Motion Forcing: Decoupling Ego and Object Motion via Sparse Inputs for Structured Video Generation

Tianshuo Xu[1] **Zhifei Chen**[1] **Leyi Wu**[1] **Hao Lu**[1] **Ying-cong Chen**[1,2]*
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
{txu647,zchen379}@connect.hkust-gz.edu.cn

Project Page: https://tianshuo-xu.github.io/Motion-Forcing/



Figure 1: **Motion Forcing** takes user input motion and the first frame to generate motion-coherent future frames. (a-d) show reactive ego-trajectories responding to dangerous scenarios initiated by other vehicles. (e) shows a physics scene where different actions lead to different collision results. (f) shows embodied AI, where directional inputs control a robotic hand to move an object in corresponding directions.

Abstract

Structured video generation (e.g., autonomous driving) faces a fundamental domain gap between simple human control and dense physical reality. Text-to-Video models accept accessible inputs but struggle with precise spatial guidance, while dedicated autonomous driving world models require expensive, dense 3D annotations that are difficult for users to provide. To bridge this gap, we introduce **Motion Forcing**, a structured video generation framework that achieves precise multi-agent control using extremely sparse 2D inputs. Our key insight is to explicitly bridge human intent and visual synthesis via a hierarchical **"Point-Shape-Appearance"** paradigm. This approach decomposes generation into verifiable stages: modeling

*Corresponding author.

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

user intent as sparse geometric points, expanding them into dynamic depth maps to explicitly resolve 3D geometry and decouple ego-motion from object dynamics, and finally rendering high-fidelity textures. Furthermore, to elevate the model from passive instruction-following to active physical reasoning, we employ a **Masked Point Recovery** strategy. By forcing the reconstruction of dynamic depth from occluded trajectories, the model learns to infer causal continuations from partial observations. Extensive experiments demonstrate that Motion Forcing significantly outperforms state-of-the-art baselines on large-scale autonomous driving benchmarks. Additional validation in rigid-body physics and robotic manipulation confirms the structural integrity and generalizability of our framework.

1 Introduction

The development of realistic "world models" has become a critical foundation for embodied AI and autonomous driving Wang et al. [2023b], Zhao et al. [2024], Li et al. [2023], Hu et al. [2023], Yang et al. [2024a]. Unlike open-domain video generation designed for entertainment Zhuang et al. [2024], Zeng et al. [2023], Zhang et al. [2024b], world models operate in safety-centric domains where generating visually plausible textures is insufficient. These models must strictly adhere to fundamental physical laws—such as inertia, multi-agent collision dynamics, and object permanence—to support reliable decision-making and simulation Kang et al. [2025], Bansal et al. [2024], Meng et al. [2024]. However, despite achieving remarkable perceptual fidelity, recent foundation models Chen et al. [2024b], Hong et al. [2022], Blattmann et al. [2023], Yang et al. [2024b], Yin et al. [2023] frequently suffer from physical hallucinations when generating complex interactive traffic scenarios, highlighting a profound gap between visual realism and rigorous 3D physical consistency.

We attribute this structural fragility to a massive generative domain gap between simple human intent and dense physical reality. Current methodologies cluster at two extremes, failing to bridge this void. On one end, text-to-video foundation models OpenAI [2024], Yang et al. [2024b], Team [2025b] provide accessible interfaces (close to human intent) but lack the fine-grained spatial guidance required to orchestrate precise multi-agent interactions. On the other end, dedicated autonomous driving world models (e.g., MagicDrive3D Gao et al. [2024a], DriveDreamer Wang et al. [2023b]) achieve precise control but rely on dense, expensive 3D bounding boxes and HD maps. While these dense conditions closely approximate the final 3D output, they represent a prohibitive barrier for simple human interaction. Furthermore, attempting to bridge this gap using 2D motion intermediates, such as predicting dense optical flow from sparse trajectories (e.g., MoFA-Video Niu et al. [2024]), fundamentally fails to capture 3D spatial structures. This lack of 3D awareness introduces severe entanglement between ego-motion and object dynamics, culminating in geometric distortions and physical collapse during complex interactions.

To explicitly bridge the semantic gap between accessible human input and complex physical output, we propose **Motion Forcing**, a decoupled structured video generation framework tailored for autonomous driving. Inspired by Diffusion Forcing Chen et al. [2024a], we bridge this generative void via a hierarchical **"Point-Shape-Appearance"** paradigm. To ensure intuitive control, we abstract human intent into extremely sparse **Points** carrying positional and scale attributes. Rather than forcing the network to hallucinate pixels directly from these 2D signals, we introduce an intermediate **Shape** stage, expanding points into dynamic depth maps. This intermediate representation serves as the perfect bridge: it explicitly encodes 3D structural geometry, remains entirely invariant to high-frequency RGB details, and provides a dense, pixel-aligned blueprint for the final output. This strictly decouples ego and object motion in a unified metric space. Finally, we render high-fidelity RGB frames (**Appearance**) rigidly conditioned on this verified structural layout. By establishing a physically grounded, pixel-aligned skeleton before rendering, we effectively eliminate entanglement artifacts and resolve the multi-agent control dilemma.

Furthermore, to elevate the model from passive instruction-following to active physical reasoning, we introduce a **Masked Point Recovery** training strategy. By randomly occluding or truncating input agent trajectories and compelling the network to reconstruct the complete underlying dynamic depth sequence, the model learns to infer causal continuations consistent with scene dynamics within its training distribution.

While designed primarily for autonomous driving (evaluated on Waymo, Driving Dojo, and YouTube), the structural integrity of our paradigm exhibits strong cross-domain robustness. In summary, our main contributions to structured video generation are fourfold:

• **Bridging the Generative Gap.** We identify and resolve the domain gap between simple human input and dense physical output. Our "Point-Shape-Appearance" hierarchy enables precise, low-barrier multi-agent control without requiring dense 3D bounding boxes.

• **Decoupling Ego and Object Motion.** By introducing dynamic depth as the intermediate "Shape" stage, we structurally decouple camera motion from dynamic objects in 3D space, eliminating the entanglement artifacts common in previous sparse-control methods.

• **Active Physical Reasoning.** We introduce Masked Point Recovery, which forces the model to reconstruct dynamic shapes from interrupted trajectories, learning to infer causal continuations from partial observations within its training distribution.

• **Robust Generalizability.** We demonstrate that our framework maintains stable, physically coherent generation not only in driving scenarios, but also in highly complex rigid-body collisions (Physion) and embodied robotic manipulation (Jaco Play).

2 Related Work

Video Diffusion Models. Early video generation primarily extended UNet-based latent diffusion models (LDMs) Rombach et al. [2022] with temporal attention mechanisms Guo et al. [2023], Wang et al. [2023a], cerspense [2023], Chen et al. [2023, 2024b], Zhang et al. [2024a, 2023]. Recently, the field shifted toward large-scale Diffusion Transformers (DiT) and Flow Matching Lipman et al. [2022], with foundational models OpenAI [2024], Yang et al. [2024b], Kong et al. [2024], Team [2025b] significantly improving spatiotemporal coherence.

Controllable Motion and World Models. Bridging intent and generation, prior works enable control via trajectories Yin et al. [2023], Wu et al. [2024], keypoints, or flow Namekata et al. [2024], Niu et al. [2024]. In parallel, autonomous driving world models Wang et al. [2023b], Zhao et al. [2024], Li et al. [2023], Hu et al. [2023], Wang et al. [2024a], Yang et al. [2024a] emerged to simulate driving. Recent advances integrate 3D geometry to enhance consistency Ren et al. [2025], Yang et al. [2026], Lu et al. [2025], Chen et al. [2025a]. Concurrent work STANCE Chen et al. [2025b] also adopts sparse-to-dense control, encoding instance cues as attention tokens for joint RGB generation. In contrast, Motion Forcing expands sparse points into an explicit depth intermediate under independent timesteps, providing a geometric scaffold that hard-constrains the RGB stage rather than soft-conditioning it.

