HiPhy: Hierarchical Alignment for Physically-Plausible Multi-Principle Video Generation

Tahira Kazimi[1] **Shubhankar Borse**[2] **Munawar Hayat**[2] **Fatih Porikli**[2] **Pinar Yanardag**[1]

[1]Virginia Tech [2]Qualcomm AI Research

Project Page: `https://hiphy-video.github.io`

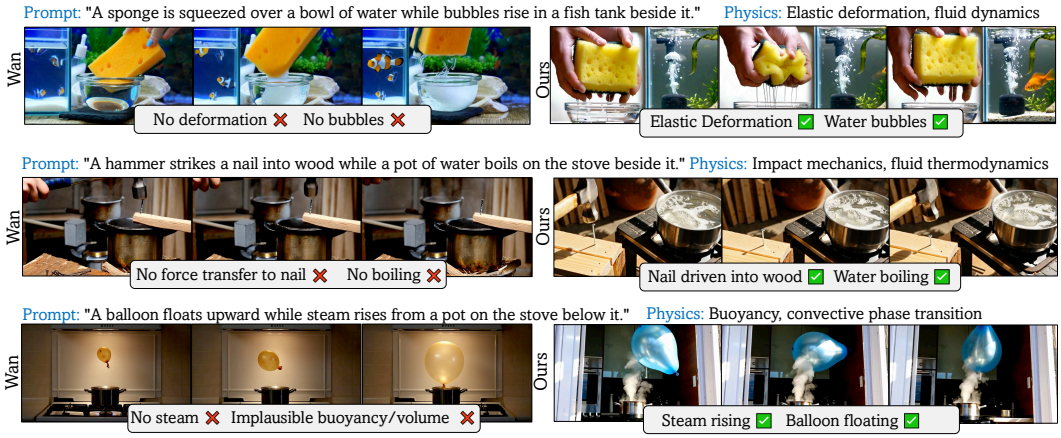

Figure 1: `HiPhy` **enables generation of physically plausible videos for scenes with multiple concurrent physical principles.** Wan (left) collapses on multi-principle prompts, failing to generate coherent motion or dropping a physical principle entirely, such as missing *boiling* or *steam* motions. Instead of collapsing physical plausibility into a single score, `HiPhy` (right) aligns generation with a hierarchical reward that tracks the temporal sub-stages of each principle as an independent learning stream, jointly recovering all concurrent processes with correct temporal progression.

Abstract

Video generation models have achieved remarkable visual fidelity and have strong potential to become general-purpose world simulators. Despite this progress, they still fail to generate videos which adhere to laws of physics. The problem becomes even more apparent in realistic settings where multiple physical principles must work together within the same video; for example, *'a balloon floating upward while steam rises from a pot'* requires buoyancy and fluid dynamics to unfold coherently and simultaneously. Yet existing methods largely ignore multi-principle interactions, focusing on a single principle per video. We propose HiPhy (**Hi**erarchical **Phy**sical Alignment), a reinforcement learning framework that grounds video generation in physical laws through a dual-level objective: locally enforcing the temporal dynamics of individual physical principles, and globally ensuring the physical and semantic coherence of the entire scene. To support multi-principle generation, we construct a 50K-prompt dataset and introduce a prompt benchmark *MultiPhyBench*, spanning a diverse range of co-occurring physical events. Our experiments show that HiPhy significantly outperforms prior methods and baselines, improving physical commonsense and semantic alignment significantly across various benchmarks, with the largest gains on scenes involving multiple concurrent physical principles where competing methods degrade most sharply.

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

1 Introduction

Video diffusion models [24, 36] have made remarkable progress in producing high-resolution, temporally coherent videos [38, 49]. This rapid progress has made video generation models promising candidates for general-purpose world simulators that are capable of synthesizing diverse physical scenarios on demand for embodied AI, robotics simulation, scientific visualization, and counterfactual reasoning [18, 23, 20, 2].

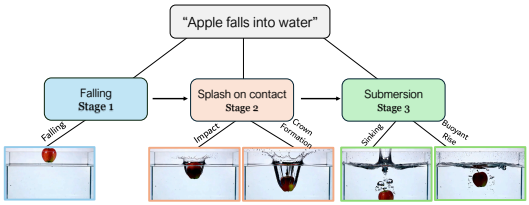

Figure 2: HiPhy explicitly decomposes complex physical events (e.g., an apple falling into water) into temporally ordered sub-stages. This allows us to recursively score the chronological progression, completeness, and alignment of the underlying physics.

Despite their visual fidelity, current video generators routinely produce clips that violate elementary physical laws, such as momentum, fluid dynamics, and material properties [23, 18]. We argue this reflects a deeper representational gap: because models are trained merely to maximize frame likelihood given text [14, 13], they reproduce plausible patterns from their vast training data without reasoning about the underlying mechanisms [31, 9]. Consider a prompt *"an apple falls into water,"* which entails two co-occurring physical principles: rigid-body and fluid dynamics (see Fig. 2). We find this representational gap manifests in two consistent failure modes on such compound prompts: (1) **Principle Omission**, where models ignore secondary physics, for instance, the apple drops realistically but the water remains undisturbed; and (2) **Temporal Shortcutting**, where unfolding dynamics are collapsed into a single visual snapshot, such as rendering the splash as a static spraying effect instead of progressing through *impact* $\rightarrow$ *crown formation*. Importantly, these failures do not reflect a deficit in visual capability; given sufficiently detailed text, modern video models generate individual physical principles with striking fidelity. The core challenge is finding a systematic mechanism to elicit this latent knowledge along two dimensions: the presence of each principle and its temporal structure. This motivates HiPhy's dual-level design: per-principle supervision targets *Principle Omission*, while fine-grained sub-stage reward design addresses *Temporal Shortcutting*.

Existing efforts to inject physical knowledge into video models largely rely on 3D engines [28, 30, 7], fine-tuning [6, 41, 12, 40], or inference-time guidance [48, 15, 45, 26, 34, 17]. While 3D engines can simulate concurrent dynamics, they are computationally heavy and constrained to rigid domains. Conversely, learning-based and guidance methods offer flexibility but typically supervise only a single physical principle per video. Consequently, they lack mechanisms to ensure compound events are jointly satisfied. Empirically, we observe these latter methods degrade sharply as concurrent processes increase, often dropping one principle entirely.

In this paper, we propose HiPhy (**Hi**erarchical **Phy**sical Alignment), a reinforcement learning framework that closes the previously mentioned gap by formulating compound physical events as a hierarchical alignment problem. By optimizing each physical principle locally, HiPhy enables physically grounded video generation in concurrent, multi-principle scenes. Our key insight is that physical plausibility is fundamentally both a temporal and a compositional property. While modern video generators struggle to reason about the temporal progression of physical events, leading to *Principle Omission* and *Temporal Shortcutting*, we model multi-principle scenes as compositions of independent, temporally structured physical principles. Consequently, we introduce a novel training objective that explicitly enforces stage-level supervision across each concurrent physical principle. HiPhy mirrors this insight through a hierarchical alignment framework. At the foundational level, we explicitly optimize for the chronological progression of each physical event by breaking it into smaller stages. At the same time, we align these stages with global video coherence and semantic alignment objectives, ensuring that multi-principle interactions compose naturally while maintaining semantic and visual coherence. Crucially, when a prompt contains multiple concurrent physical events, HiPhy treats each as an independent learning stream, preventing high-variance signals from washing out the gradient. Our contributions are summarized as follows:

- We introduce HiPhy, a reinforcement-learning framework that hierarchically grounds video generation in physical principles. Unlike prior methods that target a single physical principle per video, HiPhy is the first framework that jointly handles multiple concurrent events.
- HiPhy is a plug-and-play post-training framework that integrates with any text-to-video model without architectural modification both open-source (Wan2.1) and black-box (VEO3) backbones.
- HiPhy improves physical commonsense by up to 34% and semantic alignment by up to 60% over prior physics-aware baselines.
- We curate a 50K-prompt training set and a 1K-prompt benchmark spanning multiple concurrent physical principles, each annotated with per-principle sub-process decompositions. We will share our data, code, and checkpoints publicly.

2 Related Work

Physically Plausible Video Generation One line of work fine-tunes video backbones on large physically annotated datasets [6, 41, 12, 40]. Because such models lack explicit physical knowledge, generation quality remains tightly coupled to training-set coverage and rarely transfers to the diversity of real-world scenarios [31, 18]. A second line injects physical knowledge through 3D engines such as Blender or Unreal, depth or edge predictors, or auxiliary 3D objectives [28, 30, 22], and graphics rendering pipelines [3]. These methods are precise but constrained to fixed physical categories and incur heavy computational cost, particularly for phenomena such as fluids or steam [7]. Additionally, some works introduce auxiliary signals at inference time, including coarse intermediate representations [48, 15], off-the-shelf verifiers and judges [45], and test-time optimization [34, 17]. Verifier-based methods either require large memory budgets or incur long inference times, and test-time optimization typically refines trajectories without addressing implausible initial plans. While prior work holistically supervises a single physical principle per video, HiPhy is the first framework to enforce temporal progression in multi-principle scenes via a hierarchical stage-level objective.

RL-based Video Generation Prior work adapts preference learning to video along several axes. VideoDPO [27] extends DPO to text-to-video diffusion with an omni-preference objective balancing visual quality and text relevance. Flow-DPO [25] aligns flow-matching video generators using a multi-dimensional reward, improving motion smoothness and prompt adherence over SFT and alternative RLHF variants. More fine-grained supervision has also been explored via DenseDPO [44], which targets temporal locality by densifying video preferences across time. Complementary directions include GRPO variants tailored to video control, such as DanceGRPO [46] for learning temporally coherent policies. Additionally, methods such as [42, 8] provide global supervision over state properties. Unlike these methods, HiPhy is the first to handle multi-principle physical scenes by integrating a hierarchical objective that targets per-principle optimization in its training.

