WorldDirector: Building Controllable World Simulators with Persistent Dynamic Memory

Hanlin Wang[1,2] **Hao Ouyang**[2] **Qiuyu Wang**[2] **Wen Wang**[3] **Qingyan Bai**[1,2]
Ka Leong Cheng[2] **Yue Yu**[1,2] **Yixuan Li**[4,2] **Yihao Meng**[1,2] **Zichen Liu**[1,2]
Yanhong Zeng[2] **Yujun Shen**[2] **Qifeng Chen**[1*]
[1]HKUST [2]Ant Group [3]ZJU [4]CUHK

Figure 1: **Controllable world simulation with persistent dynamic memory via `WorldDirector`.** By decoupling 3D semantic orchestration from latent video synthesis, our framework autoregressively generates long-horizon videos via causal chunks, ensuring rigorous dynamic memory and object permanence. Please refer to the video results on our project page for intuitive demonstrations.

Abstract

We present `WorldDirector`, a highly controllable video world model framework designed for persistent dynamic object memory and unrestricted viewpoint exploration. Unlike existing world models that entangle physical dynamics with pixel rendering and rely on continuous visual observation to sustain motion, our framework explicitly decouples semantic motion orchestration from visual generation. By leveraging an LLM to coordinate 3D trajectories with camera movements and subsequently employing these orchestrated trajectories as control signals for video generation, our approach ensures strict physical logic and appearance stability, successfully preserving the exact visual identities of dynamic entities even when they re-enter the scene after prolonged periods out of view. Experimental results demonstrate that our method supports the synthesis of complex and extended events with unprecedented controllability and persistent dynamic object memory.

*Corresponding author.

Preprint.

1 Introduction

The landscape of video generation is undergoing a profound transformation, transitioning from passive pixel synthesis [7, 5, 27, 13] to interactive environment simulation [8, 9, 11, 20, 33, 38, 36]. A cornerstone of this paradigm shift is memory: the ability to maintain consistency of static scenes and the continuous movements of dynamic objects, whether they are visible in the frame or out-of-view. While recent methods have achieved remarkable success in preserving static scene consistency through memory retrieval or contextual conditioning [50, 46, 21, 35, 44], a crucial area remains largely unexplored: **"Object Permanence"** and **"Dynamic Object Memory"**. Specifically, this entails that dynamic entities persistently exist and execute their physical movements independent of camera visibility. Consequently, whenever the dynamic objects reappear in the frame, their newly updated positions and states should be accurately observed.

To achieve this, we argue that a world simulator with robust dynamic memory must be built upon two foundational pillars. First, entities must exhibit independent motion. Their trajectories should follow continuous physical logic unconstrained by camera visibility, ensuring that unobserved dynamics progress naturally. Second, the system must guarantee strict appearance consistency. When a hidden entity re-enters the frame, its visual identity and fine details must remain entirely intact without distortion. Satisfying these two criteria is the prerequisite for elevating unpredictable video generation to the level of persistent world simulation. Driven by this goal, several methods have been proposed to realize world simulators equipped with dynamic memory recently. One framework [16] introduces a monitor-based mechanism to address out-of-sight dynamics by registering explicit "monitors" that autonomously track and fast-forward the temporal progression of unobserved active entities. However, this explicit tracking system scales poorly and incurs prohibitive computational overhead in scenarios involving multiple dynamic entities. Conversely, another approach [12] tracks dynamic features but delegates trajectory extrapolation entirely to internal generative priors. While this implicit estimation might suffice for brief occlusions, it fails during prolonged camera diversions or intricate dynamic interactions. Relying on generative weights to guess continuous physical evolution without a dedicated orchestration mechanism inevitably leads to trajectory collapse, frozen states, or severe identity errors upon re-entry.

To overcome the aforementioned limitations and fulfill the two foundational pillars, we introduce `WorldDirector`. Our primary insight is to explicitly decouple the motion planning of dynamic objects from the video synthesis process. By leveraging controllable generation paradigms, we transmit semantic-level planning results as conditions to the generative model, thereby realizing a persistent world simulator equipped with robust dynamic memory. This architecture not only guarantees the independent and continuous movements of dynamic objects, but also provides high controllability, enabling users to independently dictate the specific actions and semantic behaviors of multiple distinct dynamic entities. Specifically, we employ an LLM to act as a central orchestrator, which translates user instructions into 3D bounding box and camera trajectories. These spatial plans are subsequently projected into 2D bounding box sequences, providing location conditions for video synthesis. To prevent identity distortion when a hidden entity re-enters the frame, we propose an Appearance Binding mechanism that injects RGB dynamic object features from context as visual anchors. For granular state control, a spatial-aware cross-attention mechanism [41] routes entity-specific text prompts to their corresponding regions. Integrated within a causal autoregressive architecture, these mechanisms ensure extended video generation with strict dynamic memory.

Extensive evaluations demonstrate that `WorldDirector` synthesizes highly controllable dynamic scenarios while rigorously maintaining dynamic memory across extended sequences. By ensuring object permanence and appearance consistency after prolonged out-of-view intervals, our approach transcends passive video generation and represents a significant step toward interactive and persistent world simulators with unprecedented dynamic object memory.

2 Related Works

2.1 Foundation Video Models and World Simulators

Generative video synthesis has progressed rapidly with diffusion and transformer architectures [7, 5, 27, 13]. Beyond pixel fidelity, the field is increasingly shifting toward video world models for simulating interactive environments. Pioneering works such as Sora [8], Genie [9], Oasis [15], and

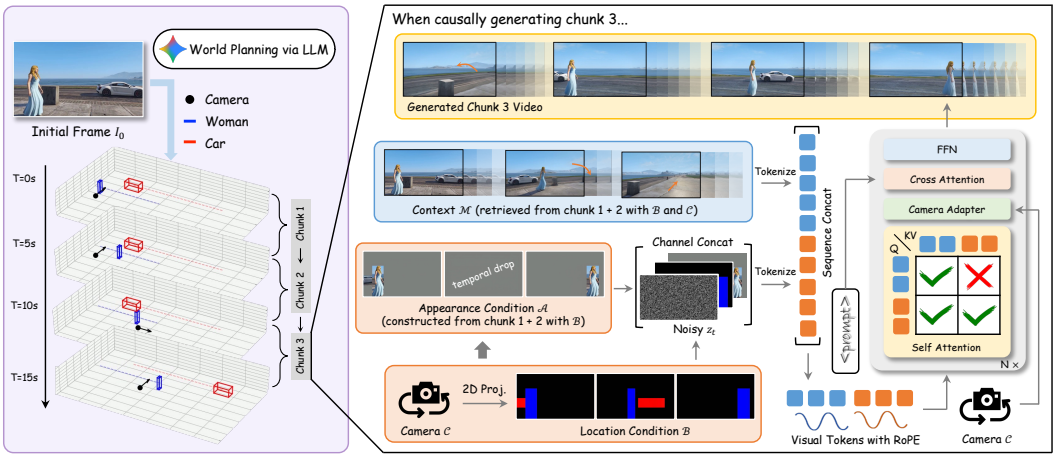

Figure 2: **Overview of `WorldDirector`.** An LLM orchestrates 3D trajectories that are projected into 2D Location Conditions for causal chunk generation. Location ($\mathcal{B}$) and Appearance ($\mathcal{A}$) conditions are channel-concatenated with the noisy latent, while historical Context ($\mathcal{M}$) is sequence-concatenated. During generation, *temporal drop* is applied and an asymmetric attention routing prevents noise from polluting the context memory.

DIAMOND [1] treat generative models as rudimentary physics engines, with further advances in game-like simulators [11, 20] and long-sequence interactive generators [33, 22, 36]. However, relying on generative models to implicitly memorize object states, actions, and appearances overloads their capacity: when active entities exit the camera's field of view, these entangled models fail to sustain dynamics, causing objects to freeze or vanish. Our work addresses this by explicitly decoupling semantic motion orchestration from visual rendering.