Physically Coherent Video Generation. Standard models often violate physical laws Motamed et al. [2025], Bansal et al. [2024]. Prior solutions integrate physics engines Liu et al. [2024b], leverage LLMs for reasoning Zhang et al. [2025a,b], or embed latent priors Kundur Satish et al. [2026]. However, entangling dynamics and appearance remains a bottleneck, as textures often override structure. We resolve this by explicitly decoupling physical reasoning from rendering.

3 Method

Generating controllable structured videos (e.g., autonomous driving) requires jointly reasoning about two intertwined aspects: *where things move* (physical dynamics) and *what things look like* (visual rendering). Mapping sparse controls $\mathbf{P}$ directly to dense pixels $\mathbf{V}$ conflates these challenges, forcing models to simultaneously resolve 3D geometry and local textures. We address this by introducing **Depth** as an intermediate representation, decomposing generation into *Physical Reasoning* ($\mathbf{P}, \mathbf{C} \rightarrow \mathbf{D}$) and *Neural Rendering* ($\mathbf{D} \rightarrow \mathbf{V}$), where $\mathbf{C}$ denotes camera motion. Crucially, sharing a single diffusion backbone enables geometric feature reuse and avoids representation misalignment.

We formalize the representations (Section 3.1), describe depth warping for camera motion (Section 3.2), the hierarchical forcing formulation (Section 3.3), the masked point recovery strategy (Section 3.4), and inference (Section 3.5).

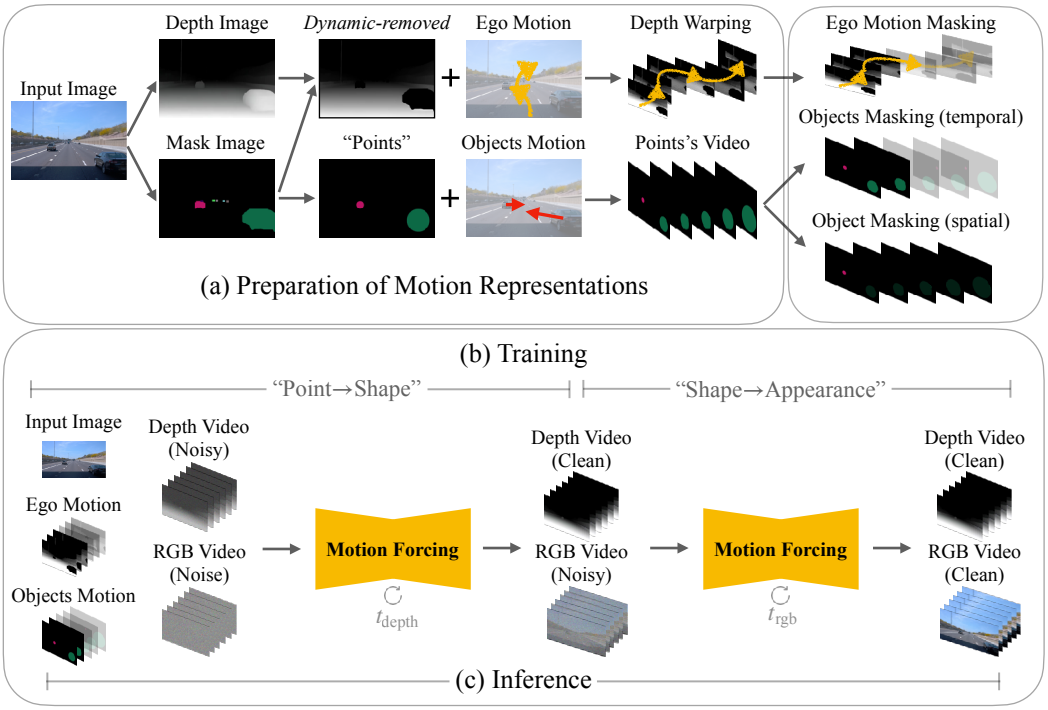


Figure 2: **Overview of the Motion Forcing framework. (a) Preparation of Motion Representations:** Input control signals are processed and subjected to spatial and temporal masking before being fed into the model. **(b) Training:** The model is trained by randomly sampling between two independent stages: "Point → Shape" (depth generation) and "Shape → Appearance" (RGB rendering). **(c) Inference:** The generation follows a complete two-stage hierarchical process, sequentially mapping sparse points to structural depth (Shape), and finally to the target RGB video (Appearance).

3.1 Preliminaries

Given a reference image $\mathbf{I}_0$ and a sequence of sparse control signals, our goal is to generate a physically plausible driving video $\mathbf{V} = \{\mathbf{I}_t\}_{t=0}^{T}$ that faithfully follows the prescribed vehicle motions and camera trajectory. We formulate this as a hierarchical mapping across three levels of scene description:

Point $\mathbf{P} = \{\mathbf{P}_t\}_{t=0}^{T}$: Sparse control signals for dynamic agents, abstracted as maximum inscribed circles (centroid and radius). The centroid governs planar motion, while the radius implicitly encodes depth via perspective projection, offering procedural flexibility. **Shape (Depth)** $\mathbf{D} = \{\mathbf{D}_t\}_{t=0}^{T}$: Intermediate structural representation of dense depth maps. Depth captures continuous 3D scene geometry—surface distances, spatial relationships, and occlusion ordering—bridging control-level reasoning and pixel-level rendering. **Appearance** $\mathbf{V} = \{\mathbf{I}_t\}_{t=0}^{T}$: The target RGB video containing high-frequency visual details (texture, lighting, materials).

In addition to agent-level control, driving scenarios require explicit camera motion specification. We denote the camera trajectory as $\mathbf{C} = \{(\mathbf{R}_t, \mathbf{t}_t, \mathbf{K}_t)\}_{t=0}^{T}$, comprising per-frame extrinsics $[\mathbf{R}_t|\mathbf{t}_t] \in \mathbb{R}^{3\times4}$ and intrinsics $\mathbf{K}_t \in \mathbb{R}^{3\times3}$. During training, these are estimated by VGGT Wang et al. [2025]. We next describe how $\mathbf{C}$ is encoded into a form compatible with our depth-centric framework.

3.2 Camera Motion via Depth Warping

A standard approach to injecting camera motion into a diffusion model is to embed the extrinsic parameters $(\mathbf{R}_t, \mathbf{t}_t)$ and intrinsics $\mathbf{K}_t$ as global conditioning vectors via cross-attention or adaptive normalization. However, this parametric strategy faces two critical challenges in the driving domain.

First, **entanglement under limited data diversity.** Injecting camera pose as a low-dimensional embedding alongside other conditions tends to entangle camera motion with scene content. This is exacerbated in driving datasets, where camera trajectories correlate strongly with scene layout and agent behavior, making factorization exceptionally difficult without massive data.

Second, **insufficient spatial precision.** Autonomous driving demands geometrically precise 3D camera manipulation. A parametric embedding compresses the 6-DoF transformation into a vector that the network must internally re-expand into dense displacements—a lossy bottleneck poorly suited for pixel-aligned processing.

To simultaneously address these issues, we propose representing camera motion as **warped depth maps**. By geometrically warping the first-frame depth $\mathbf{D}_0$ according to the target camera pose, we produce a dense, pixel-aligned conditioning signal. This representation is (i) *structurally decoupled* from appearance and layout conditions by construction, as it encodes only geometric camera effects, and (ii) spatially precise, making the full 6-DoF transformation explicit at every pixel location. Furthermore, since our intermediate representation is also depth-based, encoding camera motion in the same modality establishes a shared geometric vocabulary. This is significantly easier for the network to internalize than bridging the gap between abstract pose parameters and pixel-space effects.

