CAUSALMOTION: STRUCTURED PHYSICAL REASONING AS KEYFRAME AND TRAJECTORY GUIDANCE FOR TRAINING-FREE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

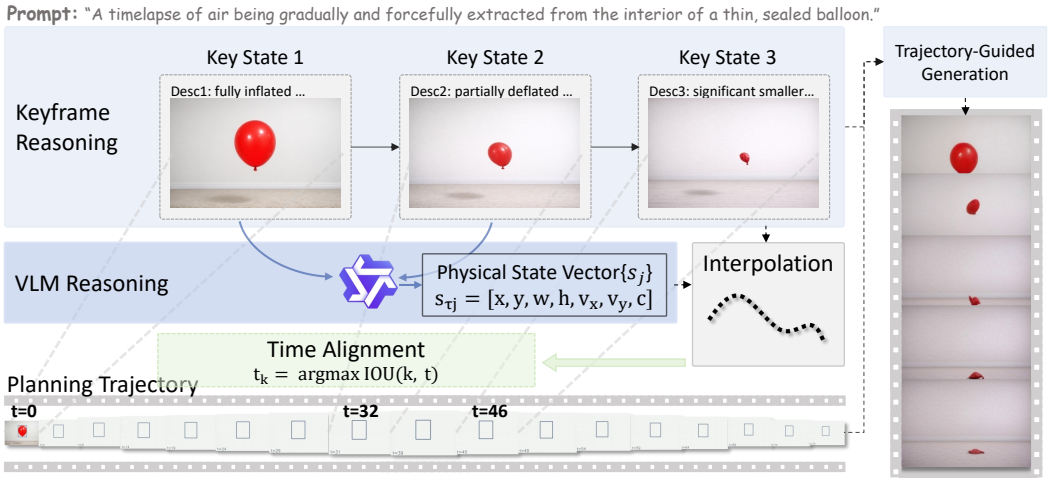


Figure 1: **Overview of CausalMotion.** Our training-free framework performs structured physical reasoning to generate keyframes and trajectories, then guides the diffusion model with these intermediate representations.

ABSTRACT

Recent advances in diffusion-based video generation have significantly improved visual quality and short-term temporal coherence. However, existing video diffusion models, while effective at modeling local structures and motion, often struggle with long-horizon physical inference, resulting in dynamics that lack physical consistency and causal plausibility. We therefore propose **CausalMotion**, a training-free framework that injects explicit physical reasoning into video generation through structured intermediate representations. Our key idea is to decouple reasoning from generation by leveraging a vision-language model to translate the text prompt into a sequence of causally consistent scene states represented by keyframes, together with object-centric motion trajectories. These representations provide complementary spatial and temporal guidance, which we align and integrate as soft constraints into a pretrained video diffusion model during inference. By explicitly planning object states and dynamics before generation, CausalMotion improves physical plausibility and temporal coherence without additional training or supervision. Extensive experiments demonstrate consistent improvements over existing baselines, particularly in dynamics-intensive scenarios, while maintaining high perceptual video quality.

1 INTRODUCTION

Video generation aims to synthesize temporally coherent and visually realistic dynamic sequences conditioned on inputs such as text prompts, reference images, or motion signals. Recent advances

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

in generative modeling, particularly diffusion-based approaches (MiniMax-AI, 2026; Yuan et al., 2026; Wan et al., 2025; Kong et al., 2024), have substantially improved visual fidelity and short-term temporal consistency. However, current video generation models still struggle to produce physically consistent and causally plausible state transitions, especially in compositional settings involving long-horizon dynamics.

A key limitation of existing approaches is their reliance on statistical correlations implicitly learned from large-scale video data, with limited explicit modeling of causal structure or physical constraints (Motamed et al., 2026). For instance, Runway Gen 3 (Runway Team, 2024) may generate a candle that appears and becomes lit as a burning match is lowered into water, producing visually plausible but physically inconsistent frames. Moreover, high-quality video datasets with rich physical interactions remain scarce and typically lack structured annotations of causality and physics (Kang et al., 2025). As a result, generated videos can exhibit several characteristic failure modes, including missing intermediate states, temporal discontinuities, and physically implausible interactions. For example, an ice cream cone may transition directly from a solid object to a puddle, skipping intermediate melting stages; an object may abruptly appear at a different location; or two objects may pass through each other during an interaction. In parallel, vision-language models (VLMs) have demonstrated strong capabilities in cross-modal understanding and structured reasoning (Hurst et al., 2024; Liu et al., 2025a; Bai et al., 2025a; Hong et al., 2025; Guo et al., 2025). With the ability to perform structured inference and chain-of-thought reasoning (Wei et al., 2022; Zhang et al., 2022; 2024; Kojima et al., 2022), these models can decompose complex events into a sequence of intermediate states and infer physically plausible state transitions. For instance, a VLM can reason that a water droplet on a sloped surface will move downward along the slope. However, VLMs lack the ability to generate temporally consistent visual content, limiting their direct applicability to video synthesis.

This gap motivates us to separate physical reasoning from visual synthesis and use the former to explicitly guide the latter without retraining the video generation model. We therefore propose **CausalMotion**, a training-free, reasoning-guided framework for physically plausible video generation. The key idea is to decouple reasoning from generation and use a VLM to structure the generation process via intermediate representations that capture both scene states and object motion. As illustrated in Figure 1, CausalMotion consists of three stages. First, a VLM translates the input prompt into causally consistent keyframes. Using object locations from the first keyframe and textual descriptions of all keyframes, the VLM predicts object-centric motion trajectories in the form of bounding-box sequences that capture object movement over time. The keyframes and trajectories are then temporally aligned and integrated as complementary spatial and temporal guidance for the latent updates of the diffusion model to synthesize the final video.

CausalMotion provides explicit guidance for causal state transitions and object dynamics without modifying the parameters of a pretrained diffusion model or requiring any supervision. This design preserves the flexibility of the underlying generator while improving physical plausibility and temporal coherence. On PhyGenBench, CausalMotion achieves an average score of 0.65, corresponding to a 67% relative improvement over the LTX-Video baseline. It also achieves a VBench quality score of 82.52%, compared with 80.57% for LTX-Video and 80.44% for Wan2.2.

We summarize our contributions as follows:

- We propose CausalMotion, a training-free framework that decouples physical reasoning from visual synthesis and incorporates VLM reasoning into video generation to improve causal consistency and physical plausibility.
- We introduce a structured intermediate representation consisting of keyframes and motion trajectories, which reduces ambiguity in temporal modeling and improves controllability.
- We develop an inference-time guidance mechanism that integrates keyframes and motion trajectories into the latent updates of a pretrained video diffusion model without additional training or annotations.

2 Related Work

Diffusion-based Video Generation. Diffusion-based video generation has become a dominant paradigm for high-quality video synthesis, with recent models (Yang et al., 2025b; HaCohen et al.,

2024; Kong et al., 2024; Wan et al., 2025; Ali et al., 2025; HaCohen et al., 2026; Yuan et al., 2026; MiniMax-AI, 2026) significantly improving visual fidelity, motion quality, and temporal consistency by learning spatiotemporal distributions from large-scale video data. However, existing video diffusion models primarily learn spatiotemporal correlations from data and do not explicitly model physical dynamics and causal state transitions. Recent benchmarks, including VideoPhy, PhyGenBench and Physics-IQ (Bansal et al., 2025; 2026; Meng et al., 2025; Motamed et al., 2026) reveal that strong perceptual quality does not necessarily translate into physically plausible behaviors. Our work addresses this gap by introducing explicit physical reasoning and structured inference-time guidance into pretrained video diffusion models.

