Spatially-Grounded Long Video Generation with Self Geometry Forcing

Chenguo Lin[1], **Panwang Pan**[2], **Bowen Xue**[3], **Ruijie Lu**[1],
Yuchen Lin[1], **Chenxin Li**[4], **Brandon Y. Feng**[5], **Yadong Mu**[1]
[1]Peking University, [2]Xiamen University, [3]Stanford University, [4]CUHK, [5]MIT
https://chenguolin.github.io/projects/jag



Figure 1: **Overview**. JAG is a spatially-grounded causal video diffusion model. Under interactive camera control, it streams long videos with coherent appearance and geometry in real time.

Abstract

Recent video generative models can produce photorealistic short clips, but struggle with long-horizon generation under dynamic camera control, often suffering from viewpoint drift, geometric inconsistency, and error accumulation. We propose **J**oint **A**ppearance-**G**eometry Diffusion Transformer (**JAG**), a spatially-grounded causal video diffusion model that explicitly integrates 3D perception for long and consistent video generation. It adopts a novel **dual-branch causal architecture** that jointly models visual appearance and scene geometry. Building upon this formulation, we introduce **Self Geometry Forcing**, a geometry-aware distillation scheme that conditions current generation on self-predicted geometric structures as persistent long-term memory and imposes self-verified geometric rewards on intermediate rollouts, effectively mitigating error accumulation over long horizons. Extensive experiments show that JAG can generate long videos with superior coherent appearance and geometry at constant per-frame computation, improving long-horizon controllability, geometric consistency, and visual fidelity over previous state-of-the-art methods.

1 Introduction

Advanced video world models [1, 2] have demonstrated the potential of video generative models as simulators for the real world. However, most video diffusion models focus on visually appealing short clips [3, 4, 5, 6, 7], rather than controllable, long-horizon, and physically grounded generation. Toward world modeling, it requires: (1) **controllability**, enabling accurate action- or camera-driven scene evolution, (2) **explorability**, supporting extendable and long-horizon generation, and (3) **consistency**, maintaining coherent appearance and geometry across frames.

Prior studies [8, 9, 10, 11, 12, 13, 14, 15] have explored incorporating spatial cues into conditional video generation with actions or camera trajectories. But they typically use bidirectional attention

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

over short clips, limiting their scalability to long sequences and streaming scenarios. For long-horizon generation, several recent works [16, 17, 18, 19, 20, 21, 22] propose to fine-tune or distill bidirectional diffusion models into causal counterparts, leveraging key-value (KV) caches and sliding window attention for explorable and efficient inference. However, these methods merely emphasize visual continuity while largely neglecting explicit 3D geometric coherence, leading to accumulated viewpoint drift and inconsistent cross-view structure over long horizons.

To address these challenges, we present **J**oint **A**ppearance-**G**eometry Diffusion Transformer (**JAG**), a spatially-grounded causal video diffusion model for consistent long video generation. When viewpoints evolve, maintaining coherent appearance depends on capturing and propagating the underlying geometric correspondences across frames. Thus, JAG adopts a **dual-branch causal architecture** that jointly models visual appearance and scene geometry within a unified auto-regressive diffusion process. Specifically, alongside a standard video diffusion transformer (DiT) branch [23], a geometry branch predicts geometric signals directly from noisy latents. To effectively leverage 3D priors, this branch is initialized from a model pretrained on large-scale geometric datasets [24] and adapted to operate on the diffusion latent space. The two branches are coupled bidirectionally at every layer, so that explicit geometry directly conditions visual denoising.

Building on the dual-branch formulation, we further introduce **Self Geometry Forcing**, which leverages explicit 3D perception in self-supervised post-training. While the self-forcing paradigm [19] primarily aligns appearance distributions via self-rollouts, we reformulate it in a geometry-grounded manner by conditioning each generation step on its own predicted historical 3D structure as persistent long-term spatial memory. Moreover, geometry predictions are used as self-verified rewards to enforce viewpoint compliance and cross-view consistency during intermediate rollouts, effectively stabilizing long-horizon generation, mitigating viewpoint drift, and preserving spatial coherence under dynamic camera motion.

Owing to the flexible and unified appearance-geometry modeling, JAG enables accurate viewpoint control in streaming video generation, while preserving high visual fidelity and long-range spatial consistency. We conduct a comprehensive evaluation of JAG using both perceptual and geometry-aware metrics, covering short and long horizon settings. Extensive quantitative and qualitative results demonstrate that JAG outperforms strong bidirectional and causal baselines across hundreds of frames, exhibiting reduced viewpoint drift and improved geometric coherence. Detailed ablation studies validate the effectiveness of each core design component.

2 Related Work

2.1 Long Video Generation

Most existing video generation models are limited to producing only a few seconds of content [3, 4, 6, 25]. To extend generation to longer horizons, recent studies have shifted from bidirectional diffusion to autoregressive modeling [16, 26]. CausVid [27] distills bidirectional video diffusion models into causal ones using distribution matching distillation (DMD) [28, 29], while Self-Forcing [19] further mitigates exposure bias by training causal generators to condition on their own imperfect past predictions. Subsequent works enhance this paradigm through reinforcement learning [30, 31], rolling denoising [22], sliding window with global sink [21], or sophisticated context selection or compression strategies [32, 33], enabling minute-level video generation. Orthogonal to causal modeling, another line of work focuses on scalable context compression for long-video conditioning. FramePack [34, 35] encodes arbitrarily many frames into fixed-dimensional representations, while TTTVideo [36] and LaCT [37] leverage test-time training [38] to adaptively encode historical context during inference. LCT [39] extends single-shot video diffusion models to multi-shot long-video generation by conditioning on all preceding shots, while Mixture-of-Context [40] further formulates long-context generation as internal information retrieval with sparse attention routing for long-term memory. Despite significant progress, these methods primarily emphasize visual continuity and temporal scalability, while largely neglecting explicit 3D geometric structure. As a result, they remain prone to error accumulation and spatial inconsistency under long-horizon generation, limiting their suitability for world modeling with dynamic camera motion.

2.2 Spatially-aware World Modeling

For controllable world modeling, early studies [9, 41, 42, 10] inject camera information into diffusion models via architectural designs, such as addition, concatenation, or ControlNet [43]. To further

encourage spatial awareness, some approaches jointly generate RGB and depth latents [44, 12, 45, 13, 46, 47] or incorporate a 3D module trained from scratch [48], while others reconstruct explicit 3D representations to maintain view consistency across frames [49, 50, 51, 52, 53]. Recent methods also explore selection-based memory mechanisms that retrieve relevant contexts via field-of-view overlap [14, 54, 55]. Geometry Forcing [56] injects geometry implicitly by aligning intermediate diffusion features with a frozen 3D reconstruction model [57]. Long-term spatial memory [53] and GEN3C [51] instead keep explicit point-cloud memories, but build them with external estimators applied to generated frames. JAG differs along both axes: it predicts explicit depth and camera in-network from the same noisy latents, which keeps all chunks in one coordinate frame and turns geometry into self-verified rewards. Moreover, most of these methods operate on short bidirectional contexts, whereas JAG targets streaming generation over hundreds of frames.

3 Method

3.1 Problem Statement

Given streaming interactive signals $\{\mathbf{u}^s := (\mathbf{C}^s, \mathbf{T}^s)\}_{s=1}^{S\to\infty}$, where $\mathbf{C}^s$ and $\mathbf{T}^s$ denote the camera parameters and textual prompt at state s, we aim to generate a long video frame sequence $\{\mathbf{F}^s\}_{s=1}^{S\to\infty}$ in an auto-regressive manner, while preserving visual continuity and physically plausible spatial structure. Formally, we model the distribution:

$$p(\mathbf{F}^{1:S} \mid \mathbf{u}^{1:S}) := \prod_{s=1}^{S} p(\mathbf{F}^s \mid \mathbf{F}^{<s}, \mathbf{u}^{\leq s}). \tag{1}$$

The challenge stems from two aspects: (1) **lack of spatial awareness**, where visual-only denoising yields geometrically inconsistent and physically implausible content under dynamic camera motion, and (2) **error accumulation**, where long-horizon generation progressively suffers from historical forgetting and spatial drift.

We propose JAG with two components: (1) a **spatially-grounded causal architecture** (Section 3.2) that jointly models visual appearance and scene geometry within a unified auto-regressive diffusion process, and (2) **Self Geometry Forcing** (Section 3.3), a self-supervised post-training scheme with self-conditioned 3D memory and self-verified geometric rewards.

3.2 Spatially-grounded Causal Generation

3.2.1 Dual-branch Architecture

Given noisy video diffusion latents $\mathbf{X}_t^s$ at timestep t and state s, obtained from video frames $\mathbf{F}$ via a pretrained video VAE [6], we decompose denoising into two coupled sub-tasks: (i) recovering visual appearance and (ii) inferring the underlying 3D structure that explains the observation. Instead of a single denoising function $f_\theta(\mathbf{X}_t^s, t, \mathbf{u}^s)$, we adopt a dual-branch formulation:

$$\hat{\mathbf{X}}_0^s = f_\theta^{\text{app}}(\mathbf{X}_t^s, t, \mathbf{u}^s, \mathbf{Z}_t^s), \quad \hat{\mathbf{Z}}_0^s = f_\phi^{\text{geo}}(\mathbf{X}_t^s, t, \mathbf{u}^s), \tag{2}$$

where the appearance branch $f_\theta^{\text{app}}(\cdot)$ predicts denoised visual latents, and the geometry branch $f_\phi^{\text{geo}}(\cdot)$ estimates explicit geometric attributes $\hat{\mathbf{Z}}_0^s$ (e.g., depth and camera rays) from the same noisy inputs.