Formally, given the first-frame depth $\mathbf{D}_0$ and camera poses $\{(\mathbf{R}_t, \mathbf{t}_t, \mathbf{K}_t)\}_{t=0}^{T}$, we construct the camera motion conditioning $\mathbf{W} = \{\mathbf{W}_t\}_{t=0}^{T}$ by unprojecting the first-frame depth into 3D space, transforming it to each target camera's perspective, and forward-splatting the depth values onto the target image plane. Because this geometric warping assumes a static scene, dynamic agents in $\mathbf{D}_0$ may leave localized "ghosts" at their initial positions in $\mathbf{W}_t$. However, we provide the concurrent sparse point trajectories $\mathbf{P}$ as an overriding control signal. The attention mechanism of the DiT effectively learns to utilize $\mathbf{W}_t$ exclusively as a rigid background constraint while attending to $\mathbf{P}$ to guide dynamic agent rendering, successfully resolving this potential spatial conflict. The detailed mathematical formulation for this projection and splatting process is provided in Appendix A.1.

3.3 Motion Forcing: Unified Hierarchical Diffusion

Rather than cascading separate models for reasoning and rendering, we propose a **single unified diffusion model** ϵ_θ that handles both stages within a shared 3D DiT backbone Yang et al. [2024b]. The key mechanism enabling this is **dual independent diffusion timesteps**: τ_d controls the noise level of the depth latents, and τ_v controls the noise level of the video latents.

Let $\mathbf{z}^d_{\tau_d}$ and $\mathbf{z}^v_{\tau_v}$ denote the VAE-encoded latents of $\mathbf{D}$ and $\mathbf{V}$, perturbed by Gaussian noise at levels τ_d and τ_v respectively. These latent streams are concatenated temporally and processed jointly by the backbone. While doubling the sequence length T increases the self-attention computational cost during training, this overhead is isolated to the joint training phase. At inference time, because the generation is strictly decoupled into two distinct passes (first predicting depth, then rendering RGB), the model reverts to the standard $O(T^2)$ sequence length per stage, ensuring highly scalable generation. Each stream is modulated by its specific timestep embedding via adaptive layer normalization (AdaLN), enabling disentangled denoising control. The unified objective is:

$$\mathcal{L} = \mathbb{E}_{\mathbf{D},\mathbf{V},\tau_d,\tau_v,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\big(\mathbf{z}^d_{\tau_d} \oplus \mathbf{z}^v_{\tau_v},\ \tau_d,\ \tau_v,\ \mathbf{I}_0,\ \mathbf{P},\ \mathbf{W}\big)\right\|^2\right], \tag{1}$$

where $\oplus$ denotes temporal concatenation. The conditioning inputs—reference image $\mathbf{I}_0$, point control $\mathbf{P}$, and camera depth $\mathbf{W}$—are each VAE-encoded and concatenated along the channel dimension of the noisy input.

3.3.1 Dual Adaptive Layer Normalization

To enable a single set of transformer blocks to simultaneously serve two denoising tasks, we extend the standard AdaLN mechanism. For each block, the depth and video halves of the hidden state sequence receive *different* scale and shift parameters derived from their respective timestep embeddings τ_d and τ_v:

$$\mathbf{h}'_{1:T} = \mathrm{LN}(\mathbf{h}_{1:T}) \odot [\gamma(\tau_d) \| \gamma(\tau_v)] + [\beta(\tau_d) \| \beta(\tau_v)], \tag{2}$$

where $\|$ denotes concatenation along the temporal (sequence) axis, and γ, β are learned affine functions. At the output layer, norm_out is applied sequentially with both embeddings to produce the final prediction. This design allows the backbone attention layers to share representations while the normalization provides task-specific modulation.

3.3.2 Forcing Strategy

Inspired by Diffusion Forcing Chen et al. [2024a], we employ a **stochastic mode-switching** strategy. Each training iteration forces the model to solve exactly one sub-problem by fixing the non-target timestep:

Mode I: Physical Reasoning (Point + Camera $\rightarrow$ Shape). Set $\tau_v = T_{\max}$ (pure noise) and sample $\tau_d \sim \mathcal{U}\{0, \dots, T_{\max}-1\}$. Deprived of RGB cues, the model reconstructs $\mathbf{D}$ from the sparse control $\mathbf{P}$, camera depth $\mathbf{W}$, and reference $\mathbf{I}_0$. This mode cultivates physical reasoning: inferring 3D structure evolution, resolving occlusions, and predicting dynamic depth changes from minimal control inputs. The loss is $\mathcal{L}_{\text{reason}} = \mathbb{E}[\|\epsilon_d - \hat{\epsilon}_d\|^2]$ s.t. $\tau_v = T_{\max}$.

Mode II: Neural Rendering (Shape $\rightarrow$ Appearance). Set $\tau_d = 0$ (ground truth) and sample $\tau_v \sim \mathcal{U}\{0, \dots, T_{\max}-1\}$. Conditioned on perfect geometry, the model focuses on hallucinating consistent textures, lighting, and material properties. The loss is $\mathcal{L}_{\text{render}} = \mathbb{E}[\|\epsilon_v - \hat{\epsilon}_v\|^2]$ s.t. $\tau_d = 0$.

By alternating between modes, ϵ_θ learns to function as both a physics engine and a renderer within a shared latent space.

3.4 Masked Point Recovery for Physical Reasoning

To explicitly enforce causal reasoning and active physical understanding, we introduce **Masked Point Recovery**. A critical secondary objective of this strategy is to bridge the inherent gap between dense training trajectories and the sparse, partial inputs typical of real-world inference. By systematically corrupting input motion controls while strictly supervising on the complete intermediate depth video $\mathbf{D}$, we compel the model to internalize physical laws (e.g., inertia, object permanence) and hallucinate missing geometries. As shown in Figure 2, we employ three distinct masking strategies:

Temporal Ego Masking: We randomly truncate ego-motion conditioning. Camera signals after frame $t = \lfloor \tau_{\text{ego}} T \rfloor$, with a cutoff ratio $\tau_{\text{ego}} \sim \mathcal{U}(0.3, 1.0)$, are masked. The model must extrapolate the subsequent ego-trajectory from initial momentum and scene layout. **Temporal Object Masking:** Similarly, we apply a temporal mask to object control points $\mathbf{P}$ beyond a sampled cutoff $\tau_{\text{obj}} \sim \mathcal{U}(0.3, 1.0)$, compelling the model to predict ongoing object motion from initial velocities. **Spatial Object Masking:** To model object permanence and implicit interactions, we independently drop the entire trajectory of each object i with probability p_{drop}. The model must deduce the missing agent's physical presence and trajectory from the target depth $\mathbf{D}$ and the reactive behaviors of visible agents.

3.5 Inference: Two-Stage Hierarchical Generation

In inference, the dual-timestep formulation enables sequential chaining of the learned capabilities. Note that the target camera poses $\mathbf{C}$ required for depth warping can be flexibly extracted from reference videos or procedurally generated, as detailed in Appendix A.2:

Stage 1: Depth Generation (Physical Reasoning). Initialize $\mathbf{z}^d, \mathbf{z}^v \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Execute the DDIM sampling loop targeting only the depth stream (τ_v fixed at $T_{\max}$). The model operates in Mode I, synthesizing a clean depth latent $\hat{\mathbf{z}}^d$ from the control inputs $\mathbf{P}$ and camera motion $\mathbf{W}$. Throughout this stage, the RGB noise $\mathbf{z}^v$ is held constant.