2.2 Controllable Video Generation

To move beyond random generation, controllable synthesis has been widely explored. Image control mechanisms [51, 34, 25, 2] have been extended to video [52]. For spatial and motion control, Boximator [42] and GLIGEN [30] leverage bounding boxes, while others target camera trajectories [19] or motion tracking [49, 43]. More recently, dense or point-trajectory guidance has emerged as a flexible interface for fine-grained, entity-level motion control [53, 45, 18, 28, 40, 14]. Though effective in short clips, these methods lack the autoregressive memory required for long-horizon simulation. Our framework adopts the spatial-aware cross-attention of GLIGEN [30] within a persistent memory architecture, enabling consistent control across extended temporal windows.

2.3 Memory Mechanisms in Video World Models

Memory underpins temporal coherence beyond the immediate context window. Long video generators [21, 35] use sliding windows but struggle with extended occlusion. Prior work preserves static-scene consistency via FOV retrieval [50, 46] or 3D representations [29], yet assumes a static world. Object permanence, defined as objects persisting and evolving when unobserved, remains a core challenge in physical reasoning [48, 4, 6], and is even harder for active entities in complex scenes. Recent approaches employ implicit hybrid memory tokens [12] or external monitors for out-of-sight dynamic simulation [16]. However, internal priors risk trajectory collapse during prolonged diversions, while external monitors are computationally prohibitive. Coupled with an Appearance Binding mechanism, our LLM-orchestrated approach delivers controllable generation and robust dynamic object memory, offering a scalable path to object permanence in world exploration.

3 Method

This section outlines our data curation pipeline (Section 3.1), model design (Section 3.2), training objective and inference workflow (Section 3.4). An overview of `WorldDirector` is illustrated in Figure 2. We use the LingBot-World-Base model [38] as our foundation model.

3.1 Data Curation Pipeline

We introduce a tailored data curation pipeline to address the foundational requirements outlined in Section 1. Specifically, this pipeline constructs comprehensive training tuples that encapsulate: 2D bounding boxes for dynamic spatial grounding, appearance references for fine-grained visual conditioning, object-centric captions detailing behavioral dynamics, and contextual signals to facilitate causal generation. The proposed pipeline comprises the following key components:

Dynamic Object Tracking and Entity-Based Captioning. To address the scarcity of real-world data featuring dynamic entities exiting and re-entering the field of view (FOV), we developed a game-based platform to generate 15-second videos with precise camera parameters, deliberately scripted to induce target disappearances and reappearances. We employ SAM3 [10] to extract 2D bounding box trajectories; its robust re-identification seamlessly tracks objects despite temporary FOV exits, ensuring highly reliable annotations. For training, we sample a contiguous 5-second window from each video that maximizes the number of newly visible objects (absent in the first frame but appearing later) to specifically capture critical re-entry scenarios. The remaining 10 seconds function as a candidate pool to provide historical appearance and spatiotemporal context. Finally, we superimpose unique color-coded bounding boxes onto the source frames to preserve instance identities and feed these visually augmented sequences into Qwen2.5-VL-72B [3] to generate fine-grained textual captions of each entity's action dynamics.

Dual-Conditioning Preparation for Dynamic Objects. We construct two conditioning videos for each training sequence. First, to encode spatio-temporal trajectories and provide positional priors, we generate a spatial location condition video by filling each dynamic object's 2D bounding box with a unique color identifier against a zero-initialized background. Second, to ensure appearance consistency for re-entering objects regardless of their absence duration, we introduce an appearance conditioning video. Specifically, for an object a at frame t with bounding box $\text{box}_{a,t}$, we retrieve a reference $\text{box}_{a,t'}$ of the identical object from the 10-second candidate pool that minimizes aspect ratio divergence relative to $\text{box}_{a,t}$. The image region within $\text{box}_{a,t'}$ is then cropped, spatially resampled, and mapped onto the coordinates of $\text{box}_{a,t}$ in the appearance video, directly equipping the model with exact visual features at designated spatio-temporal indices.

Static and Dynamic Context Retrieval. To better support robust causal inference and preserve spatio-temporal consistency, we retrieve context through a dual-perspective approach. For static scenes, we follow [50] to retrieve the top-K frames maximizing Field of View (FoV) overlap. For dynamic objects, we introduce a greedy algorithm prioritizing frames with active object identities within the current temporal chunk. To ensure uniform spatio-temporal distribution, we enforce a minimum temporal stride of four frames between selected frames. Ranked lists derived from both strategies are then interleaved and deduplicated to yield the final N memory frame indices. The detailed selection procedure is outlined in Appendix B.

3.2 Building Controllable World Simulator with Persistent Dynamic Memory

Building upon our established data curation pipeline, we present `WorldDirector`, a framework that reconceptualizes video generation as a controllable world simulation with persistent dynamic memory. We formulate this objective as a conditional denoising process guided by multi-modal structural priors. Formally, let $V \in \mathbb{R}^{T\times 3\times H\times W}$ denote a training video sequence comprising T frames. The generative process is conditioned on a composite tuple $\mathcal{T} = \{\mathcal{B}, \mathcal{A}, \mathcal{P}, \mathcal{M}\}$:

- **Location Condition** $\mathcal{B} \in \mathbb{R}^{T\times 3\times H\times W}$ encodes the precise spatiotemporal trajectories of 2D bounding boxes for all entities, rendered as identity-preserving, color-coded masks.
- **Appearance Condition** $\mathcal{A} \in \mathbb{R}^{T\times 3\times H\times W}$ provides sparse RGB features derived from contextual frames to maintain dynamic object appearance consistency across the sequence.
- **Multi-Granularity Prompts** $\mathcal{P} = \{p_{\text{global}}, p_1, p_2, \ldots, p_k\}$ consists of a global prompt p_{global} summarizing the overall video narrative, coupled with fine-grained textual descriptions $\{p_i\}_{i=1}^{k}$ detailing the specific semantic behaviors of k dynamic entities.
- **Contextual Memory Frames** $\mathcal{M}$ represents contextual frames retrieved via a dual-stream selection strategy, paired with their corresponding location and appearance conditioning to align feature dimensions with the current generation window, thereby anchoring the generated content within the broader global scene.

In this section, we elaborate on how these multi-modal conditioning priors are leveraged to accomplish a dynamic memory-augmented world simulation.

3.2.1 Control of Spatial Location and Visual Appearance

To achieve a controllable world model with strict dynamic consistency, we extend the LingBot-World-Base architecture with auxiliary feature channels for spatial ($\mathcal{B}$) and appearance ($\mathcal{A}$) constraints, enabling high-fidelity free-exploration simulations with high physical fidelity. Specifically, $\mathcal{B}$ employs instance-specific color-coded masks to explicitly distinguish multiple entities, serving as a deterministic geometric prior for their trajectories, shapes, and orientations. Simultaneously, $\mathcal{A}$ anchors historical visual features to ensure identity coherence, preventing visual degradation when entities re-enter the camera view. To prevent the model from over-relying on $\mathcal{A}$ and generating unnatural sliding artifacts where entities merely translate without exhibiting proper articulated motion, we introduce a *Temporal Drop Mechanism*. For each dynamic entity, we preserve a dense sequence of the initial 16 frames immediately following its entry into the view. Subsequently, we employ a sparse sampling strategy, retaining only one reference frame per six-frame interval. This information bottleneck compels the model to synthesize natural object movements driven by trajectories and semantic captions, utilizing $\mathcal{A}$ strictly as an identity anchor.