Dual-branch Interaction As the DiT backbone progressively removes noise across layers, to expose the geometry branch to meaningful features, the two branches share early transformer layers of the DiT backbone and then specialize in appearance denoising and geometric reasoning, respectively. They further interact at each layer through lightweight bidirectional residual connections:

$$\mathbf{X}_i' = \mathbf{X}_i + f_{\xi_{\text{app}}}([\mathbf{X}_i, \mathbf{Z}_i]), \ \mathbf{Z}_i' = \mathbf{Z}_i + f_{\xi_{\text{geo}}}([\mathbf{X}_i, \mathbf{Z}_i]), \tag{3}$$

where i indexes network layers and $[\cdot,\cdot]$ denotes dimension-wise concatenation. t and s are omitted here for simplicity. The interaction module enables geometric cues to guide visual denoising while appearance features provide contextual grounding for geometry estimation.

Geometric Priors Different from previous studies that either compress geometric attributes into latent space [12, 13, 47] or train 3D modules from scratch [48], to leverage strong geometric priors, we initialize the geometry branch with an advanced 3D foundation model, Depth-Anything-3 (DA3) [24].

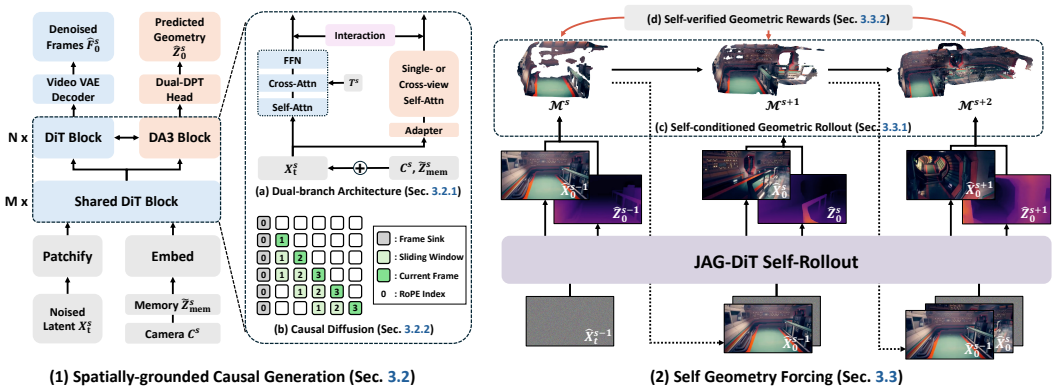

Figure 2: **JAG Framework** includes two components: (1) **Spatially-grounded Causal Generation** (Section 3.2) that jointly models visual appearance and scene geometry within a unified causal diffusion process, and (2) **Self Geometry Forcing** (Section 3.3), a self-supervised post-training scheme with self-conditioned persistent spatial memory and self-verified geometric rewards.

A linear layer adapts diffusion features to the geometry encoder input, with the encoder fine-tuned and the original decoding heads frozen. Thus, the geometry branch learns to map noisy diffusion latents into the original geometric feature space, enabling reliable depth and camera prediction without re-learning 3D perception from scratch. Note that *joint* refers to modeling appearance and geometry within one causal process: appearance is denoised, while geometry is regressed from the same latents, which yields clean depth and camera after every causal step for memory and rewards.

3.2.2 Causal Diffusion

To support explorable and interactive video generation, we adopt a causal diffusion formulation that factorizes the video distribution along the temporal order (Equation 1). Each conditional state $p(\mathbf{X}^s \mid \mathbf{X}^{<s}, \mathbf{Z}^{<s}, \mathbf{u}^{\leq s})$ and $p(\mathbf{Z}^s \mid \mathbf{Z}^{<s}, \mathbf{X}^{<s}, \mathbf{u}^{\leq s})$ is generated progressively by the proposed dual-branch diffusion process and conditioned only on previously generated content. During training, block-wise sparse attention masks are applied to all self-attention layers in both branches, where each noised latent $\mathbf{X}_t^s$ and $\mathbf{Z}_t^s$ attends only to itself and historical observations $\mathbf{X}_{t'}^{\leq s}$ and $\mathbf{Z}_{t'}^{\leq s}$ at random noise levels t', following the diffusion forcing paradigm [16, 26].

Sliding Window and Frame Sink To handle the growing context in long-horizon generation, we replace causal attention with *sliding window attention* in both branches, which bounds memory and computation while retaining local temporal continuity. Similar to LLMs [58, 59], we find that the first generated block plays a critical role in maintaining global coherence, encoding scene layout, appearance statistics, and illumination. Moreover, for the geometry branch, it additionally defines the canonical coordinate system and spatial scale. Accordingly, we retain all tokens of the first block in the attention window as *frame sink*, providing a stable global reference throughout generation.

Re-RoPE To avoid extrapolation issues when combining the frame sink and window attention with rotary positional embeddings (RoPE) [60], we discard absolute temporal indices and dynamically re-assign positional encodings within the attention window, coined as *re-RoPE*. This ensures that **both** training and inference operate within a fixed positional range, preserving the influence of critical past frames and enabling robust long-horizon extrapolation.

3.3 Self Geometry Forcing

3.3.1 Preliminary: Self Forcing

Parallel training for auto-regressive generation suffers from a pronounced train-test gap [61], as the context during training differs from the self-generated history encountered at inference time. Self Forcing [19] mitigates this issue by conditioning each step on self-rollout predictions, thereby exposing the model to its own generation distribution. Combined with Distribution Matching Distillation (DMD) loss [29, 27], it aligns the holistic video-level distribution of causal generations

with that of a bidirectional teacher:

$$\nabla_\theta \mathcal{L}_{\text{DMD}}(\theta) \approx -\mathbb{E}_{\hat{\mathbf{X}}_0,t}\left[\left(S_{\text{real}}(\hat{\mathbf{X}}_0,t) - S_{\text{fake}}(\hat{\mathbf{X}}_0,t)\right)\frac{\partial \hat{\mathbf{X}}_0}{\partial \theta}\right], \tag{4}$$

where S_{real} and S_{fake} denote the score functions [62] of the frozen bidirectional teacher and its replica trained online with outputs of the student generator f_θ.

Thanks to the dual-branch design in Section 3.2, spatial awareness can be explicitly endowed into the distillation process, a direction largely unexplored by prior causal distillation methods. Based on this insight, we introduce *Self Geometry Forcing*,which extends Self Forcing from purely visual alignment to geometry-aware distillation from two complementary perspectives: (i) long-term spatial memory (Section 3.3.2) and (ii) accurate long-horizon control (Section 3.3.3).

3.3.2 Self-conditioned Geometric Rollout

A key limitation of Self Forcing is that its historical context is implicitly encoded only through recent visual tokens within a limited context window, which is insufficient for preserving global spatial structure over long horizons. Leveraging the explicit geometry predicted via the dual-branch architecture, we introduce a *persistent spatial memory* that enables geometry-conditioned self-rollouts. Specifically, we maintain a memory state $\mathcal{M}^S$ that aggregates 3D structures inferred from previous self-rollouts as a set of geometry primitives $\mathbf{G}^s$ throughout inference:

$$\mathcal{M}^S \coloneqq \{\mathbf{G}^s\}_{s=1}^{S-1} = \left\{\pi\Big(h_{\text{depth}}(\hat{\mathbf{Z}}_0^s),\, h_{\text{cam}}(\hat{\mathbf{Z}}_0^s)\Big)\right\}_{s=1}^{S-1}, \tag{5}$$

where $\pi(\cdot)$ denotes a confidence-filtered unprojection into a canonical world coordinate system, and h_{depth} and h_{cam} are the frozen depth and camera heads of the geometry branch.

Memory Read At the rollout step s, we retrieve relevant spatial context by rendering the persistent memory into the current target viewpoint $\mathbf{C}^s$:

$$\tilde{\mathbf{Z}}_{\text{mem}}^s \coloneqq \mathcal{R}(\mathcal{M}^s, \mathbf{C}^s), \tag{6}$$

where $\mathcal{R}(\cdot)$ denotes a differentiable projection operator. The retrieved memory $\tilde{\mathbf{Z}}_{\text{mem}}^s$ conditions both the appearance and geometry branches during self-rollout generation:

$$(\hat{\mathbf{X}}_0^s, \hat{\mathbf{Z}}_0^s) = f_{\theta,\phi}(\mathbf{X}_t^s, \mathbf{Z}_t^s \mid \mathbf{X}_0^{<s}, \tilde{\mathbf{Z}}_{\text{mem}}^s, \mathbf{u}^s). \tag{7}$$

Memory Update The new predicted geometry $\hat{\mathbf{Z}}_0^s$ is unprojected into a candidate point set $\mathcal{P}^s = \{(\mathbf{x}_i, \mathbf{c}_i, \alpha_i)\}$ (3D position, RGB tuple from the appearance branch, confidence) and merged with $\mathcal{M}^s$ via a voxelized confidence-weighted update:

$$\mathcal{M}^{s+1} \coloneqq \mathcal{U}(\mathcal{M}^s, \hat{\mathbf{Z}}_0^s) = \left\{\rho\cdot\alpha_v\cdot[\![\alpha_v > \tau]\!]\, p_v \;\middle|\; p_v = \arg\max_{p\in\mathcal{B}_v} \alpha(p)\right\}, \tag{8}$$

where $\mathcal{B}_v$ is the set of candidate and existing points falling into voxel v, the inner $\arg\max$ keeps the highest-confidence sample at each voxel, $\rho \in (0,1)$ exponentially decays older points to bound memory size, and $[\![\alpha_v > \tau]\!]$ prunes low-confidence outliers. When new and stale points fall into the same voxel, the update keeps the more confident observation, which suppresses ghost trails.

Complementing sliding-window attention for local temporal coherence, the persistent memory preserves a holistic and compact global scene representation beyond the window range, mitigating forgetting and spatial drift during long-horizon generation.

3.3.3 Self-verified Geometric Rewards

Beyond providing persistent spatial memory, the geometry branch produces physically interpretable signals and naturally serves as a self-verified reward model. Quantitatively evaluating viewpoint compliance and cross-view consistency of online generations, it provides feedback that can be leveraged via reward feedback learning [63, 64, 65] to reinforce controllability during self-rollouts.