3 Background

Group Relative Policy Optimization GRPO [33] is a reinforcement learning framework that computes advantages through group-based normalization of sampled outputs' rewards, eliminating the need for a separate critic network. Given a query q, the method samples G responses $o_1, \ldots, o_G$ from the old policy $\pi_{\theta_{\text{old}}}$ and updates the current policy π_θ to maximize $\mathcal{J}_{\text{GRPO}}(\theta)$ given as:

$$\mathbb{E}_{q,\{o_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}A_i,\ \ \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\epsilon, 1+\epsilon\right)A_i\right) - \beta\,\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\right] \quad (1)$$

Advantages A_i are computed by normalizing rewards r_i within the group:

$$A_i = \frac{r_i - \text{mean}(r_{1:G})}{\text{std}(r_{1:G})}. \quad (2)$$

The standard GRPO objective (Eq. 1) maximizes the expected reward, with advantage normalization that upweights samples with higher-than-average reward within the group. While effective for single-reward optimization, this formulation collapses distinct reward combinations into identical advantages in multi-reward optimization [29].

4 Method

Given a user prompt q describing n concurrent physical principles and a frozen text-to-video model $\mathcal{G}$, HiPhy learns a policy π_θ such that the video $v = \mathcal{G}(p)$ generated from $p \sim \pi_\theta(\cdot \mid q)$ depicts all n

principles with the correct temporal progression of their sub-stages. Two insights shape this design: **(i)** Modern video models already possess strong visual priors over individual physical phenomena but lack explicit temporal structure for how a physical event should unfold; surfacing this structure at the prompt level is therefore more sample-efficient than retraining the backbone, **(ii)** Generating complex physical scenes demands compositional reasoning; a video may generate one physical principle faithfully yet fail to compose multiple principles together. We train a reasoning policy π_θ via GRPO [33] using a hierarchical compositional objective. By explicitly isolating each concurrent physical principle and optimizing its temporal progression through hierarchical rewards, our policy improves the physical coherence of compound scenes.

4.1 Multi-Principle Dataset Curation

We construct a 50K-prompt training dataset and introduce MultiPhyBench, a 1000-prompt evaluation suite in which each prompt describes up to three simultaneous physical events. To generate these, we extract single-principle processes and their domains from VideoPhy2 [5] and WISA-80k [39], sample combinations of 1–3 events, and use an LLM [47] to filter out implausible co-occurrences. For each valid combination, we generate a unified video prompt and a chronologically ordered hierarchical tree for each constituent principle (Fig.2). These hierarchical annotations serve as pseudo-labels for our local reward (Sec. 4.2). Full curation details are provided in the SM A.

4.2 HiPhy: Hierarchical Physical Alignment

We introduce `HiPhy`, a hierarchical physical alignment framework that scores the temporal progression of each physical principle in a fine-grained fashion (see Fig. 3 for an overview). Consider an input prompt p describing n physical principles $\{p^{(i)}\}_{i=1}^{n}$. During data curation, each physical principle $p^{(i)}$ is paired with an ordered tree $\mathcal{T}^{*(i)}$ of depth $D^{(i)}$ in which each node represents a stage within that physical principle. For *"an apple falls into water and submerges,"* the outer level nodes are *falling* $\rightarrow$ *splash on contact* $\rightarrow$ *submerging*. The stage *splash on contact* then further expands into *impact* $\rightarrow$ *crown formation* (Fig.2). Nodes in each level are temporally ordered, meaning that the completion of one node depends on the completion of the nodes before it. During training, we generate a video, $\hat{v}$, using a backbone video diffusion model, $\mathcal{G}$ noted as $\hat{v} = \mathcal{G}(p)$. To evaluate how well the generated video, $\hat{v}$, depicts each physical principle $\{p^{(i)}\}_{i=1}^{n}$, we extract a similar hierarchical structure as $\mathcal{T}^{*(i)}$ from $\hat{v}$ for each physical principle $p^{(i)}$. This procedure is done in the same way as detailed in SM Sec.A. Formally, given a VLM extractor, $\mathcal{R}$, we derive an observed tree, $\hat{\mathcal{T}}^{(i)} = \mathcal{R}(\hat{v}, p^{(i)})$, that decomposes the observed stages in the generated video into a hierarchically ordered tree.

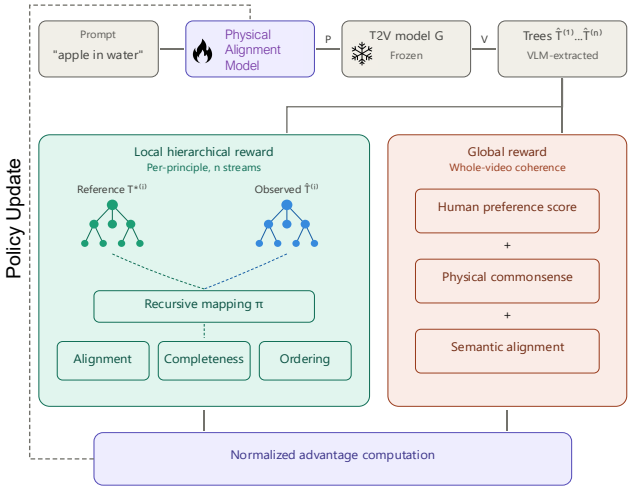

Figure 3: Overview of the `HiPhy` framework. We introduce a hierarchical optimization framework aimed to elicit physically grounded video generation in scenes with multiple physical principles.

To measure how plausibly each physical principle, $\{p^{(i)}\}_{i=1}^{n}$, is depicted in the generated video, $\hat{v}$, we align the reference hierarchically ordered tree $\mathcal{T}^{*(i)}$ with the observed tree $\hat{\mathcal{T}}^{(i)}$. We align the two trees by calculating a one-to-one mapping π using the Hungarian assignment [21], applied level by level. For each level L_j in $\mathcal{T}^{*(i)}$, we build a similarity matrix $\mathbf{C_j}$, where an element c_{km} in $\mathbf{C_j}$ measures how similar is node k in $\mathcal{T}^{*(i)}$ to node m in $\hat{\mathcal{T}}^{(i)}$. Each entry in $\mathbf{C_j}$ blends a sentence encoder ϕ with the VLM judge $\mathcal{J}_{\text{stage}}$ (for computational efficiency, during training, $\mathcal{J}_{\text{stage}}$ is the same

as $\mathcal{R}$).

$$\mathbf{C}_{km} = \alpha \cdot \cos\big(\phi(s_k^*), \phi(\hat{s}_m)\big) + (1-\alpha) \cdot \mathcal{J}_{\text{stage}}\big(s_k^*, \hat{s}_m\big), \tag{3}$$

where the cosine term handles paraphrase and the VLM judge catches cases where wording is similar but the underlying event differs. If node a in $\mathcal{T}^{*(i)}$ has sub-stages in level L_{j+1}, we recursively build $\mathbf{C_{j+1}}$ for the nodes in L_{j+1} and calculate our mapping π for each level L_j from its associated $\mathbf{C_j}$. Please refer to SM 1 for full pseudocode.

Using the mapping π and the similarity matrices, we evaluate how well the video depicts the physical principle along three dimensions: *Alignment* $r_{\text{align}}^{(i)}$, *completeness* $r_{\text{comp}}^{(i)}$, and *ordering* score $r_{\text{order}}^{(i)}$.

First, let $t^{(i)}$ denote the total number of leaf nodes (the most granular sub-stages) in the reference tree, and let $t_{\text{matched}}^{(i)}$ be the subset of successfully matched leaves. A leaf is in $t_{\text{matched}}^{(i)}$ if its similarity score and all of its parent nodes in the similarity matrix $\mathbf{C}$ are above a threshold τ. This logical dependency ensures that if a high-level stage fails to generate (e.g., missing the *splash*), all of its dependent sub-stages (e.g., *crown formation*) are strictly treated as missing. *Alignment* calculates the average similarity of the successfully matched leaves to assess how well the observed stages in $\hat{v}$ match the expected reference stages in $\mathcal{T}^{*(i)}$, while *completeness* calculates the fraction of expected leaves that passed the threshold to assess how much of the physical process was captured:

$$r_{\text{align}}^{(i)} = \frac{1}{t^{(i)}} \sum_{k \in t_{\text{matched}}^{(i)}} \mathbf{C}_{k,\pi(k)}, \quad r_{\text{comp}}^{(i)} = \frac{1}{t^{(i)}} \sum_{k \in t_{\text{matched}}^{(i)}} \mathbf{1}\big[\mathbf{C}_{k,\pi(k)} \geq \tau\big]. \tag{4}$$

To ensure the generated sub-stages unfold in the correct chronological sequence, we compute an *ordering* score $r_{\text{order}}^{(i)}$. We evaluate this locally for each successfully matched parent node $\mathbf{u} \in \mathcal{N}^{(i)}$. Letting $\mathcal{M}_{\mathbf{u}} = (m_1, \ldots, m_{|\mathcal{M}_{\mathbf{u}}|})$ be its sequence of matched children ordered by their expected reference position, the local ordering score $O_{\mathbf{u}}^{(i)}$ calculates the fraction of these children that appear in the correct relative temporal sequence ($\succ_{\mathbf{u}}$) within the generated video:

$$O_{\mathbf{u}}^{(i)} = \frac{1}{|\mathcal{M}_{\mathbf{u}}| - 1} \sum_{k=2}^{|\mathcal{M}_{\mathbf{u}}|} \mathbf{1}[\pi(m_k) \succ_{\mathbf{u}} \pi(m_{k-1})],$$

where $\pi(m_k) \succ_{\mathbf{u}} \pi(m_{k-1})$ indicates that the observed sibling matched to m_k appears later than the one matched to m_{k-1}, and $O_{\mathbf{u}}^{(i)} = 1$ when $|\mathcal{M}_{\mathbf{u}}| \leq 1$. The principle-level ordering score is then the geometric mean across all matched internal nodes, given by $r_{\text{order}}^{(i)} = \big(\prod_{\mathbf{u} \in \mathcal{N}^{(i)}} O_{\mathbf{u}}^{(i)}\big)^{1/|\mathcal{N}^{(i)}|}$. Finally, the hierarchical reward for the physical principle is the mean of these three sub-scores:

$$R_{\text{hier}}^{(i)}(\hat{v}) = \frac{1}{3}\big(r_{\text{comp}}^{(i)} + r_{\text{align}}^{(i)} + r_{\text{order}}^{(i)}\big). \tag{5}$$

Notably, if a physical principle lacks hierarchical depth ($D^{(i)} = 1$), the recursion bottoms out after a single assignment at the root, naturally recovering a standard flat-stage reward formulation as a special case.

