scaling Up Video Diffusion Models. *arXiv preprint arXiv:2501.08453*.
- Feng, Tianrui, Li, Zhi, Yang, Shuo, Xi, Haocheng, Li, Muyang, Li, Xiuyu, Zhang, Lvmin, Yang, Keting, Peng, Kelly, Han, Song, Agrawala, Maneesh, Keutzer, Kurt, Kodaira, Akio, & Xu, Chenfeng. 2025. *StreamDiffusionV2: A Streaming System for Dynamic and Interactive Video Generation*.
- Fridman, Rafail, Abecasis, Amit, Kasten, Yoni, & Dekel, Tali. 2024. SceneScape: Text-Driven Consistent Scene Generation. *Advances in Neural Information Processing Systems*, **36**.
- Guo, Junliang, Ye, Yang, He, Tianyu, Wu, Haoyu, Jiang, Yushu, Pearce, Tim, & Bian, Jiang. 2025a. *MineWorld: a Real-Time and Open-Source Interactive World Model on Minecraft*.
- Guo, Yanjiang, Shi, Lucy Xiaoyang, Chen, Jianyu, & Finn, Chelsea. 2025b. *Ctrl-World: A Controllable Generative World Model for Robot Manipulation*.
- HaCohen, Yoav, Chiprut, Nisan, Brazowski, Benny, Shalem, Daniel, Moshe, Dudu, Richardson, Eitan, Levin, Eran, Shiran, Guy, Zabari, Nir, Gordon, Ori, *et al.* 2024. LTX-Video: Realtime Video Latent Diffusion. *arXiv preprint arXiv:2501.00103*.
- HailuoAI. 2024. *Hailuo*. <https://hailuoai.video/>. Accessed: 2025-02-24.
- Höllein, Lukas, Cao, Ang, Owens, Andrew, Johnson, Justin, & Nießner, Matthias. 2023 (October). Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models. *Pages 7909–7920 of: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Hong, Yicong, Mei, Yiqun, Ge, Chongjian, Xu, Yiran, Zhou, Yang, Bi, Sai, Hold-Geoffroy, Yannick, Roberts, Mike, Fisher, Matthew, Shechtman, Eli, Sunkavalli, Kalyan, Liu, Feng, Li, Zhengqi, & Tan, Hao. 2025. *RELIC: Interactive Video World Model with Long-Horizon Memory*.
- Hu, Mengkang, Chen, Tianxing, Zou, Yude, Lei, Yuheng, Chen, Qiguang, Li, Ming, Mu, Yao, Zhang, Hongyuan, Shao, Wenqi, & Luo, Ping. 2025. *Text2World: Benchmarking Large Language Models for Symbolic World Model Generation*.
- Huang, Nan, Zheng, Wenzhao, Xu, Chenfeng, Keutzer, Kurt, Zhang, Shanghang, Kanazawa, Angjoo, & Wang, Qianqian. 2025a. Segment Any Motion in Videos. *In: Proceedings of the Computer Vision and Pattern Recognition Conference*.
- Huang, Tianyu, Zheng, Wangguandong, Wang, Tengfei, Liu, Yuhao, Wang, Zhenwei, Wu, Junta, Jiang, Jie, Li, Hui, Lau, Rynson, Zuo, Wangmeng, & Guo, Chuncho. 2025b. Voyager: Long-Range and World-Consistent Video Diffusion for Explorable 3D Scene Generation. *ACM Trans. Graph.*, **44**(6).
- Huang, Xun, Li, Zhengqi, He, Guande, Zhou, Mingyuan, & Shechtman, Eli. 2025c. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. *CoRR*.
- Jin, Bu, Gu, Songen, Hu, Xiaotao, Zheng, Yupeng, Guo, Xiaoyang, Zhang, Qian, Long, Xiaoxiao, & Yin, Wei. 2025. *OccTENS: 3D Occupancy World Model via Temporal Next-Scale Prediction*.
- Li, Jiachen, Feng, Weixi, Fu, Tsu-Jui, Wang, Xinyi, Basu, Sugato, Chen, Wenhui, & Wang, William Yang. 2024. T2V-Turbo: Breaking the Quality Bottleneck of Video Consistency Model with Mixed Reward Feedback. *arXiv preprint arXiv:2405.18750*.

Liao, Minghui, Zou, Zhisheng, Wan, Zhaoyi, Yao, Cong, & Bai, Xiang. 2022. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE transactions on pattern analysis and machine intelligence*.