Stage 2: Appearance Synthesis (Neural Rendering). Fix the depth latent to $\hat{\mathbf{z}}^d$ (setting $\tau_d = 0$) and re-initialize $\mathbf{z}^v$ with fresh noise. Run a second DDIM loop targeting the video stream, with the model operating in Mode II. The generated depth serves as a fixed geometric blueprint for rendering.

This hierarchical pipeline offers **interpretability**: the intermediate depth $\hat{\mathbf{D}}$ provides a verifiable 3D scene layout that users can inspect before committing to the expensive rendering stage. It also provides a natural editing interface—users may modify the predicted depth (e.g., removing or repositioning agents) and re-render with consistent appearance.



Figure 3: **Comparison with state-of-the-art models in synthesizing cut-in and evasive driving scenarios.** While the leading text-conditioned models (Seed Dance2.0, Wan2.6, Vista) rely on text prompts, trajectory-controllable baselines like MOFA-Video require object trajectories coupled with scene trajectories to simulate camera movement. Our method **Motion Forcing**, however, effectively captures these complex dynamics given only the initial ego and object motion trajectories.

Table 1: Quantitative results on the Waymo dataset. Although our geometry-first approach yields a higher FVD compared to massively pretrained closed-source models, this is a deliberate tradeoff that significantly improves motion consistency (FVMD) and physical plausibility (Physics-IQ). One-stage* refers to the single-stage version of Motion Forcing, which directly generates RGB frames without intermediate depth.

Methods	MOFA-Video	Seed Dance2.0	Wan2.6	Vista	One-stage*	**Ours**
FVD↓	272.6	**112.5**	118.3	147.2	152.4	157.8
FVMD↓	421.3	345.6	316.2	323.1	218.4	**205.2**
Physics-IQ↑	21.6	30.5	31.2	28.4	30.7	**33.2**

4 Experiments

4.1 Experimental Setup

Training Details. We fine-tune CogVideoX1.5-5B-I2V Yang et al. [2024b]. Comprehensive training details and hyperparameters are provided in Appendix A.3. **Datasets.** We primarily evaluate on **autonomous driving** using sequences from Waymo Sun et al. [2020], Driving Dojo Wang et al. [2024b], and YouTube. To validate generality, we additionally evaluate on **Physion** Bear et al. [2021] (rigid-body physics) and **Jaco Play** Dass et al. [2023] (robotic manipulation). **Evaluation Metrics.** We report results on 100 Waymo Sun et al. [2020] test videos using: (1) FVD Unterthiner et al. [2019] for distributional similarity, (2) FVMD Liu et al. [2024a] for temporal motion coherence, and (3) Physics-IQ Motamed et al. [2025] for physical plausibility.

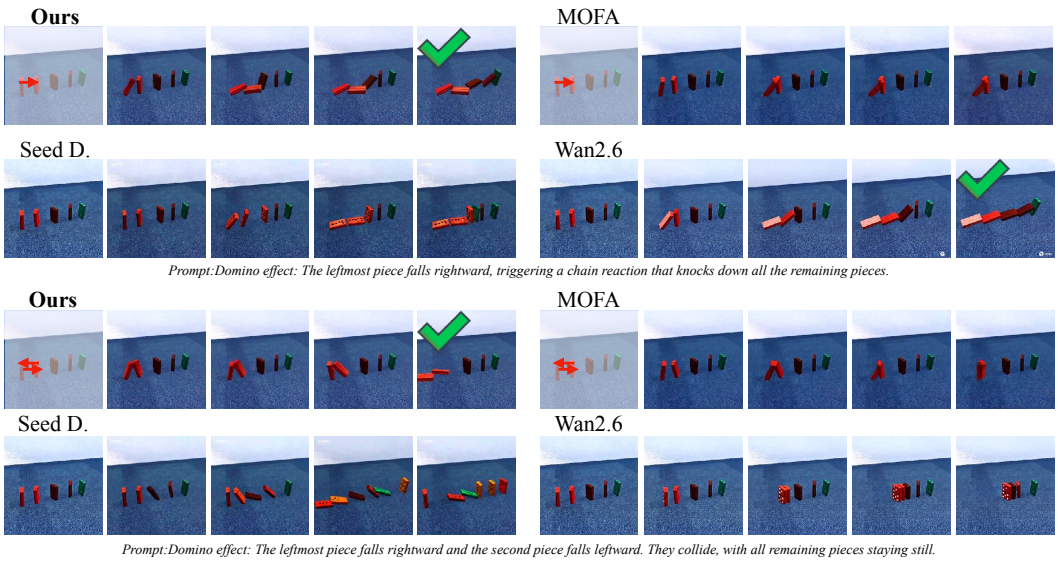


Figure 4: **Validation of physical capabilities.** ✓ indicate physically coherent. MOFA-Video is fine-tuned on our exact dataset for fairness. As the control commands become more complex, only our method maintains stable and physically coherent generation.

4.2 Comparisons with the State-of-the-art Methods

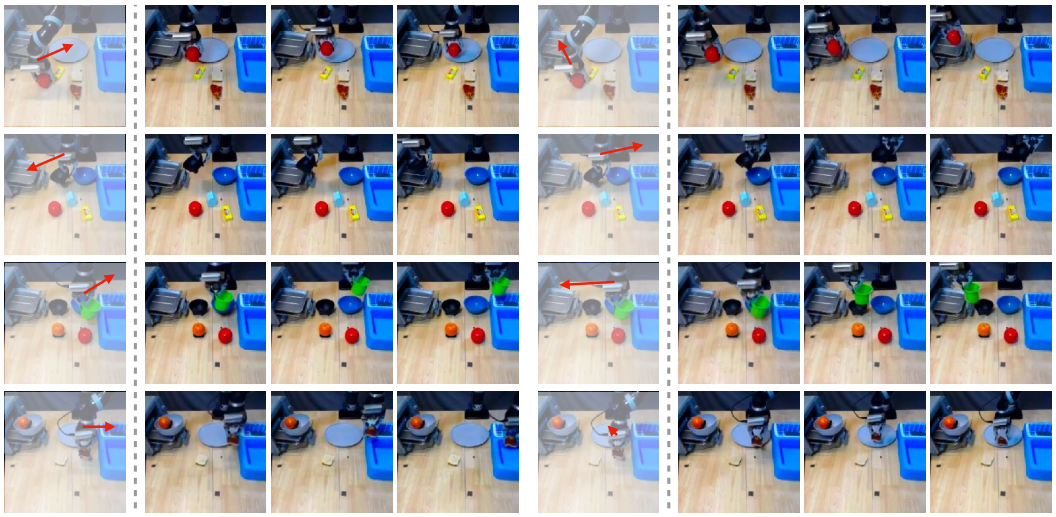


Figure 5: **Controllable robotic hand manipulation.** By providing different motion control inputs, our model can control the movement of the robotic arm and the grasped object in specified directions.

