PhyMotion: Structured 3D Motion Reward for Physics-Grounded Human Video Generation

Yidong Huang[1*] **Zun Wang**[1*] **Han Lin**[1] **Dong-Ki Kim**[2] **Shayegan Omidshafiei**[2]
Jaehong Yoon[3] **Jaemin Cho**[4,5] **Yue Zhang**[1] **Mohit Bansal**[1]

[1]UNC Chapel Hill [2]FieldAI [3]NTU Singapore [4]AI2 [5]Johns Hopkins University

https://phy-motion.github.io/

Abstract

Generating realistic human motion is a central yet unsolved challenge in video generation. While reinforcement learning (RL)-based post-training has driven recent gains in general video quality, extending it to human motion remains bottlenecked by a reward signal that cannot reliably score motion realism. Existing video rewards primarily rely on 2D perceptual signals, without explicitly modeling the 3D body state, contact, and dynamics underlying articulated human motion, and often assign high scores to videos with floating bodies or physically implausible movements. To address this, we propose PhyMotion, a structured, fine-grained motion reward that grounds recovered 3D human trajectories in a physics simulator and evaluates motion quality along multiple dimensions of physical feasibility. Concretely, we recover SMPL body meshes from generated videos, retarget them onto a humanoid in the MuJoCo physics simulator, and evaluate the resulting motion along three axes: kinematic plausibility, contact and balance consistency, and dynamic feasibility. Each component provides a continuous and interpretable signal tied to a specific aspect of motion quality, allowing the reward to capture which aspects of motion are physically correct or violated. Experiments show that PhyMotion achieves stronger correlation with human judgments than existing reward formulations. These gains carry over to RL-based post-training, where optimizing PhyMotion leads to larger and more consistent improvements than optimizing existing rewards, improving motion realism across both autoregressive and bidirectional video generators under both automatic metrics and blind human evaluation (+68 Elo gain). Ablations show that the three axes provide complementary supervision signals, while the reward preserves overall video generation quality with only modest training overhead.

1 Introduction

Recent video generation models [Ho et al., 2022, OpenAI, 2024, Kong et al., 2024, Team Wan et al., 2025] produce increasingly photorealistic outputs, yet generating realistic human motion remains one of the most significant open challenges. While scene textures, lighting, and camera motion have reached impressive fidelity [Zheng et al., 2025, Liu et al., 2024, Gao et al., 2025], modeling human motion remains fundamentally challenging due to the tight coupling between body articulation and physical constraints. As a result, generated videos frequently exhibit artifacts like floating feet, limb penetrations, anatomical distortions, and movements that defy basic laws of physics [Zheng et al., 2025, Yang et al., 2026]. As human-centric content constitutes one of the most important application domains, from entertainment and education to virtual communication [Zhu et al., 2024b, Sui et al., 2026], improving human motion quality is a pressing goal.

*Equal contribution.

