MOTION DREAMER: REALIZING PHYSICALLY COHERENT VIDEO GENERATION THROUGH SCENE-AWARE MOTION REASONING

Anonymous authors
Paper under double-blind review

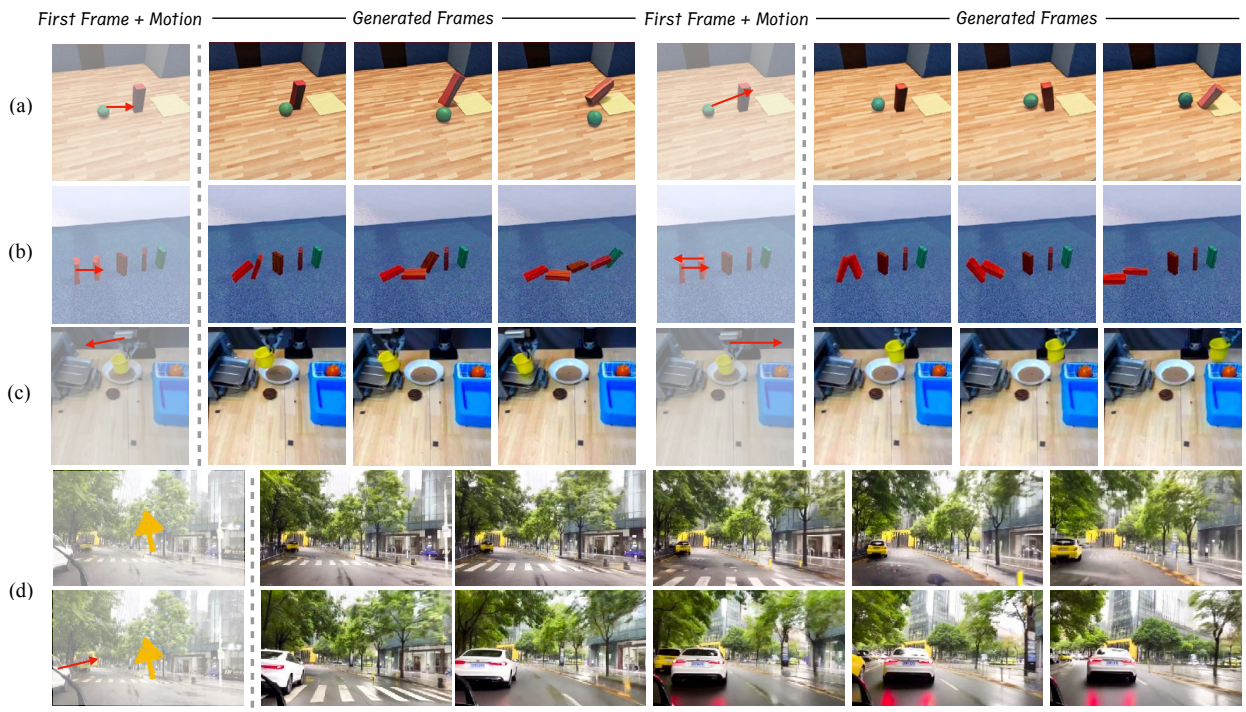

Figure 1: Our framework, **Motion Dreamer**, taking user input motion and the first frame, can successfully generate the motion-coherent future frames. (a) and (b) showcasing the different degrees of motion results in different object contact times and the momentum it carries. (c) demonstrates that providing different directional inputs allows the robotic hand to grasp an object and move it in the corresponding directions. (d) is the autonomous driving case, where the yellow arrow indicates the camera motion. Given the right arrow, the white car gradually leans towards the right.

ABSTRACT

Current video generation models often fail to produce logically and physically coherent future scenarios, a critical weakness for applications in autonomous driving and robotics. This stems from a fundamental conflict in end-to-end training: the pursuit of perceptual fidelity diverts capacity from modeling long-range temporal structure, while architectural priors fail to enforce physical laws. We introduce Motion Dreamer, a two-stage framework that resolves this conflict by explicitly decoupling motion reasoning from visual synthesis. Our approach is designed to generate complex scenes from an initial frame and sparse motion cues. To achieve this, we introduce instance flow, a novel sparse-to-dense motion representation, and a motion inpainting training strategy. Together, these techniques allow the model to robustly infer a complete, coherent motion field from partial inputs. This motion-aware representation then guides a synthesis model to generate high-fidelity video grounded in plausible dynamics. Across extensive experiments on robotics, physics, and a large-scale driving dataset, Motion Dreamer significantly outperforms leading methods in both motion coherence and visual realism.

1 INTRODUCTION

Video generation (Chen et al., 2024; Hong et al., 2022; Blattmann et al., 2023; Yang et al., 2024b; Yin et al., 2023) has emerged as a pivotal area in computer vision and artificial intelligence, impacting entertainment (Zhuang et al., 2024; Zeng et al., 2023; Hu et al., 2023b), virtual reality (Zhang et al., 2024b; Cai et al., 2024; Liu et al., 2024a), autonomous driving (Wang et al., 2023b; Zhao et al., 2024; Li et al., 2023; Hu et al., 2023a; Yang et al., 2024a), and robotics (Wu et al., 2024; Escontrela et al., 2023; Ko et al., 2023). Advances in deep learning have led to powerful video generation frameworks known as "world models" (Yang et al., 2024a; Wang et al., 2023b; Li et al., 2023; Hu et al., 2023a; Wang et al., 2024), which predict future scenarios based on current observations to enhance decision-making in dynamic settings. Despite notable progress, recent studies (Kang et al., 2024; Bansal et al., 2024; Meng et al., 2024) highlight ongoing challenges, particularly in maintaining logical and physical coherence. For instance, Kang *et al.* (Kang et al., 2024) showed that models frequently struggle with physical plausibility in out-of-distribution conditions, while Bansal *et al.* (Bansal et al., 2024) noted violations of fundamental physical laws, such as Newtonian mechanics, during interactions involving diverse materials.

These challenges raise a central question: *Why do current video generators struggle with logical coherence?* End-to-end models that map conditions to future frames face two issues: (i) **conflicting objectives**—pursuing perceptual fidelity often diverts capacity from long-range temporal structure; and (ii) **absent motion constraints**—architectural priors (e.g., 3D convolutions, temporal attention) model correlations rather than enforce physics. The result is motion that appears plausible at a glance yet violates inertia, contact causality, and object permanence.

To address these fundamental challenges, we propose a novel two-stage video generation framework called **Motion Dreamer**. In the first stage, we explicitly construct a motion-coherent intermediate representation from an initial frame and sparse motion cues, effectively generating and "inpainting" motion trajectories and spatial relationships into future frames. In the second stage, we condition a video synthesis model on this representation to generate visually detailed frames. By explicitly decoupling motion reasoning from visual synthesis, our approach alleviates the optimization conflicts inherent in end-to-end methods and incorporates explicit motion constraints, resulting in improved logical coherence and visual fidelity.