Architecturally, both $\mathcal{B}$ and $\mathcal{A}$ are encoded by a 3D VAE into latent tokens and concatenated with the noisy latent sequence along the feature dimension:

$$z_{\text{in}} = \text{Conv3D}\Big(z_t \oplus \mathcal{E}(\mathcal{B}) \oplus \mathcal{E}\big(\mathcal{D}_\tau(\mathcal{A})\big)\Big), \tag{1}$$

where z_t denotes the noisy latent, and $\mathcal{E}(\cdot)$ represents the pre-trained 3D VAE encoder. $\oplus$ denotes channel concatenation. $\mathcal{D}_\tau(\cdot)$ formulates our *Temporal Drop Mechanism*, defined as:

$$\mathcal{D}_\tau\big(\mathcal{A}_t^{(i)}\big) = \begin{cases} \mathcal{A}_t^{(i)}, & \text{if } k^{(i)} < 16 \\ \mathcal{A}_t^{(i)}, & \text{if } k^{(i)} \geq 16 \text{ and } (k^{(i)} - 16) \pmod 6 = 0 \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{2}$$

Here, $\mathcal{A}_t^{(i)}$ represents the appearance condition for entity i at global frame t, and $k^{(i)} \geq 0$ is its instance-specific relative frame index (i.e., the number of frames elapsed since entity i newly entered the view). The full appearance conditioning feature $\mathcal{D}_\tau(\mathcal{A})$ aggregates these processed entity-level representations, where masked entities are replaced by null embeddings $\mathbf{0}$. The fused latent is then processed by a dedicated layer Conv3D$(\cdot)$, where the channel weights corresponding to $\mathcal{E}\big(\mathcal{D}_\tau(\mathcal{A})\big)$ are initialized from the first-frame processor for RGB identity transfer, whereas those for $\mathcal{E}(\mathcal{B})$ are zero-initialized to learn trajectory-guided generation as a residual process.

3.2.2 Contextual Integration

To preserve the consistency of both static scenes and dynamic objects during causal chunk generation, we integrate the retrieved context $\mathcal{M}$ through sequence-level concatenation by prepending the context frames to the noisy latent sequence along the temporal axis. To enforce structural alignment, the location and appearance conditioning associated with each context frame are concatenated along the feature dimension, strictly mirroring the input formulation of the training segment. Furthermore, to explicitly disentangle these historical anchors from the current generative chunk, the time steps used for Rotary Position Embedding (RoPE) of the context frames are shifted by an offset substantially exceeding the maximum training sequence length, establishing a definitive frequency boundary within the RoPE representation space. To prevent the noisy training latent from polluting the high-fidelity context $\mathcal{M}$, we impose an asymmetric attention mask where context tokens exclusively self-attend to remain stable, noise-free references. This allows the model to leverage historical priors without compromising contextual integrity. Finally, to equip the model with the capability to directly generate the initial sequence chunk from scratch, we randomly discard the contextual information $\mathcal{M}$ with a probability of 30% during the training phase.

3.3 Camera Injection and Spatial-Aware Text Control

To accurately model perspective variations and effectively leverage contextual camera information during generation, we first convert the camera poses of all context frames and the current video chunk

into relative camera poses with respect to the first frame of the current generated chunk. Following Wan [39], we utilize Plücker coordinates to encode these relative intrinsic and extrinsic parameters. Next, we apply a spatial downsampling to this representation, followed by a series of convolutional modules to extract multi-level camera motion embeddings. These embeddings are subsequently injected into each Diffusion Transformer (DiT) block via an adaptive normalization layer.

For textual condition injection, it is crucial to guarantee that entity-specific captions are precisely grounded in their corresponding spatial regions. To achieve this, we adopt the Spatial-Aware Weighted Cross-Attention mechanism from [41]. Rather than computing cross-attention uniformly across the entire frame, this scheme identifies the visual tokens encompassed by each entity's 2D bounding box trajectory. We then apply a targeted spatial weight bias to the pre-softmax attention logits between these localized visual tokens and the specific text tokens describing that entity. By doing so, it effectively mitigates semantic leakage and facilitates fine-grained control over multiple dynamic objects within the synthesized scene.

3.4 Training and Inference

We follow the flow matching framework [31, 17] to perform post-training using the mean squared error (MSE) loss. The training objective is applied exclusively to the current target segment, while the historical context remains non-noisy and serves solely as a reference. Formally, let x_1 denote the ground-truth latent of the target video chunk and $x_0 \sim \mathcal{N}(0, I)$ be the random noise. At a sampled timestep $t \in [0, 1]$, the training input for the target portion is $x_{tgt,t} = tx_1 + (1-t)x_0$, with the corresponding ground-truth velocity defined as $v_t = x_1 - x_0$. As described in Section 3.2, the model u receives a concatenated sequence $[x_{ctx}, x_{tgt,t}]$, where x_{ctx} represents the clean context tokens. The training objective is formulated as:

$$\mathcal{L} = \mathbb{E}_{x_0, x_1, t, \Omega} \left[\sum_{i \in \mathcal{I}_{tgt}} \left\| u(x_t, t, \Omega; \theta)_i - v_{t,i} \right\|^2 \right], \tag{3}$$

where $\Omega = \{\mathcal{B}, \mathcal{A}, \mathcal{P}, \mathcal{M}\}$ is the union of all location, appearance, text, and contextual conditions, and $\mathcal{I}_{tgt}$ denotes the set of token indices belonging to the current video segment. By restricting the loss calculation to $\mathcal{I}_{tgt}$, we ensure that the model learns to synthesize new content anchored by the clean memory of previous frames without attempting to reconstruct the already-determined context. During inference, our method operates in two primary stages as described below: *World Planning via LLM* and *Causal Chunk-Based Generation*. Further details regarding the inference implementation are provided in Appendix C.

World Planning via LLM. We first estimate the 3D bounding boxes of target dynamic objects in the given initial image to provide a foundational spatial context for the LLM, which then forecasts continuous 3D box trajectories—comprising both spatial coordinates and orientations—alongside our designed camera path. This trajectory planning encompasses not only the entities present in the initial frame but also those that appear later. Objects absent from the initial frame are synthesized based on their captions when they first enter the camera view. Subsequent generations are then conditioned on these initial outputs to maintain appearance consistency. These 3D trajectories are then projected onto the 2D image plane to yield a sequence of 2D bounding boxes, formulating a spatial condition $\mathcal{B}$ that strictly aligns with the location conditioning format employed during our training phase.

Causal Chunk-Based Generation. To facilitate computationally efficient long-horizon world exploration, we introduce an autoregressive chunk-based generation strategy (detailed in Appendix C). The projected 2D location condition $\mathcal{B}$ is partitioned into contiguous temporal chunks. During the first chunk generation, the process relies exclusively on the initial reference frame for the appearance condition $\mathcal{A}$, with an empty memory context $\mathcal{M}$. For all subsequent chunks, we recursively construct $\mathcal{A}$ and retrieve $\mathcal{M}$ from the continuously updated pool of previously generated chunks. This causal loop explicitly preserves entity identities and spatiotemporal consistency throughout the dynamic simulation, ultimately facilitating arbitrary-length world exploration.

4 Experiments

Implementation Details We build `WorldDirector` on the pre-trained LingBot-World-Base model [38]. All training videos are pre-processed to a fixed resolution of 832×480 pixels at

Table 1: **Quantitative results**. The best and runner-up are in **bold** and underlined.

Method	PSNR↑	SSIM↑	LPIPS↓	Subject Consistency↑	Background Consistency↑	DSC_DINO↑	DSC_CLIP↑
Yume1.5	14.391	0.455	0.425	0.898	0.919	0.765	0.898
HY-World	14.782	0.418	0.398	0.923	**0.931**	0.758	0.911
Infinite-World	14.574	0.431	0.406	**0.934**	0.908	**0.773**	0.913
LingBot-World	14.116	0.409	0.412	0.887	0.911	0.736	0.891
HyDRA	13.421	0.352	0.439	0.855	0.902	0.632	0.877
Ours	**18.127**	**0.502**	**0.359**	0.891	0.909	0.769	**0.917**

16 fps. For conditioning, the context length is set to $N = 10$ frames, with each frame independently encoded via the pre-trained 3D VAE. Our model is trained for 3,000 steps utilizing a global batch size of 64 and a constant learning rate of 1×10^{-5}. During inference, we leverage Gemini [37] as the orchestrator to plan 3D trajectories and states for all dynamic entities. Subsequently, the full-length video is partitioned into five-second segments and generated chunk by chunk in an autoregressive manner. Comprehensive prompt templates for the LLM are detailed in the supplementary material.