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

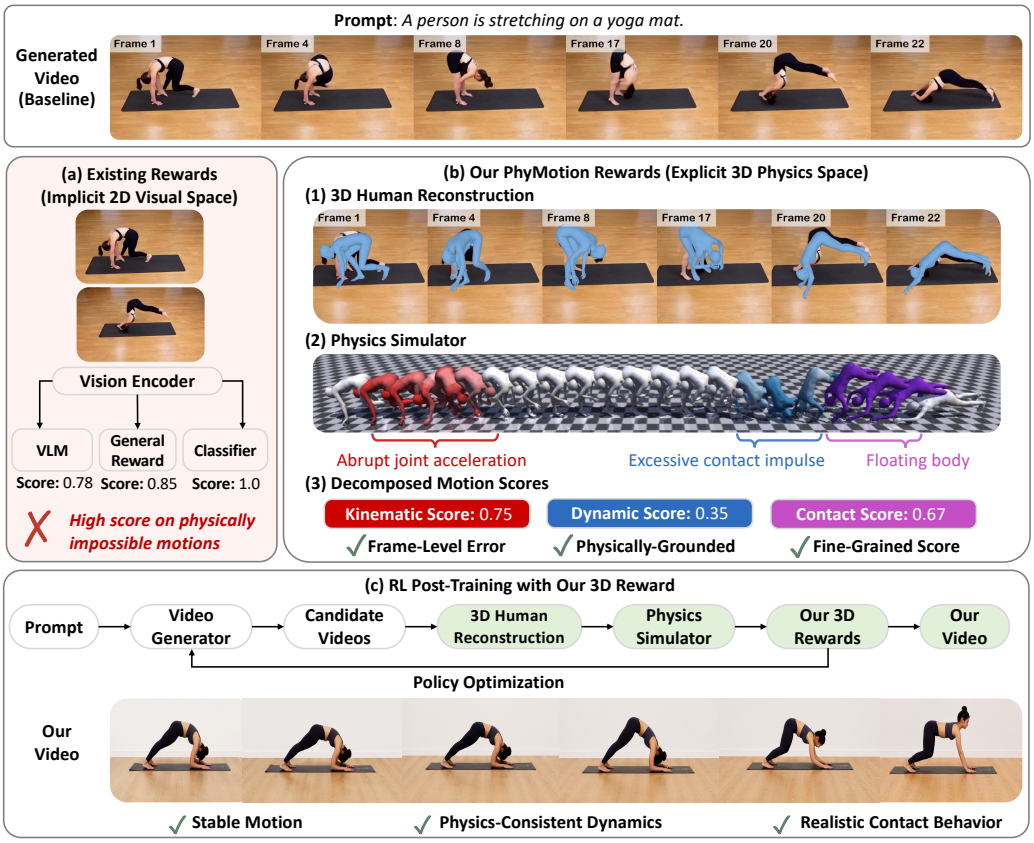


Figure 1: **Overview of PhyMotion.** Existing metrics operate in 2D pixel space (VLMs [Liu et al., 2025b], frame-level classifiers [Zheng et al., 2025]), producing misleading scores that are structurally blind to articulated motion. We recover SMPL meshes from generated videos and place them in a physics simulator, yielding decomposed, physically-grounded motion scores. The structured reward drives RL-based post-training (bottom), consistently improving human motion quality.

A promising direction is reinforcement learning (RL)-based post-training, which has recently demonstrated clear gains on general visual quality and text alignment [Black et al., 2024, Xue et al., 2025, Zhang et al., 2026]. The success of RL, however, hinges on the reward signal, and for human motion, existing rewards remain inadequate. Specifically, they broadly fall into three categories, each with a distinct failure mode: *(i) frame-level 2D classifiers* [Huang et al., 2024, Zheng et al., 2025, Motamed et al., 2026] detect localized artifacts but miss trajectory-level failures; *(ii) vision-language evaluators* [Meng et al., 2025, Chow et al., 2025, Hong et al., 2025] provide coarse semantic judgments unsuitable as dense RL rewards; and *(iii) general preference rewards* [Bansal et al., 2025, Liu et al., 2025b, He et al., 2024, Ma et al., 2025] can hallucinate quality from appearance or prompt alignment rather than motion plausibility. Despite their differences, these rewards primarily rely on pixel-space perceptual representations and learned visual features. While effective for appearance quality and semantic alignment, we empirically observe that they still struggle to reliably capture key aspects of articulated human motion and physical feasibility, such as kinematic consistency, anatomical constraints, and contact dynamics (see Sec. 3.1). Consequently, they can assign high scores to videos with clear physical violations (Fig. 1(a) and Fig. 2).

To address this limitation, we propose PhyMotion, a structured, fine-grained motion reward for human video generation that grounds 3D human trajectories in a physics simulator to evaluate motion quality across multiple dimensions of physical feasibility (Fig. 1(b)). We lift each generated video into a structured 3D representation by recovering an SMPL [Loper et al., 2015] body mesh and grounding it in a physically consistent human model in the MuJoCo simulator [Todorov et al., 2012]. This enables access to structured physical signals (e.g., contact consistency, joint-level kinematics, and motion-driving torques) that are not explicitly captured in pixel-based representations, allowing evaluation based on whether the motion satisfies physical constraints, which directly underlie many

visually implausible artifacts. From these observables, we derive three structured feasibility scores: **kinematic** (joint consistency), **contact/balance** (interaction correctness), and **dynamic** (force and motion consistency). Each score targets a distinct failure mode identified in Sec. 3.1. To test whether our physics-grounded scores better capture perceived motion quality, we conduct a pairwise human study in which raters compare pairs of generated videos according to motion realism, and measure how well each metric predicts these preferences. We compare PHYMOTION against prior 2D perceptual metrics and learned video reward models. We further use the same scores as a dense and interpretable reward for RL-based post-training (Fig. 1(c)), examining whether optimizing physical feasibility leads to improved human motion generation. Because the reward decomposes into physically meaningful components, it provides transparency into which aspects of motion are optimized, enabling fine-grained diagnosis of motion failures and revealing improvements across different physical dimensions.