Controllable video generation still lacks control modalities that are both intuitive for users and effective as conditioning signals (Tulyakov et al., 2018; Niu et al., 2024). Inspired by MoFA-Video (Niu et al., 2024), we introduce **instance flow**, a sparse-to-dense, pixel-aligned motion cue: during training we compute per-instance average optical flow and propagate it within each segmentation mask; at inference users provide sparse arrows (as average-flow proxies) and obtain masks via SAM (Kirillov et al., 2023). To further strengthen reasoning from partial cues, we adopt **motion inpainting**: we randomly mask portions of the instance flow during training and require the model to reconstruct the dense field, which encourages inference of missing motion, improves generalization, and yields more temporally coherent videos from sparse inputs.

We validate our framework across multiple domains using **Jaco Play** (Dass et al., 2023), **Physion** (Bear et al., 2021), and a large-scale autonomous driving corpus curated from YouTube, comprising over 8,000 interactive driving clips (more than 200 hours). We conduct extensive experiments to assess performance. In side-by-side comparisons with state-of-the-art video editing models (e.g., MoFA-Video (Niu et al., 2024)) and a leading driving video generation model (VISTA (Gao et al., 2024)), our method yields more reasonable and coherent results in both qualitative and quantitative evaluations. These findings indicate that our framework not only produces high-quality videos but also substantially improves motion coherence and physical plausibility in complex scenarios.

In summary, our contributions are threefold:

- **Motion Dreamer: a two-stage framework.** We decouple *motion reasoning* from *visual synthesis* by first constructing a motion-coherent intermediate representation from a keyframe and sparse cues, then conditioning a generator on this representation. This separation mitigates objective conflicts in end-to-end training and strengthens temporal/interaction coherence.
- **Instance flow and motion inpainting.** We introduce *instance flow*, a sparse-to-dense, pixel-aligned control that aggregates per-instance average flow during training and is driven at inference by user arrows plus SAM masks. We further adopt *motion inpainting*—randomly masking instance flow and training the model to reconstruct the dense field—to improve reasoning from partial cues and generalization to sparse inputs.
- **Cross-domain validation.** Extensive experiments on PHYSION, JACOPLAY, and a large YouTube driving dataset (8k clips) show consistent gains over strong baselines (e.g., MoFA-Video, VISTA) in both qualitative and quantitative evaluations, improving motion coherence and physical plausibility while maintaining high visual quality.

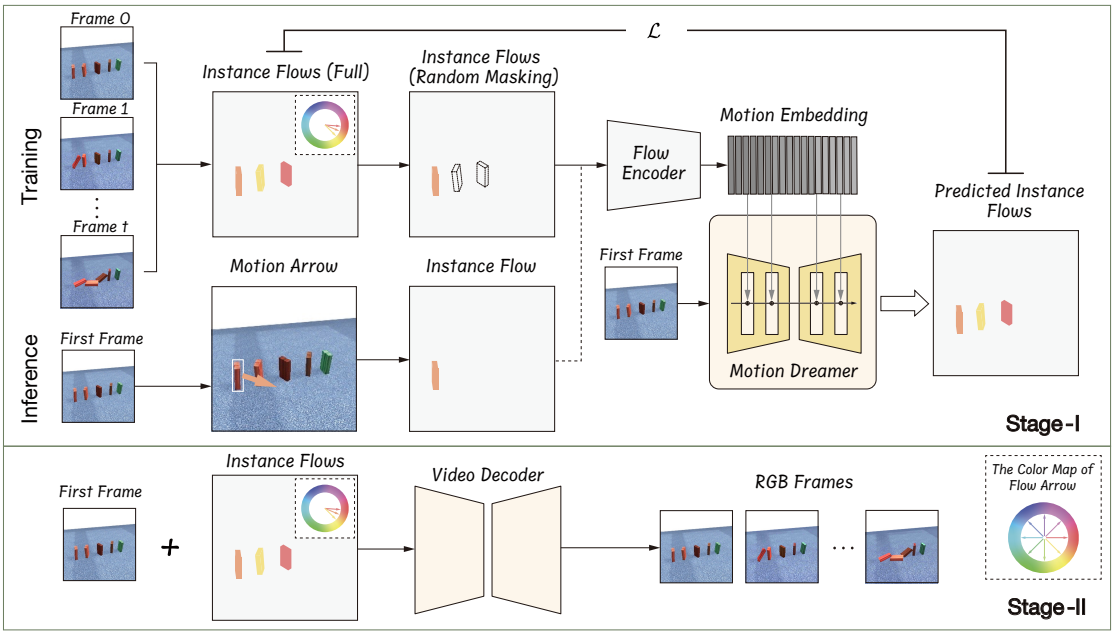

Figure 2: **Overview of the Motion Dreamer pipeline.** The "*Instance Flows (Full)*" shown in the figure incorporates our proposed instance flow along with several intermediate motion representations, such as segmentation maps. The symbol $\mathcal{L}$ denotes the loss function computed between the predicted instance flow and the ground-truth instance flow. The different color of the *Flow Arrow* indicates the different direction of the flow.

2 Related Work

2.1 Video Diffusion Models

Video generation models predominantly extended Unet-based latent diffusion models (LDMs) from text-to-image frameworks like Stable Diffusion (Rombach et al., 2022) to accommodate video applications. For instance, AnimateDiff (Guo et al., 2023) introduced a temporal attention module to enhance temporal consistency across frames. Building upon this, subsequent video generation models (Wang et al., 2023a; cerspense, 2023; Chen et al., 2023; 2024; Zhang et al., 2024a; 2023) adopted alternating attention mechanisms that combine 2D spatial attention with 1D temporal attention. Notable examples include ModelScope, VideoCrafter, Moonshot, and Show-1, which have demonstrated significant improvements in video generation tasks.

2.2 Controllable Motion Generation

Controllable motion generation in video synthesis aims to produce videos that not only exhibit high visual quality but also adhere to specified motion patterns and dynamics. (Yin et al., 2023; Zhu et al., 2023) allow users to specify motion trajectories directly, enabling fine-grained control over the path an object takes in the video. Others utilize keypoints or motion fields to guide animations, translating abstract motion representations into realistic movements (Cheng et al., 2023; Niu et al., 2024). By integrating these control mechanisms, these approaches aim to bridge the gap between user intent and the generated content.