Global Video Reward. The local reward captures fine-grained temporal progression within each process but does not guarantee that the video is semantically faithful to the prompt, visually coherent, or consistent with broader physical commonsense. We introduce a global reward term R_{global} with three complementary terms. First, HPSv2 [43] captures visual quality and aesthetic alignment, averaged over T uniformly sampled frames to obtain R_{HPS}. Second, a physics VLM judge $\mathcal{J}$ (VideoPhy2 [5]) rates physical commonsense R_{PC}, assessing whether the primary action follows real-world physical rules, and semantic alignment R_{SA}, measuring whether events described in q actually appear in $\hat{v}$. All three terms are linearly normalized to $[0, 1]$ and combined into a single scalar:

$$R_{\text{global}}(\hat{v}, q) = w_h \tilde{R}_{\text{HPS}} + w_p \tilde{R}_{\text{PC}} + w_s \tilde{R}_{\text{SA}}, \tag{6}$$

Because all three terms characterize whole-video quality, we treat R_{global} as a single reward stream. Together, R_{PC} guards against gross physical implausibility that local rewards may overlook, R_{SA} anchors generation to prompt semantics, and R_{HPS} favors visually coherent outputs.

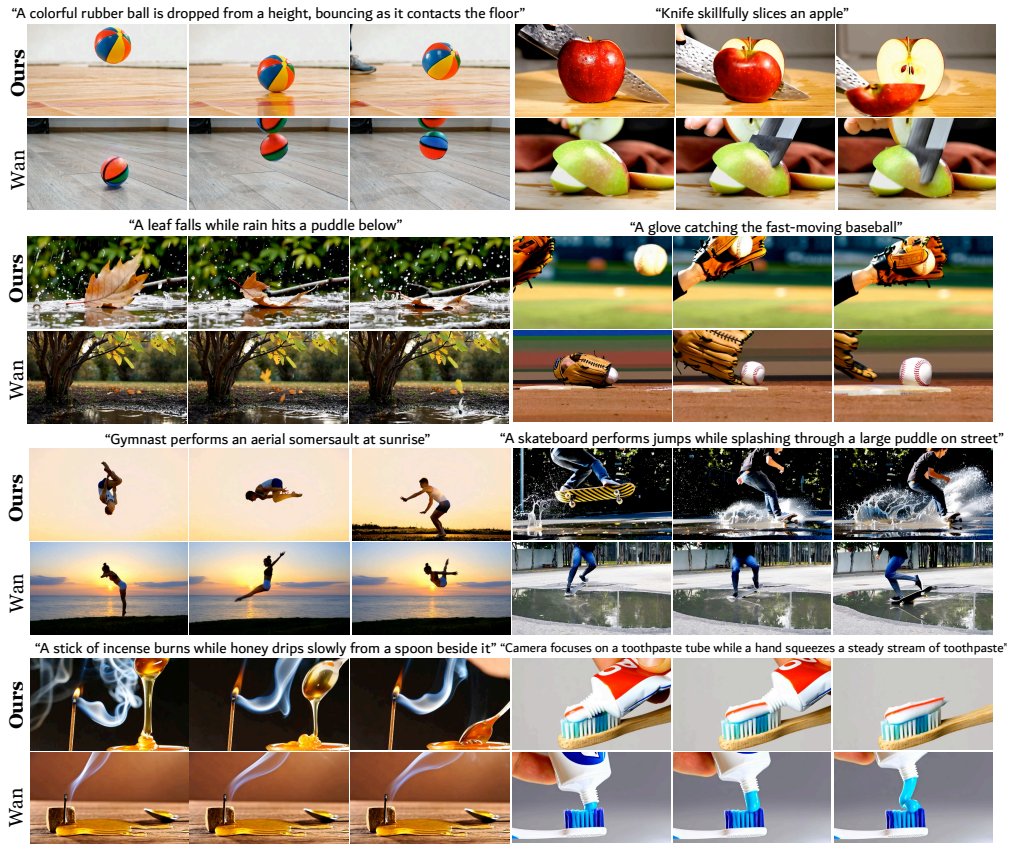

Figure 4: **Qualitative results.** `HiPhy` (top row) produces coherent motion and correct temporal progression on prompts spanning single- and multi-principle physical scenes, while Wan2.1 (bottom row) frequently drops a principle, breaks contact dynamics, or generates static scenes.

Training For a prompt with n principles, we sample G rollouts $\{\hat{v}^{(j)}\}_{j=1}^{G}$, yielding n hierarchical rewards $R_{\text{hier}}^{(i,j)}$ and one global reward $R_{\text{global}}^{(j)}$. To prevent high-variance streams from dominating the GRPO [33] gradient, we independently normalize each stream within the group [29]:

$$A_{\text{hier}}^{(i,j)} = \frac{R_{\text{hier}}^{(i,j)} - \mu_i}{\sigma_i + \epsilon}, \quad A_{\text{global}}^{(j)} = \frac{R_{\text{global}}^{(j)} - \mu_g}{\sigma_g + \epsilon}, \tag{7}$$

where μ and σ denote the per-stream mean and standard deviation across the G rollouts. Summing these unit-variance local streams directly would scale the overall variance by $\mathcal{O}(n)$, disproportionately skewing the gradient for prompts with more principles. To maintain consistent variance regardless of n, we rescale the local sum by $1/\sqrt{n}$:

$$A_{\text{sum}}^{(j)} = A_{\text{global}}^{(j)} + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} A_{\text{hier}}^{(i,j)}. \tag{8}$$

Finally, we apply batch-wise normalization $\hat{A}^{(j)} = (A_{\text{sum}}^{(j)} - \mu_b)/(\sigma_b + \epsilon)$ to absorb residual scale variation from across-prompt heterogeneity [29]. This $\hat{A}^{(j)}$ serves as the final advantage in the clipped GRPO objective, utilizing a per-token KL penalty against a frozen reference policy.

5 Experiments

We showcase the effectiveness of our method on state-of-the-art T2V models Wan2.1 [38], as well as black-box T2V model VEO3 [10]. We compare our method against finetuning methods [39, 50],

Table 1: Comprehensive evaluation across VideoPhy2, PhyWorldBench, Multi-concept, and VBench, and ablation experiments over different reward terms

	VideoPhy2			**PhyWorldBench**			**MultiPhyBench**			**VBench**	
Method	PC↑	SA↑	PC & SA ↑	PC↑	SA↑	PC & SA ↑	PC↑	SA↑	PC & SA↑	Dyn. ↑	App.↑
Wan2.1	49.10	43.60	30.5	47.7	40.9	23.7	53.0	35.0	27.0	0.808	0.573
PnP	80.1	68.0	60.14	56.1	48.5	31.2	57.0	37.0	29.0	0.844	0.549
VideoRepa	70.9	53.5	43.8	51.7	33.3	19.3	49.0	22.0	15.0	0.863	0.550
PhyT2V	79.0	65.0	55.0	40.3	47.5	22.7	50.0	26.0	15.0	0.872	0.500
WISA	68.0	59.7	44.3	56.3	50.7	30.7	52.5	43.0	24.5	0.859	0.567
GPT-5	69.2	60.9	43.1	56.9	52.5	31.8	51.5	44.0	24.5	0.882	0.531
HiPhy (Ours)	**88.6**	**79.2**	**68.8**	**74.9**	**76.3**	**61.6**	**76.7**	**68.85**	**52.7**	**0.905**	**0.592**
Qwen 2.5	60.4	62.9	41.5	58.6	58.3	44.5	50.3	39.2	19.1	0.874	0.517
HiPhy (global)	72.7	**75.3**	56.3	62.3	**64.3**	54.3	61.0	53.0	40.0	0.871	**0.576**
HiPhy (local, sum)	72.2	70.1	48.5	61.9	57.3	37.1	67.0	57.9	42.4	0.865	0.556
HiPhy (local, normalized)	**79.2**	70.4	**57.3**	**65.6**	60.8	**52.5**	**68.5**	**58.3**	**44.5**	**0.893**	0.567

inference optimization [17], prompt refinement methods [45, 35]. For video generation, we utilize Wan2.1 [38] as our backbone and Qwen2.5-7b-Instruct [47] as our alignment model. In Eq. 3, α is set 0.5 and τ is set 0.65 through a hyperparameter search. The sentence encoder ϕ is SBERT [32] and the VLM judge is Qwen2.5-VL-7B [4].

5.1 Experimental Setup

All experiments are conducted on two NVIDIA H200 GPUs. We first perform supervised finetuning for 50 iterations, followed by GRPO training for approximately 2000 iterations. The learning rate starts from 2×10^{-5} for SFT and 2×10^{-7} during GRPO optimization, with w_h, w_p, and w_s set to 0.33. Our dataset comprises of $\approx 50K$ prompts with 1–3 concepts equally distributed. Validation and test data are %10 and %5 of the data, respectively. We use a group size of 4 and a training batch size of 8, optimizing with Adam optimizer [19], and TRL [1] library. Unless otherwise specified, we adopt a guidance scale of 6, with 40 inference steps and 81 output frames per video.

5.2 Qualitative Comparisons

Figure 4 illustrates the diverse physical principles our method handles compared to the Wan [38] backbone. Scenes span rigid-body dynamics (*bouncing ball, sliced apple, catching a baseball*), fluid-rigid interactions (*leaf falling into a puddle*), athletic motion (*somersault, skateboarding*), and concurrent multi-principle events (*smoke* beside *dripping honey*, *squeezing toothpaste*). In these scenarios, Wan frequently fails: balls bounce implausibly or multiply randomly, leaves disappear into water rather than floating, and athletic motions collapse. Conversely, our method generates each process with accurate temporal progression.