Real-world Driving Video Generation. As shown in Table 1, we evaluate against MOFA-Video Niu et al. [2024], Vista Gao et al. [2024c], a one-stage variant, and text-conditioned I2V models (Seed Dance2.0 Team [2025a], Wan2.6 Team [2025b]). Map-conditioned generators MagicDrive Gao et al. [2024b] and DriveDreamer Wang et al. [2023b] are excluded as they require dense HD maps and 3D bounding boxes, incompatible with our sparse-control premise. Text-conditioned baselines achieve lower FVD via massive pretraining scale, yet fall significantly behind on FVMD and Physics-IQ, confirming text prompts cannot substitute for explicit trajectory control. The one-stage variant likewise yields inferior physical metrics, confirming intermediate depth is crucial. Qualitatively (Figure 3), text-conditioned models struggle with precise dynamics, and MOFA-Video degrades during multi-agent interactions due to coarse flow splatting, whereas our method faithfully synthesizes

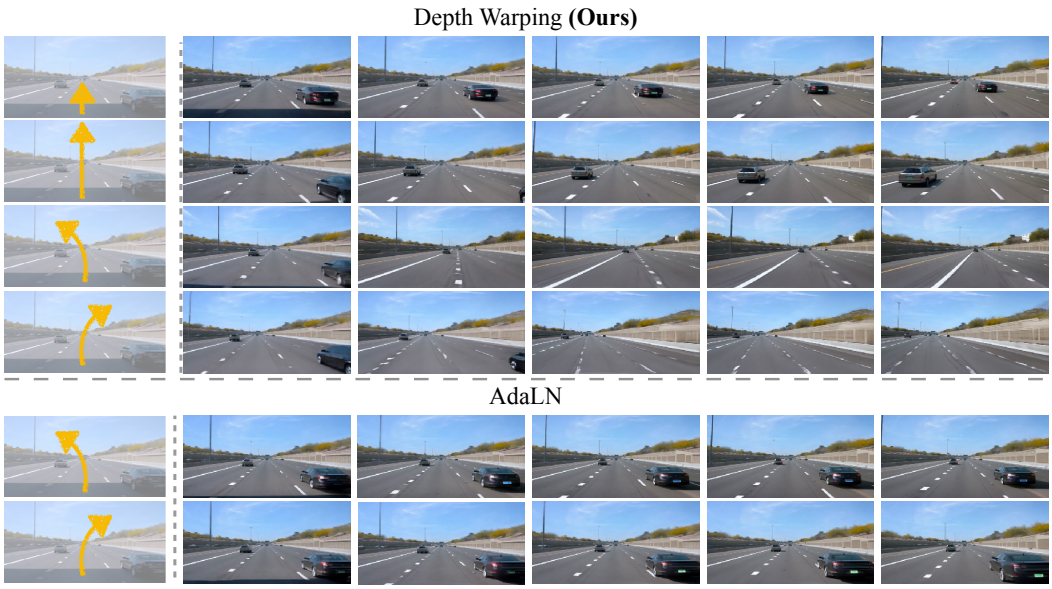


Figure 6: **Qualitative comparison of ego motion control.** Compared to the baseline AdaLN approach, our **Depth Warping-based** method demonstrates significantly superior accuracy and flexibility in controlling both the direction and speed.

lane changes and evasive maneuvers. An AD-specific evaluation on nuScenes clips (Table 3) further confirms our geometry-first approach achieves significantly lower trajectory error and collision rates.

Generalization to Physics Scenarios. To validate the generality of our framework beyond driving, we evaluate on the Physion Bear et al. [2021] dataset. As shown in Figure 4, when physical interactions grow more complex (*e.g.*, multi-object collisions in domino scenarios), MOFA-Video—even fine-tuned on our dataset—fails to maintain physical coherence, whereas Motion Forcing consistently produces plausible collision dynamics, confirming that our hierarchy effectively transfers to general physics scenarios.

Table 2: **Ablation studies** on the Waymo test set. Each row replaces one component of Motion Forcing with the specified alternative.

Methods	**FVD**↓	**FVMD**↓	**Physics-IQ**↑
Motion Forcing	157.8	**205.2**	**33.2**
w/o Mask Point Recovery	**142.0**	233.8	27.5
Motion Representation			
Segmentation	167.0	228.8	29.7
Optical Flow	173.0	224.3	28.4
Object Motion Embedding			
Softmax Splatting	160.3	251.7	28.5
Ego Motion Encoding			
AdaLN	159.6	243.8	29.1

Table 3: **AD Proxy Evaluation** on a 50-video nuScenes subset. We compare our geometry-first model against the text-conditioned Vista baseline. Our approach yields lower trajectory tracking errors (L2) and agent collision rates.

Metric	**Vista**	**Ours**
L2(1s)↓	0.176	**0.071**
L2(2s)↓	0.389	**0.194**
L2(3s)↓	0.467	**0.289**
Collision↓	0.581	**0.498**

Generalization to Robotic Manipulation. We further validate on the Jaco Play Dass et al. [2023] dataset by providing directional inputs to control robotic hand motions for grasping and object manipulation. As shown in Figure 5, our model flexibly guides the robotic arm and grasped objects in specified directions, confirming the flexibility of our point-based control primitive across domains.

4.3 Ablation Study

We conduct ablations on the Waymo test set to validate key design choices (Table 2). The full two-stage model incurs a modest FVD increase over the one-stage variant (157.8 vs. 152.4): the depth-first

decomposition geometrically constrains RGB generation, slightly trading distributional similarity for motion coherence and physical plausibility—a cost reflected in the substantially better FVMD and Physics-IQ. **Active Physical Reasoning:** Removing Masked Point Recovery modestly improves FVD by allowing passive overfitting to visual textures, but severely degrades FVMD and Physics-IQ, confirming its necessity for learning causal dynamics. **Intermediate Motion Representation:** Depth outperforms segmentation and optical flow by encoding continuous 3D geometry indispensable for motion reasoning. **Object Motion Embedding:** Replacing instance flow with softmax splatting Niu et al. [2024] degrades FVMD and Physics-IQ, as coarse pixel-level warping struggles with multi-body dynamics. **Ego Motion Encoding:** Replacing depth warping with AdaLN markedly degrades motion coherence, confirming that spatially explicit, pixel-aligned 6-DoF conditioning is essential.

5 Conclusion and Limitations

We presented **Motion Forcing**, a unified diffusion backbone for controllable video generation via a Point $\rightarrow$ Shape $\rightarrow$ Appearance hierarchy using dual timesteps, depth-warped ego conditioning, and Masked Point Recovery. This depth-first decomposition separates physical reasoning from visual rendering, allowing the shared backbone to reuse geometric features. Across driving, physics, and robotics, it significantly improves motion coherence and physical plausibility under a sparse control.

Broader Impacts. While beneficial for driving simulation, robotics, and content creation, it could be misused to generate misleading synthetic footage; responsible deployment is advised.

Limitations. We do not surpass large closed-source models in raw FVD due to massive pretraining scale. Severe occlusion, dense traffic, and ID switches can disrupt sparse point control or depth ordering. **See supplementary for representative failure cases.**

References

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *International Conference on Machine Learning*, 2024.

Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. 2021.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

cerspense. zeroscope_v2. `https://huggingface.co/cerspense/zeroscope_v2_576w`, 2023. Accessed: 2023-02-03.

Anthony Chen, Wenzhao Zheng, Yida Wang, Xueyang Zhang, Kun Zhan, Peng Jia, Kurt Keutzer, and Shanghang Zhang. Geodrive: 3d geometry-informed driving world model with precise action control. *arXiv preprint arXiv:2505.22421*, 2025a.

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

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Zhifei Chen, Tianshuo Xu, Leyi Wu, Luozhou Wang, Dongyu Yan, Zihan You, Wenting Luo, Guo Zhang, and Yingcong Chen. Stance: Motion coherent video generation via sparse-to-dense anchored encoding. *arXiv preprint arXiv:2510.14588*, 2025b.

Shivin Dass, Jullian Yapeter, Jesse Zhang, Jiahui Zhang, Karl Pertsch, Stefanos Nikolaidis, and Joseph J. Lim. Clvr jaco play dataset, 2023. URL `https://github.com/clvrai/clvr_jaco_play_dataset`.

Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *arXiv preprint arXiv:2405.14475*, 2024a.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In *International Conference on Learning Representations*, 2024b.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. 2024c.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. 2023.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *International Conference on Machine Learning*, 2025.

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.

Siddarth Nilol Kundur Satish, Devesh Jaiswal, Hongyu Chen, and Abhishek Bakshi. Physvideogenerator: Towards physically aware video generation via latent physics guidance. *arXiv preprint arXiv:2601.03665*, 2026.

Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. *arXiv preprint arXiv:2310.07771*, 2023.