Physics-Consistent Video Generation. Recent work has explored physical consistency in video generation through simulation, training-based adaptation, and inference-time reasoning. Simulator-based methods (Liu et al., 2024a; Tan et al., 2026; Chen et al., 2025; Li et al., 2025b; Liu et al., 2026; Foo et al., 2026) explicitly model object dynamics using physical simulators, providing strong physical grounding but requiring scene reconstruction, physical parameter estimation, or domain-specific simulation. Meanwhile, simulators rely on simplified assumptions and introduce substantial computational overhead, limiting scalability to open-world scenarios. Training-based approaches instead incorporate physical priors through post-training, reward optimization, or reinforcement learning, including PISA (Li et al., 2025a), Force Prompting (Gillman et al., 2026), NewtonRewards (Le et al., 2025), PhysRVG (Zhang et al., 2026), VideoREPA (Zhang et al., 2025b), ProPhy (Wang et al., 2026a), PHANTOM (Shen et al., 2026), and PhyCo (Narayanan et al., 2026). While effective, these approaches require additional training data, optimization, or carefully designed supervision. More recently, reasoning-guided approaches exploit vision-language models or video foundation models to provide physical priors during inference. Think Before You Diffuse (Zhang et al., 2025a) uses multimodal reasoning to infer physical context, while VLIPP (Yang et al., 2025a) predicts coarse motion trajectories for inference-time guidance. PhyRPR (Zhao et al., 2026) further separates physical reasoning, motion planning, and visual refinement in a training-free pipeline. CoECT (Wang et al., 2026b) further represents physical evolution as causally connected events with temporally aligned visual states, while DeforM (Li et al., 2026) uses VLM-guided reasoning to identify physics-critical regions and construct spatial-temporal guidance. These methods demonstrate the benefit of structured reasoning, but primarily focus on physical context, trajectories, causal events, or localized regions individually. CausalMotion instead combines causally consistent keyframes with object-centric motion trajectories, providing complementary spatial and temporal guidance without modifying the pretrained video generator.

Vision-Language Models for Reasoning. Vision-language models have increasingly been used to provide structured intermediate representations for visual generation and video planning. Early works employ language or vision-language models to generate layouts, semantic plans, or other structured representations for downstream generation (Lian et al., 2024; Wu et al., 2024; Pan et al., 2024). VideoDirectorGPT (Lin et al., 2023) uses an LLM to construct structured video plans containing scene descriptions, entity layouts, and consistency information. More recent approaches extend this paradigm to temporal and physical reasoning. VChain (Huang et al., 2026) generates sparse critical keyframes to represent future visual states and uses them to guide a pretrained video generator, while related methods explore structured event representations and motion planning for controllable video synthesis. These works demonstrate the effectiveness of multimodal reasoning for structured video generation, but primarily emphasize scene planning, keyframe reasoning, event-centric representations, or trajectory prediction individually. In contrast, CausalMotion jointly models causal scene states and object-centric motion, temporally aligns the two representations, and integrates them as soft spatial-temporal constraints during diffusion sampling.

3 METHOD

We propose CausalMotion, a training-free framework for physically plausible video generation that improves temporal coherence and physical plausibility without any additional training or fine-tuning. Our key idea is to decouple video generation into structured state reasoning and trajectory-guided synthesis. Instead of relying on implicit dynamics learned by diffusion models or depending on existing physics simulators, we explicitly represent video evolution as a sequence of causally con-

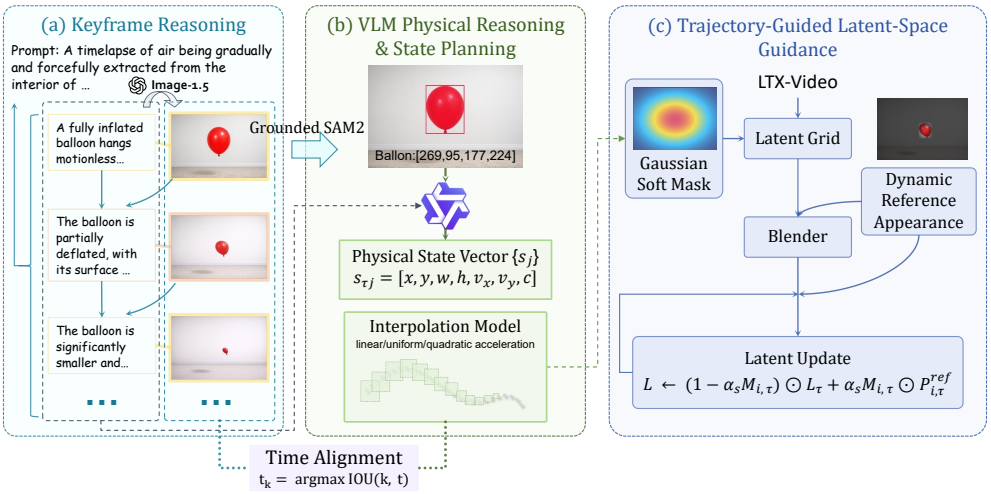


Figure 2: **The architecture of our method.** Given a text prompt, CausalMotion first performs iterative visual reasoning to decompose the described scenario into causally consistent key states and generates corresponding keyframes that capture critical transitions. It then localizes key objects, predicts sparse physical state vectors, and constructs dense motion trajectories through physics-aware interpolation, followed by temporal alignment with the keyframes. Finally, the trajectories are projected into the diffusion latent space, where keyframe-based appearance anchors and localized latent updates jointly guide denoising.

sistent states and object-centric motion trajectories, which are then used to guide the generation process.

Formally, we formulate video generation as a structured state evolution problem, where video evolution is guided by two representations: (i) scene states represented by keyframes, and (ii) object-centric motion trajectories describing the temporal evolution of objects. CausalMotion, whose architecture is overviewed in Figure 2, consists of three stages: (1) keyframe generation via iterative visual reasoning, (2) physical state and trajectory planning, and (3) trajectory-guided latent-space video generation.

3.1 Keyframe Generation via Visual Thought Reasoning

Text prompts often implicitly encode a sequence of state transitions, including object motion, interaction, and transformation. However, such structure is not explicitly represented in standard video generation pipelines, which can lead to missing intermediate states and inconsistent long-horizon dynamics.

To explicitly represent these transitions, we leverage a VLM to perform Visual Thought Reasoning, which decomposes the input prompt into a sequence of visual states. Given a prompt p, the VLM first performs a global scene understanding step and infers the *consequence*, i.e., the expected final outcome of the described scenario. This consequence serves as a global constraint guiding the generation of subsequent key states.

CausalMotion consists of three main stages, as illustrated in Fig. 2. Starting from the inferred initial state and consequence, the VLM generates a textual description txt_0 of the first key state and produces the corresponding image img_0 via a text-to-image model (GPT-image-1.5 in our experiments). At each subsequent step i, the VLM takes the previous keyframe img_{i-1}, its description txt_{i-1}, and the global consequence as input, and predicts the next textual instruction txt_i describing the next state transition. The corresponding keyframe img_i is then generated from txt_i. This autoregressive process continues until all inferred state transitions implied by *consequence* are covered.

The final output is a Chain of Visual Thoughts and Captions:

$$\{\text{img}_0, \text{img}_1, \ldots, \text{img}_N\} \text{ and } \{\text{txt}_0, \text{txt}_1, \ldots, \text{txt}_N\},$$

which provide structured visual states and semantic descriptions for subsequent physical state reasoning, trajectory planning, and video generation. A complete iterative procedure is summarized in Algorithm 1.

3.2 Physical State and Trajectory Planning

Given the generated keyframes and their descriptions, we model object dynamics as structured physical states and derive continuous motion trajectories. This stage explicitly models the spatial evolution of key objects between these states.

We first identify the objects involved in the described physical changes. The VLM reasons over the prompt, keyframe descriptions, and generated keyframes to determine the relevant objects and their physical interactions. We then localize these objects in the initial keyframe using a segmentation-based grounding model (Ravi et al., 2025; Liu et al., 2024b; Ren et al., 2024a;b; Kirillov et al., 2023; Jiang et al., 2024). For each object o_k, we obtain its initial bounding box:

$$b_{0,k} = S_{\text{seg}}(\text{img}_0, o_k), \quad k = 1, \ldots, K, \tag{1}$$

where S_{seg} denotes the grounding model and $b_{0,k}$ represents the initial spatial extent of object o_k. These initial locations serve as the geometric anchors for subsequent trajectory planning.