Recently, a novel category of video generation models, referred to as "world models" (Yang et al., 2024a; Wang et al., 2023b; Zhao et al., 2024; Li et al., 2023; Hu et al., 2023a; Wang et al., 2024). Pioneering models like DriveDreamer (Wang et al., 2023b) and DriveDiffusion (Li et al., 2023) employ diffusion models conditioned on layout and ego-action to generate controllable driving videos. GAIA-I (Hu et al., 2023a) further expands this paradigm by incorporating multiple conditioning inputs, such as video, text, layouts, and actions, enabling the generation of realistic and diverse driving scenarios with fine-grained control over vehicle behavior and scene features. GenAD (Yang et al., 2024a) advances these efforts by scaling both the video prediction model and the dataset, thereby effectively managing complex driving scene dynamics and demonstrating superior zero-shot generalization across a range of unseen data.

Despite these advancements, we observe a common limitation persists: While they enhance controllability by introducing motion guidance, they do not fully resolve the issue of logical motion coherence, particularly in complex

scenes involving multiple interacting objects or scenarios requiring adherence to physical laws. For instance, while a trajectory might dictate where an object should move, it doesn't ensure that the movement adheres to principles like inertia or collision dynamics. Studies have shown that although models can generate visually appealing content, they frequently fail to maintain consistency with fundamental physical principles such as object permanence, conservation of mass, or Newtonian mechanics (Kang et al., 2024; Bansal et al., 2024). To address this limitation, we propose decoupling the tasks of motion reasoning and high-fidelity visual generation. By separating these two aspects, we simplify the generation process, making it more feasible to achieve both objectives.

3 METHOD

In this section, we present a comprehensive overview of our proposed two-stage framework, **Motion Dreamer**. The overall pipeline is depicted in Fig. 2. In Section 3.1, we define the problem and introduce the fundamental notations used throughout the framework. Section 3.2 introduces the concept of **instance flow**, a novel motion representation. Section 3.3 details the first stage, which generates intermediate motion representations given the input of the initial frame and motion prompts. Subsequently, Section 3.4 describes the second stage, which synthesizes the final RGB video based on these intermediate representations.

3.1 PRELIMINARIES

Let $\{\mathbf{I}_t\}_{t=0}^{T}$ denote a sequence of video frames, where t indexes the time steps, and T is the total number of frames. Our objective is twofold: first, to generate future intermediate motion representations $\{\mathbf{R}_t\}_{t=1}^{T}$ conditioned on the initial frame $\mathbf{I}_0$ and user-provided motion prompts; and second, to generate the future video frames $\{\mathbf{I}_t\}_{t=1}^{T}$ conditioned on the initial frame $\mathbf{I}_0$ and the generated multi-frame intermediate motion representations $\{\mathbf{R}_t\}_{t=1}^{T}$.

In this case, the intermediate motion representation $\mathbf{R}_t$ is composed of three key components: optical flow $\mathbf{O}_t \in \mathbb{R}^{2\times H\times W}$, instance segmentation map $\mathbf{S}_t \in \mathbb{R}^{H\times W}$, and depth map $\mathbf{D}_t \in \mathbb{R}^{1\times H\times W}$,

$$\mathbf{R}_t = (\mathbf{O}_t, \mathbf{S}_t, \mathbf{D}_t)\,.$$

The choice of these intermediate motion representations was deliberate, this unified representation $\mathbf{R}$ offers a comprehensive description of scene dynamics by combining optical flow, which captures the overall motion of the scene and camera movements; the instance segmentation map, which isolates and represents movements of individual objects; and the depth map, which encodes spatial relationships and distances between objects and the camera. By integrating these components, $\mathbf{R}$ facilitates a detailed understanding of both object-specific movements and their spatial interactions within the scene.

The effectiveness of this unified intermediate motion representation and the two-stage generation process is further demonstrated through experiments presented in subsequent sections, showcasing their ability to capture and reproduce complex scene dynamics accurately.

3.2 INSTANCE FLOW

To more effectively represent object motion, we propose a novel sparse-to-dense motion modality termed *instance flow*, which is well-suited for both human-in-the-loop input and video generation tasks. In the following, we provide a detailed description of the instance flow computation during both training and inference.

Training. During training, we assume that all intermediate motion representations are available within the dataset. Specifically, given optical flow fields $\{\mathbf{O}_t\}_{t=0}^{T-1}$ and the instance segmentation map $\mathbf{S}_0$, we calculate the instance flow for each instance $i \in \mathcal{I}$. The instance mask for instance i is defined as follows:

$$\mathbf{M}^{(i)}(x,y) = \begin{cases} 1, & \text{if } \mathbf{S}_0(x,y) = i, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The instance flow $\mathbf{F}^{(i)} \in \mathbb{R}^{2\times H\times W}$ for instance i is obtained by averaging the optical flow vectors within the instance mask over the temporal window T:

$$\mathbf{F}^{(i)} = \frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbf{M}^{(i)} \odot \mathbf{O}_t\right), \tag{2}$$

where $\odot$ denotes element-wise multiplication, with the instance mask $\mathbf{M}^{(i)}$ broadcasted over the flow dimensions.

The overall instance flow field $\mathbf{F} \in \mathbb{R}^{2\times H\times W}$ is then constructed by aggregating the instance flows for all instances:

$$\mathbf{F}(x,y) = \sum_{i\in\mathcal{I}} \mathbf{M}^{(i)}(x,y) \cdot \mathbf{F}^{(i)}(x,y). \tag{3}$$

Inference. During inference, users can provide sparse motion cues, such as arrows representing the desired average motion vectors for specific objects. Instance segmentation masks can be generated using advanced models like the Segment Anything Model (SAM) (Kirillov et al., 2023). These inputs are combined to generate a sparse instance flow $\mathbf{F}_{\text{user}}$:

$$\mathbf{F}_{\text{user}}(x,y) = \sum_{i\in\mathcal{I}_{\text{user}}} \mathbf{M}^{(i)}(x,y) \cdot \mathbf{v}^{(i)} \cdot \delta, \tag{4}$$

where $\mathcal{I}_{\text{user}}$ denotes the set of instances for which the user provides motion cues, $\mathbf{v}^{(i)}$ is the user-specified motion vector (e.g., an arrow) for instance i, δ is a scaling factor, and $\mathbf{M}^{(i)}$ is the instance mask for object i.

This formulation enables effective user guidance for instance-specific motion, offering an intuitive framework for both video generation and downstream tasks.

3.3 Stage I: Reasoning Motion Generation

In Stage I, Motion Dreamer, we employ a diffusion-based video generation model built upon the pre-trained CogVideoX model (Yang et al., 2024b). To emphasize low-frequency motion representations and improve temporal coherence, we adopted the $\mathbf{x}_0$ prediction parameterization.