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.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fréchet video motion distance: A metric for evaluating motion consistency in videos. *arXiv preprint arXiv:2407.16124*, 2024a.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pages 360–378. Springer, 2024b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Hao Lu, Zhuang Ma, Guangfeng Jiang, Wenhang Ge, Bohan Li, Yuzhan Cai, Wenzhao Zheng, Yunpeng Zhang, and Yingcong Chen. 4d driving scene generation with stereo forcing. *arXiv preprint arXiv:2509.20251*, 2025.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? 2025.

Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. 2024.

Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. *arXiv preprint arXiv:2405.20222*, 2024.

OpenAI. Video generation models as world simulators. *OpenAI Blog*, 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.

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*, pages 6121–6132, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Kennedy, Charles Todd, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

ByteDance Seed Team. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025a.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025b. URL `https://arxiv.org/abs/2503.20314`.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. In *Deep Generative Models Workshop at ICLR*, 2019.

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*, pages 5294–5306, 2025.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023a.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023b.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985*, 2024a.

Yuqi Wang, Ke Cheng, Jiawei He, Qitai Wang, Hengchen Dai, Yuntao Chen, Fei Xia, and Zhaoxiang Zhang. Drivingdojo dataset: Advancing interactive and knowledge-enriched driving world model. In *Advances in Neural Information Processing Systems*, 2024b.

Weijia 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*, 2024.

Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14662–14672, 2024a.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026.

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. *arXiv preprint arXiv:2408.06072*, 2024b.

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.

Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. *arXiv:2311.10982*, 2023.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023.

David Junhao Zhang, Dongxu Li, Hung Le, Mike Zheng Shou, Caiming Xiong, and Doyen Sahoo. Moonshot: Towards controllable video generation and editing with multimodal conditions. *arXiv preprint arXiv:2401.01827*, 2024a.

Ke Zhang, Cihan Xiao, Yiqun Mei, Jiacong Xu, and Vishal M Patel. Think before you diffuse: Llms-guided physics-aware video generation. *arXiv preprint arXiv:2505.21653*, 2025a.

Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. PhysDreamer: Physics-based interaction with 3d objects via video generation. 2024b.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025b.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024.

Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. *arXiv preprint arXiv:2401.09414*, 2024.

A Additional Implementation Details

A.1 Mathematical Formulation of Depth Warping

As introduced in the main text, we construct the camera motion conditioning $\mathbf{W} = \{\mathbf{W}_t\}_{t=0}^{T}$ by unprojecting the first-frame depth into 3D, transforming it to each target camera, and forward-splatting:

$$\mathbf{W}_t = \text{Splat}\big(\mathbf{D}_0,\ \Pi_t \circ \Pi_0^{-1}(\mathbf{u}, \mathbf{D}_0(\mathbf{u}))\big)\,. \tag{3}$$

Concretely, for each pixel $\mathbf{u} = (u, v)$ in frame 0, we first unproject to a world-space point:

$$\mathbf{p}_{\text{world}} = \mathbf{R}_0^{\top}\left(\mathbf{D}_0(\mathbf{u}) \cdot \mathbf{K}_0^{-1}\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} - \mathbf{t}_0\right), \tag{4}$$

and then project it onto frame t's image plane:

$$\mathbf{u}_t = \pi(\mathbf{K}_t(\mathbf{R}_t\mathbf{p}_{\text{world}} + \mathbf{t}_t))\,, \tag{5}$$

where π denotes perspective division. Finally, the depth value $\mathbf{D}_0(\mathbf{u})$ is splatted to location $\mathbf{u}_t$ in the target frame.

A.2 Camera Pose Construction During Inference

During inference, the camera trajectory $\mathbf{C}$ required for Depth Warping can be constructed through multiple intuitive paradigms depending on the generation task. For video-to-video editing or style transfer scenarios where a reference video exists, we directly extract the camera poses from the source video using off-the-shelf pose estimators such as VGGT Wang et al. [2025] or DA3 Lin et al. [2025]. This provides a highly realistic, metric-accurate trajectory that mirrors the original video's ego-motion. For pure generation tasks from a single image, users can procedurally generate camera trajectories by defining a sequence of 6-DoF rigid body transformations. For example, a user can script a forward driving motion at a constant speed or simulate a lane change by mathematically computing the consecutive camera matrices $\mathbf{K}_t, \mathbf{R}_t, \mathbf{t}_t$. This flexibility ensures that Motion Forcing supports both data-driven tracking and explicitly scripted camera controls, maintaining the sparse-control paradigm without relying on dense scene representations.

A.3 Training Details

We fine-tune CogVideoX1.5-5B-I2V Yang et al. [2024b] on 8 NVIDIA A100 GPUs with DeepSpeed ZeRO Stage-2 Rajbhandari et al. [2020] in `bfloat16` mixed precision. The total training process takes approximately one week. We use AdamW Loshchilov and Hutter [2019] with learning rate 1×10^{-5}, $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 1×10^{-4}, and 100 warmup steps. Each sample consists of 33 frames at 320×480 resolution (stride 2). We train for 10 epochs with an effective batch size of 8.

A.4 Point Construction and Control Interface

Motion Forcing uses the same point tensor format during training and inference. For each dynamic object, we first obtain an instance mask or tracked object region. In datasets with annotated masks, such as Physion, we use the provided mask directly; for driving videos, masks are obtained with an off-the-shelf instance segmentation model and temporally associated with object tracks when available. Each mask is converted into the maximum inscribed circle, represented by a center (x_t^i, y_t^i) and radius r_t^i. The centers specify image-plane motion, while the radius provides a coarse perspective cue that helps distinguish objects moving laterally from objects approaching or receding.

At inference time, the same representation can be produced from user sketches, tracked objects, or scripted trajectories. A user-provided arrow specifies the desired displacement of an object center, while the radius can be kept fixed or updated by a simple perspective rule. This keeps the control interface sparse: the model is not given a dense flow field or full segmentation video, and must expand sparse point controls into dynamic depth before RGB rendering.

The main sources of point-construction error are imperfect masks, ID switches in crowded scenes, and radius noise under heavy occlusion. Such errors mostly affect small non-motorized agents and

densely overlapping objects. Masked Point Recovery partially exposes the model to missing or corrupted controls, but it is not a substitute for accurate perception; the limitations section discusses failure modes caused by severe tracking and occlusion ambiguity.

A.5 Masked Point Recovery

Masked Point Recovery is applied only during the Point-to-Shape training mode. We randomly hide temporal suffixes of ego or object trajectories, or drop entire object trajectories, while keeping supervision on the complete dynamic depth sequence. The objective asks the model to explain the hidden part of the scene using visible motion, depth ordering, object persistence, and interaction cues. This design is intended to encourage causal interaction inference in structured domains, for example predicting how a nearby vehicle reacts to a cut-in or how a rigid object moves after contact. It should not be interpreted as learning general physical laws independent of the training distribution.

A.6 Shared Backbone

The unified backbone shares attention layers across the depth and RGB streams, but the two streams are separated by independent diffusion timesteps and task-specific AdaLN modulation. During Point-to-Shape training, the RGB stream is fixed at pure noise and the loss is applied to depth. During Shape-to-Appearance training, the depth stream is clean and the loss is applied to RGB. This mode switching prevents the model from solving both tasks through a single undifferentiated denoising target, while still allowing the renderer to reuse geometric features learned in the depth stage.

B Additional Qualitative Results

B.1 Collected Driving Data

The YouTube driving subset focuses on highly interactive traffic scenes in which ego motion or one vehicle's action influences surrounding vehicles or pedestrians. It contains more than 9,000 clips and roughly 200 hours of video covering diverse weather, road layouts, urban scenes, and interaction patterns. All case videos referenced in the main paper are provided in the supplementary material.