To enable explicit modeling of object dynamics, we define a physics-informed object state vector for each object i at key time τ_j:

$$\mathbf{s}^i_{\tau_j} = \left[x^i_{\tau_j},\ y^i_{\tau_j},\ w^i_{\tau_j},\ h^i_{\tau_j},\ v^i_{x,\tau_j},\ v^i_{y,\tau_j}\right], \tag{2}$$

where (x, y, w, h) denotes spatial extent, and (v_x, v_y) describes image-plane motion. This representation combines object geometry, motion, and interaction cues in a compact form.

The VLM then acts as a high-level physical state planner. Given the scene description, initial keyframe img_0, initial location $\{b_{0,k}\}$, and the generated keyframe sequence, it identifies the dominant physical factors and predicts sparse physics-informed object states $\mathbf{s}^i_{\tau_j}$ at selected timesteps τ_j. Specifically, the reasoning process consists of three steps: (1) identifying the dominant physical factors governing the scene, such as gravity or collision; (2) inferring the resulting interaction and motion trends; and (3) predicting the corresponding object positions and spatial extents in the image plane. This produces a sparse, physics-informed motion prior without requiring explicit scene reconstruction or a physics simulator.

As directly predicting dense frame-level motion is inefficient and prone to accumulated errors, the VLM predicts states only at a sparse set of key timesteps. In our experiments, we choose to use 13 key timesteps and interpolate them to the full $T = 121$ frame sequence. Specifically, the interpolation model is selected according to the inferred motion regime:

$$b^i_t = \begin{cases} \text{Interp}_{\text{linear}}\left(b^i_{\tau_j}, b^i_{\tau_{j+1}}\right), & \text{uniform or approximately uniform motion} \\ b^i_{\tau_j} + v^i_{\tau_j}\,\Delta t + \frac{1}{2}a^i(\Delta t)^2, & \text{accelerated motion} \end{cases} \tag{3}$$

where $\Delta t = t - \tau_j$. This resulting dense trajectories preserve the overall motion trend while providing smooth frame-level guidance for video generation.

3.3 Time Alignment

The generated keyframes and dense object trajectories provide complementary representations of the same scene evolution, but are defined at different temporal resolutions. We therefore establish an explicit temporal correspondence between them. Let t_k denote the assigned timestep for keyframe k. We fix the first keyframe at $t_0 = 0$ and greedily assign each subsequent keyframe to the trajectory timestep that maximizes the spatial overlap between the keyframe object locations and the corresponding trajectory bounding boxes, while preserving temporal ordering:

$$t_k = \underset{t \in (t_{k-1}, T-k)}{\arg\max}\ \text{IoU}(k, t), \quad k = 1, \ldots, N, \tag{4}$$

where $\mathrm{IoU}(k, t)$ measures the overlap between the object bounding boxes extracted from keyframe k and the corresponding trajectory bounding boxes at timestep t. This monotonic alignment establishes explicit temporal anchors between the keyframe-based appearance representation and the trajectory-based motion representation, which are subsequently integrated during latent-space guidance. Further analysis of the consistency between keyframe predictions and trajectory representations is provided in Appendix A.

3.4 Trajectory-Guided Latent-Space Guidance

We incorporate the planned object trajectories as soft spatial constraints during diffusion sampling. Rather than directly replacing the latent representation with trajectory-conditioned features, we perform localized latent updates at selected denoising steps, balancing the generative prior of the pretrained video diffusion model with the planned motion.

At each selected denoising step, we first unpatchify the video tokens into a frame-wise latent representation. For object i at latent timestep τ, the predicted trajectory provides a target bounding box $B_{i,\tau}$. We enlarge this box by a factor ρ_t and construct a soft spatial mask $M_{i,\tau}$ over the latent feature map. The enlarged region provides tolerance to trajectory prediction errors while the soft mask avoids imposing a hard spatial constraint.

$$P_{i,\tau}^{\mathrm{ref}} = \frac{w_{i,\tau}^{-} P_{i,\tau}^{-} + w_{i,\tau}^{p} P_{i,\tau}^{\mathrm{prev}} + w_{i,\tau}^{+} P_{i,\tau}^{+}}{w_{i,\tau}^{-} + w_{i,\tau}^{p} + w_{i,\tau}^{+}}, \tag{5}$$

where $P_{i,\tau}^{-}$ and $P_{i,\tau}^{+}$ are the appearance anchors from the neighboring keyframes, and $P_{i,\tau}^{\mathrm{prev}}$ is obtained by warping the preceding latent frame from the previous trajectory box to the current one.

We then update the latent representation only within the trajectory-guided region:

$$L_\tau \leftarrow (1 - \alpha_\tau M_{i,\tau}) \odot L_\tau + \alpha_\tau M_{i,\tau} \odot P_{i,\tau}^{\mathrm{ref}}, \tag{6}$$

where α_τ is the step-dependent guidance strength determined by the denoising schedule. The update is applied only at selected denoising steps, and the resulting latent is subsequently repatchified for continued diffusion sampling.

This design provides three complementary forms of guidance: the trajectory specifies *where* the object should move, the warped preceding latent provides local temporal appearance propagation, and neighboring keyframe anchors provide appearance references. The soft spatial update therefore encourages trajectory-consistent motion while preserving the flexibility of the pretrained video generator.

3.5 Implementation

We adopt LTX-Video (HaCohen et al., 2024) as the video synthesis backbone, which supports conditioning on multiple keyframes at user-specified temporal indices. To incorporate the physics-aware trajectory, we use the temporal anchors t_k obtained from Eq. 4 as the conditioning indices for the corresponding keyframe. Specifically, for each anchor t_k, we provide LTX-Video with corresponding keyframe img_k generated in Section 3.1.

For each keyframe condition, LTX-Video also expects a guidance strength λ. We use a fixed value of $\lambda = 0.9$ for all keyframes in our experiments. In addition to keyframe conditioning, we apply the trajectory-guided latent update described in Section 3.4 during diffusion sampling. The trajectory guidance is controlled by the step-dependent strength α_τ, which determines the magnitude of localized latent update at each denoising step. Thus, keyframe conditioning provides explicit visual anchors at selected temporal positions, while trajectory guidance encourages object motion to follow the planned trajectories between these anchors.

4 Experiments

4.1 Experiments Setup

For keyframe generation, we adopt GPT-image-1.5, a high-fidelity text-to-image model with a strong instruction-following capability. For trajectory reasoning, we employ Qwen-VL-2.5-72B (Bai et al.,

2025b) as the vision-language model to infer object states, motion trends, and physical interactions from the input prompt and generated keyframes. All experiments are conducted at a spatial resolution of 720×480 and a frame rate of 30 FPS. We follow the default configuration of LTX-Video and generate 121-frame videos, corresponding to approximately 4 seconds. Unless otherwise specified, we use the same backbone and generation configuration for LTX-Video and its ablated variants.

We evaluate CausalMotion from two complementary perspectives: physical plausibility and general video quality. We first use PhyGenBench (Meng et al., 2025) to evaluate physical reasoning and then use VBench (Huang et al., 2024) to assess perceptual video quality. We further conduct a VLM-based evaluation of higher-level properties, including physical plausibility, temporal consistency, and semantic alignment.

4.2 Evaluation

Quantitative comparisons We first quantitatively evaluate our method on PhyGenBench using the PhyGenEval protocol. As shown in Table 1, CausalMotion achieves an average score of 0.65, outperforming the strongest baseline, VLIPP (Yang et al., 2025a), by 0.05. CausalMotion achieves the highest scores in mechanics (0.61), thermal (0.68), and material (0.61), and matches VLIPP in optics (0.71). The largest gains over VLIPP occur in mechanics (+0.06) and thermal (+0.08), which involve object motion, interactions, and state transitions. These results demonstrate the benefit of explicitly incorporating structured scene-state reasoning and object-centric motion planning into the generation process.

Table 1: Quantitative comparison on PhyGenBench. We report scores across mechanics, optics, thermal, and material phenomena. Bold and underlined numbers denote the best and second-best results, respectively.