As discussed in many diffusion approaches (Ho et al., 2020; Rombach et al., 2022), the training objective can be written as

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon},t}\left[\left\|\mathbf{x}_0 - \hat{\mathbf{x}}_0\left(\mathbf{x}_t, t, \mathbf{c}\right)\right\|^2\right], \tag{5}$$

where $\hat{\mathbf{x}}_0\left(\mathbf{x}_t, t, \mathbf{c}\right)$ is the model's prediction of the original intermediate motion representation $\mathbf{x}_0$, composed of the optical flow $\{\mathbf{O}_t\}_{t=0}^{T}$, instance segmentation map $\mathbf{S}_t$, and depth map $\mathbf{D}_t$, conditioned on the noisy input $\mathbf{x}_t$, timestep t, and conditioning information $\mathbf{c}$ (which includes the instance flow $\mathbf{F}$).

Conditions Incorporation. To effectively integrate the instance flow into the model, we prepare multi-scale versions of the instance flow $\mathbf{F}$ to align with the feature maps at different scales within the network. Specifically, for each scale $s \in \{8, 16, 32, 64\}$ used in the network, we resize the instance flow to match the spatial dimensions of the corresponding feature maps:

$$\mathbf{F}^{(s)} = 1/s \cdot \text{Resize}\left(\mathbf{F}, H/s, W/s\right), \tag{6}$$

where the division by s scales the flow vectors appropriately for the new resolution.

Next, we warp the feature maps $\mathbf{C}^{(s)}$ of the first-frame conditions (RGB image, instance segmentation, and depth) at each scale using the scaled instance flows. Following the approaches of DragNUWA (Yin et al., 2023) and MOFA-Video (Niu et al., 2024), we apply the Softmax Splatting function (Niklaus & Liu, 2020):

$$\mathbf{W}^{(s)} = \text{Softsplat}\left(\mathbf{C}^{(s)}, \mathbf{F}^{(s)}\right), \tag{7}$$

which warps the feature map according to the flow field by performing a differentiable splatting operation that aggregates input features onto the output grid based on the flow vectors. The Softmax Splatting function effectively distributes the features of $\mathbf{C}^{(s)}$ to new positions dictated by $\mathbf{F}^{(s)}$, allowing for seamless integration of motion information while maintaining differentiability for end-to-end training.

These warped features $\mathbf{W}^{(s)}$ are then integrated into the network by adding them to the corresponding feature maps at each scale during the encoding process:

$$\mathbf{X}^{(s)} = \mathbf{X}^{(s)} + \mathbf{W}^{(s)}. \tag{8}$$

This fusion strategy allows the network to incorporate explicit motion cues at multiple scales, enhancing its ability to generate motion-coherent intermediate representations and improving temporal consistency across frames.

Motion Inpainting. To enhance the model's capability for reasoning-based motion generation, we introduce a simple yet effective training strategy where we randomly mask out portions of the instance flow and require the model to reconstruct the same dense motion representation. Specifically, we apply a random mask to the instance flow field $\mathbf{F}$ to create a partially observed flow $\tilde{\mathbf{F}}$ in training:

$$\tilde{\mathbf{F}} = \mathbf{F} \odot \mathbf{M}^{(p)}_{\text{mask}}, \tag{9}$$

where $\mathbf{M}_{\text{mask}} \in \{0,1\}^{2\times H\times W}$ is a binary mask with each element independently set to zero with probability p (the masking ratio) and one otherwise, and $\odot$ denotes element-wise multiplication.

The model is trained to reconstruct the target intermediate motion representation $\mathbf{x}_0$ using the masked instance flow $\tilde{\mathbf{F}}$, minimizing the loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon},t}\left[\left\|\mathbf{x}_0 - \hat{\mathbf{x}}_0\left(\mathbf{x}_t, t, \tilde{\mathbf{F}}, \mathbf{c}\right)\right\|^2\right], \tag{10}$$

where $\hat{\mathbf{x}}_0\left(\mathbf{x}_t, t, \tilde{\mathbf{F}}, \mathbf{c}\right)$ is the model's prediction of the original intermediate motion representation $\mathbf{x}_0$, conditioned on the noisy input $\mathbf{x}_t$, timestep t, the masked instance flow $\tilde{\mathbf{F}}$, and other conditioning inputs $\mathbf{c}$ (e.g., textual descriptions or additional modalities).

By training with incomplete motion information, the model learns to infer missing motion cues and reason about object interactions, thereby improving its generalization and reasoning abilities. This approach enables the model to predict plausible motion trajectories even when provided with sparse inputs.

Motion Enhancement Loss. To further improve the consistency between the generated results and the initial instance flow, we propose a motion enhancement loss. Given the (possibly masked) instance flow $\tilde{\mathbf{F}}$, we define a motion mask $\mathbf{M}'$ that preserves regions with significant motion. Specifically, we compute the magnitude of each flow vector $\|\tilde{\mathbf{F}}_{ij}\|$, and set a motion threshold τ to define:

$$\mathbf{M}'^{(\tau)}_{ij} = \begin{cases} 1, & \text{if } \|\tilde{\mathbf{F}}_{ij}\| > \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

The motion enhancement loss $\mathcal{L}_{\text{motion}}$ is then defined as:

$$\mathcal{L}^{(\tau)}_{\text{motion}} = \mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon},t}\left[\left\|\left(\mathbf{x}_0 - \hat{\mathbf{x}}_0\left(\mathbf{x}_t, t, \mathbf{c}\right)\right) \odot \mathbf{M}'^{(\tau)}\right\|^2\right], \tag{12}$$

where $\odot$ denotes element-wise multiplication.

This loss term enforces the model to produce outputs that maintain spatial and temporal alignment with the high-movement regions specified by the instance flow and segmentation maps, thereby enhancing the fidelity and coherence of the generated motions.

3.4 Stage II: Video Decoder

In Stage II, Video Decoder, we transform the intermediate motion representations generated by the Motion Dreamer into high-quality RGB video frames. Building upon the pre-trained CogVideoX model (Yang et al., 2024b), the Video Decoder performs conditioned image-to-video generation by leveraging the initial frame and the generated intermediate motion representations.

Formally, let $\mathbf{x}_0$ denote the intermediate motion representation produced by the Motion Dreamer, which includes the optical flow, instance segmentation map, and depth map. The Video Decoder synthesizes the final RGB video $\{\mathbf{I}_t\}_{t=1}^T$ by conditioning on $\mathbf{x}_0$ and the initial frame $\mathbf{I}_0$:

$$\{\mathbf{I}_t\}_{t=1}^T = \text{Decoder}(\mathbf{I}_0, \mathbf{x}_0). \tag{13}$$

This two-stage approach decouples motion reasoning from high-fidelity video synthesis, simplifying the generation process and making it more feasible to achieve both rich visual details and coherent motion. The integration of the pre-trained CogVideoX model facilitates efficient and effective video generation, leveraging existing powerful image-to-video capabilities while enhancing them with our novel motion reasoning framework.