Figure 5 compares our method against Wan and three physics-aware baselines: inference-time optimization (Predict-and-Perturb [17]), iterative LLM refinement (PhyT2V [45]), and finetuning (WISA [39]). The results reveal consistent baseline failure modes that compound as the physical event count increases. On two-principle prompts like "*book sliding while a ball falls*", Fig.5, **a**, baselines struggle with coherent interaction: Wan and PnP omit the *sliding book*, PhyT2V hallucinates extra objects, and WISA produces static scenes. Similarly, in Fig.5, **b**, gradual integration is often skipped. This performance gap widens on three-principle prompts (bottom). Baselines routinely drop key elements, such as the *simmering sauce* or *rain* in Fig.5, **c**, or the *boiling water* and *whisking action* in Fig.5, **d**. In contrast, our method coherently generates all constituent processes without degrading overall video quality, which we attribute to optimizing each process as a separate local reward stream during training.

5.3 Quantitative Experiments

We evaluate our method across multiple physical benchmarks such as Videophy2 and PhyWorld-Bench [5, 11]. We assess physical commonsense (PC) and semantic alignment (SA) using Gemini 3 Flash [37], which supports higher frame rates. Additionally, we report a "PC & SA" metric that measures the fraction of videos that are simultaneously physically grounded and semantically



Figure 5: **Qualitative comparisons.** As the number of physical principles per prompt grows, baselines exhibit consistent failure modes, incoherent motion, hallucinated objects, or skipping a principle entirely, and these compound on three-principle prompts (bottom) where multiple processes are dropped simultaneously. `HiPhy` (top row of each block) renders each constituent process coherently without degrading visual quality.

aligned. To assess robustness beyond physics-specific benchmarks, we also compare against baselines on VBench [16], summarizing its orthogonal dimensions into two axes: *Dynamics* (subject and background consistency, motion smoothness, dynamic degree) and *Appearance* (imaging and aesthetic quality). Table 1 reports results across two single-principle benchmarks (VideoPhy2 [5] and PhyWorldBench [11]) and our curated multi-principle benchmark of 1K prompts containing concurrent physical principles. `HiPhy` achieves the strongest performance on all physics benchmarks. Our method significantly outperforms the best baselines on VideoPhy2 (+10% physical, +16% alignment) and PhyWorldBench ($\approx$ +30% physical, +45% semantic), demonstrating the strength of our hierarchical reward. While generating concurrent physical events remains challenging across all methods, our per-principle hierarchical optimization preserves distinct learning signals, maintaining the largest margin on the multi-principle benchmark. Finally, VBench [16] evaluations confirm that physics-aware training preserves general video quality. Both *Dynamics* and *Appearance* match or exceed the Wan2.1 baseline, as our global reward terms effectively anchor overall coherence.

5.4 Human Evaluation.

We conducted a user study ($n = 50$) with 30 prompts to evaluate *physical commonsense* and *semantic alignment* using a 1–5 Likert scale. As shown in Table 2, our method outperforms all baselines across all metrics, with the most significant margin in text alignment, indicating that raters perceive a clear improvement in compositional physical fidelity. A sample survey question is provided in the SM. Furthermore, we analyzed performance on videos containing multiple concurrent physical principles (**Fig. 7, D**).

Method	PC	SA	Principle
PhyT2V	2.28	2.11	2.19
WAN	2.76	2.19	2.40
WISA	2.49	2.29	2.11
PnP	3.07	2.90	2.67
Ours	**3.89**	**4.13**	**3.81**

Table 2: User Study.

While baselines exhibit weak performance, often failing to depict specific principles entirely within a scene, HiPhy maintains steady, high ratings across all individual events. This consistency validates our per-process reward design, which ensures each physical principle receives a distinct learning signal.

5.5 Computational Efficiency.

We compare the inference time of our method against prior physics-grounded video generation baselines on a single H200 GPU. PhyT2V [45] relies on iterative VLM-based prompt refinement and re-generation, incurring substantial overhead on top of the base video model. PnP [17] incurs extra computation for its extra optimization steps during inference. WISA [39] introduces moderate cost through fine-tuned LoRA. In contrast, HiPhy runs once before generation and adds only $\sim$2.4 seconds of overhead, making the total inference time effectively identical to the base model while achieving stronger physical grounding.

6 Ablation

6.1 Reward Composition

We ablate the contribution of each reward component in Table 1. *Qwen 2.5* denotes the base model with no RL training; HiPhy (global) trains with only the global reward; HiPhy (local, sum) trains with only the local hierarchical rewards summed as a single scalar; and HiPhy (local, normalized) trains the hierarchical reward with normalized streams as in Eq. 8. The two reward families turn out to be complementary. Training with only the global reward yields strong semantic alignment and visual quality but weaker physical commonsense, since the global reward provides no fine-grained signal about how individual processes unfold over time. Local-only training improves physical commonsense and semantic alignment on multi-principle scenes but yields lower semantic alignment on single-principle benchmarks. Its effect on video dynamics varies with reward aggregation. Combining local and global rewards with per-stream normalization achieves the strongest overall performance, highlighting their complementary roles in preserving physical fidelity and scene-level coherence.

Method	Time (min)	Overhead
PhyT2V	8.79	+170s
WISA	6.66	+36s
PnP	8.86	+175s
Wan	5.96	—
Ours	**6.00**	**+2.4s**

Table 3: Computational efficiency comparison. HiPhy achieves minimal overhead.

6.2 Generalization across video backbones.

Figure 6 shows that both Wan [38] and VEO [10] produce near-static coins on the prompt *"a coin dances in a spiral on the table,"* despite the motion being well within their generative capacity. Pairing each backbone with HiPhy recovers the spiral trajectory in both cases, supporting our claim that the failure to render physically plausible scenes reflects a representational gap rather than a deficit of visual capability. HiPhy transfers across any backbone with sufficient underlying physical priors without retraining.

6.3 Performance vs. principle count.

Figure 7 **(A–C)** illustrates performance as the number of concurrent principles per prompt increases. While all methods degrade with rising complexity, baselines exhibit a sharp collapse; PnP and PhyT2V lose approximately half their PC scores between 1- and 3-process prompts, with joint PC & SA scores falling below 20. In contrast, `HiPhy` maintains a joint score of approximately 50, widening the performance gap as complexity grows. We attribute this robustness to our per-process reward stream, which provides distinct learning signals for each event. This prevents the gradient dilution seen in baselines that attempt to collapse multi-process correctness into a single scalar.



Figure 6: **Ablation on different backbones.**

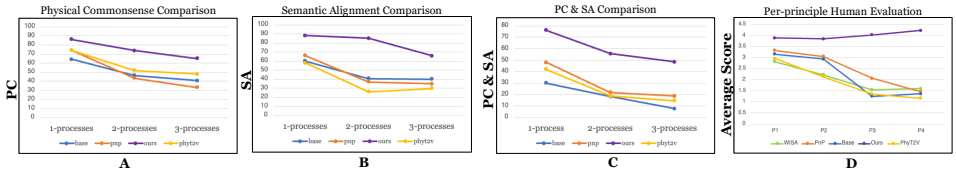

Figure 7: **Performance scaling and human evaluation.** (A–C) Automatic metrics for PC, SA, and joint PC&SA across process counts. `HiPhy` (purple) maintains a significantly wider margin as complexity increases. (D) Averaged human scores per principle (P1–P4). While baselines show constant degradation in multi-event scenes, `HiPhy` achieves consistent physical fidelity across all concurrent principles.

7 Limitations and Broader Impacts

`HiPhy` uses a frozen backbone T2V model, so its fidelity is upper-bounded by the visual capacity of the underlying generator, similar to other existing work [45, 48]. More broadly, physics-aware generation supports applications in education, scientific visualization, and embodied AI, but more plausible video raises the stakes for misinformation and non-consensual content; we build on public backbones and datasets and encourage provenance tools and responsible-release practices.

8 Conclusion

We presented a reinforcement learning approach to physics-grounded video generation that decomposes physical correctness into per-principle local rewards over temporal sub-stages and a global reward for semantic and visual coherence. Across VideoPhy2, PhyWorldBench, and our multi-principle benchmark, `HiPhy` achieves the strongest physical commonsense and semantic alignment, with the largest margins on prompts containing multiple concurrent principles, where prior methods degrade most sharply. These results suggest that explicit temporal decomposition, combined with a learned refinement signal, is an effective approach to closing the gap between video generation and physical realism.

9 Acknowledgments

This work is supported by a Qualcomm Gift and National Science Foundation under Grant No. 2543524. We also gratefully acknowledge fal.ai for providing compute support.

References

[1] TRL: Transformer Reinforcement Learning (2020)

[2] Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.W., et al.: World simulation with video foundation models for physical ai. arXiv preprint arXiv:2511.00062 (2025)

[3] An, Z., Kupyn, O., Uscidda, T., Colaco, A., Ahuja, K., Belongie, S., Gonzalez-Franco, M., Tintore Gazulla, M.: Vggrpo: Towards world-consistent video generation with 4d latent reward. In: European Conference on Computer Vision. pp. 305–322. Springer (2026)

[4] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)

[5] Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. arXiv preprint arXiv:2503.06800 (2025)

[6] Cai, Y., Li, K., Jia, M., Wang, J., Sun, J., Liang, F., Chen, W., Juefei-Xu, F., Wang, C., Thabet, A., et al.: Phygdpo: Physics-aware groupwise direct preference optimization for physically consistent text-to-video generation. arXiv preprint arXiv:2512.24551 (2025)

[7] Chen, D., Lao, Z., Guo, Y., Yu, H.: Fluid composer: Fluid detail composition and rendering using video diffusion models. In: Computer Graphics Forum. p. e70300. Wiley Online Library (2025)

[8] Chen, H.H., Huang, H., Chen, Q., Yang, H., Lim, S.N.: Hierarchical fine-grained preference optimization for physically plausible video generation. Advances in Neural Information Processing Systems **38**, 133919–133951 (2026)

[9] Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)

[10] Google DeepMind: Veo: a text-to-video generation system (veo 3 tech report) (2025), accessed: 2025-11-12

[11] Gu, J., Liu, X., Zeng, Y., Nagarajan, A., Zhu, F., Hong, D., Fan, Y., Yan, Q., Zhou, K., Liu, M.Y., et al.: " phyworldbench": A comprehensive evaluation of physical realism in text-to-video models. arXiv preprint arXiv:2507.13428 (2025)