Mao, Xiaofeng, Lin, Shaoheng, Li, Zhen, Li, Chuanhao, Peng, Wenshuo, He, Tong, Pang, Jiangmiao, Chi, Mingmin, Qiao, Yu, & Zhang, Kaipeng. 2025. Yume: An Interactive World Generation Model. *arXiv preprint arXiv:2507.17744*.

Millon, Erwann. 2025. *Krea Realtime 14B: Real-time Video Generation*.

Moritz, Philipp, Nishihara, Robert, Wang, Stephanie, Tumanov, Alexey, Liaw, Richard, Liang, Eric, Paul, William, Jordan, Michael I., & Stoica, Ion. 2017. Ray: A Distributed Framework for Emerging AI Applications. *CoRR*, **abs/1712.05889**.

NVIDIA, :, Agarwal, Niket, Ali, Arslan, Bala, Maciej, Balaji, Yogesh, Barker, Erik, Cai, Tiffany, Chattopadhyay, Prithvijit, Chen, Yongxin, Cui, Yin, Ding, Yifan, Dworakowski, Daniel, Fan, Jiaojiao, Fenzi, Michele, Ferroni, Francesco, Fidler, Sanja, Fox, Dieter, Ge, Songwei, Ge, Yunhao, Gu, Jinwei, Gururani, Siddharth, He, Ethan, Huang, Jiahui, Huffman, Jacob, Jannaty, Pooya, Jin, Jingyi, Kim, Seung Wook, Klár, Gergely, Lam, Grace, Lan, Shiyi, Leal-Taixe, Laura, Li, Anqi, Li, Zhaoshuo, Lin, Chen-Hsuan, Lin, Tsung-Yi, Ling, Huan, Liu, Ming-Yu, Liu, Xian, Luo, Alice, Ma, Qianli, Mao, Hanzi, Mo, Kaichun, Mousavian, Arsalan, Nah, Seungjun, Niverty, Sriharsha, Page, David, Paschalidou, Despoina, Patel, Zeeshan, Pavao, Lindsey, Ramezanali, Morteza, Reda, Fitsum, Ren, Xiaowei, Sabavat, Vasanth Rao Naik, Schmerling, Ed, Shi, Stella, Stefaniak, Bartosz, Tang, Shitao, Tchapmi, Lyne, Tredak, Przemek, Tseng, Wei-Cheng, Varghese, Jibin, Wang, Hao, Wang, Haoxiang, Wang, Heng, Wang, Ting-Chun, Wei, Fangyin, Wei, Xinyue, Wu, Jay Zhangjie, Xu, Jiashu, Yang, Wei, Yen-Chen, Lin, Zeng, Xiaohui, Zeng, Yu, Zhang, Jing, Zhang, Qinsheng, Zhang, Yuxuan, Zhao, Qingqing, & Zolkowski, Artur. 2025. *Cosmos World Foundation Model Platform for Physical AI*.

Runway. 2024. *Introducing Gen-3 Alpha: A New Frontier for Video Generation*. <https://runwayml.com/research/introducing-gen-3-alpha>. Accessed: 2025-02-24.

Tang, Junshu, Liu, Jiacheng, Li, Jiaqi, Wu, Longhuang, Yang, Haoyu, Zhao, Penghao, Gong, Siruis, Yuan, Xiang, Shao, Shuai, & Lu, Qinglin. 2025. *Hunyuan-GameCraft-2: Instruction-following Interactive Game World Model*.

Team, HunyuanWorld, Wang, Zhenwei, Liu, Yuhao, Wu, Junta, Gu, Zixiao, Wang, Haoyuan, Zuo, Xuhui, Huang, Tianyu, Li, Wenhuan, Zhang, Sheng, Lian, Yihang, Tsai, Yulin, Wang, Lifu, Liu, Sicong, Jiang, Puhua, Yang, Xianghui, Guo, Dongyuan, Tang, Yixuan, Mao, Xinyue, Yu, Jiaao, Yu, Junlin, Zhang, Jihong, Chen, Meng, Dong, Liang, Jia, Yiwen, Zhang, Chao, Tan, Yonghao, Zhang, Hao, Ye, Zheng, He, Peng, Wu, Runzhou, Chen, Minghui, Li, Zhan, Qin, Wangchen, Wang, Lei, Sun, Yifu, Niu, Lin, Yuan, Xiang, Yang, Xiaofeng, He, Yingping, Xiao, Jie, Tao, Yangyu, Zhu, Jianchen, Xue, Jinbao, Liu, Kai, Zhao, Chongqing, Wu, Ximeng, Liu, Tian, Chen, Peng, Wang, Di, Liu, Yuhong, Linus, Jiang, Jie, Wang, Tengfei, & Guo, Chunhao. 2025. *HunyuanWorld 1.0: Generating Immersive, Explorable, and Interactive 3D Worlds from Words or Pixels*.