Baselines. We compare `WorldDirector` with state-of-the-art causal interactive world models: Yume 1.5 [32], which uses uniform temporal downsampling for memory; HY-World 1.5 [24], applying FOV-based attention on mixed data to achieve memorization; Infinite World [44], which achieves memorization through hierarchical context compression; LingBot-World-Fast [38], leveraging causal attention for infinite generation; and HyDRA [12], which utilizes spatiotemporal retrieval for maintaining off-screen character motion.

Evaluation Protocol. To evaluate our method, we use our data pipeline to construct a test set of 100 video samples featuring novel scenes and subjects that are unseen during training. Following HyDRA [12], we evaluate our model using PSNR, SSIM, and LPIPS to measure overall reconstruction fidelity via pixel-wise analysis, along with VBench's [23] Subject and Background Consistency for frame-level coherence. We also adopt Dynamic Subject Consistency (DSC) by cropping YOLO-detected bounding boxes of dynamic objects and computing their average DINO and CLIP similarities with their contextual counterparts. This metric effectively captures dynamic object consistency, especially for off-screen reappearance.

4.1 Comparisons

Quantitative Results. As reported in Table 1, `WorldDirector` achieves state-of-the-art performance across all three reconstruction metrics. This stems from our location conditioning, which captures continuous object positions and reflects camera poses, facilitating more accurate generation that aligns with the ground truth. For the VBench results, Yume, HY-World, and Infinite-World attain the best performance. However, analyzing the generated videos indicates that this is largely because they generate less subject or camera motion, giving them an inherent advantage when calculating these metrics. Even though these methods also have an inherent advantage on the DSC metric due to their limited motion, our method still attains superior results. This proves our method's strong capability in preserving dynamic consistency while producing highly dynamic generations.

Qualitative Results. We show a qualitative comparison result in Figure 3. Since HyDRA requires a reference video for motion extraction, we use the first 10s of our result to prompt its subsequent 5s generation. We leveraged Gemini to script a specific scenario: a man stands stationary and then walks away; concurrently, the camera pans left (moving the man out of frame) and later pans back to reveal his reappearance. Comparisons against baselines yield the following observations: (1) Limited dynamic generation: Yume, HY-World, and Infinite-World render the man stationary even though the prompt specifies that the man walks into the distance. (2) Identity inconsistency: While LingBot-World and HyDRA capture the man's movement, they struggle with identity preservation. Lingbot-World exhibits slight appearance degradation despite keeping the man in-frame, while HyDRA generates a completely new identity upon the man's reappearance. (3) Insufficient control: Due to the lack of Location Condition, all baselines fail to properly synchronize camera and object dynamics with the user's design. Lingbot-World automatically generates camera translation to ensure the man remains in the shot; Infinite-World executes camera controls correctly but misses object motion; HyDRA directly ignores the man in the distance and generates a new man walking in front of the camera. In contrast, by explicitly conditioning on location and appearance conditions, our method

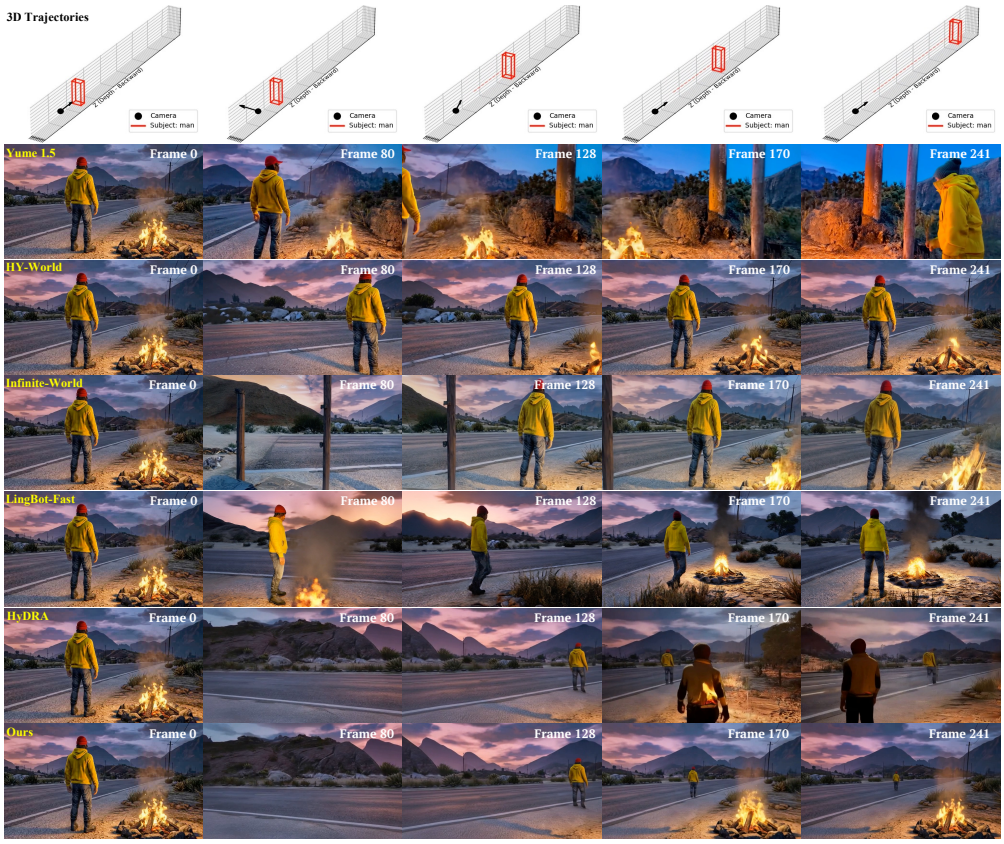

Figure 3: **Qualitative comparison with baselines.** Note that HyDRA uses the initial 10s of our results as a reference video for its generation. Please refer to the video results on our project page for intuitive demonstrations.

Table 2: **Quantitative ablation results on Appearance Condition**.

Method	PSNR↑	SSIM↑	LPIPS↓	Subject Consistency↑	Background Consistency↑	DSC_DINO↑	DSC_CLIP↑
No $\mathcal{A}$	16.764	0.469	0.385	0.878	0.898	0.693	0.882
No $\mathcal{A}$ + routing	17.461	0.486	0.372	0.881	0.901	0.686	0.886
Ours	**18.127**	**0.502**	**0.359**	**0.891**	**0.909**	**0.769**	**0.917**

accurately generates the user-expected scene and maintains the consistency of the man reappearing after a long period of disappearance. We show more qualitative comparison results in Appendix F.

4.2 Ablation Studies and Promptable World Events

Ablation Studies. We investigate whether the model can implicitly maintain visual consistency without the explicit Appearance Condition. Assuming the unique color-coded masks in the Location Condition could guide appearance retrieval from the contextual frames, we observe that the model fails to autonomously leverage this context, causing severe identity loss for re-entering dynamic objects (Figure 4, second row). Attempting to resolve this without introducing new condition channels, we applied a heuristic self-attention routing strategy to amplify the attention weights between current and contextual dynamic object tokens sharing the same identity. Although this explicit bias captures general styles (the color of people's apparel remains consistent), it fundamentally disrupts the pre-trained latent distribution, inducing severe artifacts, blurring, and the loss of fine-grained textures (Figure 4, third row). We attribute these failures to imbalances in pixel distributions in the training data. Since static backgrounds account for the majority of pixels and dominate the MSE loss, the model struggles to implicitly learn the complex mappings required for high-fidelity consistency in small dynamic regions. Results in Table 2 also show that all metrics drop without the Appearance Condition. This confirms explicitly injecting Appearance Condition is necessary for dynamic memory. We also conduct ablations on *Dynamic Context* and *Appearance Condition Drop* in Appendix D.