[12] Gupta, A., Yu, L., Sohn, K., Gu, X., Hahn, M., Li, F.F., Essa, I., Jiang, L., Lezama, J.: Photorealistic video generation with diffusion models. In: European Conference on Computer Vision. pp. 393–411. Springer (2024)

[13] Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)

[14] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv:2204.03458 (2022)

[15] Huang, Y., Wang, Z., Lin, H., Kim, D.K., Omidshafiei, S., Yoon, J., Zhang, Y., Bansal, M.: Planning with sketch-guided verification for physics-aware video generation. arXiv preprint arXiv:2511.17450 (2025)

[16] Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., 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)

[17] Jang, S., Ki, T., Jo, J., Xie, S., Yoon, J., Hwang, S.J.: Self-refining video sampling. arXiv preprint arXiv:2601.18577 (2026)

[18] Kang, B., Yue, Y., Lu, R., Lin, Z., Zhao, Y., Wang, K., Huang, G., Feng, J.: How far is video generation from world model: A physical law perspective. arXiv preprint arXiv:2411.02385 (2024)

[19] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[20] Krüger, M., Gilbert, D., Kuhlen, T.W., Gerrits, T.: Game engines for immersive visualization: Using unreal engine beyond entertainment. PRESENCE: Virtual and Augmented Reality **33**, 31–55 (2024)

[21] Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)

[22] Kupyn, O., Uscidda, T., Gazulla, M.T., Manhardt, F., Tombari, F., Rupprecht, C.: Epipolar geometry improves video generation models. arXiv preprint arXiv:2510.21615 (2025)

[23] Li, C., Michel, O., Pan, X., Liu, S., Roberts, M., Xie, S.: Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. arXiv preprint arXiv:2503.09595 (2025)

[24] Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)

[25] Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Xia, M., Wang, X., et al.: Improving video generation with human feedback. arXiv preprint arXiv:2501.13918 (2025)

[26] Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. arXiv preprint arXiv:2206.01714 (2022)

[27] Liu, R., Wu, H., Zheng, Z., Wei, C., He, Y., Pi, R., Chen, Q.: Videodpo: Omni-preference alignment for video diffusion generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8009–8019 (2025)

[28] Liu, S., Ren, Z., Gupta, S., Wang, S.: Physgen: Rigid-body physics-grounded image-to-video generation. In: European Conference on Computer Vision. pp. 360–378. Springer (2024)

[29] Liu, S.Y., Dong, X., Lu, X., Diao, S., Belcak, P., Liu, M., Chen, M.H., Yin, H., Wang, Y.C.F., Cheng, K.T., et al.: Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. arXiv preprint arXiv:2601.05242 (2026)

[30] Lv, J., Huang, Y., Yan, M., Huang, J., Liu, J., Liu, Y., Wen, Y., Chen, X., Chen, S.: Gpt4motion: Scripting physical motions in text-to-video generation via blender-oriented gpt planning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1430–1440 (2024)

[31] Motamed, S., Culp, L., Swersky, K., Jaini, P., Geirhos, R.: Do generative video models understand physical principles? In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 948–958 (2026)

[32] Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)

[33] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)

[34] Shaulov, A., Hazan, I., Wolf, L., Chefer, H.: Flowmo: Variance-based flow guidance for coherent motion in video generation. arXiv preprint arXiv:2506.01144 (2025)

[35] Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)

[36] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

[37] Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (2024)

[38] Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)

[39] Wang, J., Ma, A., Cao, K., Zheng, J., Zhang, Z., Feng, J., Liu, S., Ma, Y., Cheng, B., Leng, D., et al.: Wisa: World simulator assistant for physics-aware text-to-video generation. arXiv preprint arXiv:2503.08153 (2025)

[40] Wang, X., Zhu, Z., Huang, G., Wang, B., Chen, X., Lu, J.: Worlddreamer: Towards general world models for video generation via predicting masked tokens. arXiv preprint arXiv:2401.09985 (2024)

[41] Wang, Z., Hu, P., Wang, J., Zhang, T.J., Cheng, Y., Chen, L., Yan, Y., Jiang, Z., Li, H., Liang, X.: Prophy: Progressive physical alignment for dynamic world simulation. arXiv preprint arXiv:2512.05564 (2025)

[42] Wu, S., Xu, C., Xia, Z., Li, W., Lu, L., Maneriker, P., Du, F., Li, M., Liu, H.: Phyprompt: Rl-based prompt refinement for physically plausible text-to-video generation. arXiv preprint arXiv:2603.03505 (2026)

[43] Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)

[44] Wu, Z., Kag, A., Skorokhodov, I., Menapace, W., Mirzaei, A., Gilitschenski, I., Tulyakov, S., Siarohin, A.: Densedpo: Fine-grained temporal preference optimization for video diffusion models. arXiv preprint arXiv:2506.03517 (2025)

[45] Xue, Q., Yin, X., Yang, B., Gao, W.: 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)

[46] Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)

[47] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)

[48] Yang, X., Li, B., Zhang, Y., Yin, Z., Bai, L., Ma, L., Wang, Z., Cai, J., Wong, T.T., Lu, H., 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 (2025)

[49] Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)

[50] Zhang, X., Liao, J., Zhang, S., Meng, F., Wan, X., Yan, J., Cheng, Y.: Videorepa: Learning physics for video generation through relational alignment with foundation models. arXiv preprint arXiv:2505.23656 (2025)

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are directly supported by the results and contributions presented throughout the paper, including both quantitative and qualitative evaluations.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Justification: The paper discusses several limitations in a dedicated section. It highlights the capacity of underlying generators.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A]

Justification: The paper does not include any formal theoretical results, assumptions, or proofs. Instead, it introduces a novel post-training framework for physically plausible video generation, which is validated through extensive qualitative and quantitative experiments. As the contribution is methodological rather than theoretical, this item is not applicable.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: Authors provide details about hyperparameters, models used, and embedding choices, as well as training data size and how the dataset is curated, and the number of training steps. Details are found in the supplementary materials, Experimental Setup section.

Guidelines:

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

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: The dataset curation sections in the main paper and the appendix provide detailed information about how dataset is constructed using existing open-source data. The experimental setup section includes detailed instructions for reproducing the experiments and results, including environment setup, base models, prompts, and hyper-parameters as well. Our code will be released publicly, with a README, license file, and usage instructions to ensure reproducibility and clarity. These resources enable faithful replication of the results and support further research in physically plausible video generation.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Experimental setup section in the supplementary materials provide hyperparameters, optimizers, data splits and details related to reproducibility.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: Yes, the paper reports error bars (mean $\pm$ standard deviation) for all major quantitative results. These are shown in the tables of the supplementary material due to space constraints in the main paper and are discussed in the corresponding experimental sections. The use of standard deviation is consistent across evaluations, allowing readers to assess result variability and robustness.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We provide information about our GPUs used which are two NVIDIA-H200 GPUs. We also specify training parameters like group size, time complexity, inference-time efficiency, and data batch size in our experiment setup section.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics. It does not involve sensitive human subjects, personal data, or dual-use risks. Any potential model biases (e.g., VLM bias toward Western aesthetics) are explicitly acknowledged and discussed in the Limitations section. The paper promotes transparency by releasing the code and detailing all training parameters, encouraging reproducibility and responsible development of generative models.

Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The paper addresses both benefits and risks of the proposed method. It supports applications in education, scientific visualization, and embodied AI, but more plausible video raises the stakes for misinformation and non-consensual content; we build on public backbones and datasets and encourage provenance tools and responsible-release practices.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: Our contribution is a policy network and a text-prompt dataset describing physical phenomena (e.g., fluid dynamics, rigid-body collisions). We do not release a new generative backbone; `HiPhy` operates on top of frozen, already-public models (Wan2.1, VEO3). The dataset contains no human subjects, faces, identifying information, or scraped web content — only generated descriptions of physical scenes which are built on top of previously public dataset. We provide a sample of our dataset in our supplementary materials. The misuse risk is therefore upper-bounded by that of the underlying backbones, which are governed by their original release licenses.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All external assets used in this work are credited via citation in the main paper. Specifically: we use Wan2.1 [38] as open-source video backbones and VEO3 [10] as a black-box backbone; Qwen2.5-7B-Instruct [47] and Qwen2.5-VL-7B [4] as the alignment and judge models; SBERT [32] as the sentence encoder; HPSv2 [43] as the visual quality reward; and the VideoPhy2 [5] judge for physical commonsense and semantic alignment scoring. Our training prompts are derived from VideoPhy2 [5] and WISA-80k [39], both of which are publicly released for research use. Evaluations are conducted on VideoPhy2 [5], PhyWorldBench [11], and VBench [16]. All models and datasets are used under their original licenses (Apache 2.0, MIT, or research-only terms as specified by each release), and our use is consistent with their intended research purposes.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: We will release our trained policy checkpoints, training and inference code, and the curated multi-principle prompt dataset (including per-principle sub-stage annotations). Dataset documentation specifying the prompt curation procedure, physics domains is provided in the supplementary materials. Documentation will be provided alongside the assets, including: (i) a README describing the training pipeline and reward model configuration; (ii) inference scripts with example commands for reproducing results on VideoPhy2, PhyWorldBench, and our multi-principle benchmark; and (iii) a license file specifying terms of use consistent with the underlying backbone licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [Yes]

    Justification: We conducted a user study with 50 participants on Prolific. The paper describes the evaluation procedure and questions shown to participants. Screenshots of the interface and full instructions are included in the supplemental material. Participants were compensated in accordance with Prolific's fair pay policies.

    Guidelines:

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

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

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

    Answer: [Yes]

    Justification: Our user study involved 50 participants on Prolific. Based on our institution's guidelines, this study was reviewed and determined to fall under exempt or minimal-risk research. Participants provided informed consent and were compensated fairly. No personally identifiable information was collected.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [Yes]

Justification: LLM usage is described in detail in the main paper. Specifically: (i) we train a policy π_θ initialized from Qwen2.5-7B-Instruct [47] as described in method section; (ii) we use Qwen2.5-VL-7B [4] as the VLM judge within the reward model (Section 4.2); and (iii) we use an LLM during dataset curation as described in method section and Appendix. All model versions, prompts, and hyperparameters are documented in the main paper and appendix.