Teng, Hansi, Jia, Hongyu, Sun, Lei, Li, Lingzhi, Li, Maolin, Tang, Mingqiu, Han, Shuai, Zhang, Tianning, Zhang, W. Q., Luo, Weifeng, Kang, Xiaoyang, Sun, Yuchen, Cao, Yue, Huang, Yunpeng, Lin, Yutong, Fang, Yuxin, Tao, Zewei, Zhang, Zheng, Wang, Zhongshu, Liu, Zixun, Shi, Dai, Su, Guoli, Sun, Hanwen, Pan, Hong, Wang, Jie, Sheng, Jiexin, Cui, Min, Hu, Min, Yan, Ming, Yin, Shucheng, Zhang, Siran, Liu, Tingting, Yin, Xianping, Yang, Xiaoyu, Song, Xin, Hu, Xuan, Zhang, Yankai, & Li, Yuqiao. 2025. MAGI-1: Autoregressive Video Generation at Scale. *CoRR*.

Wang, Ang, Ai, Baole, Wen, Bin, Mao, Chaojie, Xie, Chen-Wei, Chen, Di, Yu, Feiwu, Zhao, Haiming, Yang, Jianxiao, Zeng, Jianyuan, Wang, Jiayu, Zhang, Jingfeng, Zhou, Jingren, Wang, Jinkai, Chen, Jixuan, Zhu, Kai, Zhao, Kang, Yan, Keyu, Huang, Lianghua, Meng, Xiaofeng, Zhang, Ningyi, Li, Pandeng, Wu, Pingyu, Chu, Ruihang, Feng, Ruili, Zhang, Shiwei, Sun, Siyang, Fang, Tao, Wang, Tianxing, Gui, Tianyi, Weng, Tingyu, Shen, Tong, Lin, Wei, Wang, Wei, Wang, Wei, Zhou, Wenmeng, Wang, Wente, Shen,