4 Experiments

Implementation Details. Both our Motion Dreamer and Video Decoder models are based on the CogVideoX-5b-I2V (Yang et al., 2024b) architecture, with additional fine-tuning implemented through LoRA. In the Motion Dreamer, we concatenate the RGB frame of the initial frame and its intermediate motion representation along the channel dimension. The instance flow is incorporated via softmax-splatting, producing the completed instance flow and intermediate motion representations for subsequent frames. In the Video Decoder, we concatenate the initial RGB frame with the intermediate motion representations of future frames to generate subsequent RGB frames.

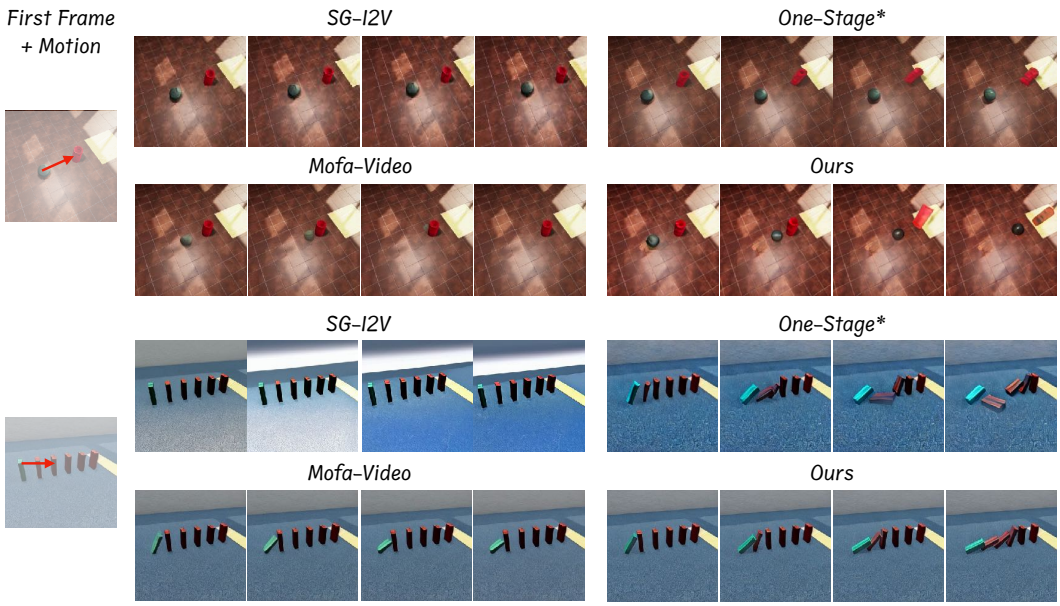

Figure 3: Comparisons with state-of-the-art video editing approaches on the Physion (Bear et al., 2021) dataset. One-stage* refers to the simplified one-stage version of our method, where the model directly generates RGB video results from the input without intermediate motion representations. For a fair comparison, Mofa-Video is also finetuned under the same dataset. Our model demonstrates the ability to generate physically coherent results.

Table 1: Quantitative results on the Physion dataset. One-stage* refers to the single-stage version of Motion Dreamer, which does not generate intermediate motion representations. For a fair comparison, Mofa-Video is also finetuned under the same dataset.

Methods	SG-I2V	DragAnything	One-stage*	Mofa-Video	**Ours-SVD**	**Ours-Cogvideo**
FVD↓	272.6	291.3	173.0	170.1	**166.3**	**157.8**
FVMD↓	424.7	398.0	224.3	226.2	**210.7**	**205.2**
Physics-IQ↑	21.3	18.7	28.4	29.1	**31.6**	**33.2**

Evaluation Metrics. To assess the performance of Motion Dreamer in generating physically coherent videos, we utilize 100 samples from the test set of Physion (Bear et al., 2021) and calculate two widely recognized metrics: (1) Frechet Video Distance (FVD) (Unterthiner et al., 2019), measures the statistical similarity between generated and real video distributions. (2) Frechet Video Motion Distance (FVMD) (Liu et al., 2024b), evaluates motion coherence by comparing the temporal consistency of generated frames with that of ground-truth videos. (3) Physics-IQ (Motamed et al., 2025), assesses the physical plausibility of generated videos by testing their adherence to fundamental principles such as gravity, object permanence, and collision dynamics.

For real-world driving video generation, we use 100 test videos from our collected driving dataset and compute the same two metrics. These evaluations provide a comprehensive analysis of our model's ability to generate high-quality, temporally consistent, and physically plausible videos across diverse scenarios.

4.1 Comparions with the State-of-the-art Methods

Physical Coherent Video Generation. We evaluate the proposed Motion Dreamer on the Physion (Bear et al., 2021) test set, shown in Figure 3, comparing it with state-of-the-art video editing models, including SG-I2V (Cheng et al., 2023) and Mofa-Video (Niu et al., 2024), for generating physically coherent motion videos. Additionally, we implement a simplified "one-stage" version of Motion Dreamer, which shares the same input and output structure but omits the generation of intermediate motion representations. This ablation study allows us to validate the effectiveness and superiority of our two-stage pipeline.

Table 1 presents the quantitative results for physical video generation on the Physion dataset. The one-stage version of Motion Dreamer (One-stage*) exhibits inferior performance compared to the full two-stage pipeline, highlighting