Guidelines:

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

Table of Contents

A Multi-Process Dataset Curation **23**

B Hierarchical alignment algorithm **23**

C More Qualitative **24**

D Extracted Stage validation **24**

E Hierarchical vs. flat rewards. **24**

F Ablation on individual alignment, completeness, and ordering **25**

G Ablation on Normalization vs. scaling **25**

H User Study **26**

I Error Bars for the quantitative comparisons **27**

J Stage Generation Prompt **28**

K Video Prompts **29**

A Multi-Process Dataset Curation

Existing physical video datasets and benchmarks evaluate physical plausibility along a single process per prompt (e.g., a ball rolling down an incline or water pouring into a glass). However, real-world scenes routinely involve multiple physical processes unfolding simultaneously. To model physical realism in multi-process settings, we construct a dataset in which each prompt describes multiple concurrent physical events. The number of processes per prompt, n, scales with the video length; for instance, at 81 frames we support up to three simultaneous processes.

Physical Process Extraction. We collect single-process physical concepts from existing datasets, including VideoPhy2 [5] and WISA-80k [39]. For each source prompt, we use a large language model to extract the underlying physical process along with its physics domain. For instance, from the prompt *"a bowling ball rolls off a table and hits the ground,"* we extract the process *"ball rolling and falling"* with the domain *rigid-body dynamics*; from *"steam rises from a boiling pot,"* we extract *"water boiling"* with the domain *thermodynamics*.

Combinatorial Sampling and Plausibility Filtering. We sample subsets of 1, 2, or 3 physical concepts per prompt. To capture the true complexity of real-world scenes, we allow combinations of both inter-domain events (e.g., fluid dynamics paired with rigid-body physics) and simultaneous intra-domain events (e.g., multiple concurrent rigid-body interactions). Because not all sampled combinations describe events that can realistically co-occur within a single scene, we employ an LLM to judge whether a given set of processes is physically compatible, discarding combinations marked as implausible.

Prompt Generation and Stage Annotation. For each plausible combination, we generate a natural-language video description that coherently integrates all constituent processes within a single scene. In addition, for each physical process, we generate a sequence of chronologically ordered stages that describe the progression of that process over time (e.g., *contact* $\rightarrow$ *momentum transfer* $\rightarrow$ *deceleration* for a collision). These stage annotations serve as pseudo-labels references for the local stage-aware reward described in the next section. Figure 2 the structure of the resulting annotations.

B Hierarchical alignment algorithm

We provide pseudocode for the recursive alignment procedure used to compute the hierarchical reward (Sec. 4.2). Given a reference tree $\mathcal{T}^{*(i)}$ and an observed tree $\hat{\mathcal{T}}^{(i)}$ extracted by the VLM, the procedure is invoked at the roots and descends both trees in parallel, solving a local assignment problem at each matched node and recursing into matched pairs whose similarity exceeds the threshold τ.

Algorithm 1 HIPHYALIGN$(s^*, \hat{s}, \Pi, \mathcal{S}, \mathcal{N})$ — recursive tree alignment

Require: reference node s^*, observed node $\hat{s}$; accumulators Π (assignment), $\mathcal{S}$ (matched-leaf similarities), $\mathcal{N}$ (matched internal nodes)

if s^* is a leaf **then**
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{(s^*, \mathbf{C}_{s^*,\hat{s}})\}$
  return
end if
$\mathcal{N} \leftarrow \mathcal{N} \cup \{s^*\}$
$S^* \leftarrow \mathrm{ch}(s^*) = (s_1^*, \ldots, s_K^*)$, $\hat{S} \leftarrow \mathrm{ch}(\hat{s}) = (\hat{s}_1, \ldots, \hat{s}_M)$
Build $\mathbf{C} \in \mathbb{R}^{K \times M}$ with entries from Eq. 3; pad with zeros to a square matrix
$\pi_{\text{local}} \leftarrow \arg\max_\pi \sum_k \mathbf{C}_{k,\pi(k)}$ ▷ Hungarian assignment
for $k = 1, \ldots, K$ **do**
  $\Pi \leftarrow \Pi \cup \{(s_k^*, \hat{s}_{\pi_{\text{local}}(k)})\}$
  if $\mathbf{C}_{k,\pi_{\text{local}}(k)} \geq \tau$ **then**
    HIPHYALIGN$(s_k^*, \hat{s}_{\pi_{\text{local}}(k)}, \Pi, \mathcal{S}, \mathcal{N})$ ▷ recurse on reliable matches only
  end if
end for

The procedure is invoked once per principle as HIPHYALIGN$(\mathrm{root}(\mathcal{T}^{*(i)}), \mathrm{root}(\hat{\mathcal{T}}^{(i)}), \emptyset, \emptyset, \emptyset)$, after which Π contains the global assignment π pairing every reference node with its observed counterpart,

Physics Domain	#Videos	Stage Quality *mean ± 95% CI*	Ordering *mean ± 95% CI*	Overall *mean*
Rigid-body / collision	11	4.42 ± 0.21	4.61 ± 0.23	4.46
Fluid dynamics	12	4.31 ± 0.23	4.55 ± 0.25	4.36
Thermodynamics	8	4.05 ± 0.30	4.28 ± 0.32	4.10
Deformation	5	4.24 ± 0.34	4.40 ± 0.36	4.27
Optics	8	3.88 ± 0.32	4.12 ± 0.34	3.92
Combustion	6	4.18 ± 0.31	4.45 ± 0.33	4.24
Overall	**50**	$\mathbf{4.21 \pm 0.10}$	$\mathbf{4.42 \pm 0.11}$	**4.25**

Table 4: Human validation of extracted sub-stages across $n = 50$ videos and 5 raters. Each video was rated by all 5 raters on a 1–5 Likert scale. We report mean ratings with 95% confidence intervals for individual sub-stage rendering quality (averaged across all sub-stages of a video) and temporal ordering.

$\mathcal{S}$ holds the leaf-level similarities $\mathbf{C}_{k,\pi(k)}$ used by $r^{(i)}_{\text{align}}$ and $r^{(i)}_{\text{comp}}$, and $\mathcal{N} = \mathcal{N}^{(i)}$ is the set of matched internal nodes used by $r^{(i)}_{\text{order}}$. Descendants of a sub-threshold match are never recursed into and count as unmatched: their leaves enter the denominators of $r^{(i)}_{\text{align}}$ and $r^{(i)}_{\text{comp}}$ but not the numerators. The Hungarian step has cost $O((K + M)^3)$ per node; the branching factor at every level is small in our datasets (typically ≤ 5), so alignment cost is negligible compared to video generation.

C More Qualitative

Please refer to our supplementary webpage (.html) file in our supplementary material for more qualitative examples and video samples mentioned in the main paper.

D Extracted Stage validation

To validate that the sub-stages extracted by our pipeline correspond to events that humans actually perceive in the underlying videos, we conducted a small-scale human study on 50 randomly sampled videos spanning six physics domains. Each video was rated by 5 annotators on a 1–5 Likert scale along two axes: how clearly each sub-stage was depicted and whether the sub-stages occurred in the listed chronological order. Table 4 reports the results. The mean stage-quality rating of 4.21/5 and ordering rating of 4.42/5 indicate that our extracted decompositions are faithful to the videos at both the level of individual sub-stages and their temporal arrangement. Scores are consistent across domains, with optics events scoring lowest (3.88), reflecting the difficulty of extracting discrete temporal stages from scenes dominated by events such as reflection and scattering. Furthermore, We conduct a human evaluation survey on whether extracted stages are physically valid. Each of the 50 videos was rated by all 5 raters, giving 250 individual ratings, and the resulting estimates are precise with mean pairwise percent agreement of 0.853 , with inter-rater agreement of 0.861 using Gwet's AC2.

Additionally, to validate physical structure, 10 physics graduate raters evaluated decompositions for 20 canonical events, with 10 corrupted trees interspersed and unmarked; 5 with stages shuffled out of chronological order, 5 with a physically impossible stage inserted. Raters were told only that some decompositions contained errors. Table 5 shows the average score collected from these 10 raters. As seen, the extracted trees score comparably and the corrupted variants scores are low, with the largest gap on ordering, where shuffling the stages is the most direct violation of physical sequence. This confirms that raters were evaluating physical validity, and that our reference decompositions reflect the physical structure of each event.

E Hierarchical vs. flat rewards.

We remove the sub-stages from each reference tree, keeping only the top-level stages. This gives a flat reward and is the depth-one special case of our formulation. We evaluate both variants on a

	Physical correctness	Ordering	Completeness
Extracted tree	**4.12**	**3.78**	**3.86**
Corrupted tree	2.23	1.80	2.40

Table 5: Human evaluation results for extracted versus corrupted stage trees.

Benchmark	Variant	PC	SA
VideoPhy2	Hierarchical	66.0	52.0
VideoPhy2	Flat	63.0	47.0
MultiPhyBench	Hierarchical	59.0	44.0
MultiPhyBench	Flat	47.0	29.0

Table 6: Ablation on reward structure

subset of 100 prompts in table 6. The flat reward barely affects VideoPhy2 but substantially degrades performance on MultiPhyBench.

F Ablation on individual alignment, completeness, and ordering

We ablate each term of the hierarchical reward (Eq. 5) and evaluate on MultiPhyBench using three diagnostics, each targeting one term. Omission rate measures the fraction of physical principles absent from the video, ordering accuracy is the fraction of matched sub-stages in correct relative sequence, and stage match is the mean similarity between reference and observed stages. Reported scores are over 30 videos per variant, evaluated by Gemini Flash. To confirm these automatic diagnostics are reliable, five human raters independently scored the videos drawn evenly across variants; agreement between the human and automatic scores are also reported at the bottom of Table 7. Removing a term degrades its own diagnostic far more than the others, confirming that the three terms supervise

Reward Variant	Omission rate (%) ↓	Ordering accuracy (%) ↑	Stage match (%) ↑
Full	**6.1**	**86.0**	**81.0**
(−) ordering	6.8	61.0	72.0
(−) completeness	17.4	84.0	79.0
(−) alignment	9.7	83.0	68.0

Table 7: Gemini–human agreement is computed as the fraction of items on which the automatic and human labels match: 0.84 (omission), 0.77 (ordering), 0.82 (stage match).

distinct properties. Two smaller cross-effects are expected. Out-of-order processes produce less coherent extracted stage descriptions, which lowers stage match, and alignment is normalized by the total reference leaf count, so it carries a soft completeness signal.