Method	Mech. ↑	Opt. ↑	Therm. ↑	Mat. ↑	Avg.
Text-to-Video					
CogVideoX-T2V-5B (Yang et al., 2025b)	0.43	0.55	0.40	0.42	0.45
LTX-Video-T2V (HaCohen et al., 2024)	0.35	0.45	0.36	0.38	0.39
OpenSora (Zheng et al., 2024)	0.43	0.50	0.44	0.37	0.44
Image-to-Video / Text-Image-to-Video					
CogVideoX-I2V-5B	0.48	0.69	0.43	0.41	0.52
SVD-XT (Blattmann et al., 2023)	0.46	0.68	0.48	0.41	0.52
LTX-Video-I2V	0.47	0.65	0.46	0.37	0.50
SG-I2V (Namekata et al., 2025)	0.52	0.69	0.51	0.39	0.54
Wan2.2-TI2V-5B (Wan et al., 2025)	0.47	0.68	0.48	0.44	0.54
Helios-Base (Yuan et al., 2026)	0.46	0.63	0.37	0.35	0.47
Physics- or Reasoning-Guided					
LLM-Grounding Video Diffusion (Lian et al., 2024)	0.32	0.41	0.26	0.24	0.31
PhyT2V (Xue et al., 2025)	0.49	<u>0.61</u>	0.49	0.47	0.52
VideoDPO (Liu et al., 2025b)	0.48	0.60	0.47	<u>0.58</u>	0.54
Diffphy (Zhang et al., 2025a)	0.53	0.59	0.58	0.46	0.54
PhyGDPO (Cai et al., 2026)	<u>0.55</u>	0.60	0.58	0.47	0.55
VLIPP (Yang et al., 2025a)	<u>0.55</u>	**0.71**	<u>0.60</u>	0.53	<u>0.60</u>
CausalMotion	**0.61**	**0.71**	**0.68**	**0.61**	**0.65**

We next evaluate perceptual quality and high-level consistency using VBench and VLM-as-judge. For a fair comparison with VChain (Huang et al., 2026), we use the same 20 prompts for VBench evaluation. As shown in Table 2, CausalMotion achieves a VBench Quality Score of 82.52%, slightly outperforming Wan2.2-TI2V-5B (80.44%), LTX-Video (80.57%), and Helios-Base (77.34%). This result indicates that introducing structured physical guidance does not compromise the perceptual quality of the generated videos.

Beyond standard metrics, we employ a vision-language model (Gemini 2.5 Flash (Comanici et al., 2025)) as a judge to evaluate higher-level properties, including *physical plausibility*, *temporal consistency*, and *semantic alignment*. Our method achieves a substantially higher overall score from

Table 2: Evaluation of perceptual quality and high-level consistency. VBench reports the quality score (%), while VLM-as-judge evaluates physical plausibility (Phys.), temporal consistency (Temp.), semantic alignment (Sem.), and their overall score.

Method	VBench ↑	Phys. ↑	Temp. ↑	Sem. ↑	Overall
Wan2.2-TI2V-5B	80.44	3.20	4.05	3.25	3.50
LTX-Video	80.57	2.53	3.89	1.68	2.70
Helios-Base	77.34	2.90	4.10	3.25	3.42
VChain (Huang et al., 2026)	78.49	–	–	–	–
Ablation variants					
-w/o keyframe reasoning	80.91	2.60	4.65	2.35	3.20
-w/o trajectory physical state	82.44	3.05	3.65	3.40	3.37
-w/o latent trajectory guidance	82.46	3.20	3.90	3.45	3.55
CausalMotion	**82.52**	**4.05**	**4.65**	**4.40**	**4.37**

2.70 to 4.37 compared with the baseline model of LTX-Video. In particular, the gain in temporal consistency (3.89 → 4.65) indicates that trajectory guidance effectively stabilizes motion across frames, while the large improvement in semantic alignment (1.68 → 4.40) suggests that keyframe reasoning helps better ground the generated content to the input prompt.

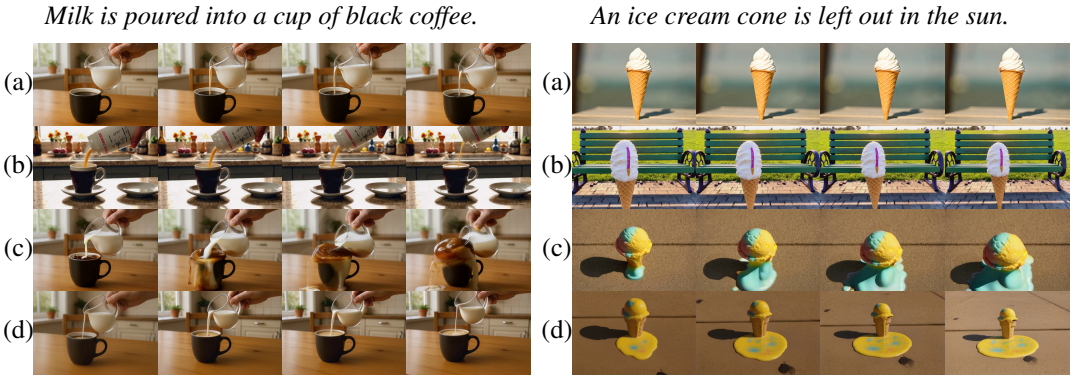


Figure 3: Qualitative comparisons on physical video generation tasks. **Left:** liquid interaction scenario, where milk is poured into black coffee. **Right:** thermal state transition, where an ice cream cone is exposed to sunlight. (a) Wan2.2-TI2V-5B; (b) LTX-Video; (c) VLIPP; (d) CausalMotion.

Qualitative comparisons Figure 3 presents qualitative comparisons on two representative physical video generation scenarios. In the liquid interaction example, existing methods exhibit physically implausible liquid motion or limited state transitions, while CausalMotion produces continuous changes in liquid height and appearance. In the thermal transition example, baseline methods show limited or unrealistic melting dynamics, whereas CausalMotion generates a more gradual and consistent melting process. These results qualitatively demonstrate the benefit of explicit state and trajectory guidance for modeling physical transitions.

4.3 Ablation Studies

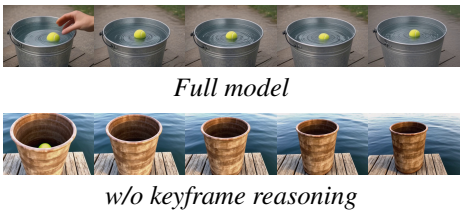
We conduct ablation studies by removing keyframe reasoning, trajectory-aware physical state planning, and latent trajectory guidance. Results are reported in Table 3.

Effect of trajectory-aware physical state planning. Removing trajectory planning and uniformly interpolating keyframes reduces the average score from 0.654 to 0.642, with notable drops in mechanics and material. This indicates that explicit motion and interaction modeling provides additional physical guidance.

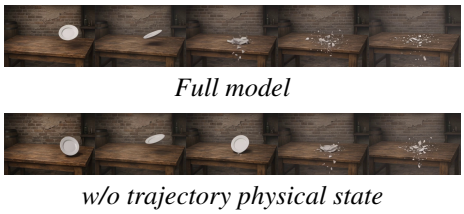
Effect of latent trajectory guidance. Disabling latent trajectory guidance further reduces the average score to 0.645, particularly in optics and mechanics. Although aligned keyframes provide coarse temporal guidance, localized trajectory updates improve fine-grained motion consistency.

Table 3: Quantitative results for ablation study on PhyGenBench

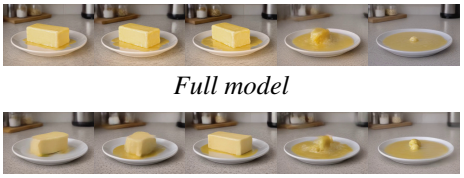
Model Variant	Mech. ↑	Opt. ↑	Therm. ↑	Mat. ↑	Avg.
CausalMotion	**0.608**	**0.713**	0.678	0.608	**0.654**
-w/o keyframe reasoning	0.425	0.600	0.467	0.358	0.471
-w/o trajectory physical state	0.583	0.700	**0.689**	0.592	0.642
-w/o latent trajectory guidance	0.594	0.680	0.678	**0.625**	0.645



(a) A tennis ball is gently placed on the surface of a bucket filled with water. (Force)



(b) A weak, frail porcelain plate is flung with significant speed at a robust, wooden table, where it collides upon impact. (Physical Properties)



(c) A timelapse captures the gradual transformation of butter as the temperature rises significantly. (Heat)



(d) A timelapse captures the reaction as concentrated sulfuric acid is poured onto a cotton ball. (Chemical Properties)

Figure 4: Qualitative ablation results. Removing keyframe reasoning (a) leads to missing or incorrect states, while removing trajectory physical state (b) results in unnatural collision dynamics of object interactions. Disabling latent trajectory guidance (c, d) leads to less smooth or abrupt transitions: in (d), the cotton ball disappears abruptly instead of gradually darkening and shrinking as the reaction proceeds. The full model maintains gradual transformations and physical consistency across all cases.