Experiments demonstrate that PHYMOTION serves as both a reliable motion evaluator and an effective reward for RL-based video post-training. As an evaluator, PHYMOTION achieves 80% average pairwise agreement with human judgments and the highest aggregate Spearman correlation (ρ=0.376) on $1{,}200$ video pairs, across multiple aspects of motion quality including body structure, balance, and motion naturalness. This substantially outperforms existing perceptual, preference-based, and physics-aware reward baselines, which typically achieve 50–66$\%$ agreement and only weak Spearman correlations ($\rho = 0$–0.25). As a reward, PHYMOTION further enables effective RL-based post-training across both autoregressive [Zhu et al., 2026] and bidirectional [Zhang et al., 2025] video generators. Additionally, our reward also improves scores on external evaluators, including VBench metrics [Huang et al., 2024], VideoAlign [Liu et al., 2025b], and VideoPhy-PC [Bansal et al., 2025], by an average of 7.1%, while achieving consistent gains across all three physical feasibility dimensions. Human preference eval further confirms these improvements: our post-trained model achieves the highest Elo scores on body structure, balance, motion naturalness, and overall preference, outperforming all baselines, including the larger Wan2.2 14B model. Meanwhile, our ablation studies show that each reward component improves its corresponding physical dimension, while the combined reward achieves the best overall trade-off across all dimensions. Finally, we further show that training with PHYMOTION preserves general video generation quality on VBench and VBench-2.0 [Huang et al., 2024, Zheng et al., 2025], while introducing only modest overhead through pipelined reward computation.

2 Related Work

RL-based post-training and video reward models. Reinforcement learning has become a central paradigm for post-training generative models [Ouyang et al., 2022, Christiano et al., 2017, Shao et al., 2024b, Guo et al., 2025]. For diffusion, DDPO [Black et al., 2024], DPOK [Fan et al., 2023], and Diffusion-DPO [Wallace et al., 2024] optimize policy gradients or preference objectives on reverse sampling, while DanceGRPO [Xue et al., 2025], Flow-GRPO [Liu et al., 2025a], and DiffusionNFT [Zheng et al., 2026] extend on-policy RL to video generators; recent work further scales RL to distilled autoregressive video models [Zhang et al., 2026, Wang et al., 2026a, Lu et al., 2026, He et al., 2025]. The reward itself is typically a learned scalar over general human preferences, e.g., HPSv3 [Ma et al., 2025], VideoReward [Liu et al., 2025b], VisionReward [Xu et al., 2026], and VideoScore [He et al., 2024], or an aggregate of VBench-derived feature-model scores [Huang et al., 2024, 2026]. Such rewards are structurally agnostic to articulated human motion and provide no dimension-specific feedback. In contrast, we derive the reward from a structured 3D body model, enabling the RL optimization signal to incorporate explicit physical priors on human kinematics, balance, and dynamics.

Human motion in video generation and evaluation. A growing line of work improves human motion by directly modifying the generator or incorporating motion priors. Pose-conditioned methods [Wang et al., 2024, Ma et al., 2024, Xu et al., 2024, Hu, 2024, Zhou et al., 2024, Zhu et al., 2024a, Shao et al., 2024a, Wang et al., 2026b,c] condition on external 2D poses or 3D SMPL priors, while VideoJAM [Chefer et al., 2025] and EchoMotion [Yang et al., 2026] co-predict motion representations with video. However, these methods typically require architectural changes, external pose conditioning, or additional motion data supervision. On the evaluation side, beyond distribution-level statistics [Salimans et al., 2016, Heusel et al., 2017, Unterthiner et al., 2019], recent benchmarks [Huang et al., 2024, 2026, Zheng et al., 2025, Liu et al., 2024, 2023, Ling et al., 2025] decompose evaluation into interpretable dimensions, but still rely largely on perceptual video cues, providing only indirect