G Ablation on Normalization vs. scaling

The normalization and scaling serve two distinct purposes in our formulation, isolated in Table 8 below. Per-stream normalization controls scale and variance differences among reward streams across the rollouts for a prompt. Each stream is standardized against its own mean and standard deviation over the group, so every principle contributes comparably to the advantage. Without it, a stream that varies more widely dominates and the policy maximizes that one term while ignoring the rest, degrading alignment and multi-process scenes (second row). The $\frac{1}{\sqrt{n}}$ rescaling addresses a different problem. Summing more terms inflates the variance of the advantage, so a three-principle sample pushes the gradient harder than a one-principle sample; the rescaling holds this variance constant. It also preserves the balance between the two reward families, since the advantage combines one global stream with a sum of n local ones. Without the factor the local sum overwhelms the global term and the policy loses the signal anchoring generation to prompt semantics and visual coherence, visible as the SA drop in the third row. At $N = 1$ the factor is inactive and per-stream normalization has little

Reward Configuration	PC (N=1)	PC (N=2)	PC (N=3)	Overall SA
Full	72.1	**63.0**	**48.0**	**58.3**
w/o per-stream norm	70.3	38.0	22.3	21.6
w/o $1/\sqrt{n}$ rescaling	72.8	41.3	37.6	35.0

Table 8: Ablation results on normalization and rescaling components.

to balance, with only one local stream alongside the global one; the differences emerge as principles co-occur. The evaluation is done over 100 prompts.

H User Study

To complement the automatic evaluations, we analyze survey results for videos containing multiple concurrent physical principles. We report the average ratings from 50 participants who evaluated the depiction of each physical event; a representative question from this study is shown in Fig. 8. The user study included 30 distinct prompts across 5 different methods. As illustrated in Fig. 7, A–C, our method consistently achieves the highest scores across all principles. Notably, while baseline performance is often unstable, dropping specific principles entirely within a scene or the inability to generate all of them plausibly, our approach maintains balanced adherence across the board. This validates our per-process reward design, which successfully allocates a distinct learning signal to each physical event.

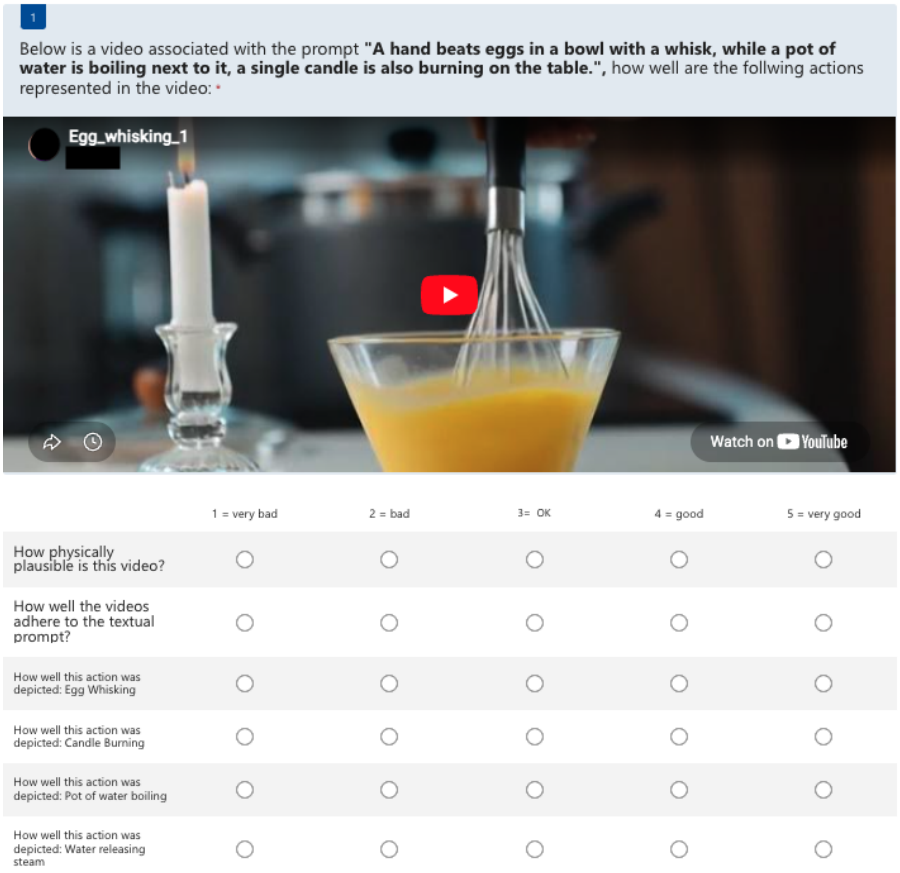

Figure 8: **User study sample question**

Table 9: Comprehensive evaluation across VideoPhy2, PhyWorldBench, Multi-concept, and VBench along with their standard deviations.

	VideoPhy2			**PhyWorldBench**			**Multi-concept**			**VBench**	
Method	PC↑	SA↑	Both↑	PC↑	SA↑	Both↑	PC↑	SA↑	Both↑	Dyn.↑	App.↑
Wan2.1	49.10±2.06	43.60±2.04	30.50±1.90	47.7±2.91	40.9±2.86	23.7±2.48	53.0±2.88	35.0±2.75	27.0±2.56	0.808±0.014	0.573±0.003
PnP	80.1±1.64	68.0±1.92	60.14±2.02	56.1±2.89	48.5±2.91	31.2±2.70	57.0±2.86	37.0±2.79	29.0±2.62	0.844±0.013	0.549±0.003
VideoRepa	70.9±1.87	53.5±2.05	43.8±2.04	51.7±2.91	33.3±2.74	19.3±2.30	49.0±2.89	22.0±2.39	15.0±2.06	0.863±0.012	0.550±0.003
PhyT2V	79.0±1.68	65.0±1.96	55.0±2.05	40.3±2.86	47.5±2.91	22.7±2.44	50.0±2.89	26.0±2.53	15.0±2.06	0.872±0.012	0.500±0.003
WISA	68.0±1.92	59.7±2.02	44.3±2.05	56.3±2.89	50.7±2.91	30.7±2.69	52.5±2.88	43.0±2.86	24.5±2.48	0.859±0.012	0.567±0.003
GPT-5	69.2±1.90	60.9±2.01	43.1±2.04	56.9±2.88	52.5±2.91	31.8±2.71	51.5±2.89	44.0±2.87	24.5±2.48	0.882±0.011	0.531±0.003
Ours (T2V)	**88.6**±1.31	**79.2**±1.67	**68.8**±1.91	**74.9**±2.52	**76.3**±2.48	**61.6**±2.83	**76.7**±2.44	**68.85**±2.67	**52.7**±2.88	**0.905**±0.010	**0.592**±0.003
Qwen 2.5	60.4±2.01	62.9±1.99	41.5±2.03	58.6±2.87	58.3±2.87	44.5±2.89	50.3±2.89	39.2±2.82	19.1±2.27	0.874±0.012	0.517±0.003
Ours (global)	72.7±1.83	**75.3**±1.78	56.3±2.04	62.3±2.82	**64.3**±2.79	54.3±2.90	61.0±2.82	53.0±2.88	40.0±2.83	0.871±0.012	**0.576**±0.003
Ours (local, sum)	72.2±1.84	70.1±1.88	48.5±2.06	61.9±2.83	57.3±2.88	37.1±2.81	67.0±2.71	57.9±2.85	42.4±2.85	0.865±0.012	0.556±0.003
Ours (local, normalized)	**79.2**±1.67	70.4±1.88	**57.3**±2.04	**65.6**±2.77	60.8±2.84	**52.5**±2.91	**68.5**±2.68	**58.3**±2.85	**44.5**±2.87	**0.893**±0.011	0.567±0.003

I Error Bars for the quantitative comparisons

Table 9 includes all the standard deviations as the error bar for the main quantitative comparisons and ablations. To enhance readability, we present the error bars in the supplementary rather than the main table.

J Stage Generation Prompt

We use the following prompt to instruct an LLM to decompose each video description into distinct physical processes and per-process temporal stages. The output is parsed as JSON and used as the supervision target $T^{*(i)}$ for our hierarchical training (Sec. 4.2).

Listing 1: Stage generation prompt used to produce reference stage trees $T^{*(i)}$.

```
{Stage Generation Prompt}
You are a physics expert analyzing a video. Given the description and
    ↪ physics context below, decompose the video into distinct
    ↪ physical processes, and break down each process into its own
    ↪ ordered stages.

**Video Description:**
{prompt}

**Category:** {category}



**Instructions:**
1. Identify all distinct physical processes or interacting entities
    ↪ occurring in the video (e.g., "steel_marble_kinematics" and
    ↪ "viscous_honey_displacement").
2. For EACH process, decompose its temporal progression into
    ↪ chronological stages.
3. Stages within a process must represent physically meaningful
    ↪ transitions (e.g., initial state, momentum transfer,
    ↪ deformation, settling).
4. Provide 2-5 stages per process, depending on its complexity.
5. Each stage name should be a concise identifier (2-4 words,
    ↪ snake_case).
6. The description must specify the observable forces, energy
    ↪ transfers, or material interactions unique to that specific
    ↪ process.

**Respond in this exact JSON format and nothing else:**

{
    "processes": [
        {
            "process_name": "identifier_for_process_1",
            "physics_domain": "rigid_body | fluid_dynamics |
                ↪ thermodynamics | deformation | optics | acoustics
                ↪ | biological | electromagnetic | multi_domain",
            "stages": [
                {"stage": 1, "name": "stage_name", "description":
                    ↪ "What physically happens and what is visually
                    ↪ observable for this entity"},
                {"stage": 2, "name": "stage_name", "description":
                    ↪ "..."}
            ]
        },
        {
            "process_name": "identifier_for_process_2",
            "physics_domain": "...",
            "stages": [
                {"stage": 1, "name": "...", "description": "..."},
                {"stage": 2, "name": "...", "description": "..."}
            ]
        }
    ]
}
```

Table 10: Video Prompts

A sponge is squeezed over a bowl of water while bubbles rise in a fish tank beside it.