Viewpoint Compliance Reward At each rollout step s, the generation is conditioned on camera parameters $\mathbf{C}^s = (\mathbf{K}^s, \mathbf{R}^s, \mathbf{t}^s)$, comprising intrinsics K^s and extrinsics $[\mathbf{R}^s|\mathbf{t}^s]$. Meanwhile, the geometry branch also predicts the corresponding camera parameters $\hat{\mathbf{C}}_0^s$ from noisy diffusion latents. This enables a self-consistent score that penalizes deviations from the intended control signal:

$$r_{\text{cam}}^s \coloneqq -\|\hat{\mathbf{C}}_0^s - \mathbf{C}^s\|_2, \tag{9}$$

encouraging accurate viewpoint following and mitigating long-horizon camera drift.

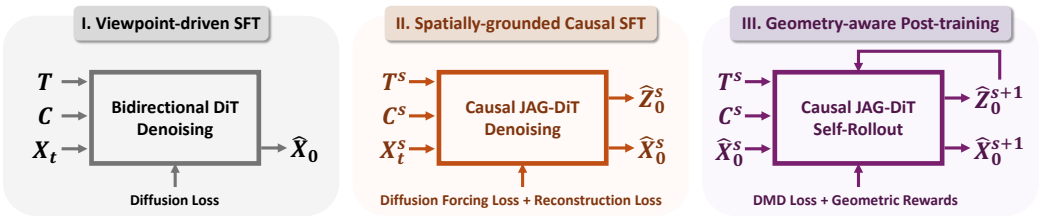

Figure 3: **Training Procedure of JAG** involves three stages: (1) **Viewpoint-driven SFT** that endows controllability to a basic text-to-video bidirectional diffusion model, (2) **Spatially-grounded Causal SFT** that incorporates geometric priors by a dual-branch design and enables auto-regressive inference (Section 3.2), and (3) **Geometry-aware Post-training** that mitigates train-test gap and long-horizon drift by Self Geometry Forcing (Section 3.3). More training details are provided in Appendix A.

Cross-view Consistency Reward The predicted depth $\hat{\mathbf{D}}_0^s$ and camera $\hat{\mathbf{C}}_0^s$ unproject the source frame $D(\hat{\mathbf{X}}_0^s)$ into a colored 3D point cloud $\hat{\mathbf{P}}_0^s$ in the canonical world coordinate, where each 3D point carries the RGB value of its source pixel. Rendering this colored point cloud into a different conditioning viewpoint $\mathbf{C}^{s'}$ yields a reprojected image $\hat{\mathbf{I}}_0^{s\to s'} := \pi(\hat{\mathbf{P}}_0^s, \mathbf{C}^{s'})$ that should agree with the appearance-branch prediction at s', giving a photometric reward:

$$r_{\text{cross}}^{s\to s'} := -\left\|\hat{\mathbf{I}}_0^{s\to s'} - D(\hat{\mathbf{X}}_0^{s'})\right\|_2, \tag{10}$$

where $D(\cdot)$ denotes the video VAE decoder. Since cross-view consistency does not hold for moving objects across timestamps, we compute this reward only over high-confidence depth regions predicted by the geometry branch, which predominantly correspond to static scene structure.

Reward-driven Distillation Together, the combined geometric reward $r_{\text{geo}} := r_{\text{cam}} + \lambda \cdot r_{\text{cross}}^{s\to s'}$ serves as self-verified feedback during self-forced rollouts. This reward is integrated into the distillation process via reward feedback learning (ReFL) [64, 65], reinforcing generations that are both visually plausible and geometrically consistent. Importantly, all rewards are computed from the model's own predictions without external supervision.

3.4 Training Procedure

As shown in Figure 3, the training procedure of JAG includes three stages that progressively transition the model from bidirectional short-video generation to geometry-aware long-horizon causal synthesis: (1) **viewpoint-driven SFT** that introduces Plücker camera control to a T2V base, (2) **spatially-grounded causal SFT** that adds the geometry branch and chunk-wise causal attention with explicit depth and camera supervision, and (3) **geometry-aware post-training** that applies Self Geometry Forcing using the Stage I model as the bidirectional teacher. Per-stage objectives, design rationale, and training details are provided in Appendix A.

4 Experiments

4.1 Experimental Settings

Implementation Wan2.1-T2V-1.3B [6] is adopted as the base model for fine-tuning and the teacher model in distillation. Geometry branch in JAG is initialized from the pretrained Depth-Anything-3-LARGE [24], with a fine-tuned backbone and frozen decoding heads. We choose it over VGGT [57] for its stronger camera and depth accuracy, its single-ViT interface that attaches to the DiT through a lightweight adapter, and its training data spanning static and dynamic scenes. As described in Section 3.3, we perform block-wise causal modeling, where each block consists of 3 latent frames for both branches. Consequently, each auto-regressive step generates three latent frames with geometric attributes, and causality is enforced between blocks. Attention sink and sliding window size are set to 3 and 9 respectively, resulting in the KV cache holding tokens for 12 latent frames. All models are trained with the AdamW optimizer [66], and full hyperparameters and training compute are listed in Appendix B. To improve GPU memory and computational efficiency, we adopt several techniques, including gradient checkpointing [67], DeepSpeed [68], and BF16 mixed precision. Inference efficiency is reported in Table 5.

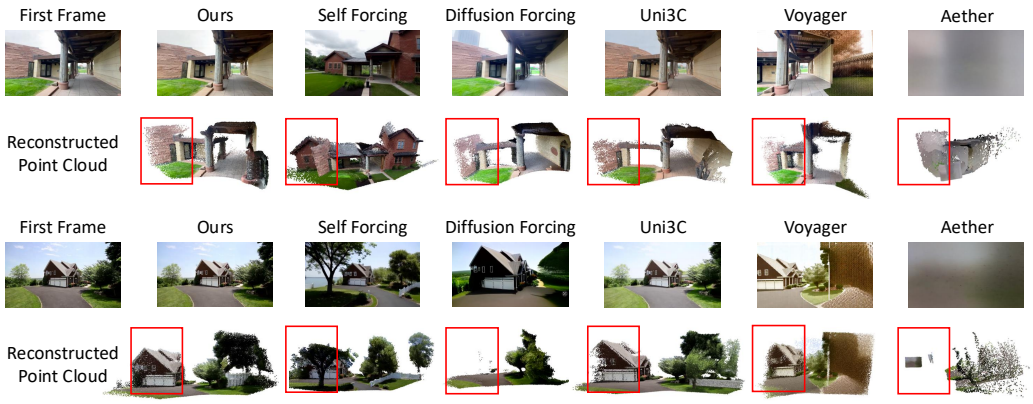

Figure 4: **Qualitative Comparisons** on 201-frame cycle trajectories. Top: frames generated when the camera returns to the first viewpoint. Bottom: point clouds reconstructed from the generated videos, with red boxes marking the first-frame region. More results are in Appendix E.

Datasets We curate a high-quality and diverse training dataset by combining RealEstate10K [69], DL3DV [70], and MiraData [71], comprising about 80K, 10K, and 10K videos, respectively. These datasets cover a wide range of real-world scenes with high visual quality, featuring both complex camera motions and dynamic object movements. To support text-conditioned generation and geometry-aware supervision, we construct a unified annotation pipeline for the entire dataset. All videos are captioned using Qwen2.5-VL [72], while dense metric depth maps and camera trajectories are annotated using a nested Depth-Anything-3-GIANT-LARGE-1.1 [24] framework, producing pixel-wise depth maps with confidence estimates and accurate camera intrinsics and extrinsics. Evaluation uses held-out test scenes from the same three datasets, disjoint from training at the scene level. Overall, this yields a curated dataset of nearly 100K videos with aligned visual, textual, and metric 3D annotations, supporting viewpoint-driven and spatially-grounded long-horizon video generation.

Baselines For controlled comparison, the causal baselines (Diffusion Forcing† [16], Self Forcing† [19]) are *re-implemented* by us under identical settings as JAG: same Wan-T2V-1.3B base model, training corpus, Plücker camera conditioning, optimizer, and total compute. Concretely, Diffusion Forcing† corresponds to our Stage II checkpoint (causal SFT before distillation), Self Forcing† corresponds to our Stage III without the geometry branch (appearance-only DMD distillation), and JAG denotes the full Stage III with Self Geometry Forcing. Tables 1 and 2 therefore double as a stage-wise comparison: DF† $\rightarrow$ SF† isolates the contribution of distillation, while SF† $\rightarrow$ JAG isolates the contribution of the geometry branch and Self Geometry Forcing. The spatially-aware bidirectional baselines, Aether [12], Voyager [13], and Uni3C [45], use their official checkpoints.

4.2 Visual Fidelity Evaluation

Setting We evaluate the visual fidelity of JAG on VBench [73] against strong baselines, including spatial-aware bidirectional methods: Aether [12], Voyager [13], and Uni3C [45], as well as causal methods enhanced with camera conditioning, Diffusion Forcing [16] and Self Forcing [19]. These bidirectional methods are trained on 41, 49, and 81 frames, respectively, while causal methods, including ours, are trained on 81 frames. For fair comparison, all methods are provided with identical viewpoint trajectories, text prompts, and the same first-frame image. Evaluations are conducted under three horizons: short (maintaining the training length), medium (201 frames), and long (405 frames).

Comparison Quantitative results are reported in Table 1, with qualitative visualizations in Figure 4. Bidirectional methods all fail to scale beyond training horizons: their generation length is inflexible, and memory usage grows rapidly with sequence length, causing severe degradation or out-of-memory error (OOM). In contrast, causal methods support long-horizon inference via KV caching, but Diffusion Forcing suffers from pronounced error accumulation, leading to quality degradation at mid and long lengths. Equipping Self Forcing with the same causal inference strategy (Section 3.2.2) prevents catastrophic collapse, yet its generations still exhibit spatial inconsistency and viewpoint drift. JAG gains mainly in subject and background consistency, with the largest margin at 405 frames, while the other dimensions stay close to Self Forcing†. Although spatial inconsistency and

Table 1: **Evaluation on Visual Fidelity** on VBench under short (training length), medium (201 frames), and long (405 frames) horizons. The **best**, second-best, and third-best results are highlighted in each column.