The slightly higher material score without trajectory guidance suggests that strong motion constraints can be less beneficial for appearance-dominated transformations with limited spatial motion.

Figure 4 qualitatively supports the complementary roles of the three components. Keyframe reasoning provides explicit intermediate states, trajectory-aware physical state planning improves object motion and interactions, and latent trajectory guidance enhances frame-to-frame transition smoothness. Together, these components enable more gradual and physically plausible transformations.

5 Conclusion

In this work, we proposed CausalMotion, a training-free physical grounding framework that integrates VLM reasoning ability into video generation. By decoupling causal reasoning from visual synthesis, our approach introduces structured intermediate representations and physically grounded trajectories to guide diffusion models toward more coherent and physically plausible outputs. Extensive experiments demonstrate that our method consistently improves physical consistency across diverse scenarios, particularly in mechanics and long-horizon dynamics, while preserving or even enhancing perceptual quality. Our method suggests a new paradigm for video generation, opening up a flexible and scalable direction for integrating higher-level cognition into generative models.

AI USE STATEMENT

In this work, generative AI models were used for both research purposes and manuscript preparation. As part of the proposed CausalMotion framework, a vision-language model was used for visual state reasoning, physical state and trajectory planning, and a text-to-image model (GPT-Image-1.5) was used to generate intermediate keyframe images. A vision-language model was also used as an auxiliary evaluator to assess physical plausibility, temporal consistency, and semantic alignment in our experiments. These uses are described in the corresponding method 3.1,3.2, and evaluation sections 4.2.

For manuscript preparation, generative AI tools were used for language editing, proofreading, and improving clarity and readability. They were not used to generate or alter experimental data, quantitative results, or scientific conclusions. We have reviewed all AI-assisted work. Experimental descriptions and quantitative results were checked against the original implementation and experiment logs. AI-generated text was edited and verified by the authors for accuracy, consistency, and originality. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed framework, including keyframe generation, physical state and trajectory planning, temporal alignment, and trajectory-guided latent-space guidance in the main paper. Additional implementation details, trajectory mapping analysis, intermediate visualizations, computational cost, ablation studies, limitations, and qualitative results are provided in the supplementary material. We report the video generation backbone, key experimental settings, and evaluation protocols used in our experiments. These details, together with the released implementation and configuration files, are intended to facilitate reproduction of our results.

REFERENCES

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

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 Learning Representations*, volume 2025, pp. 102075–102121, 2025.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations*, volume 2026, pp. 118456–118470, 2026.

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.

Yuanhao Cai, Kunpeng Li, Menglin Jia, Jialiang Wang, Junzhe Sun, Feng Liang, Weifeng Chen, Felix Juefei-Xu, Chu Wang, Ali Thabet, Xiaoliang Dai, Xuan Ju, Alan Yuille, and Ji Hou. Phygdpo: Physics-aware groupwise direct preference optimization for physically consistent text-to-video generation. In *ECCV*, 2026.

Boyuan Chen, Hanxiao Jiang, Shaowei Liu, Saurabh Gupta, Yunzhu Li, Hao Zhao, and Shenlong Wang. Physgen3d: Crafting a miniature interactive world from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6178–6189, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Lin Geng Foo, Mark He Huang, Alexandros Lattas, Stylianos Moschoglou, Thabo Beeler, and Christian Theobalt. Physical simulator in-the-loop video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4301–4311, June 2026.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. *Advances in Neural Information Processing Systems*, 38:103174–103201, 2026.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Ziqi Huang, Ning Yu, Gordon Chen, Haonan Qiu, Paul Debevec, and Ziwei Liu. VChain: Chain-of-visual-thought for reasoning in video generation. In *Annual Meeting of the Association for Computational Linguistics (ACL Findings), 2026*, 2026.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy. In *European Conference on Computer Vision*, pp. 38–57. Springer, 2024.

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*, pp. 28991–29017. PMLR, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

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.

Minh-Quan Le, Yuanzhi Zhu, Vicky Kalogeiton, and Dimitris Samaras. What about gravity in video generation? post-training newton's laws with verifiable rewards. *arXiv preprint arXiv:2512.00425*, 2025.

Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. In *International Conference on Machine Learning*, pp. 35685–35709. PMLR, 2025a.

Yunyi Li, Yu Qiao, Yaohui Wang, and Xinyuan Chen. Deform: Reasoning-guided physics-aware video generation via spatial-temporal masking, 2026.

Zizhang Li, Hong-Xing Yu, Wei Liu, Yin Yang, Charles Herrmann, Gordon Wetzstein, and Jiajun Wu. Wonderplay: Dynamic 3d scene generation from a single image and actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9080–9090, 2025b.

Long Lian, Baifeng Shi, Adam Yala, Boyi Li, et al. Llm-grounded video diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 50207–50227, 2024.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025a.

Runtao Liu, Haoyu Wu, Ziqiang Zheng, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8009–8019, 2025b.

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024b.

Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. Realwonder: Real-time physical action-conditioned video generation. *arXiv preprint arXiv:2603.05449*, 2026.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 43781–43806, 2025.

MiniMax-AI. Minimax h3, 2026. URL `https://github.com/MiniMax-AI/MiniMax-H3`.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. In *International Conference on Learning Representations*, volume 2025, pp. 38483–38505, 2025.

Sriram Narayanan, Ziyu Jiang, Srinivasa Narasimhan, and Manmohan Chandraker. Phyco: Learning controllable physical priors for generative motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41892–41902, June 2026.

Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14760–14769, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the" edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024a.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024b.

Runway Team. Runway. `https://runwayml.com`, 2024. Platform for AI-powered video editing and generative media creation.

Ying Shen, Jerry Xiong, Tianjiao Yu, and Ismini Lourentzou. Phantom: Physics-infused video generation via joint modeling of visual and latent physical dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11185–11194, June 2026.

Xiyang Tan, Ying Jiang, Xuan Li, Tianyi Xie, Zeshun Zong, Yin Yang, and Chenfanfu Jiang. Physmotion: Physics-grounded dynamics from a single image. In *Thirteenth International Conference on 3D Vision*, 2026.

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

Zijun Wang, Panwen Hu, Jing Wang, Terry Jingchen Zhang, Yuhao Cheng, Long Chen, Yiqiang Yan, Zutao Jiang, Hanhui Li, and Xiaodan Liang. Prophy: Progressive physical alignment for dynamic world simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14492–14501, 2026a.

Zixuan Wang, Yixin Hu, Haolan Wang, Feng Chen, Yan Liu, Wen Li, and Yinjie Lei. Chain of event-centric causal thought for physically plausible video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38122–38131, June 2026b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6327–6336, 2024.

Qiyao Xue, Xiangyu Yin, Boyuan Yang, and Wei Gao. Phyt2v: Llm-guided iterative self-refinement for physics-grounded text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18826–18836, 2025.

Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, et al. Vlipp: Towards physically plausible video generation with vision and language informed physical prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12360–12370, 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025b.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026.

Ke Zhang, Cihan Xiao, Yiqun Mei, Jiacong Xu, and Vishal M Patel. Think before you diffuse: Llms-guided physics-aware video generation. *arXiv e-prints*, pp. arXiv–2505, 2025a.