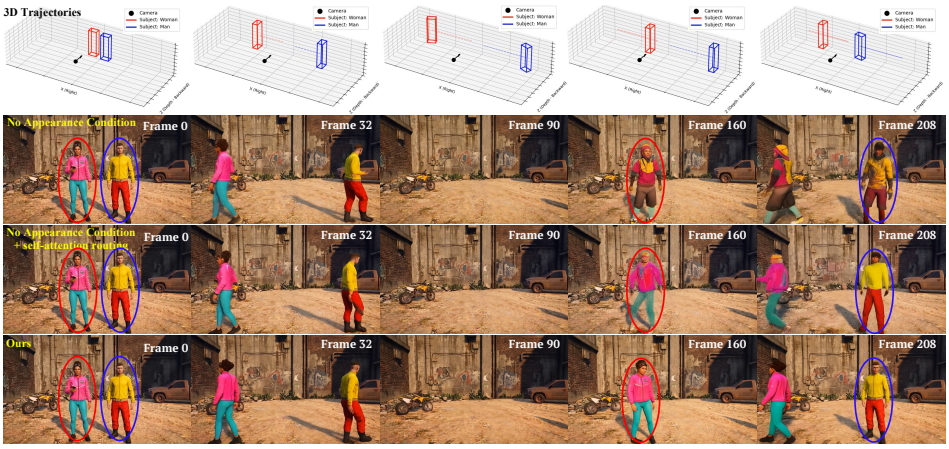

Figure 4: Ablation on Appearance Condition. We conduct experiments on a case involving complex character movements and multiple pose changes. The findings highlight the significance of the Appearance Condition for preserving dynamic consistency.

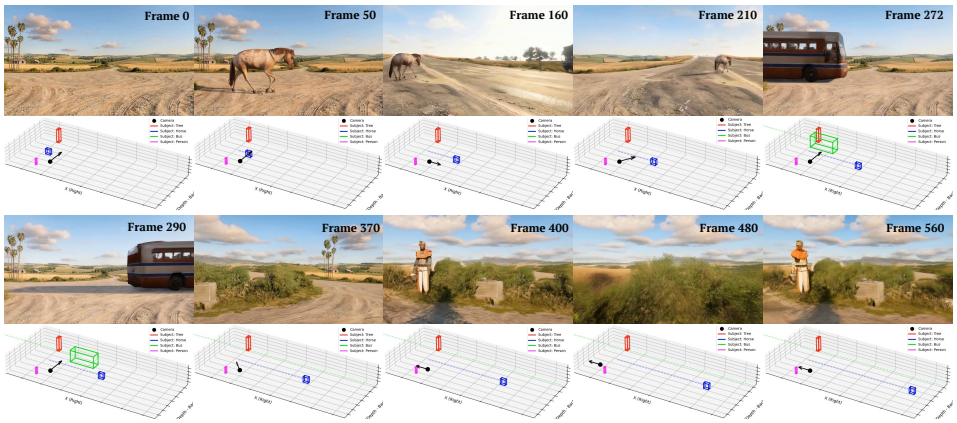

Figure 5: A generation example of Promptable World Events.

Promptable World Events. Our framework is not constrained by the entities present in the initial frame. The LLM can freely populate the simulated world by defining identities, entrance timings, and 3D motion trajectories for novel objects. Upon first entering the camera view, their appearance and movements are synthesized directly from text prompts and appended to the Appearance Condition pool to ensure subsequent temporal consistency. As shown in Figure 5, this paradigm enables the simultaneous choreography of multiple emerging entities alongside unconstrained camera exploration. Consequently, rather than merely extrapolating existing video content, our approach provides a highly controllable mechanism for open-ended scene generation and dynamic environment simulation.

5 Conclusion

We present `WorldDirector`, a novel framework for free exploration and flexible event design in video world models while preserving rigorous dynamic memory. By decoupling semantic orchestration from latent synthesis, `WorldDirector` empowers LLMs to plan complex 3D trajectories and open-world events. Abstract planning is visually realized via causal chunk-based context routing, utilizing spatial and appearance conditioning. Experiments confirm our approach maintains rigorous dynamic consistency, establishing a highly controllable paradigm for future video world models.

Limitation. Relying on synthetic game data introduces a domain gap that occasionally restricts visual fidelity (e.g., unnatural locomotion or blurry faces). Future work will incorporate real-world datasets to bridge this gap and enhance overall visual realism.

References

[1] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[2] Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, et al. Scaling instruction-based video editing with a high-quality synthetic dataset. *arXiv preprint arXiv:2510.15742*, 2025.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[5] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, 2024.

[6] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research (TMLR)*, 2024.

[7] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

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

[9] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, 2024.

[10] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[11] Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *International Conference on Learning Representations (ICLR)*, 2025.

[12] Kaijin Chen, Dingkang Liang, Xin Zhou, Yikang Ding, Xiaoqiang Liu, Pengfei Wan, and Xiang Bai. Out of sight but not out of mind: Hybrid memory for dynamic video world models. *arXiv preprint arXiv:2603.25716*, 2026.

[13] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *International Conference on Learning Representations (ICLR)*, 2024.

[14] Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, Bin Xia, Dingdong Wang, Hongwei Yi, Xihui Liu, Hengshuang Zhao, Yu Liu, Yingya Zhang, and Yujiu Yang. Wan-Move: Motion-controllable video generation via latent trajectory guidance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[15] Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. Project page, 2024. URL `https://oasis-model.github.io/`.

[16] Zicheng Duan, Jiatong Xia, Zeyu Zhang, Wenbo Zhang, Gengze Zhou, Chenhui Gou, Yefei He, Feng Chen, Xinyu Zhang, and Lingqiao Liu. Liveworld: Simulating out-of-sight dynamics in generative video world models. *arXiv preprint arXiv:2603.07145*, 2026.

[17] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

[18] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[19] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

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

[21] Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[22] 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.

[23] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[24] Team HunyuanWorld. Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. *arXiv preprint*, 2025.

[25] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2023.

[27] Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Joshua V. Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A Ross, Bryan Seybold, and Lu Jiang. VideoPoet: A large language model for zero-shot video generation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[28] Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. MagicMotion: Controllable video generation with dense-to-sparse trajectory guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[29] Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[30] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[31] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

[32] Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025.

[33] 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.

[34] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, 2024.

[35] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023.

[36] Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

[37] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[38] Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

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

[40] Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Qifeng Chen, Yujun Shen, and Limin Wang. Levitor: 3d trajectory oriented image-to-video synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[41] Hanlin Wang, Hao Ouyang, Qiuyu Wang, Yue Yu, Yihao Meng, Wen Wang, Ka Leong Cheng, Shuailei Ma, Qingyan Bai, Yixuan Li, et al. The world is your canvas: Painting promptable events with reference images, trajectories, and text. *arXiv preprint arXiv:2512.16924*, 2025.

[42] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024.

[43] Zhouxia Wang, Ziyang Yuan, Xintao Wang, 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*, 2024.

[44] Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, et al. Infinite-world: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. *arXiv preprint arXiv:2602.02393*, 2026.

[45] Wejia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. DragAnything: Motion control for anything using entity representation. In *European Conference on Computer Vision (ECCV)*, 2024.

[46] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[47] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[48] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020.

[49] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

[50] 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.

[51] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[52] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023.

[53] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

A Training and Compute Details.