	Causal?	↑ Background Cons.			↑ Subject Cons.			↑ Aesthetic Qual.			↑ Imaging Qual.			↑ Motion Smooth.		
		Short	Medium	Long	Short	Medium	Long	Short	Medium	Long	Short	Medium	Long	Short	Medium	Long
Aether	✗	94.13	79.47	OOM	91.24	56.47	OOM	52.02	29.54	OOM	56.30	36.15	OOM	**98.77**	**98.94**	OOM
Voyager	✗	94.01	93.64	OOM	92.88	90.10	OOM	51.21	50.26	OOM	61.44	61.22	OOM	98.37	98.78	OOM
Uni3C	✗	94.63	93.73	OOM	92.84	91.47	OOM	52.68	51.87	OOM	67.37	65.95	OOM	98.03	97.95	OOM
Diffusion Forcing†	✔	93.72	92.82	90.87	91.55	89.39	85.48	50.74	49.41	45.97	67.98	66.18	65.87	98.12	97.73	97.90
Self Forcing†	✔	**94.71**	93.94	92.80	94.48	92.33	89.88	52.64	**52.03**	**51.37**	72.93	72.39	**72.15**	98.63	98.77	**98.89**
Ours	✔	94.68	**94.45**	**93.66**	**94.97**	**93.09**	**92.73**	**52.85**	52.01	51.34	**73.01**	**72.45**	71.90	98.47	98.76	98.55

Table 2: **Quantitative Evaluation on Controllability and Consistency** under short (training length), medium (201 frames), and long (405 frames) horizons. Camera poses are estimated by Depth-Anything-3-GIANT at the short and medium horizons, and by VGGT*, outside the Depth-Anything-3 family, at the long horizon. Consistency uses revisit PSNR, SSIM, and LPIPS between two visits of the same viewpoint. The **best**, second-best, and third-best results are highlighted.

	Controllability									Consistency (Revisit)								
	↓ ATE			↓ RPE_{trans}			↓ RPE_{rot}			↑ PSNR			↑ SSIM			↓ LPIPS		
	Short	Medium	Long*	Short	Medium	Long*	Short	Medium	Long*	Short	Medium	Long	Short	Medium	Long	Short	Medium	Long
Aether	**0.0651**	0.1437	OOM	0.0062	0.0052	OOM	0.3232	3.9987	OOM	15.05	11.46	OOM	0.5042	0.4831	OOM	0.4914	0.7282	OOM
Voyager	0.0692	0.1317	OOM	0.0061	**0.0049**	OOM	0.3076	0.3535	OOM	12.57	10.91	OOM	0.4153	0.2730	OOM	0.5634	0.6759	OOM
Uni3C	0.1050	0.1413	OOM	0.0056	0.0051	OOM	**0.2323**	0.2386	OOM	15.15	14.64	OOM	0.5456	0.5061	OOM	0.3773	0.4360	OOM
Diffusion Forcing†	0.1052	0.1421	0.2617	0.0057	0.0051	0.0063	0.2656	1.0044	2.4130	14.69	13.29	12.05	0.4442	0.4012	0.3695	0.4487	0.5062	0.5648
Self Forcing†	0.1054	0.1432	0.1904	0.0058	0.0054	0.0053	0.2558	0.3361	0.4817	11.46	11.00	10.62	0.3355	0.3102	0.2884	0.5821	0.6146	0.6533
Ours	0.0711	**0.1150**	**0.1385**	**0.0055**	0.0050	**0.0048**	0.2405	**0.2333**	**0.2691**	**16.14**	**14.68**	**14.02**	**0.5704**	**0.5137**	**0.4871**	**0.3395**	**0.4333**	**0.4522**

viewpoint drift are weakly reflected in VBench (Table 1), they are clearly revealed by controllability and consistency metrics (Table 2), and in the qualitative results (Figure 4).

4.3 Controllability Evaluation

Setting Beyond visual metrics, we estimate per-frame camera poses from the generated videos with a feed-forward reconstruction model [24] and compare them against the target trajectories using absolute translation error (ATE), relative translation error (RPE_{trans}), and relative rotation error (RPE_{rot}), with translation alignment to remove scale ambiguity. At the training length and 201 frames, we use Depth-Anything-3-GIANT as the evaluator, a different checkpoint from the Depth-Anything-3-LARGE that initializes the geometry branch. At the long horizon of 405 frames, we switch to VGGT [57], which lies outside the Depth-Anything-3 family and is used nowhere in the training, memory, or rewards of JAG, so the long-horizon gains cannot stem from a geometric prior shared by model and evaluator.

Comparison Quantitative results are reported in Table 2, with visualizations in Figure 4. Similar to visual fidelity, spatially-aware bidirectional methods follow the trajectory well on short clips, but they degrade sharply beyond their training length (e.g., the RPE_{rot} of Aether jumps from 0.32 to 4.00 at 201 frames) and run out of memory at 405 frames. Diffusion Forcing and Self Forcing, lacking explicit geometric constraints, already trail the best spatially-aware methods on short clips, and their errors grow quickly with the horizon. JAG achieves the lowest pose error among causal methods in every metric and horizon. Against all methods, it stays competitive on short clips and takes the lead as the horizon grows: it is best in ATE and RPE_{rot} at 201 frames, and at 405 frames it reduces ATE by 27% and RPE_{rot} by 44% relative to Self Forcing†.

4.4 Consistency Evaluation

Setting To assess long-term consistency under dynamic camera motion, we evaluate all methods using cycle camera trajectories, where the camera revisits previously observed viewpoints. Ideally, long-horizon generation should preserve both scene geometry and appearance, such that revisited frames match their first visits up to minor stochastic variations. We compare each revisited frame with the frame generated earlier at the same viewpoint, and report *revisit* PSNR, SSIM, and LPIPS [74]. This setting is particularly challenging for causal generation, as small geometric or appearance errors accumulated over time can cause irreversible drift, preventing accurate recovery. Evaluations are conducted at the training length, 201, and 405 frames.

Comparison Results are summarized in Table 2, with qualitative examples in Figure 4. For bidirectional methods, reconstructing the first and last frames is relatively easy on short clips, since

Table 3: **Ablation Study of Design Choices** evaluated at two horizons of 81 and 201 frames. **Best** and second-best results within each component group are highlighted in bold and underline.

	↑ Aesthetic Qual.		↑ Background Cons.		↓ ATE		↓ RPE_{trans}		↓ RPE_{rot}		↑ PSNR	
	81	201	81	201	81	201	81	201	81	201	81	201
Dual Branch												
w/o geometry branch	52.64	**52.03**	**94.71**	93.94	0.1054	0.1432	0.0058	0.0054	0.2558	0.3361	11.46	11.00
w/o geometric priors	52.46	51.95	93.40	92.31	0.0988	0.1441	0.0056	0.0053	0.2511	0.3016	15.84	14.64
Ours	**52.85**	52.01	94.68	**94.45**	**0.0711**	**0.1150**	**0.0055**	**0.0050**	**0.2405**	**0.2333**	**16.14**	**14.68**
Sink and Re-RoPE												
w/o sink	52.88	46.83	94.73	90.35	0.0710	0.1347	0.0057	0.0055	**0.2402**	0.3193	15.40	10.22
w/o re-RoPE	**52.96**	49.38	**94.74**	91.35	**0.0705**	0.1244	**0.0054**	0.0052	0.2410	0.2485	15.87	14.03
Ours	52.85	**52.01**	94.68	**94.45**	0.0711	**0.1150**	0.0055	**0.0050**	0.2405	**0.2333**	**16.14**	**14.68**
Persistent Spatial Memory												
w/o memory	**52.88**	**52.85**	94.66	92.90	0.0799	**0.1149**	0.0055	0.0051	0.2444	0.2369	11.04	10.83
Ours	52.85	52.01	**94.68**	**94.45**	**0.0711**	0.1150	0.0055	**0.0050**	**0.2405**	**0.2333**	**16.14**	**14.68**
Geometric Rewards												
w/o r^{s}_{cam}	52.87	52.00	94.62	94.39	0.0813	0.1399	0.0056	0.0058	0.2486	0.2356	16.08	14.47
w/o $r^{s\to s'}_{\text{cross}}$	**52.94**	**52.04**	94.21	94.18	0.0804	0.1347	0.0056	0.0054	0.2462	0.2361	16.06	13.62
Ours	52.85	52.01	**94.68**	**94.45**	**0.0711**	**0.1150**	**0.0055**	**0.0050**	**0.2405**	**0.2333**	**16.14**	**14.68**

Table 4: **In-network versus External Geometry** at 405 frames, with poses estimated by VGGT. Latency denotes the per-chunk cost of obtaining geometry.

	↓ Latency	↓ ATE	↓ RPE_{rot}	↑ PSNR
Self Forcing† (no geometry)	0 ms	0.1904	0.4817	10.62
Ours w/ External DA3 Memory	214 ms	0.1663	0.3852	12.38
Ours (in-network geometry)	**60 ms**	**0.1385**	**0.2691**	**14.02**

Table 5: **Inference Efficiency** on a single H20 at 288×512, with peak GPU memory measured at 405 frames.

	Steps	↑ FPS	↓ Memory
Diffusion Forcing†	50	1.4	19.4 GB
Self Forcing†	4	16.8	19.4 GB
Ours	4	15.1	23.7 GB

the full sequence is generated jointly and enhanced with spatial awareness, so they exhibit strong consistency unless generation collapses. In contrast, causal methods rely on sliding-window attention at inference time. Despite sink mechanisms, they lack global memory beyond the window, making it difficult to maintain long-term consistency, especially for content outside the context window. Our method addresses this limitation by maintaining a self-generated persistent spatial memory that is continuously updated online and records previously generated content in 3D space, complementing local windowed information. By conditioning generation on this persistent memory and enforcing self-verified geometric rewards, the model preserves a coherent global scene representation throughout long rollouts, enabling accurate recovery of previously observed views. JAG ranks first in all three metrics at every horizon, with the margin persisting at 405 frames.