Qiyuan Zhang, Biao Gong, Shuai Tan, Zheng Zhang, Yujun Shen, Xing Zhu, Yuyuan Li, Kelu Yao, Chunhua Shen, and Changqing Zou. Physrvg: Physics-aware unified reinforcement learning for video generative models. *ECCV 2026*, 2026.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

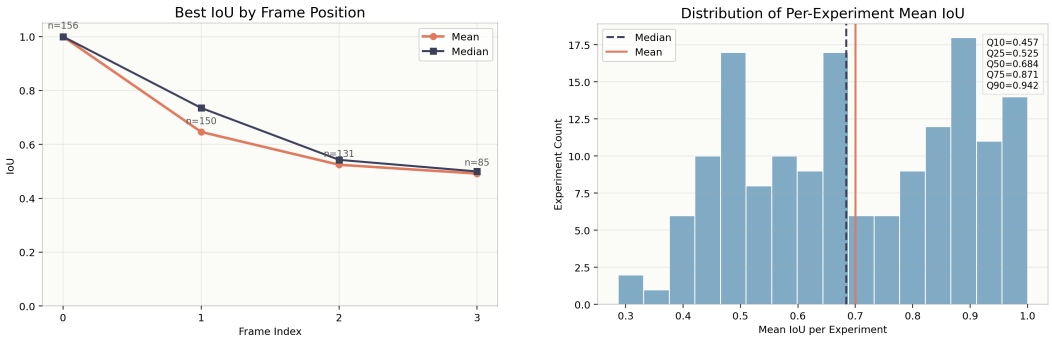
Yibo Zhao, Hengjia Li, Xiaofei He, and Boxi Wu. PhyRPR: Training-free physics-constrained video generation. In *The First Workshop on Efficient Spatial Reasoning*, 2026.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

A Discussion of Trajectory Mapping

In our framework, keyframe generation (Section 3.1) and trajectory reasoning (Section 3.2) are performed in different representational spaces. Keyframes are generated through VLM-based visual reasoning, while trajectory planning operates on object-centric spatial representations initialized from the first frame img_0. As a result, the two processes are not explicitly aligned in either coordinate system or temporal parametrization.

To analyze the consistency between these two modules, we compute the Intersection-over-Union (IoU) between object bounding boxes from trajectory prediction and those extracted from generated keyframes during temporal alignment (Section 3.3). The frame-wise IoU statistics are shown in Fig. 5(a).



(a) Frame-wise IoU across time.

(b) Distribution of per-experiment mean IoU.

Figure 5: Alignment analysis between trajectory predictions and keyframe object states. (a) shows frame-wise IoU degradation over time, while (b) illustrates variance across different prompts.

Despite being generated independently and potentially under different implicit coordinate systems, we observe that the predicted trajectories and keyframe object states exhibit consistent motion trends. In particular, objects follow similar spatial evolution patterns across both representations, indicating that the VLM reasoning process captures coherent physical dynamics even without explicit spatial supervision.

We further observe a gradual decrease in IoU as the frame index increases. This degradation is expected due to accumulated prediction errors and increasing temporal distance from the initial state. To better understand this effect, we analyze the zero-IoU rate across frame positions. As shown in Fig. 6, the proportion of complete mismatches increases steadily over time, suggesting that the performance drop is primarily caused by alignment failures rather than gradual spatial drift.

In addition, we examine the distribution of per-experiment mean IoU values, as illustrated in Fig. 5(b). While the overall average IoU remains relatively high, the distribution exhibits noticeable variance across prompts, indicating that alignment quality depends on scene complexity and motion patterns.

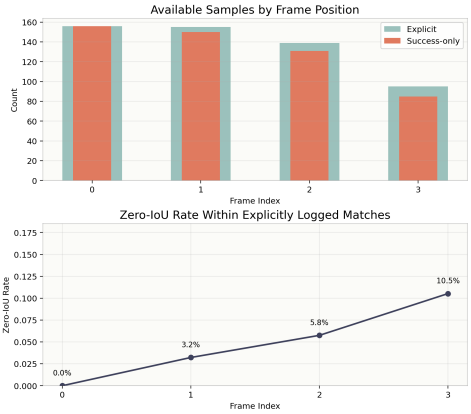


Figure 6: Zero-IoU rate across frame positions.

These observations suggest that the discrepancy between keyframe reasoning and trajectory planning lies mainly in representation rather than underlying dynamics. To bridge this gap during generation, our latent guidance mechanism (Section 3.4) integrates trajectory constraints with appearance information from keyframes. This design enables the model to reconcile differences between the two sources and produce videos that are both visually consistent and physically plausible.

Overall, our analysis shows that although the two modules operate independently, they converge to consistent motion dynamics in most cases, with failures primarily occurring in long-horizon or complex interaction scenarios.

B Intermediate Visualization

To improve the interpretability of our pipeline, we provide intermediate visualizations for representative examples, including generated keyframes, trajectory alignment, and final synthesis results.

Figures 7 and 8 present two examples of our trajectory alignment process under different physical scenarios: object-water interaction and collision dynamics. Each figure is organized as a 2×2 grid, where each sub-panel corresponds to one alignment step between sparse keyframes and the final video timeline. Within each sub-panel, the **left** image visualizes intermediate motion segments between consecutive keyframes, where overlaid bounding boxes represent the predicted object trajectories over time. The **right** image shows the corresponding aligned keyframe at the selected timestamp, where blue dashed boxes indicate the aligned object states used for trajectory guidance.

Figure 7 demonstrates a gradual object-water interaction process, where the glass ball progressively approaches and contacts the water surface. Figure 8 shows a fast collision scenario, where the basketball follows a physically plausible bouncing trajectory after impacting the ground. These examples illustrate how our alignment module bridges sparse keyframe reasoning with dense motion guidance, enabling temporally smooth and physically consistent video generation.

C Time Cost

Table 4 shows the average runtime of each stage in our method, providing a detailed breakdown of the overall computational cost.

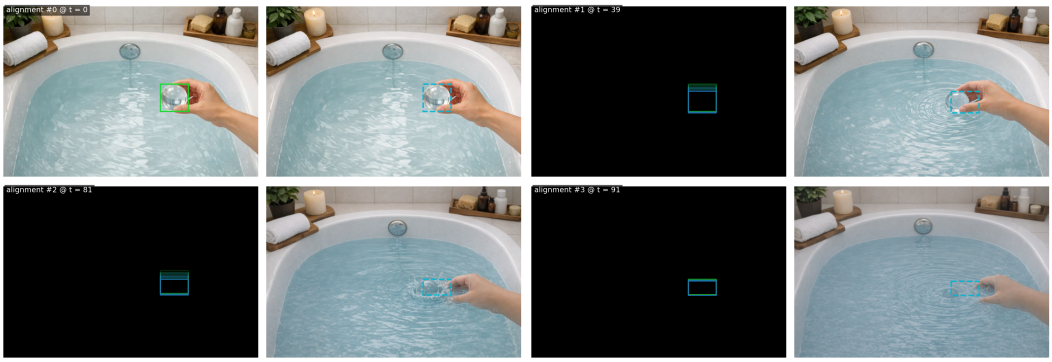


Figure 7: Trajectory alignment visualization for the prompt: *"A glass ball is gently placed on the surface of a bathtub filled with water."* Each sub-panel represents one alignment step. The left image shows intermediate trajectory transitions between adjacent keyframes, while the right image shows the aligned keyframe. Blue dashed boxes denote aligned object states used for dense trajectory guidance.

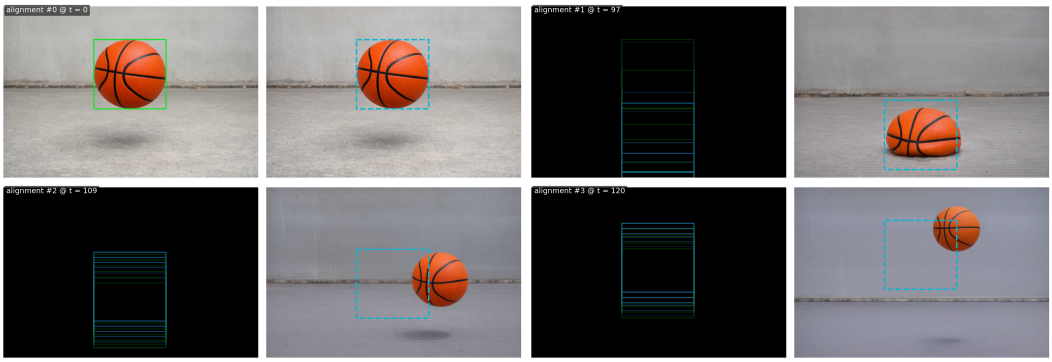


Figure 8: Trajectory alignment visualization for the prompt: *"A vibrant, elastic basketball is thrown forcefully toward the ground."* The example illustrates how our method aligns sparse keyframes with dense trajectories to model physically consistent collision and rebound dynamics.