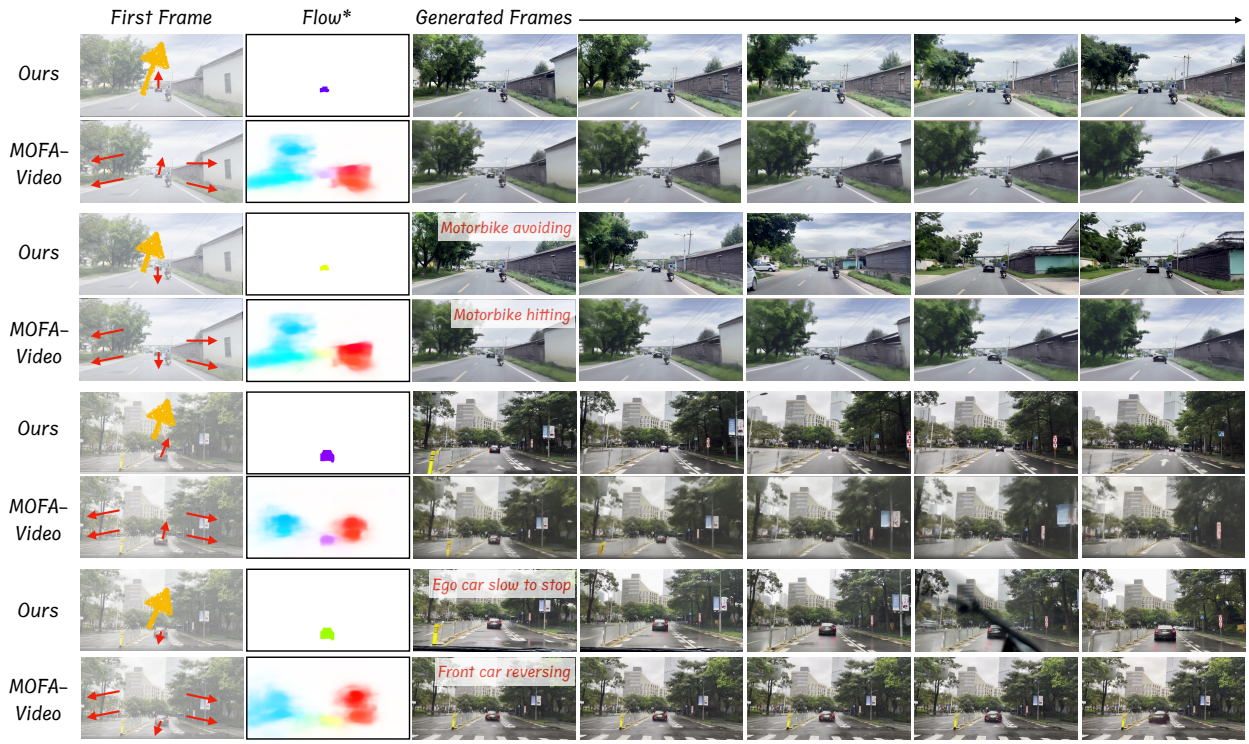

Figure 4: Illustration of reasoning-based motion generation in a driving scenario. We control the forward and backward movements of the lead car in various cases. Compared with MOFA-Video (Niu et al., 2024), our model produces more realistic and reasonable outcomes. Flow* in MOFA-Video represents the optical flow generated by the sparse-to-dense module, while in our work it denotes Instance Flow. **(Zoom in for optimal viewing)**.

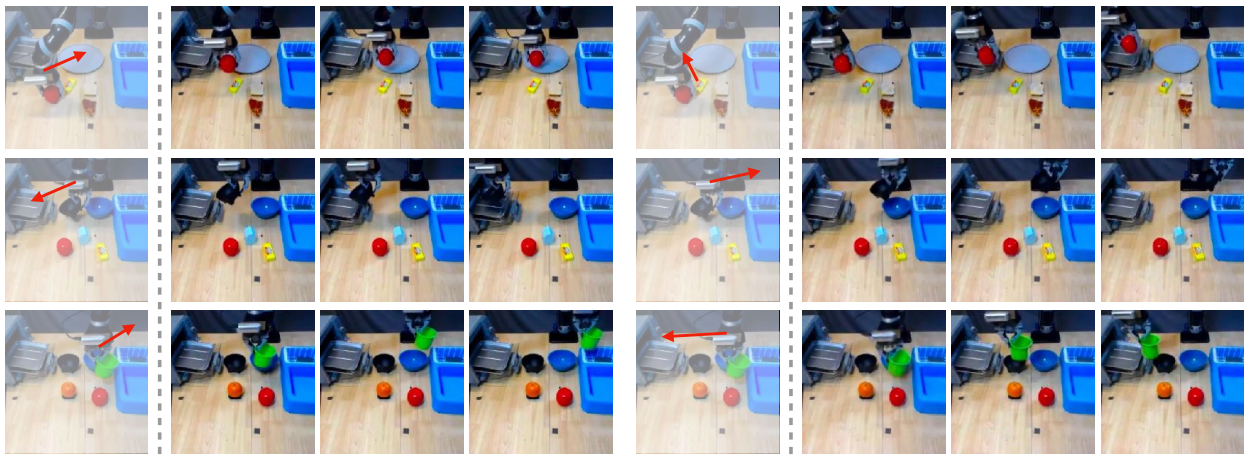

Figure 5: Visual comparison of our robotic hand manipulation on the Jaco Play dataset. In each row, we apply two different directional inputs to the same initial scene, resulting in distinct motions shown in the left and right columns.

the importance of generating intermediate motion representations. Furthermore, our method outperforms both Mofa-Video and the one-stage model in terms of FVD, FVMD, and Physics-IQ, demonstrating its ability to produce more physically plausible and temporally coherent results.

Real-world driving video generation. In Table 2, we evaluate the proposed Motion Dreamer on the test set of our collected dataset, comparing it with the state-of-the-art driving generation model, Vista (Gao et al., 2024) and state-of-the-art video editing model, Mofa-Video (fine-tuned on our data) (Niu et al., 2024), for generating real-world driving videos. The visual comparisons can be found in Figure 4.

Table 2: Quantitative comparison with Mofa-Video (fine-tuned on our data) and Vista on unconditional image-to-video generation for driving scenarios. Note that our method is conditioned only on camera motion, while Mofa-Video uses discrete background optical flow points (obtained via regional filtering) as its camera motion representation.

Methods	Mofa-Video	Vista	**Ours-SVD**	**Ours-Cogvideo**
FVD↓	309.7	285.8	**276.9**	**272.2**
FVMD↓	7176	3557	**3212**	**2913**
Physics-IQ↑	15.2	21.8	**22.4**	**24.1**

Table 3: **Ablation Studies.** We validate the effectiveness of each intermediate component, as well as the two functional parts. IMR* is the intermediate motion representation, where † indicates replacing the instance flow with sparse optical flow, the control signal in Mofa-Video.

Methods	**FVD**↓	**FVMD**↓	**Physics-IQ**↑
Motion Dreamer	157.8	205.2	33.2
IMR*			
w/o segmentation map	243.7	259.0	18.4
w/o depth map	167.0	228.8	29.7
w/o optical flow	173.0	224.3	28.4
Functional Part			
w/o motion enhancement loss	165.1	210.1	31.8
sparse optical flow†	169.1	217.5	30.2

Robotic hand manipulation. We conduct experiments on the Jaco Play (Dass et al., 2023) dataset to evaluate our model's ability to control a robotic hand. By providing a directional input, the model can generate motions to guide the robotic hand to a target for grasping, or to move an already-grasped object in the specified direction. The visual results for these manipulation tasks are presented in Figure 5.