4.5 Ablation and Analysis

Extensive ablation studies are conducted to analyze the contribution of each component in JAG. Quantitative results are reported in Table 3.

Dual Branch (Section 3.2.1) We first ablate the geometry branch, which collapses all geometry-grounded designs and reduces the model to a standard causal DiT with Self Forcing, resulting in pronounced viewpoint drift and spatial inconsistency. Moreover, training the geometry branch from scratch, without geometric priors, leads to slow convergence and inferior performance under the same training budget. These results jointly confirm that both DA3's pretrained 3D priors and the dual-branch architectural separation are necessary, and neither is sufficient on its own.

Sink and Re-RoPE (Section 3.2.2) We validate the frame sink and re-RoPE strategy by reverting to standard positional encoding and attention mechanisms. Without these adaptations, the model exhibits reduced extrapolation performance and degraded attention alignment in long sequences. The degradation is most pronounced past the original training horizon, where untrained positional indices push the attention pattern out-of-distribution and the absence of a stable global anchor leaves the model without a reliable canonical reference.

Persistent Spatial Memory (Section 3.3.2) To assess the role of persistent spatial memory, we disable memory updates during self-rollouts. This variant shows severe forgetting of previously observed scene regions, especially under cyclic camera trajectories. The gap is largest on cyclic

trajectories, indicating that the memory specifically addresses long-range identity preservation rather than local temporal smoothness.

Geometric Rewards (Section 3.3.3) We further ablate each self-verified geometric reward during post-training. Removing either reward leaves visual quality nearly unchanged but raises ATE at both horizons, and removing r_{cross} also lowers revisit PSNR at 201 frames. This validates both rewards as complementary internal feedback for viewpoint accuracy and cross-view consistency.

In-network versus External Geometry A natural alternative is to build the same memory from an off-the-shelf estimator. Ours w/ External DA3 Memory keeps everything else unchanged and replaces the geometry branch with a frozen Depth-Anything-3 applied to decoded RGB frames. As shown in Table 4, it recovers part of the gain over Self Forcing†, yet JAG further reduces ATE by 17% and RPE_{rot} by 30% at $3.6\times$ lower per-chunk cost. Beyond speed, the shared trunk and frame sink keep all chunks in one coordinate frame and scale, whereas per-chunk external estimates must be registered to the memory and accumulate errors. In-network geometry also lets the rewards be optimized without backpropagating through the VAE decoder and an external estimator.

Inference Efficiency As reported in Table 5, JAG streams at 15.1 FPS on a single H20, an order of magnitude faster than Diffusion Forcing†. Compared with Self Forcing†, explicit geometry and persistent memory cost only 10% of the throughput and 4.3 GB of peak memory, while the external alternative drops to 12.7 FPS.

5 Conclusion

We present **JAG**, a **spatially-grounded causal diffusion framework** for long-horizon and controllable video generation. By disentangling appearance and geometry in a dual-branch auto-regressive architecture, JAG enables explicit spatial reasoning and streaming extrapolation under a fixed-size context. With this design, we further introduce **Self Geometry Forcing**, which generalizes self-forcing paradigms from visual alignment to geometry-aware distillation. With persistent spatial memory and self-verified geometric rewards, it mitigates viewpoint drift and spatial inconsistency during long-horizon rollouts. Overall, JAG suggests that explicit geometry can serve as a critical internal signal for stabilizing causal video generation, paving a promising path toward scalable and physically grounded world modeling. Limitations and future work are discussed in Appendix C.

Acknowledgment

This work was supported by the Peking University Education Foundation (No. 050621_001).

References

[1] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model, 2024.

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

Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models, 2025.

[3] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations (ICLR)*, 2025.

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

[5] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations (ICLR)*, 2025.

[6] 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.

[7] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

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

[9] 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.

[10] Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. *International Conference on Computer Vision (ICCV)*, 2025.

[11] Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[12] Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, et al. Aether: Geometric-aware unified world modeling. *International Conference on Computer Vision (ICCV)*, 2025.

[13] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025.

[14] 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.

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

[16] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[17] Deyu Zhou, Quan Sun, Yuang Peng, Kun Yan, Runpei Dong, Duomin Wang, Zheng Ge, Nan Duan, and Xiangyu Zhang. Taming teacher forcing for masked autoregressive video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[18] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhiheng Xu, Yuzhe Jin, Yupeng Liang, Yubing Song, Peng Zhao, Boyuan Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. Skyreels-v2: Infinite-length film generative model, 2025.

[19] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[20] Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Cyrus Wu, Wei Li, Xuchen Song, Yang Liu, Eric Li, and Yahui Zhou. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

[21] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

[22] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

[23] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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

[25] Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, et al. Longcat-video technical report. *arXiv preprint arXiv:2510.22200*, 2025.

[26] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *Forty-second International Conference on Machine Learning*, 2025.

[27] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[28] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[29] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[30] Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025.

[31] Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *arXiv preprint arXiv:2512.04678*, 2025.

[32] 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.

[33] 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.

[34] Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[35] Lvmin Zhang and Maneesh Agrawala. Packing input frame contexts in next-frame prediction models for video generation. *arXiv preprint*, 2025.

[36] Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[37] Tianyuan Zhang, Sai Bi, Yicong Hong, Kai Zhang, Fujun Luan, Songlin Yang, Kalyan Sunkavalli, William T Freeman, and Hao Tan. Test-time training done right. *arXiv preprint arXiv:2505.23884*, 2025.

[38] Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): Rnns with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024.

[39] Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[40] Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, et al. Mixture of contexts for long video generation. *arXiv preprint arXiv:2508.21058*, 2025.

[41] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024.

[42] Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[43] 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.

[44] Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista Martin, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[45] Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, 2025.

[46] Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, et al. Deepverse: 4d autoregressive video generation as a world model. *arXiv preprint arXiv:2506.01103*, 2025.

[47] Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 4dnex: Feed-forward 4d generative modeling made easy. *arXiv preprint arXiv:2508.13154*, 2025.

[48] Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. *arXiv preprint arXiv:2509.21657*, 2025.

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

[50] Baorui Ma, Huachen Gao, Haoge Deng, Zhengxiong Luo, Tiejun Huang, Lulu Tang, and Xinlong Wang. You see it, you got it: Learning 3d creation on pose-free videos at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[51] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[52] Luxi Chen, Zihan Zhou, Min Zhao, Yikai Wang, Ge Zhang, Wenhao Huang, Hao Sun, Ji-Rong Wen, and Chongxuan Li. Flexworld: Progressively expanding 3d scenes for flexiable-view synthesis. *arXiv preprint arXiv:2503.13265*, 2025.

[53] Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[54] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

[55] 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.

[56] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025.

[57] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025.

[58] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

[59] Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Zero-shot extreme length generalization for large language models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (ACL)*, 2024.

[60] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.

[61] Florian Schmidt. Generalization in generation: A closer look at exposure bias. *arXiv preprint arXiv:1910.00292*, 2019.

[62] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

[63] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

[64] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[65] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[66] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2018.

[67] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

[68] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020.

[69] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018.

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

[71] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[72] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[73] 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.

[74] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[75] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

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

[77] Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *International Conference on Learning Representations (ICLR)*, 2023.

[78] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[79] Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. In *International Conference on Learning Representations (ICLR)*, 2024.

A Training Details

As illustrated in Figure 3, the training of JAG consists of three stages: (1) **viewpoint-driven supervised fine-tuning**, (2) **spatially grounded causal supervised fine-tuning**, and (3) **geometry-aware post-training**. The first two stages rely on labeled videos with captions and calibrated camera intrinsics and extrinsics, and are therefore conducted via supervised fine-tuning (SFT). In contrast, the final stage is self-supervised, where the model is optimized from its own rollouts without requiring external annotations. We detail the training objective of each stage below for reproducibility.

A.1 Stage I: Viewpoint-driven SFT

We initialize from a bidirectional text-to-video diffusion model, i.e., Wan-T2V-1.3B [6], and aim to enable explicit and accurate viewpoint controllability. Given a noisy video latent $\mathbf{X}_t \coloneqq (1-t)\cdot\mathbf{X}_0 + t\cdot\boldsymbol{\epsilon}$, a text prompt $\mathbf{T}$, and a camera trajectory $\mathbf{C}$, the diffusion transformer (DiT) $v_\theta(\cdot)$ is trained using the standard flow-matching objective [75, 76]:

$$\mathcal{L}_{\text{diffusion}}(\theta) \coloneqq \omega(t)\cdot\|v_\theta(\mathbf{X}_t, t, \mathbf{C}, \mathbf{T}) - (\boldsymbol{\epsilon} - \mathbf{X}_0)\|_2^2, \tag{11}$$

where $t \in [0,1]$ denotes the noise level, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is standard Gaussian noise. Text prompts are encoded using a pretrained text encoder [77] and injected into the DiT via cross-attention. Viewpoint trajectories are represented using Plücker embeddings [78] of camera rays, and linearly projected to match the spatial resolution and channel dimension of video latents. The resulting camera embeddings are added directly to the latent tokens after the DiT patch embedding layer. Since video latents are temporally compressed by a 3D VAE [79], we subsample camera parameters according to the VAE temporal downsampling ratio (4 in this work), ensuring temporal alignment between the camera trajectory $\mathbf{C}$ and the latent sequence.

After this stage, the model supports camera control for both static and dynamic scenes in a text-to-video setting, and serves as the base model for Stage II and the bidirectional teacher in Stage III.