Table 4: Average runtime breakdown of different components in our framework.

Breakdown	**Time**	**Implementation**
Keyframe Generation	271 sec	GPT-Image-1.5 API inference
Physical State and Trajectory Planning	69 sec	Qwen2.5-VL-72B API inference
Trajectory-Guided Latent-Space Guidance	108 sec	LTX-Video inference on 1× NVIDIA A800 GPU
Total	**448 sec**	

In total, our method requires approximately 448 seconds per sample. The main computational bottleneck comes from API-based keyframe generation rather than the trajectory guidance module itself. We note that this overhead can be further reduced by replacing external API calls with locally deployed VLMs or more efficient keyframe generation strategies.

D Detailed Inference Algorithms

In this section, we provide two pseudocode algorithms corresponding to Sections 3.1 and 3.4 respectively, to better illustrate the key steps of our method.

Algorithm 1 Iterative Visual Thought Reasoning for Keyframe Generation

Require: Text prompt p
Ensure: Keyframes $\{\text{img}_0, \text{img}_1, \ldots, \text{img}_N\}$ and descriptions $\{\text{txt}_0, \text{txt}_1, \ldots, \text{txt}_N\}$
consequence $\leftarrow \text{VLM}(p)$
$\text{txt}_0 \leftarrow \text{VLM}(p, \textit{consequence})$
$\text{img}_0 \leftarrow$ GPT-Image-1.5-generate(txt_0)
$i \leftarrow 1$
Flag $\leftarrow true$
while Flag **do**
  $\text{txt}_i \leftarrow \text{VLM}\big(\text{img}_{i-1}, \text{txt}_{i-1}, \textit{consequence}\big)$
  $\text{img}_i \leftarrow$ GPT-Image-1.5-edit$(\text{img}_{i-1}, \text{txt}_i)$
  $i \leftarrow i + 1$
  Flag $\leftarrow$ VLM-Judge$(\text{txt}_i, \textit{consequence})$
end while
$N \leftarrow i - 1$
return $\{\text{img}_0, \ldots, \text{img}_N\}$, $\{\text{txt}_0, \ldots, \text{txt}_N\}$

Algorithm 2 Trajectory-Guided Latent-Space Guidance

Require: Video latent tokens $\mathbf{L}$, object trajectories b_t^i, anchor memory $\mathcal{P}$, anchor frames $\mathcal{A}$, denoising step s
if s is not a selected guidance step **then**
  return $\mathbf{L}$
end if
$\alpha_s \leftarrow \text{StepGate}(s)$
$\mathbf{L}^{\text{vid}} \leftarrow \text{Unpatchify}(\mathbf{L})$
for $\tau = 1, \ldots, T_{\text{lat}} - 1$ **do**
  for each object i **do**
    $B_{i,\tau} \leftarrow b_\tau^i$
    $\widetilde{B}_{i,\tau} \leftarrow \text{Expand}(B_{i,\tau}, \rho_t)$
    $M_{i,\tau} \leftarrow \text{SoftMask}(\widetilde{B}_{i,\tau})$
    $(a^-, a^+) \leftarrow \text{NearestAnchors}(\tau, \mathcal{A})$
    $(w^-, w^p, w^+) \leftarrow \text{TriWeight}(\tau; a^-, a^+)$
    Obtain $P_{i,\tau}^-$ and $P_{i,\tau}^+$ from $\mathcal{P}$ at anchors a^- and a^+
    $P_{i,\tau}^{\text{prev}} \leftarrow \text{Warp}\big(\mathbf{L}_{\tau-1}^{\text{vid}}, b_{\tau-1}^i, B_{i,\tau}\big)$
    $P_{i,\tau}^{\text{ref}} \leftarrow \dfrac{w^- P_{i,\tau}^- + w^p P_{i,\tau}^{\text{prev}} + w^+ P_{i,\tau}^+}{w^- + w^p + w^+}$
    $\mathbf{L}_\tau^{\text{vid}} \leftarrow (1 - \alpha_s M_{i,\tau}) \odot \mathbf{L}_\tau^{\text{vid}} + \alpha_s M_{i,\tau} \odot P_{i,\tau}^{\text{ref}}$
  end for
end for
$\mathbf{L} \leftarrow \text{Repatchify}(\mathbf{L}^{\text{vid}})$
return $\mathbf{L}$

E PROMPT OF VLM AS JUDGE

We use the following prompt for VLM-based evaluation in Table 2. For each generated video, we repeat the evaluation five times and report the average score across the five evaluations. The overall score is computed as the average of the three aspect-level scores.

> You are a strict, consistent, and unbiased video quality evaluator.
>
> You will be given: 1. A text prompt describing a target video. 2. A generated video.
>
> Your task is to evaluate how well the video satisfies the prompt along THREE aspects. Each aspect must be scored on a 1–5 integer scale.
>
> Text prompt: description
>
> Evaluation aspects:
>
> 1. Physical Plausibility: Evaluate whether object behaviors, motions, and interactions obey the physical or logical rules implied by the prompt. Penalize physically impossible dynamics, incorrect causality, or implausible artifacts.
>
> 2. Temporal Consistency: Evaluate object permanence and identity over time. Penalize flickering, sudden appearance or disappearance, unprompted morphing, or inconsistent motion across frames.
>
> 3. Semantic Alignment: Evaluate whether the video matches both the explicit prompt content and its implied requirements. Penalize missing key objects, incorrect actions, wrong relationships, or failure to satisfy the prompt's intent.
>
> Scoring rules:
>
> - 5 = Perfect, no noticeable issues
> - 4 = Minor issues, mostly correct
> - 3 = Noticeable issues, partially correct
> - 2 = Major issues, largely incorrect
> - 1 = Completely incorrect
>
> Output format:
>
> Return ONLY a valid JSON object with the following fields:

```
{
  "physical_plausibility": {
    "score": int,
    "justification": string
  },
  "temporal_consistency": {
    "score": int,
    "justification": string
  },
  "semantic_alignment": {
    "score": int,
    "justification": string
  },
  "overall_score": float
}
```

> The overall_score should be the average of the three scores (rounded to two decimals).
>
> Do NOT include any text outside the JSON.

F FAILURE CASES AND LIMITATIONS

Despite its effectiveness, our approach has some limitations. First, trajectory planning relies on the underlying VLM to infer plausible object motions from the given context. In complex scenes with multiple interacting objects, occlusions, or ambiguous motion cues, the predicted trajectories may deviate from the actual motion and consequently affect the generated results. Second, trajectory



Figure 9: A failure case for a static appearance transformation.

guidance is more effective for scenes involving explicit spatial motion and may provide limited benefits for transformations dominated by appearance changes.

For example, in a scene depicting gradual mailbox rusting, the object remains largely static while its appearance changes over time. In such cases, trajectory guidance may introduce minor local artifacts, such as blur or jitter, as illustrated in Fig. 9.