In this section, we provide a more comprehensive breakdown of the training configuration and computational resources.

The training process for `WorldDirector` is conducted on a high-performance cluster utilizing 8 compute nodes. Each node is equipped with 8 NVIDIA A100 (80GB) GPUs, amounting to a total of 64 GPUs. To maximize memory efficiency and training throughput across this large-scale distributed setup, we employ Fully Sharded Data Parallel (FSDP) alongside activation checkpointing. This system-level optimization ensures that the memory footprint of the 3D VAE encodings, diffusion transformer blocks, and multi-modal conditioning channels is efficiently distributed, preventing out-of-memory bottlenecks when processing high-resolution video chunks and extended context memory.

As formulated in our method, the model processes training videos at a resolution of 832×480 pixels at 16 fps, with the context length explicitly set to $N = 10$ frames. We optimize the flow matching objective utilizing the AdamW optimizer with a constant learning rate of 1×10^{-5}. To accelerate computation while maintaining numerical stability during the denoising process, the training is conducted using BFloat16 (BF16) mixed precision. Operating with a global batch size of 64, the model undergoes 3,000 optimization steps. Under the 64-GPU distributed configuration, the entire post-training pipeline takes approximately 72 hours (3 days) to fully converge.

B Details of Static and Dynamic Context Retrieval.

Algorithm 1 Static and Dynamic Context Retrieval

Input: Candidate Context Frames $\mathcal{F}$, Training Frames $\mathcal{V}$, Camera Poses $\mathcal{C}$,
2D Bounding Boxes $\mathcal{B}$, Context Length N
Output: Static and Dynamic Context $\mathcal{M}$

procedure STATIC_CONTEXT($\mathcal{F}$, $\mathcal{V}$, $\mathcal{C}$)
  for each candidate $c \in \mathcal{F}$ **do**
    $\mathrm{score}(c) \leftarrow \max_{v \in \mathcal{V}} \mathrm{FoV\_Overlap}(\mathcal{C}_c, \mathcal{C}_v)$
  end for
  return $\mathcal{F}$ sorted by $\mathrm{score}(\cdot)$ descending
end procedure

procedure DYNAMIC_CONTEXT($\mathcal{F}$, $\mathcal{B}$, N, $\mathcal{V}$)
  Initialize coverage $\mathrm{cnt}[i] \leftarrow 0$ for each dynamic entity i appears in $\mathcal{V}$
  while $|\mathrm{selected}| < N$ **do**
    Find entity $i^* = \arg\min_i \mathrm{cnt}[i]$ ▷ *least-covered entity so far*
    Select frame $f^* \in \mathcal{F}$ with largest $\mathrm{area}(\mathcal{B}_f^{(i^*)})$ ▷ *best visible context for entity i^**
    Add f^* to selected. Remove f^* from $\mathcal{F}$
    $A_{\max} \leftarrow \max_e \mathrm{area}(\mathcal{B}_{f^*}^{(e)})$ ▷ *largest bbox area in f^*, used as normalizer*
    for each entity j in f^* **do**
      $\mathrm{cnt}[j]$ += $\mathrm{area}(\mathcal{B}_{f^*}^{(j)}) / A_{\max}$ ▷ *normalized bbox area as visibility weight*
    end for
  end while
  return selected
end procedure

$\mathcal{P}_{\mathrm{cam}} \leftarrow$ STATIC_CONTEXT($\mathcal{F}$, $\mathcal{V}$, $\mathcal{C}$)
$\mathcal{P}_{\mathrm{box}} \leftarrow$ DYNAMIC_CONTEXT($\mathcal{F}$, $\mathcal{B}$, N)
$\mathcal{M} \leftarrow [\,]$
for $k = 0, 1, \ldots$ **until** $|\mathcal{M}| = N$ **do**
  Append $\mathcal{P}_{\mathrm{cam}}[k]$ to $\mathcal{M}$ if not already in $\mathcal{M}$
  Append $\mathcal{P}_{\mathrm{box}}[k]$ to $\mathcal{M}$ if not already in $\mathcal{M}$
end for
return $\mathcal{M}$ sorted by temporal order

In this section, we elaborate on the specific implementation details of the static and dynamic context retrieval mechanism (as outlined in Algorithm 1). This algorithm aims to select N memory frames from a candidate frame set $\mathcal{F}$ to construct the final context set $\mathcal{M}$. The inputs primarily include the candidate frames $\mathcal{F}$, the current training frames $\mathcal{V}$, the camera poses $\mathcal{C}$, and the 2D bounding boxes of dynamic entities $\mathcal{B}$. The retrieval process consists of two parallel scoring modules and a subsequent interleaving fusion module:

Static Context Retrieval. The static retrieval module (STATIC_CONTEXT) aims to find contextual frames with the highest viewpoint overlap to provide comprehensive static background information. For each frame c within the candidate set $\mathcal{F}$, the algorithm calculates the Field of View (FoV) overlap between its camera pose $\mathcal{C}_c$ and all training frame poses $\mathcal{C}_v$, taking the maximum overlap value as the candidate's score. Subsequently, all candidate frames are sorted in descending order based on these scores to generate the static context candidate list $\mathcal{P}_{\text{cam}}$.

Dynamic Context Retrieval. To ensure balanced spatio-temporal coverage of dynamic objects, we maintain a coverage counter $\text{cnt}[i]$ (initialized to 0) for each dynamic entity $i \in \mathcal{V}$. In each greedy iteration, we identify the least-covered entity i^* and retrieve the frame $f^* \in \mathcal{F}$ that maximizes its visible 2D bounding box area. We then update the coverage for all entities j present in f^* by adding their bounding box areas, normalized by the maximum bounding box area $A_{\max}$ in f^*. This process repeats until sufficient frames are gathered, yielding the dynamic list $\mathcal{P}_{\text{box}}$.

Interleaving and Fusion. We alternately append frames from $\mathcal{P}_{\text{cam}}$ and $\mathcal{P}_{\text{box}}$ to the final memory set $\mathcal{M}$. During this step, duplicates are discarded, and a minimum temporal stride of four frames is rigidly enforced to guarantee a uniform distribution. Interleaving terminates once $|\mathcal{M}| = N$, and the context frames are returned in chronological order.

C Details of Inference System

Our inference system comprises two main components: World Planning via LLM and Causal Chunk-Based Generation. In this section, we elaborate on the specific details of these operations.

C.1 World Planning via LLM

We employ Gemini [37] as the core semantic engine for world planning. Specifically, given an initial frame, we first select the dynamic objects of interest. We leverage SAM [26] and DepthAnything v2 [47] to roughly estimate the 3D bounding boxes of these objects and establish the initial orientations for both the entities and the camera.

This structured information is then fed into the LLM, prompting it to analytically plan the corresponding 3D trajectories based on our customized narrative design. An example of the prompt we utilize is structured as follows:

> *You are an expert with strong 3D spatial imagination capabilities. Given the following information:*

```
{
  "coordinate_system": "OpenGL (X-right, Y-up, Z-backward)",
  "camera_position": {"position": [0.0, 0.0, 0.0]},
  "camera_intrinsics": {
    "fx": 565.4046, "fy": 565.4046, "cx": 416.0, "cy": 240.0,
    "image_width": 832, "image_height": 480, "fov_v_deg": 46
  },
  "ground_height_y": [-1.248],
  "bboxes_3d": [
    {
      "bbox_3d": {
        "center": [-1.7175, -0.4237, -3.4715],
        "dimensions": [0.6799, 1.8, 0.442],
        "rotation_yaw_deg": 90,
        "prompt": "A woman walks on the road."
      }
    },
    {
```

```
      "bbox_3d": {
        "center": [3.227, -0.8014, -7.1893],
        "dimensions": [1.2749, 1.1321, 3.1872],
        "rotation_yaw_deg": 90,
        "prompt": "A car first keep still, then starts driving on the road."
      }
    }
  ]
}
```


Here, "coordinate_system" indicates the 3D coordinate system; "camera_position" refers to the initial camera location; "camera_intrinsics" specifies the camera parameters; "ground_height_y" is the y-coordinate of the ground; "bboxes_3d" contains information for multiple subjects. For each "bbox_3d", "center" represents the initial 3D center coordinates, "dimensions" denotes the actual width, height, and length of the object, and "rotation_yaw_deg" is the initial yaw angle. The "prompt" provides the textual description for the trajectory generation.