A.2 Stage II: Spatially-grounded Causal SFT

Starting from the viewpoint-driven model obtained in Stage I, we further convert the model into a causal and auto-regressive formulation to support streaming and explorable video generation with flexible inference length. As described in Section 3.2.2, we enforce chunk-wise causal attention during training, such that future content is predicted solely based on past observations within a fixed-size context window. The training objective follows the standard flow-matching formulation, augmented with a causal attention mask $\mathbf{M}$ and the Diffusion Forcing noising strategy [16, 26], which injects random noise levels across different causal states s:

$$\mathcal{L}_{\text{causal-diffusion}}(\theta) \coloneqq \omega(t)\cdot\left\|v_\theta(\mathbf{X}_t^s, t, \mathbf{C}^s, \mathbf{T}^s, \mathbf{M}) - (\boldsymbol{\epsilon} - \mathbf{X}_0^s)\right\|_2^2. \tag{12}$$

The causal mask $\mathbf{M}$ restricts information flow across tokens in a chunk-wise auto-regressive manner:

$$\mathbf{M}_{ij} = \begin{cases} 1, & \text{if } j \le R, \\ 1, & \text{if } \left\lfloor \frac{j}{K} \right\rfloor \le \left\lfloor \frac{i}{K} \right\rfloor \text{ and } |i-j| \le W, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Here, i and j index the query and key frames in the attention operation. R denotes the number of frame sinks that are globally visible to all positions to stabilize long-context attention. K specifies the chunk size, enforcing auto-regressive generation at the chunk level, while W defines a local sliding window that permits limited short-range attention across adjacent chunks. Together, this design balances strict causality, temporal coherence, and computational efficiency.

To further ground causal generation in explicit 3D structure, we equip the causal diffusion backbone with a geometry branch. The original diffusion transformer is referred to as the appearance branch with parameters θ, while the geometry branch with parameters ϕ is initialized from a pretrained feed-forward visual reconstruction model [24]. As detailed in Section 3.2.1, both branches operate on shared latent representations and interact via lightweight learnable adapters with parameters ξ, enabling joint appearance generation and geometry prediction. The overall objective of this stage is:

$$\mathcal{L}_{\text{dual-branch}}(\theta, \phi, \xi) \coloneqq \mathcal{L}_{\text{causal-diffusion}}(\theta, \phi, \xi) + \mathcal{L}_{\text{recon}}(\theta, \phi, \xi), \tag{14}$$

$$\mathcal{L}_{\text{recon}} \coloneqq \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{camera}}. \tag{15}$$

The depth supervision term is a confidence-weighted, per-pixel loss combining absolute and gradient-domain agreement, with a single confidence regularizer:

$$\mathcal{L}_{\text{depth}} \coloneqq \left\langle \boldsymbol{\alpha} \odot \left| h_{\text{depth}}(\hat{\mathbf{Z}}_0^s) - D_{\text{gt}} \right| \right\rangle + \left\langle \boldsymbol{\alpha} \odot \left| \nabla h_{\text{depth}}(\hat{\mathbf{Z}}_0^s) - \nabla D_{\text{gt}} \right| \right\rangle - \lambda_{\boldsymbol{\alpha}} \cdot \left\langle \log(\boldsymbol{\alpha}) \right\rangle, \tag{16}$$

where $h_{\text{depth}}(\cdot)$ predicts a per-pixel depth map together with a self-estimated confidence map $\boldsymbol{\alpha}$ from the denoised geometry latents $\hat{\mathbf{Z}}_0^s$, D_{gt} denotes the ground-truth depth, $\odot$ denotes the element-wise product, and $\langle \cdot \rangle$ averages over spatial pixels. The first two terms drive depth and gradient agreement weighted per pixel by confidence, the regularizer with weight $\lambda_{\boldsymbol{\alpha}}$ prevents the model from collapsing $\boldsymbol{\alpha}$ to zero, and gradient-level supervision improves boundary alignment.

In addition, camera supervision enforces accurate viewpoint estimation:

$$\mathcal{L}_{\text{camera}} \coloneqq \| h_{\text{cam}}(\hat{\mathbf{Z}}_0^s) - \mathbf{C}_{\text{gt}} \|_2, \tag{17}$$

where $h_{\text{cam}}(\cdot)$ regresses camera intrinsics and extrinsics from the same latent representation. The ground-truth camera $\mathbf{C}_{\text{gt}}$ is parameterized by field of view $\mathbf{f} \in \mathbb{R}^2$, rotation quaternion $\mathbf{q} \in \mathbb{R}^4$, and translation $\mathbf{t} \in \mathbb{R}^3$. Since video latents are temporally downsampled by the 3D VAE, the geometry branch does not perform temporal upsampling. Accordingly, ground-truth depth maps and camera parameters are temporally subsampled to align with the latent sequence. All parameters θ, ϕ, and ξ are optimized jointly, except the depth and camera heads of the geometry branch, which stay frozen.

Through this stage, the model is transformed from a purely appearance-driven generator into a geometry-aware, causal video model that explicitly reasons about 3D structure while supporting efficient long-horizon inference.

A.3 Stage III: Geometry-aware Post-training

The final stage performs geometry-aware post-training based on the proposed Self Geometry Forcing as described in Section 3.3. Unlike the previous supervised stages, this stage requires no additional annotations and optimizes from its own rollouts.

We adopt a self-forcing paradigm [19] where the causal dual-branch model acts as a student, while the bidirectional viewpoint-driven model obtained in Stage I serves as a teacher. Although the teacher is limited to short clips and lacks explicit geometry awareness, its bidirectional attention yields strong visual quality on short horizons, making it a reliable reference distribution. This teacher thus regularizes post-training and prevents visual degradation when incorporating geometric rewards. Following prior work [29, 27], we employ Distribution Matching Distillation (DMD) to align the distribution of self-rollout samples generated by the student with that of the teacher. The gradient of the DMD objective is defined as:

$$\nabla \mathcal{L}_{\text{DMD}}(\theta, \phi, \xi) \coloneqq -\mathbb{E}_{\hat{\mathbf{X}}_0^{1:S}, t} \left[\left(S_{\text{real}}(\hat{\mathbf{X}}_0^{1:S}, t) - S_{\text{fake}}(\hat{\mathbf{X}}_0^{1:S}, t) \right) \cdot \frac{\partial \hat{\mathbf{X}}_0^{1:S}}{\partial(\theta, \phi, \xi)} \right], \tag{18}$$

where $\hat{\mathbf{X}}_0^{1:S}$ denotes videos generated online by the causal student model, S_{real} is the score function estimated by the frozen bidirectional teacher, and S_{fake} is estimated by a score model trained on student rollouts. This objective can be interpreted as minimizing the KL divergence $\mathbb{E}_t[D_{\text{KL}}(p_t^{\text{gen}} \| p_t^{\text{teacher}})]$, thereby aligning the overall video distribution without annotations.

In parallel, the geometry branch produces self-verified rewards that explicitly measure geometric fidelity and controllability. Specifically, the geometric reward is defined as:

$$r_{\text{geo}} \coloneqq r_{\text{cam}} + \lambda \cdot r_{\text{cross}}^{s \to s'}, \tag{19}$$

where r_{cam} penalizes deviations between predicted and conditioned camera parameters, and $r_{\text{cross}}^{s \to s'}$ enforces reprojection consistency across viewpoints using the predicted 3D structure. All reward signals are computed solely from the model's own predictions, without external supervision. These geometric rewards provide fine-grained feedback on viewpoint compliance and long-horizon spatial consistency, which are not directly captured by visual-only distillation.

The final objective integrates distribution-level alignment and geometry-aware reinforcement:

$$\mathcal{L}_{\text{post-train}}(\theta, \phi, \xi) \coloneqq \mathcal{L}_{\text{DMD}}(\theta, \phi, \xi) - \beta \cdot \mathbb{E}[r_{\text{geo}}], \tag{20}$$

where β balances visual realism preserved by DMD and geometric controllability reinforced by reward feedback learning [64, 65]. In practice, DMD stabilizes optimization and prevents visual collapse, while geometric rewards drive the model toward accurate viewpoint following and spatial consistency over long horizons.

Through this geometry-aware post-training, the causal video diffusion model acquires persistent spatial reasoning and robust long-horizon controllability, completing the transition from supervised geometry grounding to fully self-supervised geometric refinement.

B Implementation Details

Optimization All stages use AdamW with momentum coefficients $(0.9, 0.999)$ and weight decay 0.01, a 500-step linear warmup followed by a constant learning rate, and gradient accumulation to global batch sizes of 16, 8, and 8 for Stages I, II, and III. The learning rate is 1×10^{-5} for Stage I, 1×10^{-5} for Stage II with 1×10^{-4} for the newly added adapter, and 2×10^{-6} for Stage III with 4×10^{-6} for the fake score model. Inference uses 4 denoising steps.

Memory and Rewards The memory update (Section 3.3.2) uses a confidence threshold $\tau = 0.6$ and a decay factor $\rho = 0.98$, and the rendered point cloud is capped at 4M points. The reward weights are $\lambda = 1.0$ in r_{geo} and $\beta = 1.0$ in $\mathcal{L}_{\text{post-train}}$, and the confidence regularizer weight is $\lambda_\alpha = 0.2$.

Training All stages run on $8\times$NVIDIA H20 GPUs with BF16 mixed precision, the AdamW optimizer [66], and gradient checkpointing [67]. Wall-clock training time is approximately 20 hours for Stage I (viewpoint-driven SFT), 10 hours for Stage II (spatially-grounded causal SFT with the geometry branch attached), and 10 hours for Stage III (geometry-aware post-training with Self Geometry Forcing), totaling about 40 hours of wall-clock time (about 320 GPU-hours). The methodologically new component, Stage III, costs only 25% of the total budget, comparable to a single-stage Self Forcing distillation under matched settings.