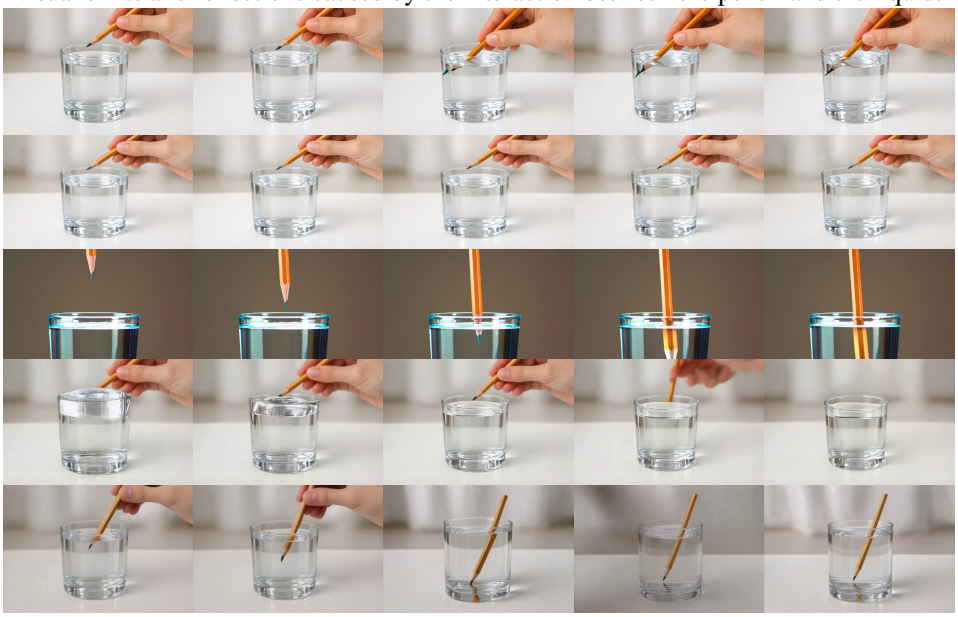
G Future Work

Although our current framework is designed for physical video generation, the proposed keyframe reasoning and trajectory-guided generation paradigm has broader potential beyond physical scenarios. Future work could extend this framework to more general video generation tasks that require long-horizon temporal planning, such as human activities, multi-agent interactions, and story-driven video generation.

H More qualitative results

To further illustrate the performance of CausalMotion, we provide additional qualitative comparisons with representative video generation methods, including Wan2.2, Helios, VLIPP, and PhyGDPO. The examples cover diverse physical phenomena and interaction scenarios, providing additional visual evidence of the physical plausibility and temporal consistency of our generated videos.

A wooden pencil is carefully dipped into a glass of crystal-clear water, showing the intriguing visual shifts and reflections caused by the interaction between the pencil and the liquid.



A small burning ball of paper was thrown into a pile of dry paper.

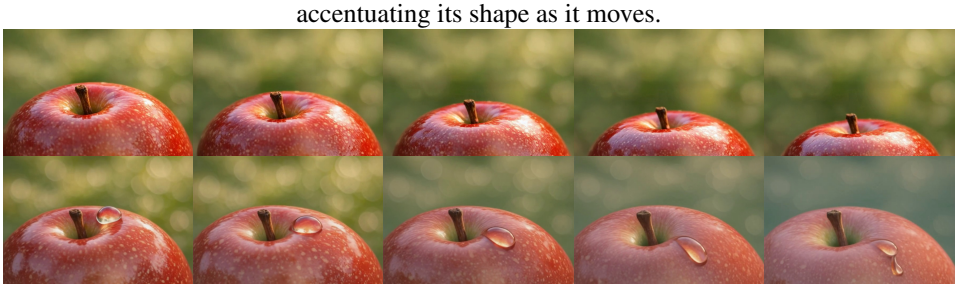


Figure 10: Additional qualitative comparisons between Wan2.2, Helios, PhyGDPO, VLIPP, and our method.

A timelapse captures the gradual transformation of butter as the temperature rises significantly.



A glistening dewdrop is sliding gracefully across the smooth surface of a waxed apple, accentuating its shape as it moves.



A weak, frail porcelain plate is flung with significant speed at a robust, wooden table, where it collides upon impact.



Figure 11: Additional qualitative comparisons between VLIPP and our method.

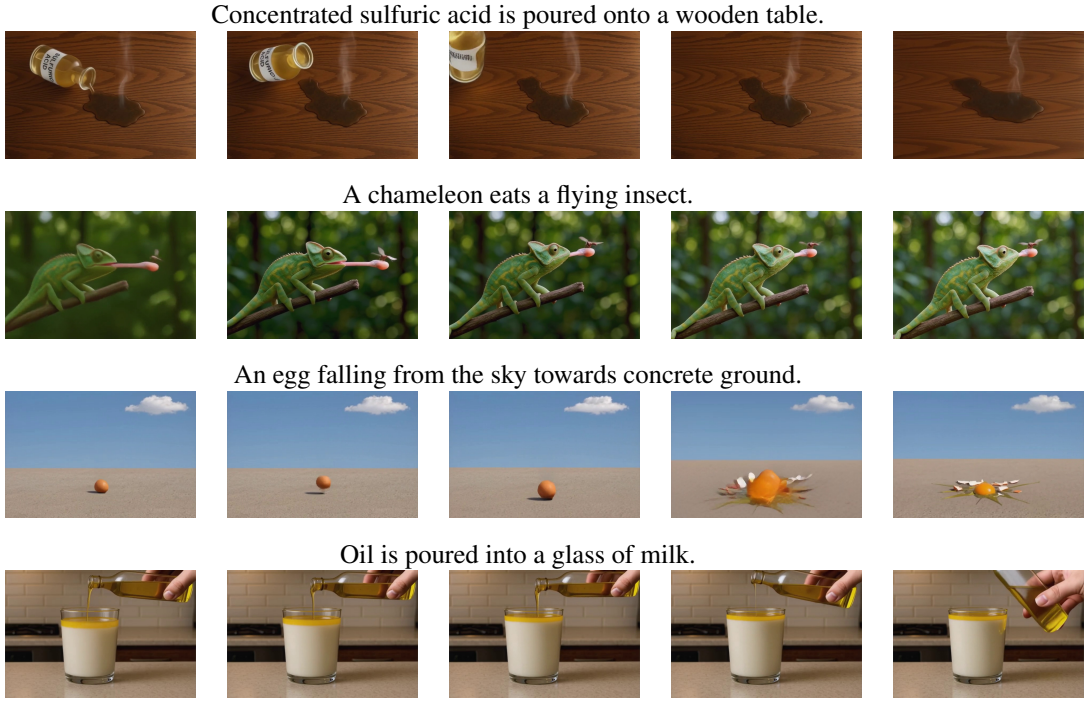


Figure 12: Additional qualitative comparisons between VLIPP and our method on diverse physical phenomena.

A small burning stick was thrown into a pile of hay.



A tennis ball is gently placed on the surface of a bucket filled with water.



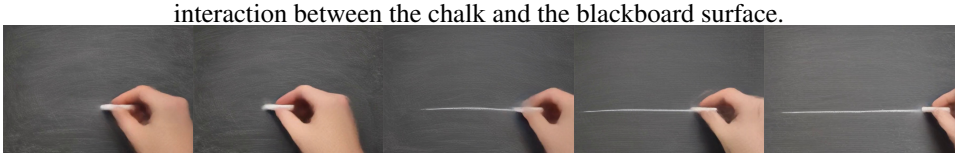
A timelapse of a water-filled soft cloth being forcefully squeezed by hand, with the pressure intensifying rapidly over time.



A ray of light is shining diagonally on a glass bottle in the dark, with the shadow of the glass bottle appearing at the bottom.



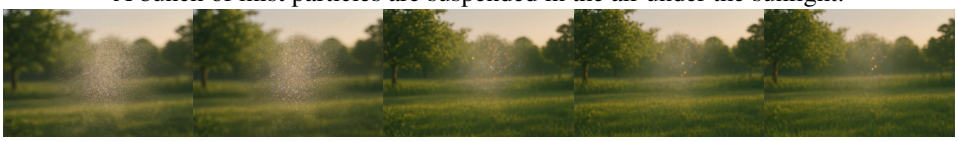
A piece of white chalk is used to write on the rough, dark surface of a blackboard,showcasing the interaction between the chalk and the blackboard surface.



A magnifying glass is gradually moving closer to a leaf, revealing the intricate details and textures of the veins and surface patterns as it approaches.



A bunch of mist particles are suspended in the air under the sunlight.



A timelapse captures the gradual transformation of ice cream as the temperature rises significantly.



Figure 13: More qualitative results generated by our method across diverse physical domains.