B.2 nuScenes AD Proxy Protocol and Depth Videos

We include an additional autonomous-driving proxy evaluation on 50 nuScenes CAM_FRONT clips. The reported trajectory metric first estimates camera poses from the generated RGB video with DA3 Lin et al. [2025], then compares the recovered camera centers with the nuScenes ego trajectory after alignment. We report both Sim(3)-aligned and scale-only L2, and use the Sim(3)-aligned L2 in the main table because monocular depth and pose are only defined up to global similarity. The collision number is a lightweight any-object image-plane overlap proxy computed with instance segmentation; it is intended as a diagnostic for visible object overlap rather than a substitute for a closed-loop driving safety metric.

B.3 Generation Pipeline and Point Control Robustness

Figures 7–9 visualize the complete two-stage generation pipeline across three driving scenes, each with three distinct point control configurations (Ctrl A/B/C) applied to the same camera motion. The top row of each figure shows the shared camera-warped depth condition $\mathbf{W}$. Below, each group of three rows presents one control variant: the sparse point input video $\mathbf{P}$, the generated intermediate depth sequence $\mathbf{D}$, and the final rendered RGB video $\mathbf{V}$. The consistent depth structure and realistic appearance across all three controls confirm that Motion Forcing reliably translates different sparse trajectories into plausible scene dynamics without requiring retraining or architectural changes.

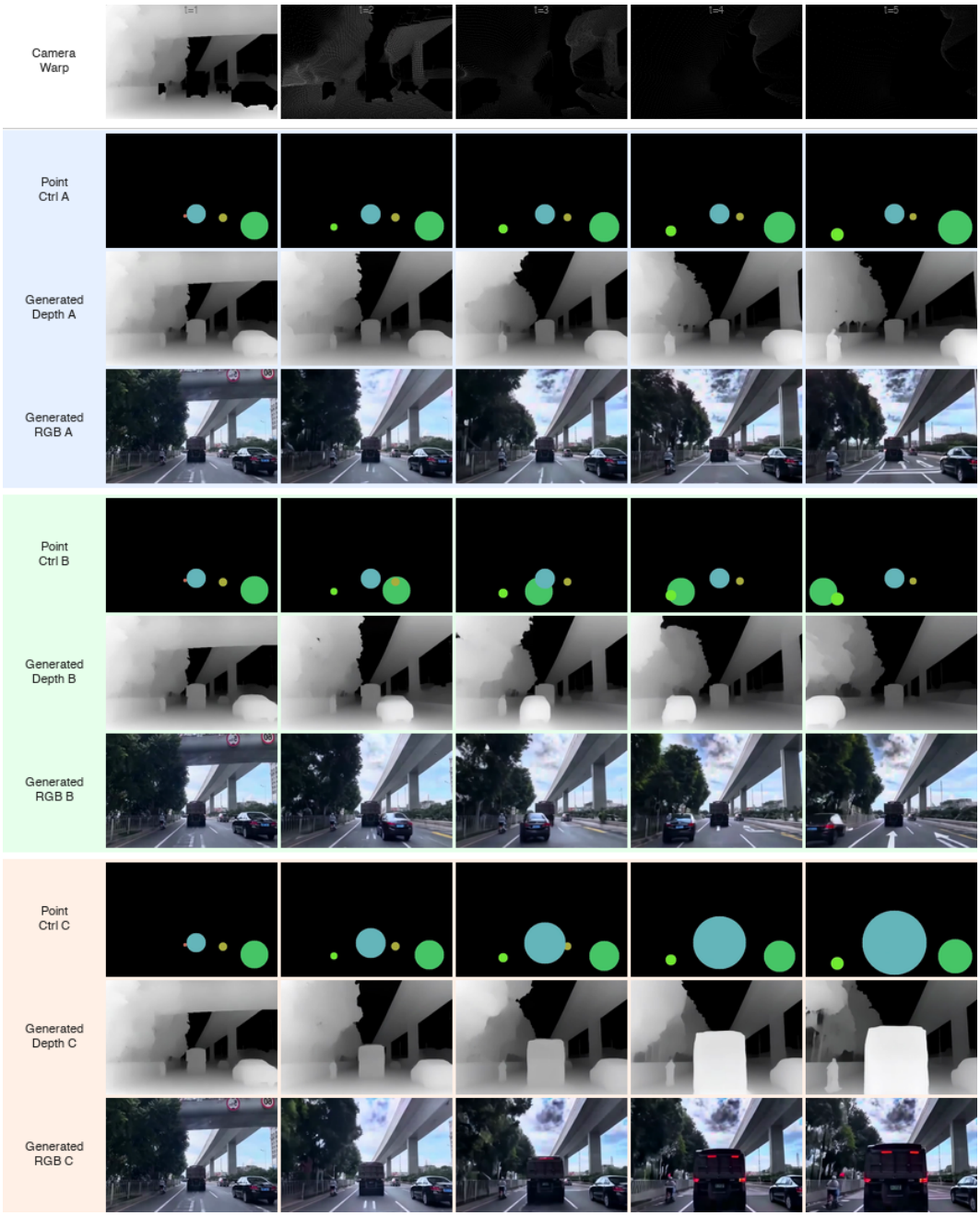


Figure 7: **Generation pipeline and point control robustness — Scene 1.** Top row: shared camera-warped depth condition. Rows 2–10: three point control variants (A/B/C), each showing point input, generated depth, and generated RGB (5 uniformly sampled frames per row).