Inference JAG runs at 15.1 FPS on a single H20 at 288×512 resolution under the chunk-causal sliding-window schedule, compared with 16.8 FPS for the appearance-only Self Forcing† variant (Table 5). The geometry branch adds about 60 ms per chunk, since it reads geometry directly from the video latents and shares the same KV cache with the appearance branch (sliding window of 9 latent frames plus 3 frame-sink latent frames, holding tokens for 12 latent frames). Per-chunk persistent-memory rendering and update add about 18 ms, dominated by point splatting onto the target view. End-to-end peak GPU memory at 405 frames is 23.7 GB, compared with 19.4 GB for Self Forcing†. The additional memory covers the weights and activations of the geometry branch, the persistent point cloud (capped at 4M points after voxel binning), and the splatting buffers.

C Limitations and Future Work

Although the geometry branch provides explicit and physically interpretable signals, the predicted depth and camera parameters are still inferred from monocular diffusion latents, and thus may contain local inaccuracies or bias, particularly in textureless, reflective, or highly dynamic regions. When such estimates are accumulated over time, small geometric errors can gradually propagate through persistent memory, leading to structural drift in extended rollouts. Moreover, the generation process operates under partial observability, where only a subset of the scene is visible at each timestep. While persistent spatial memory mitigates forgetting beyond the fixed context window, deciding what spatial information to retain, update, or discard remains heuristic, and better memory strategies may further improve robustness. Finally, the geometry branch is only initialized from Depth-Anything-3, and alternative backbones such as VGGT remain to be explored.

Looking forward, several directions may help advance spatially-grounded long video generation. Incorporating uncertainty-aware or probabilistic geometry representations could reduce error accumulation during long rollouts. The persistent memory and cross-view reward currently target static scene structure, and extending them to dynamic or non-rigid scenes may require object-centric or motion-aware representations. Beyond passive generation, integrating active perception or planning, where the model reasons about viewpoints to reduce uncertainty, could further bridge causal video generation and generative world modeling. We hope this work could provide a step toward spatially consistent generative models and help clarify the challenges ahead.

D Broader Impacts

This work presents a methodological advance in video generation and world modeling, aiming to improve the long-horizon controllability and spatial consistency. The proposed approach has potential applications in areas such as simulation, robotics, virtual content creation, and scientific visualization. As with existing video generation systems, there are also general risks related to misuse, including the generation of misleading or deceptive visual content.

The method does not introduce fundamentally new ethical concerns beyond those already associated with current video generative models, nor does it rely on sensitive personal data or target high-stakes decision-making scenarios. We consider this work primarily as a research contribution, and future work may explore additional safeguards when deploying such systems in real-world applications.

E Additional Qualitative Results

We provide additional qualitative results in two complementary forms. The frame strips collected in this section show six evenly spaced frames per clip. The accompanying supplementary archive contains the same clips played at the native frame rate, which is necessary to read off temporal coherence and viewpoint motion that static frames cannot convey. Each gallery sample is presented as a triplet of (i) the generated video, (ii) per-frame depth predicted by the geometry branch, and (iii) the *confidence-filtered point cloud* aggregated from the persistent spatial memory during streaming generation, in which points predicted with low confidence by the geometry branch are removed before rendering. The confidence-filtered point clouds are rendered without per-scene optimization or post-processing, reflecting the geometry directly emitted by the model. The side-by-side baseline comparison on identical viewpoint trajectories is reported in Figure 4 of the main paper using reconstructed 3D structure as an unbiased common ground. This section instead presents method-only behavior of JAG, starting with a controlled study where a single text prompt is rolled out under four different camera trajectories, and followed by a broader gallery across diverse samples.

Same Prompt, Different Trajectories The figures below fix a single text prompt and roll out the model under four different camera trajectories, isolating the effect of viewpoint motion under fixed textual conditioning. Each row is one trajectory rendered as six evenly-spaced RGB frames, and all four rows in a figure share the same prompt. The four prompts are drawn at random from the prompt set used for our quantitative evaluation, and we therefore omit their verbatim text in the captions.



Figure 5: **Same prompt, four camera trajectories (1).**

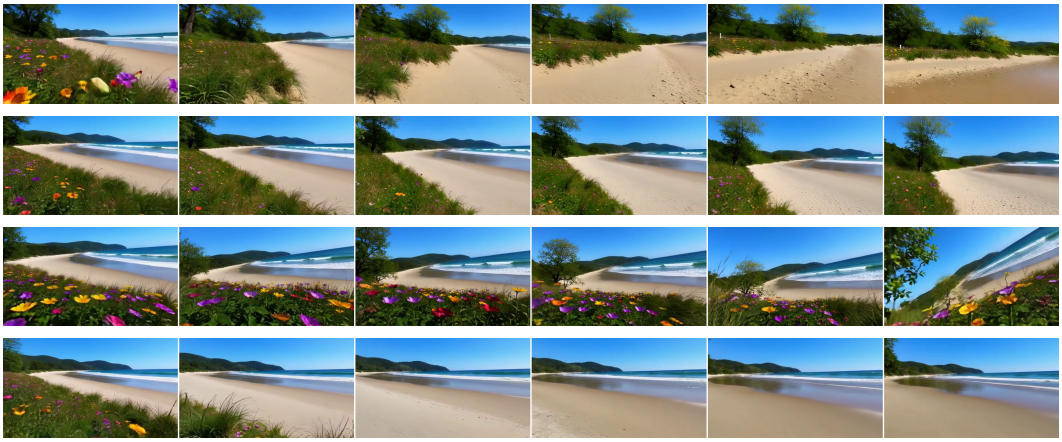

Figure 6: **Same prompt, four camera trajectories (2).**

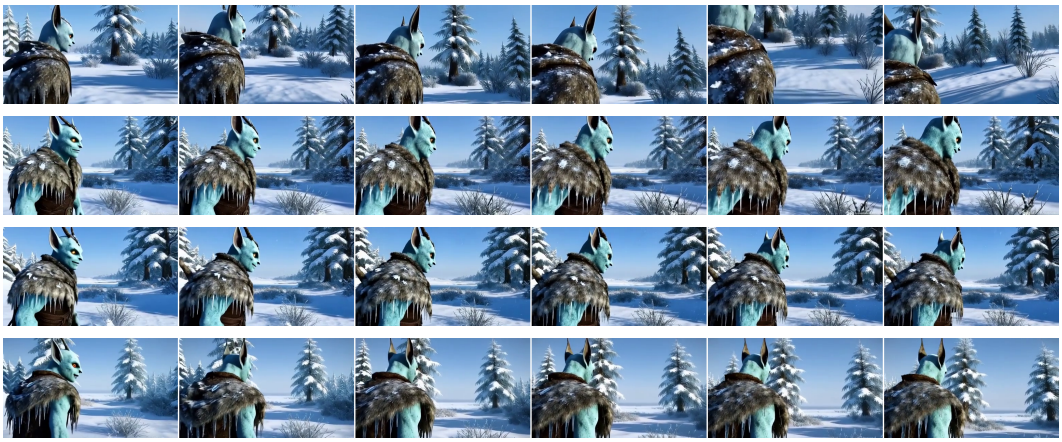

Figure 7: **Same prompt, four camera trajectories (3).**



Figure 8: **Same prompt, four camera trajectories (4).**

Gallery The figures below show six selected samples from the supplementary gallery, with rows giving the generated RGB video, the predicted depth, and the confidence-filtered point cloud.

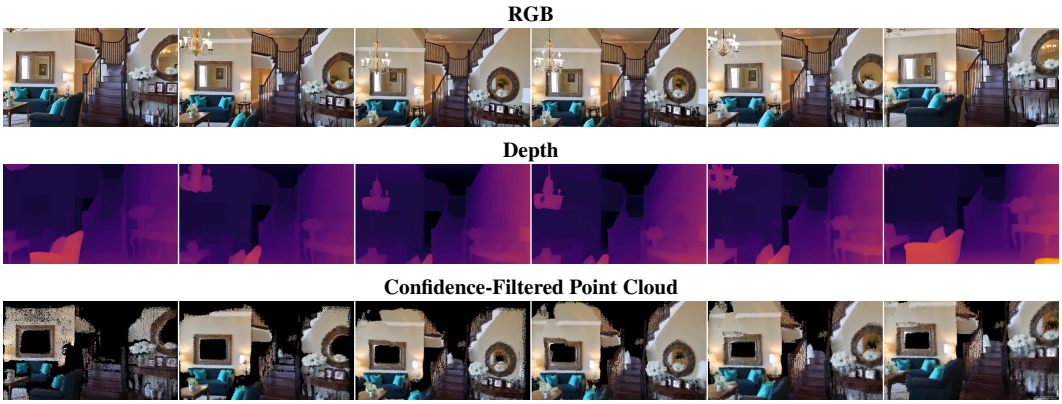

Figure 9: **Gallery Sample 1.** Top to bottom: RGB, depth, confidence-filtered point cloud.

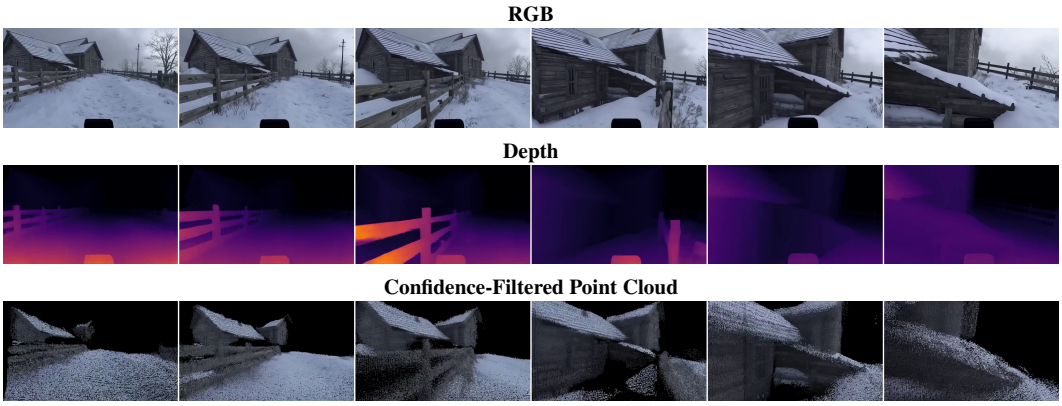

Figure 10: **Gallery Sample 2.** Top to bottom: RGB, depth, confidence-filtered point cloud.