4.2 Ablation Study

We conduct an ablation study to systematically evaluate the contribution of each component and functional part within our framework, as detailed in Table 3. Specifically, we assess how the removal or modification of intermediate motion representations (IMR) and functional parts—such as excluding the segmentation map or omitting the motion enhancement loss—impacts overall performance. Additionally, we examine the effect of substituting the control signal with sparse optical flow (as implemented in Mofa-Video) in place of our instance flow. This study aims to elucidate the individual and collective roles of these components in enhancing reasoning motion generation and robustness.

5 Limitation

Despite our model demonstrating strong performance in reasoning motion generation, it still exhibits limitations on some more complex cases, such as a ball strikes a tower composed of multiple blocks, causing the tower to collapse, as well as in driving scenarios with a high density of non-motorized vehicles. These issues will be discussed in detail in the appendix.

6 Conclusion

In this paper, we introduced Motion Dreamer, a two-stage framework that addresses the critical challenge of generating physically and logically coherent videos. By decoupling motion reasoning from visual synthesis, our method successfully mitigates the conflicting objectives inherent in end-to-end models. The introduction of instance flow provides an intuitive way for users to guide motion, while our motion inpainting strategy enables the model to reason about complex scene dynamics from only sparse inputs. Our extensive experiments across diverse domains—including robotics, physics simulations, and real-world autonomous driving scenarios—demonstrate that Motion Dreamer significantly outperforms existing state-of-the-art methods. The results show marked improvements in motion plausibility, temporal consistency, and overall visual quality. This work represents a significant step forward in creating controllable and realistic video generation models, opening up new possibilities for applications in autonomous systems and robotics.

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, 2024. URL https://arxiv.org/abs/2406.03520.

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. *arXiv preprint arXiv:2106.08261*, 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.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yuxin Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11618–11628, 2019. URL https://api.semanticscholar.org/CorpusID:85517967.

Shengqu Cai, Duygu Ceylan, Matheus Gadelha, Chun-Hao Huang, Tuanfeng Wang, and Gordon. Wetzstein. Generative rendering: Controllable 4d-guided video generation with 2d diffusion models. In *CVPR*, 2024.

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

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, 2024.

Hao Cheng, Zherui Wang, Rui Xu, and Jiwen Lu. SG-I2V: Self-guided trajectory control in image-to-video generation. *arXiv preprint arXiv:2307.13719*, 2023.

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.

Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee, Danijar Hafner, and Pieter Abbeel. Video prediction models as rewards for reinforcement learning. *arXiv preprint arXiv:2305.14343*, 2023.

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. *arXiv preprint arXiv:2405.17398*, 2024.

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. *arXiv preprint arXiv:2307.04725*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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*, 2023a.

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023b.

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, 2024. URL https://arxiv.org/abs/2411.02385.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to Act from Actionless Videos through Dense Correspondences. *arXiv:2310.08576*, 2023.

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.

Fangfu Liu, Hanyang Wang, Shunyu Yao, Shengjun Zhang, Jie Zhou, and Yueqi Duan. Physics3d: Learning physical properties of 3d gaussians via video diffusion. *arXiv preprint arXiv:2406.04338*, 2024a.

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

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? *arXiv preprint arXiv:2501.09038*, 2025.

Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

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.

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*, pp. 10684–10695, 2022.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1526–1535, 2018.

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, 2019. URL https://arxiv.org/abs/1812.01717.

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*, 2024.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, 2024.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13941–13958, 2023.

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*, pp. 14662–14672, 2024a.

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.

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. *arxiv*, 2024b.

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.

Jun-Yan Zhu, Jiapeng Wu, Yuxuan Shi, Tianyang Zhou, Dinghuang Yang, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. DragAnything: Interactive point-based manipulation on the generative image manifold. *arXiv preprint arXiv:2306.14435*, 2023.

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 Writing Assistance (LLM Use Disclosure)

We utilized a large language model (LLM), specifically Gemini, as a writing assistant to enhance the quality of this manuscript. The tool's role was strictly limited to improving language, including grammar, clarity, and academic tone. All scientific content—including the core ideas, methodology, analyses, and experimental results—was generated exclusively by the authors. We have carefully reviewed every modification suggested by the LLM to ensure it aligns with our original intent and maintains factual accuracy. The authors retain full responsibility for the final content of this paper.

B Implementation Details

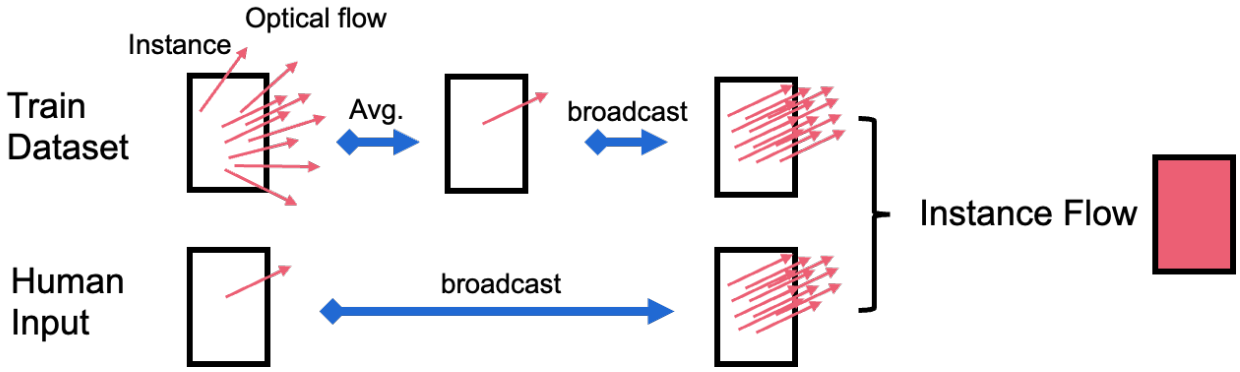

Figure 6: Schematic overview of the training and inference pipeline for instance flow extraction.

B.1 Instance Flow

The core of our work centers around the acquisition of the instance flow, as illustrated in Fig. 6. Specifically, during training, we first segment all objects throughout the video using the YOLO-v11 (Khanam & Hussain, 2024) instance segmentation model (in the Physion dataset, segmentation masks are provided directly). Subsequently, optical flow for the entire video is computed using the UniMatch (Xu et al., 2023) model. Instance flow is then defined as the mean optical flow within each object's mask. For better model interpretability, we broadcast the calculated instance flow uniformly across each respective object's mask.