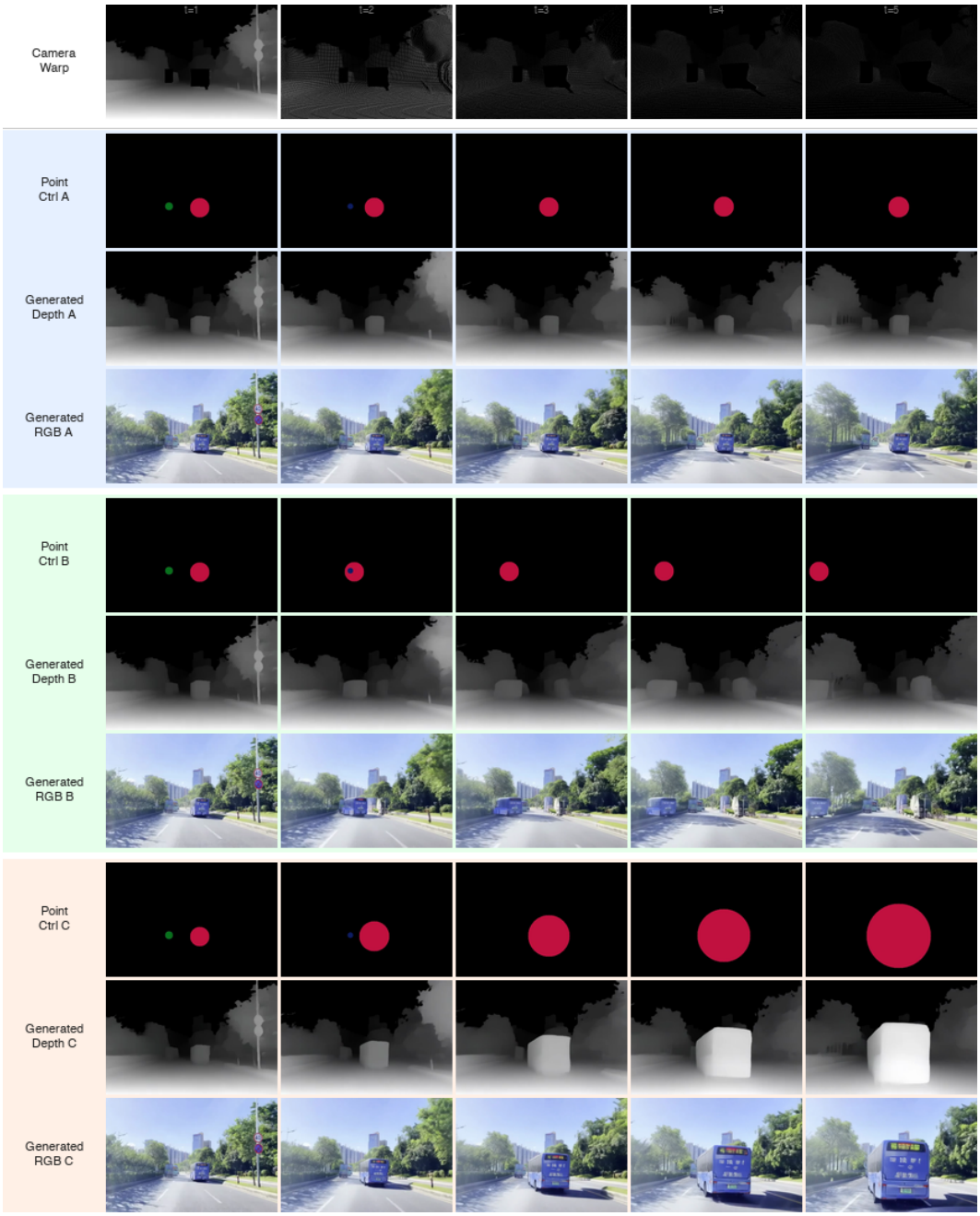


Figure 8: **Generation pipeline and point control robustness — Scene 2.** Same layout as Figure 7. The model faithfully propagates structurally distinct trajectories through the depth stage before RGB rendering.

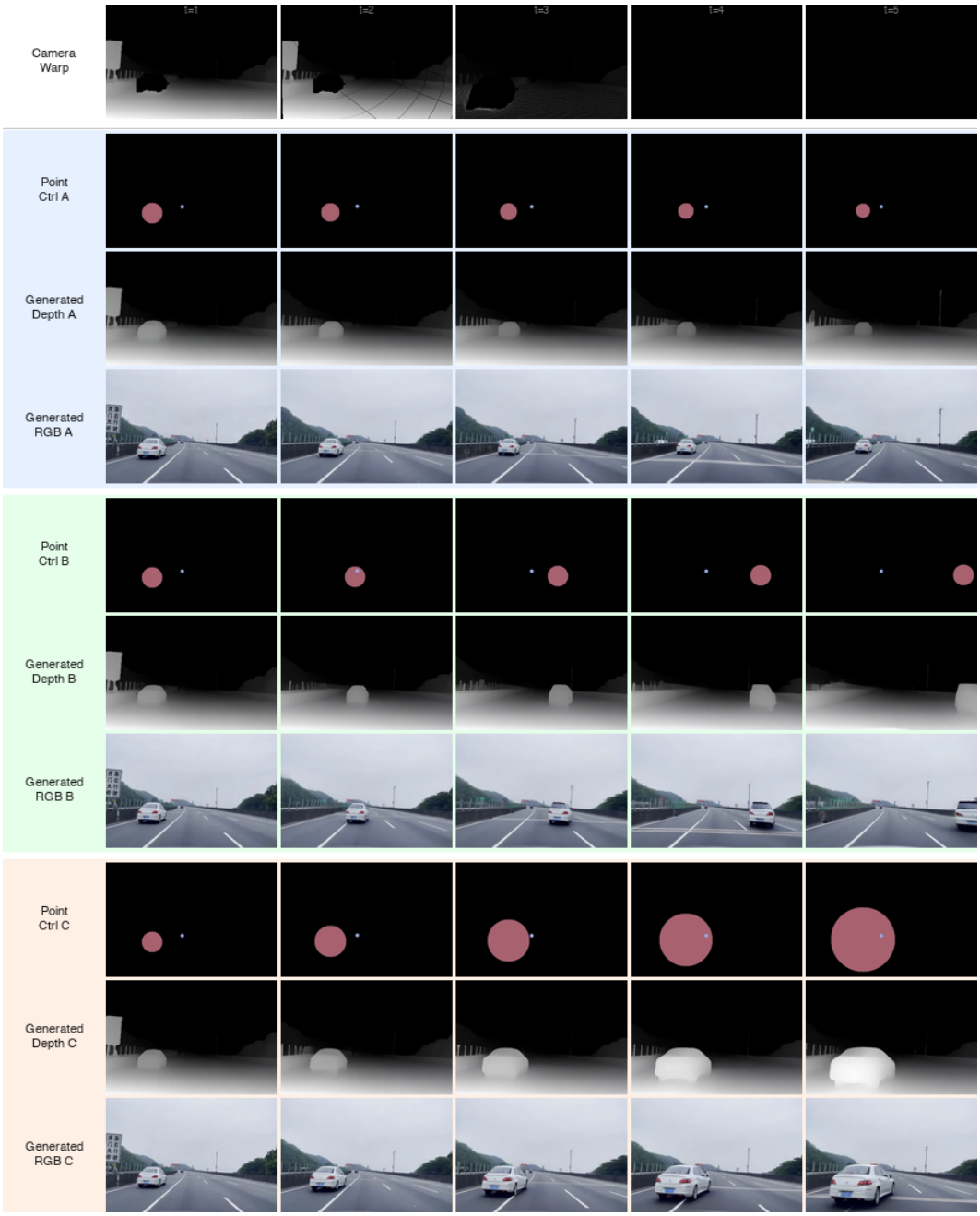


Figure 9: **Generation pipeline and point control robustness — Scene 3.** Same layout as Figure 7. Across all three scenes the depth-first hierarchy cleanly separates physical reasoning from visual rendering, enabling robust control with diverse sparse inputs.

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 introduction state the paper's intended contributions: the Point-Shape-Appearance hierarchy, Masked Point Recovery, and evaluations on autonomous driving with generalization to physics and robotics.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Sec. 5 discusses limitations in dense non-motorized traffic and highly occluded multi-agent interactions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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 proposes a model and training/evaluation protocol, but does not claim formal theoretical theorems requiring proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

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: The manuscript describes the architecture, training setup, datasets, and evaluation metrics. Furthermore, we release our inference code and pre-trained model weights to facilitate reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the 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: [Yes]

   Justification: We provide open access to our pre-trained model weights and inference code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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: Sec. 4 and Appendix A.3 specify the training setup, optimizer, datasets, evaluation metrics, and Waymo test protocol used for the reported results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

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: Our core evaluation metrics (e.g., FVD, Physics-IQ) are corpus-level distribution distances computed over a fixed test set of 100 videos. Reporting variance or standard error bars via multiple training runs is computationally prohibitive and not standard practice in large-scale video generation research.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: Appendix A.3 reports the use of 8 NVIDIA A100 GPUs and a total training time of approximately one week.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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: The paper uses publicly available datasets and models under their respective licenses, and involves no human subjects or sensitive data collection.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

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 combined Limitations & Broader Impacts paragraph in Sec. 5 discusses positive uses (driving simulation, robotic evaluation, content creation) and potential misuse risks (misleading synthetic footage).

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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 released structured video generation model and data are primarily formatted for driving simulation research and pose no substantial high risk for general-purpose misuse or malicious deepfakes.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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: We properly credit and cite all major datasets (Waymo, Driving Dojo) and models (CogVideoX, MOFA-Video) used, and our release complies with their respective academic use licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: We document the new model framework and evaluation protocol in the paper, and release the corresponding inference code and pre-trained model weights.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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 does not describe crowdsourcing experiments or new human-subject studies.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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: The paper does not involve new human-subject studies or direct interaction with study participants.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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: The core method development described in the paper does not use LLMs as an important, original, or non-standard component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.