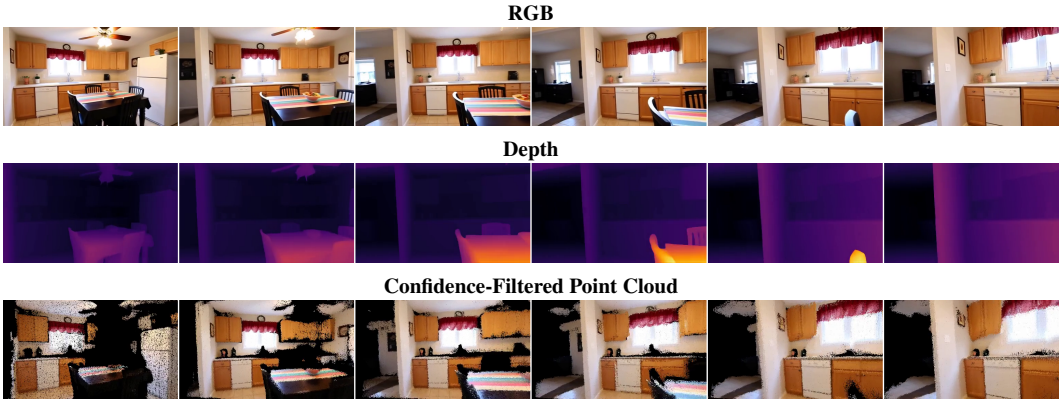

Figure 11: **Gallery Sample 3.** Top to bottom: RGB, depth, confidence-filtered point cloud.

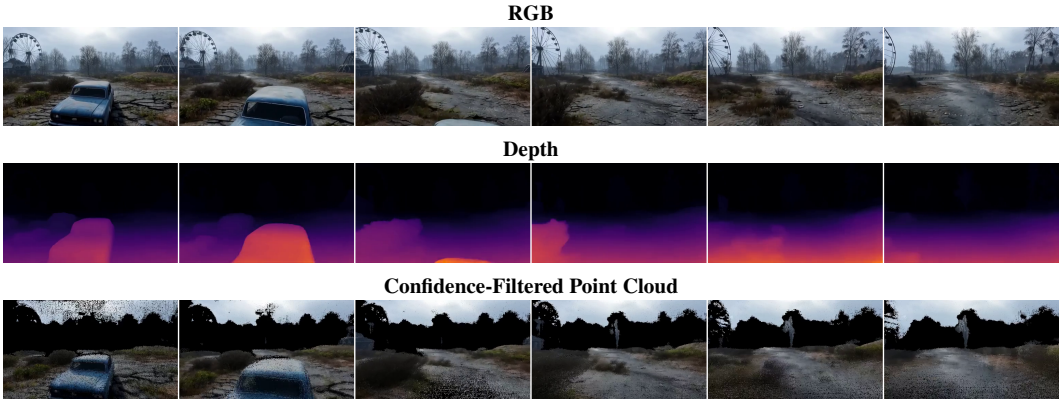

Figure 12: **Gallery Sample 14.** Top to bottom: RGB, depth, confidence-filtered point cloud.

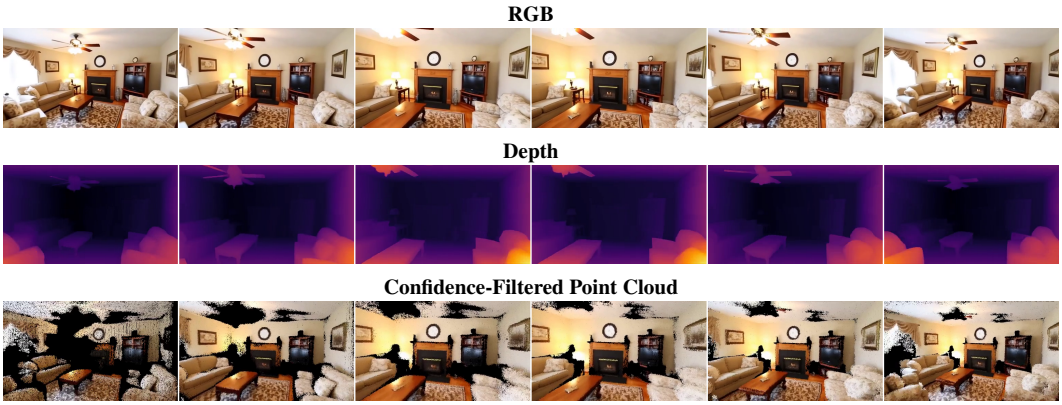

Figure 13: **Gallery Sample 9.** Top to bottom: RGB, depth, confidence-filtered point cloud.

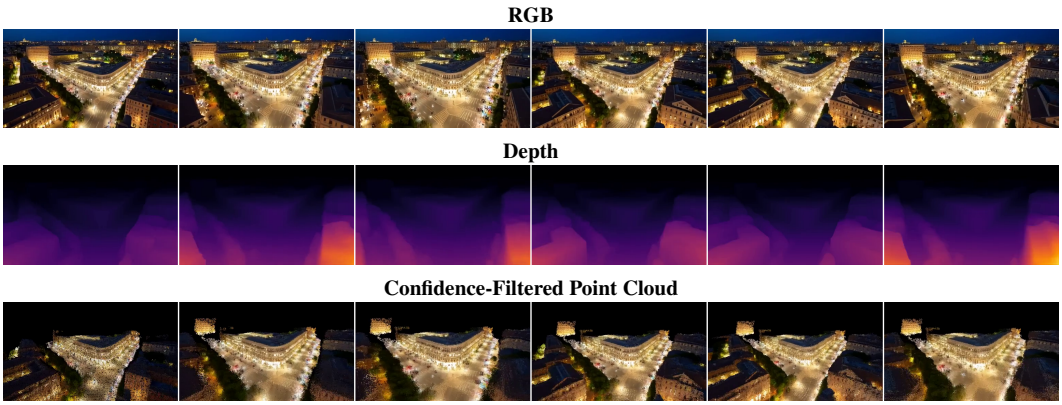

Figure 14: **Gallery Sample 11.** Top to bottom: RGB, depth, confidence-filtered point cloud.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state the main contributions of JAG, namely the dual-branch causal architecture that jointly models appearance and geometry (Section 3.2) and Self Geometry Forcing with persistent spatial memory and self-verified geometric rewards (Section 3.3). These claims are directly validated by the visual fidelity, controllability, consistency, and ablation results reported in Section 4.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations and future directions are explicitly discussed in Appendix C, including local inaccuracies of monocular depth and camera prediction in textureless, reflective, or highly dynamic regions, gradual error propagation through persistent memory under extended rollouts, the heuristic nature of the current memory update strategy, the focus of the memory and cross-view reward on static scene structure, and the unexplored choice of alternative geometry backbones.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents an empirical method without formal theorems, lemmas, or proofs. Design choices are validated through extensive ablation studies in Section 4.5.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 fully describes the dual-branch architecture and Self Geometry Forcing, while Section 4 reports the datasets, base model, branch initialization, block size, attention sink and sliding window sizes, and the optimizer. Appendix A details the per-stage training objectives, loss formulations, noising strategy, and data preprocessing, and Appendix B lists all optimization, memory, and reward hyperparameters, which together provide the information needed to reproduce the main results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: All datasets used (RealEstate10K, DL3DV, MiraData) and base or auxiliary models (Wan2.1-T2V-1.3B, Depth-Anything-3, Qwen2.5-VL) are publicly available and cited in Section 4. The method, training objectives, and all hyperparameters are described in Section 3, Appendix A, and Appendix B.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: Section 4 specifies datasets and approximate scale, the scene-level disjoint evaluation split, base model, geometry branch initialization, block size, attention sink and sliding window configurations, and the AdamW optimizer. Appendix A provides per-stage training objectives, depth and camera supervision terms, the DMD distillation objective, and the geometric reward formulation, together with the evaluation protocols at short, medium, and long horizons. Appendix B reports all hyperparameters.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Following common practice in long-horizon video generation, where each training run is highly resource-intensive, results are reported on standardized benchmarks (VBench, controllability, and consistency) without error bars across multiple seeds. Trends are instead supported by consistent improvements across short, medium, and long horizon settings and across multiple complementary metrics in Section 4.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4 reports the implementation tooling used for memory and compute efficiency, namely gradient checkpointing, DeepSpeed, and BF16 mixed precision, along with the base model size (Wan2.1-T2V-1.3B) and the block-wise causal configuration. Training the final three-stage recipe takes about 40 hours on 8 H20 GPUs (about 320 GPU-hours), and inference efficiency is reported in Table 5 and Appendix B.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and the work conforms with it. The paper relies on publicly available datasets and pretrained models, does not involve human subjects, and discusses misuse risks of generative video models in the Broader Impacts section of the appendix.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Broader Impacts section in the appendix covers both potential positive impacts (simulation, robotics, virtual content creation, and scientific visualization) and negative impacts shared with general video generative models, in particular the risk of generating misleading or deceptive visual content.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new pretrained generative models or scraped datasets that pose elevated misuse risk beyond standard text-to-video systems. The base model and datasets are already publicly available with their own licenses and usage policies.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All existing assets are credited via citations in Section 4 and Appendix A, including the datasets RealEstate10K, DL3DV, and MiraData, the base text-to-video model Wan2.1-T2V-1.3B, the geometry foundation model Depth-Anything-3, and the captioner Qwen2.5-VL. Each asset is used in a manner consistent with its publicly available license and intended research use.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper introduces a new method (JAG) trained on existing public datasets.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper involves neither crowdsourcing nor research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human subjects are involved, so IRB approval is not applicable.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as an important, original, or non-standard component of the core method. Qwen2.5-VL is invoked only as an off-the-shelf vision-language captioner during data preparation (Section 4), which is a standard data labeling step and does not affect the core methodology.