During inference, we similarly begin by obtaining segmentation masks for all objects within an image. At this stage, humans only need to provide an arrow input, which we interpret as the averaged optical flow. This flow is then broadcast onto the object at the arrow's origin to align with the instance flow generated during training. It is important to note that we randomly mask out portions of the instance flow during training, with the goal of encouraging the model to learn causal reasoning about object motion.

B.2 Model Implementations

We evaluate the performance of the proposed Motion Dreamer on two datasets: Physion (Bear et al., 2021) and a large-scale driving dataset collected from YouTube. The driving dataset consists of over 9,000 clips of interactive driving scenarios, totaling more than 200 hours of video footage. Representative examples from this dataset are shown in Figure 7. The dataset will be made publicly available. The model leverages a two-stage training approach: Stage I employs x_0-parametrization, focusing on low-frequency motion representations to capture global dynamics. Stage II adopts v-parametrization, which enhances the generation of fine-grained video details. Each stage is trained on 8 NVIDIA A800 GPUs for approximately one week, ensuring robust convergence and high-quality results.

B.3 Camera Motion

In this subsection, we detail the processing of *camera motion* and its incorporation into the model. We represent the camera motion for each sample in the batch as a vector $\mathbf{c} \in \mathbb{R}^{B\times 2}$, where B is the batch size. The camera motion

Figure 7: Driving data examples.

vectors are encoded using a motion encoder:

$$\mathbf{e}_{\text{cam}} = \text{MotionEncoder}(\mathbf{c}) \in \mathbb{R}^{B\times 1\times D},$$

where D is the dimensionality of the cross-attention features, this encoding captures the global camera movements affecting all frames in a sequence. MotionEncoder is a multi-layer perception (MLP) block, that projects these parameters into the cross-attention feature space.

The encoded camera motion $\mathbf{e}_{\text{cam}}$ is then added to the encoder hidden states $\mathbf{H}$, which are repeated across frames to match the temporal dimension:

$$\mathbf{H}' = \mathbf{H} + \mathbf{e}_{\text{cam}},$$

where $\mathbf{H}' \in \mathbb{R}^{(B\cdot L)\times 1\times D}$ and L is the number of frames per sample.

Temporal Attention Integration. We employ temporal attention layers to fuse the camera motion encoding into the model. The temporal attention mechanism captures dependencies across frames, allowing the network to consider temporal dynamics influenced by camera motion.

In each temporal attention layer, the camera motion encoding adjusts the attention weights, enhancing the model's ability to focus on relevant temporal features. This is achieved by incorporating e_{cam} into the query or key projections of the attention mechanism.

C Collected Highly Interactive Driving Data

Table 4: User study results on the Physion dataset.

Method	Video Quality (Mean ± Std)	Motion Plausibility (Mean ± Std)	Control Precision (Mean ± Std)	Physical Accuracy (Mean ± Std)	User Preference (%)
MotionDreamer	**4.3 ± 0.2**	**4.6 ± 0.3**	**4.4 ± 0.3**	**4.5 ± 0.2**	**72**
MOFA-Video	4.1 ± 0.3	3.8 ± 0.4	3.7 ± 0.4	3.6 ± 0.3	18
DragAnything	3.9 ± 0.4	3.4 ± 0.5	4.2 ± 0.3	3.1 ± 0.4	6
SG-I2V	3.7 ± 0.5	3.5 ± 0.4	3.3 ± 0.5	3.2 ± 0.5	4

Highly interactive driving data encompasses scenarios in which the movements of the ego vehicle or other vehicles significantly influence the behavior of surrounding vehicles and pedestrians. Existing publicly available driving datasets, such as nuScenes (Caesar et al., 2019), are limited in their representation of highly interactive driving scenarios. To address this limitation, we have meticulously curated a comprehensive subset of highly interactive driving data sourced from YouTube. This curated dataset comprises over 9,000 video clips, totaling nearly 200 hours of footage. It encompasses a wide range of weather conditions, diverse geographical and urban scenes, and, most critically, varying levels of vehicular and pedestrian interactions. An illustrative example of the collected data is presented in the supplemental material (supplementary/videos/Collected_highly-interactive_driving_data). The dataset will be made publicly available soon.

D Unconditional Image-to-Video Generation

In this section, we present an expanded set of visual results for unconditional image-to-video generation within driving scenarios, providing a comprehensive comparison with the Vista (Gao et al., 2024). These results are illustrated in the supplemental material (supplementary/videos/Driving_uncond_I2V), demonstrating the effectiveness of our approach in generating realistic and temporally coherent video sequences from single images. Our method exhibits enhanced visual fidelity and consistency across diverse driving environments compared to Vista.

E User Study

We conducted a comprehensive user study involving 50 participants to qualitatively evaluate the effectiveness of our proposed method compared to state-of-the-art approaches. Each participant was asked to assess 20 videos per method from the Physion dataset, scoring them on a 5-point Likert scale (1 = poor, 5 = excellent) across four evaluation dimensions: Video Quality, Motion Plausibility, Control Precision, and Physical Accuracy. As detailed in Table 4, MotionDreamer consistently outperforms existing methods across all metrics. Notably, our approach achieved the highest user preference rate of 72%, indicating a significant advantage in generating visually appealing, physically plausible, and accurately controllable video sequences compared to MOFA-Video, DragAnything, and SG-I2V. These results further validate the practical benefits and robustness of our proposed framework in boundary conditional motion reasoning scenarios.

F Limitations

Despite the effectiveness of our model in generating coherent motion sequences, it exhibits limitations when confronted with highly complex dynamic scenes involving intricate physical interactions or dense multi-agent environments. Failure cases are illustrated in the supplementary material (supplementary/videos/Failure_cases).

Specifically, in scenarios like the "tower" data from the Physion dataset—where a ball impacts a tower composed of multiple blocks causing it to collapse—our model struggles to accurately capture the resulting motion. The complex interactions among numerous blocks, involving simultaneous collisions, rotations, and translations, are difficult to predict without explicit physical modeling. As a result, the generated motion sequences lack the chaotic yet physically plausible behaviors observed in real tower collapses. This suggests that while the model handles simple object motions and interactions, it struggles to generalize to scenarios requiring detailed understanding of physics and object interdependencies.

In driving environments characterized by a high density of non-motorized vehicles, such as bicycles and pedestrians, the model's performance diminishes. The unpredictable and highly variable movements of these agents, along with frequent interactions and occlusions, pose significant challenges. Consequently, the model sometimes fails to produce realistic motion trajectories for all agents, leading to inconsistencies in crowded scenes.

These limitations highlight the difficulty of modeling complex multi-agent dynamics, where each agent's behavior is influenced by numerous factors, including other agents' actions and environmental constraints. The absence of explicit mechanisms to capture social interactions and collision avoidance behaviors contributes to the model's shortcomings in these scenarios.