- Wenting, Yu, Wenyuan, Shi, Xianzhong, Huang, Xiaoming, Xu, Xin, Kou, Yan, Lv, Yangyu, Li, Yifei, Liu, Yijing, Wang, Yiming, Zhang, Yingya, Huang, Yitong, Li, Yong, Wu, You, Liu, Yu, Pan, Yulin, Zheng, Yun, Hong, Yuntao, Shi, Yupeng, Feng, Yutong, Jiang, Zeyinzi, Han, Zhen, Wu, Zhi-Fan, & Liu, Ziyu. 2025a. Wan: Open and Advanced Large-Scale Video Generative Models. *CoRR*.
- Wang, Haonan, Zhou, Hanyu, Liu, Haoyue, & Yan, Luxin. 2025b. *4D-VGGT: A General Foundation Model with SpatioTemporal Awareness for Dynamic Scene Geometry Estimation*.
- Wang, Zeqing, Wei, Xinyu, Li, Bairui, Guo, Zhen, Zhang, Jinrui, Wei, Hongyang, Wang, Keze, & Zhang, Lei. 2025c. *VideoVerse: How Far is Your T2V Generator from a World Model?*
- Won, John, Lee, Kyungmin, Jang, Huiwon, Kim, Dongyoung, & Shin, Jinwoo. 2025. *Dual-Stream Diffusion for World-Model Augmented Vision-Language-Action Model*.
- Xiang, Xunzhi, Chen, Yabo, Zhang, Guiyu, Wang, Zhongyu, Gao, Zhe, Xiang, Quanming, Shang, Gonghu, Liu, Junqi, Huang, Haibin, Gao, Yang, et al. 2025. Macro-from-Micro Planning for High-Quality and Parallelized Autoregressive Long Video Generation. *arXiv preprint arXiv:2508.03334*.
- Xiao, Junjin, Yang, Yandan, Chang, Xinyuan, Chen, Ronghan, Xiong, Feng, Xu, Mu, Zheng, Wei-Shi, & Zhang, Qing. 2025. *World-Env: Leveraging World Model as a Virtual Environment for VLA Post-Training*.
- Xing, Jinbo, Xia, Menghan, Zhang, Yong, Chen, Haoxin, Wang, Xintao, Wong, Tien-Tsin, & Shan, Ying. 2023. Dynamicrafter: Animating Open-Domain Images with Video Diffusion Priors.
- Xu, Jiaqi, Zou, Xinyi, Huang, Kunzhe, Chen, Yunkuo, Liu, Bo, Cheng, MengLi, Shi, Xing, & Huang, Jun. 2024. EasyAnimate: A High-Performance Long Video Generation Method Based on Transformer Architecture. *arXiv preprint arXiv:2405.18991*.
- Yang, Zhongqi, Ge, Wenhang, Li, Yuqi, Chen, Jiaqi, Li, Haoyuan, An, Mengyin, Kang, Fei, Xue, Hua, Xu, Baixin, Yin, Yuyang, Li, Eric, Liu, Yang, Wang, Yikai, Guo, Hao-Xiang, & Zhou, Yahui. 2025a. *Matrix-3D: Omnidirectional Explorable 3D World Generation*.
- Yang, Zhuoyi, Teng, Jiayan, Zheng, Wendi, Ding, Ming, Huang, Shiyu, Xu, Jiazheng, Yang, Yuanming, Hong, Wenyi, Zhang, Xiaohan, Feng, Guanyu, Yin, Da, Zhang, Yuxuan, Wang, Weihan, Cheng, Yean, Xu, Bin, Gu, Xiaotao, Dong, Yuxiao, & Tang, Jie. 2025b. *CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer*.
- Yin, Tianwei, Gharbi, Michaël, Zhang, Richard, Shechtman, Eli, Durand, Fredo, Freeman, William T., & Park, Taesung. 2024. *One-step Diffusion with Distribution Matching Distillation*.
- Yin, Tianwei, Zhang, Qiang, Zhang, Richard, Freeman, William T., Durand, Frédéric, Shechtman, Eli, & Huang, Xun. 2025. From Slow Bidirectional to Fast Causal Video Generators. In: *CVPR*.
- Yu, Hong-Xing, Duan, Haoyi, Hur, Junhwa, Sargent, Kyle, Rubinstein, Michael, Freeman, William T, Cole, Forrester, Sun, Deqing, Snavely, Noah, Wu, Jiajun, & Herrmann, Charles. 2023. WonderJourney: Going from Anywhere to Everywhere. *arXiv preprint arXiv:2312.03884*.
- Yu, Hong-Xing, Duan, Haoyi, Herrmann, Charles, Freeman, William T., & Wu, Jiajun. 2025 (June). WonderWorld: Interactive 3D Scene Generation from a Single Image. Pages 5916–5926 of: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Zhao, Yanli, Gu, Andrew, Varma, Rohan, Luo, Liang, Huang, Chien-Chin, Xu, Min, Wright, Less, Shojanazeri, Hamid, Ott, Myle, Shleifer, Sam, Desmaison, Alban, Balioglu, Can, Damania, Pritam, Nguyen, Bernard, Chauhan, Geeta, Hao, Yuchen, Mathews, Ajit, & Li, Shen. 2023. *PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel*.
- Zhou, Yuan, Wang, Qiuyue, Cai, Yuxuan, & Yang, Huan. 2024. Allegro: Open the Black Box of Commercial-Level Video Generation Model. *arXiv preprint arXiv:2410.15458*.

Zhu, Zheng, Wang, Xiaofeng, Zhao, Wangbo, Min, Chen, Li, Bohan, Deng, Nianchen, Dou, Min, Wang, Yuqi, Shi, Botian, Wang, Kai, Zhang, Chi, You, Yang, Zhang, Zhaoxiang, Zhao, Dawei, Xiao, Liang, Zhao, Jian, Lu, Jiwen, & Huang, Guan. 2025. *Is Sora a World Simulator? A Comprehensive Survey on General World Models and Beyond.*

Zuo, Sicheng, Zheng, Wenzhao, Huang, Yuanhui, Zhou, Jie, & Lu, Jiwen. 2025 (June). GaussianWorld: Gaussian World Model for Streaming 3D Occupancy Prediction. *Pages 6772–6781 of: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*