User Instruction: *Please help me generate the corresponding 3D bbox trajectories and camera poses for these subjects. The total duration is 15s at 16 fps. The initial camera position is at the origin, with the pose as an identity matrix facing the -Z direction.*

0-5s: *The camera and the car remain stationary. The woman walks forward along the +X direction, reaching the edge of the camera view at 5s.* ***5-10s:*** *At 5-6s, the camera rotates from -Z to +X. From 6-10s, the camera moves forward along +X, overtaking the woman. The woman continues walking along +X, while the car remains stationary.* ***10-15s:*** *At 10s, the camera stops. From 10-11s, it rotates from +X to -Z, and from 11-15s, it remains strictly stationary. The woman walks along +X and re-enters the camera view at 11s, then continues walking within the frame. The car starts driving along +X at 10s, enters the camera view at 11s, and exits at 14s.*

Return the Python code to generate the above 3D bbox trajectories and camera poses, and visualize them. Simultaneously, for each generated frame, project the 3D bboxes onto the camera plane using the current camera pose to generate 2D bboxes. Write the projected 2D bboxes into the final output and visualize them.

Beyond planning trajectories for objects explicitly selected in the initial frame, users can also define motion paths for completely novel objects within the prompt. The LLM demonstrates remarkable capability in automatically synthesizing physically plausible kinematics for these newly introduced entities (i.e., Promptable World Events).

Consequently, we obtain the complete 3D trajectories of all dynamic objects and their corresponding 2D bounding box sequences projected onto the camera plane. These 2D projection sequences directly serve as the deterministic Spatial Location Condition ($\mathcal{B}$) applied in the subsequent generative stage.

C.2 Causal Chunk-Based Generation

Building upon the projected 2D Location Condition and camera trajectories from the planning phase, we execute the video synthesis in a causal autoregressive manner, as detailed in Algorithm 2.

Given the complete spatial location sequence $\mathcal{B}_{1:T}$ and camera poses $\mathcal{C}_{1:T}$, we first partition them into N sequential chunks of size K. The generative process maintains a global continuous video buffer V, which is initialized with the starting reference frame I_0.

During the iterative generation, the conditioning strategy adapts based on the temporal state. For the first chunk ($n = 1$), the model extracts the Appearance Condition ($\mathcal{A}$) directly from I_0 guided by the Location Condition $\mathcal{B}$, while the historical context memory ($\mathcal{M}$) remains empty as there is no preceding temporal information. For all subsequent chunks ($n > 1$), the framework dynamically constructs $\mathcal{A}$ and retrieves the historical Context ($\mathcal{M}$) from the previously generated video buffer V. This retrieval mechanism strictly leverages the location constraints $\mathcal{B}$ and camera parameters $\mathcal{C}$ to fetch precise dynamic entity identities and static background anchors.

Crucially, to guarantee temporal smoothness at the chunk boundaries, the last frame of the current buffer (V_{last}) serves as the conditional initial frame (I_{start}) for generating the next chunk $V^{(n)}$. The core diffusion model, `WorldDirector`, processes these multimodal conditions ($\mathcal{B}^{(n)}, \mathcal{P}, \mathcal{A}, \mathcal{M}, \mathcal{C}^{(n)}, I_{start}$) to synthesize the current segment. Finally, we append the generated

Algorithm 2 Causal Chunk-Based Generation

Input: Location condition sequence $\mathcal{B}_{1:T}$, Captions $\mathcal{P}$, Initial reference frame I_0, Total frames T, Chunk size K, Camera Poses $\mathcal{C}_{1:T}$
Output: Generated continuous video stream V

procedure CAUSAL_GENERATION($\mathcal{B}_{1:T}$, $\mathcal{P}$, I_0, T, K, $\mathcal{C}_{1:T}$)
  Partition $\mathcal{B}_{1:T}$ into $N = T/K$ chunks: $\{\mathcal{B}^{(1)}, \ldots, \mathcal{B}^{(N)}\}$
  Partition $\mathcal{C}_{1:T}$ into $N = T/K$ chunks: $\{\mathcal{C}^{(1)}, \ldots, \mathcal{C}^{(N)}\}$
  Initialize video buffer $V \leftarrow I_0$
  for $n = 1, 2, \ldots, N$ **do**
    if $n = 1$ **then**
      $\mathcal{A} \leftarrow$ APPEARANCE_CONDITION_GENERATION($\mathcal{B}$, I_0)
      $\mathcal{M} \leftarrow \emptyset$ ▷ *no historical context for the first chunk*
    else
      $\mathcal{A} \leftarrow$ APPEARANCE_CONDITION_GENERATION($\mathcal{B}$, V)
      $\mathcal{M} \leftarrow$ CONTEXT_RETRIEVAL($\mathcal{B}$, $\mathcal{C}$, V)
    end if
    $I_{start} \leftarrow V_{\text{last}}$
      ▷ *use the last frame of buffer V as initial frame for current chunk generation*
    $V^{(n)} \leftarrow \texttt{WorldDirector}(\mathcal{B}^{(n)}, \mathcal{P}, \mathcal{A}, \mathcal{M}, \mathcal{C}^{(n)}, I_{start})$
    $V \leftarrow V \cup V^{(n)}[1:]$ ▷ *append generated chunk without first frame to the buffer*
  end for
  return V
end procedure

frames to V—excluding the overlapping first frame to prevent redundancy—thereby progressively unrolling the long-horizon video stream without inherent length limitations.

D Ablation on Dynamic Context and Appearance Condition Drop Mechanism

We further evaluate the efficacy of the retrieved dynamic context and the Temporal Drop Mechanism. Despite the strong visual priors from the Appearance Condition, ablating the dynamic context stream confirms the necessity of retrieving dynamic objects within the contextual memory. As shown in Figure S1, relying solely on Appearance Condition for re-entering dynamic entities degrades identity preservation; the model generates semantically similar but non-identical entities. This demonstrates that dynamic context is indispensable for temporally anchoring the specific object identity across causal chunks. Furthermore, we validate the necessity of the Temporal Drop Mechanism. Removing this exposes the network to dense, frame-by-frame appearance references, which induces severe motion rigidity (e.g., characters "sliding" rather than walking naturally, as depicted in Figure S1). This evidence substantiates our design: the Temporal Drop Mechanism effectively prevents overfitting to static reference images, compelling the model to synthesize fluid, text-driven dynamics rather than executing rigid image warping.

E Flexible Viewpoint Control

By explicitly incorporating the spatial location condition, our framework intrinsically supports flexible viewpoint control, enabling seamless transitions between first- and third-person exploration paradigms. Specifically, during 3D trajectory planning, anchoring the 2D bounding box of a target dynamic entity near the center of the camera's field of view yields a third-person perspective. Conversely, decoupling the camera trajectory from dynamic objects allows for independent first-person navigation. As illustrated in Figure S2, the first scenario demonstrates pure third-person exploration, where the camera follows a running dog while simultaneously performing a continuous 360° panoramic sweep of the surrounding scene. The second scenario highlights dynamic viewpoint switching within a single sequence: the initial two temporal chunks maintain a third-person perspective following a human character, whereas the third chunk smoothly transitions to a first-person view as the camera

Figure S1: Ablation on Dynamic Context and Appearance Condition Drop Mechanism. We highly recommend viewing the video results on our project page for a more intuitive demonstration.