HiPhy Expanded: A saturated sponge is gripped and compressed between both hands, the applied pressure collapsing the open pore structure and forcing the water stored within out through the surface in multiple thin streams that converge and fall into the bowl below, the sponge body deforming visibly under the hand pressure and springing partially back as the grip relaxes and air rushes back into the reopening pores. Beside it in a fish tank an air stone releases a continuous stream of small air bubbles from the tank base, each bubble rising through the water column and expanding slightly as the water pressure decreases with decreasing depth, the bubbles reaching the surface and bursting to release small ripples that spread across the tank water surface.

Table 11: Video Prompts

A hammer strikes a nail into wood while a pot of water boils on the stove beside it.

HiPhy Expanded: A hammer is swung downward and its steel head makes hard contact with the nail head, driving the nail shaft into the wood grain with each strike as the wood fibers compress and split around the penetrating metal, the nail advancing incrementally with each blow while the hammer rebounds slightly from the elastic response of the steel on steel contact. Simultaneously on the stove beside the workbench a pot of water boils vigorously, with dense clusters of bubbles nucleating at the heated base and bursting at the surface releasing continuous steam, the vibration transmitted through the floor from each hammer blow causing the boiling water surface to ripple and the steam column above to shudder briefly with each impact.

K Video Prompts

Tables 10 to 25 present the sample prompts generated by HiPhy that are associated with the visuals in the main paper.

Table 12: Video Prompts

A balloon floats upward while steam rises from a pot on the stove below it.

HiPhy Expanded: A helium-filled balloon floats upward toward the ceiling, the buoyancy force from the less dense helium inside exceeding the downward gravitational pull on the balloon's total mass, the balloon drifting upward and rotating slowly as minor air currents in the room push it laterally. Below it on the stove a pot of boiling water releases a continuous column of steam upward from the surface, the hot water vapor rising rapidly due to its lower density than the surrounding cooler air, the rising steam column reaching the underside of the floating balloon and deflecting outward around its curved surface, slightly accelerating the balloon's upward drift as the warm rising air adds to the buoyancy force beneath it.

Table 13: Video Prompts

A colorful rubber ball is dropped from a height, showing the bounce of the ball as it makes contact with the hard floor.

HiPhy Expanded: the video captures view of one colorful rubber ball being dropped from a height. the ball is colorful, creating a vibrant and eye-catching display. as the ball descends, it bounces off the hard floor, causing to bounce lightly. the video ends with the ball lying flat on the ground, demonstrating the energy and impact of the fall.

Table 14: Video Prompts

Knife skillfully slices an apple.

HiPhy Expanded: the knife skillfully slices through the juicy apple, which is positioned upright on a wooden cutting board. the apple is bright red and glistening with moisture, indicating its freshness. as the knife cuts through the apple, it lifts off the cutting board, revealing the depth of the cut. the overall scene emphasizes the artistry involved in slicing through a ripe apple with precision.

Table 15: Video Prompts

A leaf falls from a tree while rain hits a puddle below.

HiPhy Expanded: A dry leaf detaches from a branch and falls under gravity, its flat irregular shape and low mass making it susceptible to air resistance and causing it to tilt, rotate, and drift laterally in a slow oscillating descent rather than falling straight, the leaf tumbling end over end and drifting sideways as pockets of air deflect its trajectory unpredictably. Below it rain strikes a puddle on the ground at high velocity, each raindrop impact generating a small crown splash with thin water jets ejecting upward before collapsing, and sending circular ripple rings radiating outward across the puddle surface, the falling leaf eventually landing on the puddle surface where surface tension briefly supports it before it becomes waterlogged and sinks flush with the water surface.

Table 16: Video Prompts

A glove catching the fast-moving baseball.

HiPhy Expanded: the video captures a close-up view of a glove catching a fast-moving baseball. the glove is positioned on the left hand side of the frame, while the baseball travels from right to left. as the baseball approaches, the glove extends its fingers to grip the ball firmly. the glove's movement is smooth and deliberate, ensuring a secure catch. the background is blurred, revealing a playing field with grass and possibly other sports equipment visible in the distance. the lighting suggests it might be daytime, adding to the dynamic atmosphere of the scene.

Table 17: Video Prompts

Gymnast performs an aerial somersault at sunrise.

HiPhy Expanded: A gymnast sprints forward and plants both feet simultaneously into the ground, driving both knees upward and swinging both arms overhead to convert forward momentum into vertical lift, launching the body upward into the air with no hand contact with the ground. At peak height the gymnast pulls both knees tightly into the chest and wraps both arms around the shins, tucking the body into a compact rotating mass that completes one full forward rotation around a horizontal axis through the hips. As the body descends and the rotation completes, the legs extend downward and outward to prepare for landing, both feet striking the ground simultaneously with knees bent deeply to absorb the downward impact force, the body straightening upright to hold the landing position against the warm early morning sky.

Table 18: Video Prompts

A skateboarder performs jumps while splashing through a large puddle on a street.

HiPhy Expanded: the video captures a dramatic scene of a skateboard splashing through a large puddle on a street. the skateboarder is dressed in a black shirt, blue jeans, and white sneakers. the skateboard, featuring a yellow and black striped deck, is positioned on the left side of the frame. as the video progresses, the skateboarder begins to accelerate, causing the skateboard to splash vigorously through the puddle. the splashes create a dramatic contrast against the dark background of the street. the skateboarder's movement adds to the sense of chaos and urgency in the scene. the video concludes with the skateboarder continuing to ride through the puddle, creating a continuous stream of water as it flows downstream. the background remains dark, emphasizing the vibrant colors and movement of the water and the skateboarder. there are no animals or human characters present in the video, focusing solely on the interaction between the skateboarder, the water, and the urban setting. the video captures the raw energy and fun of skateboarding in a dynamic outdoor environment.

Table 19: Video Prompts

A stick of incense burns while honey drips slowly from a spoon beside it.

HiPhy Expanded: A stick of incense burns at its tip with a small glowing ember, releasing a thin continuous thread of white smoke that rises straight upward in a slow coherent column before breaking into irregular curling wisps as it reaches the turbulent ambient air above, the smoke dispersing and spreading outward as it cools and loses buoyancy. Beside it honey drips from a spoon in a slow thick strand that stretches and thins under gravity, the high viscosity of the honey resisting the flow and causing the strand to elongate into a long thin thread before detaching and falling, the detached drop coiling slowly as it lands on the surface below while the rising smoke from the incense bends slightly toward the warmth of the honey.

Table 20: Video Prompts

Camera focuses on a toothpaste tube while a hand squeezes a steady stream of toothpaste.

HiPhy Expanded: A finger presses down on the flexible body of a toothpaste tube, deforming the plastic casing inward and forcing the viscous paste toward the nozzle where it resists briefly before yielding and extruding in a smooth continuous ribbon that lands and coils onto the bristles of a toothbrush held below. The ribbon conforms slightly to the gaps between the bristle tufts under its own weight without spreading or dripping, and when finger pressure is released the flow stops abruptly with the paste holding its coiled shape on the brush due to its yield stress preventing further movement.

Table 21: Video Prompts

A book slides off a tilted shelf and falls to the ground while a small rubber ball bounces repeatedly on the desk.

HiPhy Expanded: the video captures a dynamic scene set on a desk. the focal point is a book that slides off a tilted shelf. as the book falls, it creates a dramatic contrast against the plain white background of the desk. in the background, a small rubber ball bounces on the desk, adding a playful element to the otherwise focused scene. the overall atmosphere is one of tension and surprise, with the sudden fall of the book creating a stark contrast against the static background.

Table 22: Video Prompts

Coffee accepting a gentle pour of milk.

HiPhy Expanded: the video depicts a close-up view of a coffee mug. inside the mug, there is a small amount of coffee. the mug is being gently poured into by a hand holding a clear glass bottle. the liquid being poured appears to be milk or a similar liquid. as the liquid is poured, it creates a gentle frothy mixture inside the mug. the overall scene remains static throughout the video, focusing on the interaction between the liquid being poured and the existing contents of the mug.

Table 23: Video Prompts

Sauce simmers in a pan on the stove releasing steam, at the ironing board an iron presses down on a shirt, and through the kitchen window wind drives rain against the glass.

HiPhy Expanded: A cabin interior frames two adjacent tables and a rain streaked window. on the left, a saucepan of tomato sauce undergoes low intensity convective boiling on an active burner, with small bubbles nucleating at the base and bursting at the viscous surface, releasing intermittent steam plumes that rise and disperse into the warm air. On the right table, a cotton shirt is pressed under a heated iron driven in firm, repetitive linear strokes. through the back window, dense near vertical rainfall streaks downward under gravity, producing crown splashes and ripple rings on pooling backyard surfaces while irregular rivulets distort the gloomy view through the glass.

Table 24: Video Prompts

A hand beats egg in a bowl with a whisk, while a pot of water is boiling next to it, a single candle is also burning on the table

HiPhy Expanded: On a kitchen counter top, A hand beats egg white in a bowl with a whisk, the liquid thickening into white foam with each stroke. On the stove beside it, a pot of water is boiling with bubbles breaking at the surface and releasing steady puffs of steam. On the counter top next to the stove, a single candle burns on the table with a small steady flame.

Table 25: Video Prompts

A coin dances in a spiral on the table

HiPhy Expanded: A coin dances in a spiral on the table.": "the video captures a close-up view of a coin dancing in a spiral on the table. the coin is silver and features a portrait of queen elizabeth II on its obverse side. the background shows a plain white surface, likely a table or countertop. as the video progresses, the coin continues to dance in a spiral pattern. the spiral is tight at the beginning and gradually expands as the coin moves further away from its initial position. the movement of the coin is smooth and continuous, creating a visually appealing and mesmerizing dance.