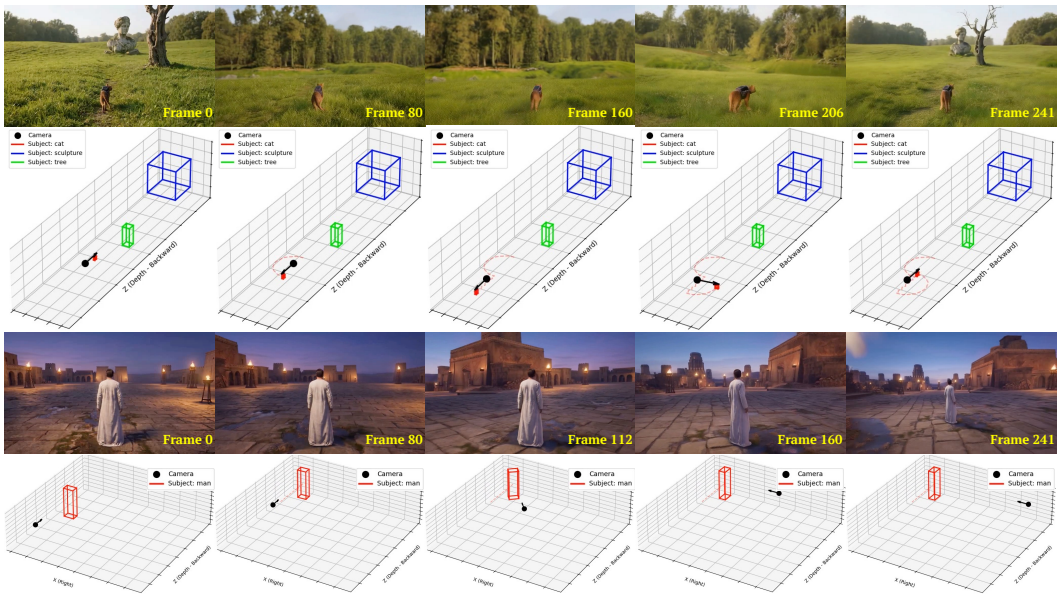

Figure S2: **Flexible Viewpoint Control.** `WorldDirector` supports diverse exploration paradigms. Top: A pure third-person view tracking a running dog with a 360° panoramic sweep. Bottom: A dynamic viewpoint switch from a third-person tracking shot to an independent first-person backward movement.

detaches and moves backward. These capabilities underscore the model's profound flexibility in directing simulated environments.

F More Qualitative Comparisons

We provide additional qualitative comparison results in Figure S3 and Figure S4 to further evaluate the baselines. The observations remain consistent with our main findings in Section 4.1. Specifically, Yume, HY-World, and Infinite-World tend to generate significantly less subject motion. LingBot-World successfully produces highly dynamic results that align well with the textual prompts. However, it lacks fine-grained interactive control precision, making it difficult to strictly match the user's specific scenario designs. HyDRA consistently exhibits a strong bias towards generating a prominent subject walking directly in front of the camera, which is likely an artifact of its training data distribution. In contrast, our method accurately executes the user's intended spatial layout and maintains precise interactive control and dynamic memory.

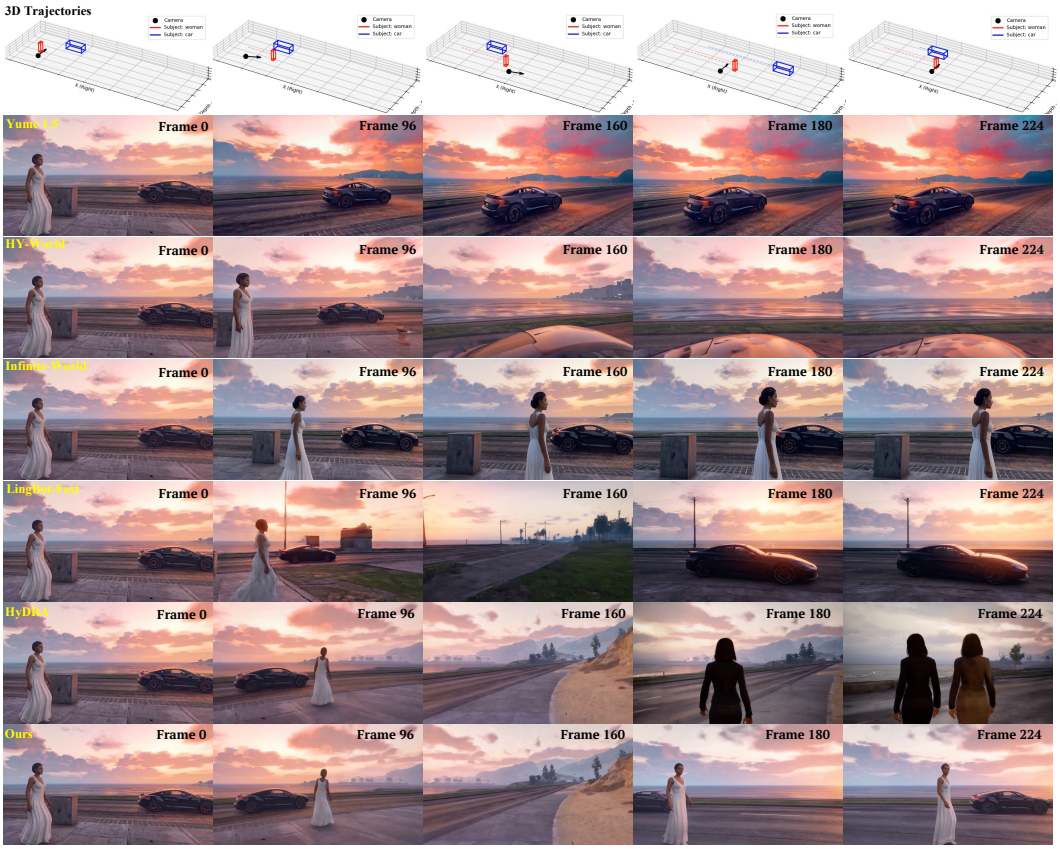

Figure S3: **Qualitative comparison with baselines.** Note that HyDRA uses the initial 10s of our results as a reference video for its generation. We highly recommend viewing the video results on our project page for a more intuitive demonstration.

G Impact Statement

This paper focuses on the technical advancements in controllable video world simulation with persistent dynamic memory. The work aims to enhance applications in virtual reality, gaming, film-making, and interactive design, which could have positive societal implications in these domains. However, this study does not directly address potential societal impacts, including possible negative consequences such as malicious or unintended uses (e.g., generating deceptive or fake video content), fairness considerations, privacy concerns, or security risks that might arise from the application of this generative technology. The paper primarily presents foundational technical research and does not discuss the commercial deployment of the technology or specific mitigation strategies for these negative impacts.

H Responsible Release and Safeguards

Because `WorldDirector` is a highly controllable generative video model driven by an LLM orchestrator, we plan a staged and documented release. We will release the inference codebase, LLM prompt templates, and pre-trained model checkpoints strictly intended for academic research and evaluation.

For downstream applications, we strongly recommend combining `WorldDirector` with established safety mechanisms that fall outside the scope of this foundational paper. These include LLM prompt safety filters to prevent malicious planning, generated-video watermarking, content provenance metadata, and deployment-time monitoring. Given that our framework facilitates the continuous generation of long-horizon events with persistent entities, it inherently lowers the barrier for creating complex, logically coherent synthetic scenarios. Therefore, these external safeguards remain an

Figure S4: **Qualitative comparison with baselines.** Note that HyDRA uses the initial 10s of our results as a reference video for its generation. We highly recommend viewing the video results on our project page for a more intuitive demonstration.

important consideration to mitigate the general risks of misuse associated with video generation technologies.