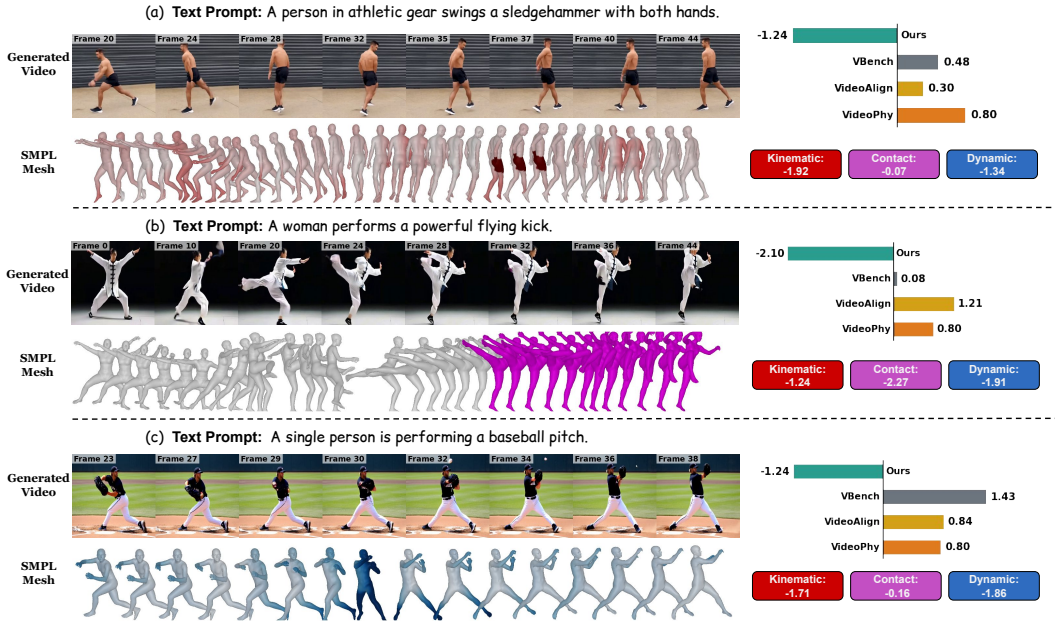


Figure 2: **PhyMotion identifies distinct physical failure modes overlooked by existing 2D metrics.** Each example shows a generated video and its recovered SMPL trajectory. Colored SMPL frames highlight the dominant motion errors: kinematic (articulated-body inconsistency), contact (unstable body–environment interaction), and dynamic (physically infeasible motion). All reported metric values are quartile-normalized z-scores.

evidence of articulated motion quality. VLM-based evaluators [Bansal et al., 2025, Huang et al., 2025, Meng et al., 2025, Wang et al., 2025b, Sun et al., 2025, Hong et al., 2025] probe physical commonsense and motion understanding, but remain limited for fine-grained human articulation. Related 3D motion evaluators such as MBench [Lin et al., 2026] and MotionCritic [Wang et al., 2025a] assess motion quality directly, but are not designed as rewards for synthetic videos. Our work connects generation and evaluation through a structured 3D motion reward for human video generation. Without modifying the generator architecture, we use PhyMotion for RL post-training, providing continuous, physics-grounded supervision that targets distinct motion failure modes.

3 Physics-Grounded 3D Motion Evaluation

We present our evaluation in two parts: we first characterize the failure modes of human motion that are fundamentally unidentifiable from pixel observations (Sec. 3.1). We then introduce a physics-grounded evaluation protocol that recovers 3D motion along three feasibility axes (Sec. 3.2).

3.1 The Failure Modes of 2D Motion Evaluation

Fig. 2 illustrates a key limitation of existing 2D motion evaluators, such as VBench [Huang et al., 2024, Zheng et al., 2025], VideoAlign [Liu et al., 2025b], and VideoPhy [Bansal et al., 2025]. Although these metrics can assign high scores based on perceptual quality, text alignment, or visual commonsense, the generated videos may still contain physically implausible human motion. These failures require reasoning about human bodies' physical quantities beyond RGB appearance: whether the 3D body configuration is anatomically valid over time, whether the body is properly supported by ground contacts, and whether the motion can be produced by plausible forces and torques. Therefore, we identify three classes of motion failures that existing 2D evaluators struggle to reliably capture.

Kinematic inconsistency over time. Frame-level visual metrics may judge individual poses as plausible, but fail to detect whether the recovered body configuration is anatomically valid in 3D. Such metrics can miss abnormal joint configurations, unrealistic joint velocities, and self-penetrations that become clear only after reconstructing the articulated body. As shown in Fig. 2(a), although the

frames appear to show a swing, the reconstructed 3D body reveals that the hand penetrates the hip, a self-penetration error that existing 2D rewards fail to detect (e.g., VideoPhy assigns a high normalized z-score of 0.8).

Physically inconsistent contact. 2D evaluators often score a video high even when the body is floating, sliding, penetrating the ground, losing balance, or changing contact state inconsistently over time. As shown in Fig. 2(b), the RGB frames appear to show a powerful flying kick, but the recovered 3D trajectory reveals an implausible support pattern: the body becomes airborne toward the end of the video. Such contact and balance failures are difficult for 2D rewards to detect (e.g. VideoAlign assigns a high normalized z-score of 1.21).

Dynamically infeasible motion. A video can appear smooth and semantically correct while still requiring physically implausible forces to execute. As shown in Fig. 2(c), the frames depict a reasonable baseball pitch, but the recovered 3D motion indicates that the body would require excessive joint torques to reproduce the observed pose transition. Such dynamic failures are difficult for existing 2D metrics to detect because they do not test whether the motion can be produced by a physically valid human body (e.g. VBench assigns a high normalized z-score of 1.43).

3.2 PhyMotion: Physics-Grounded 3D Rewards

To address the failure modes in Sec. 3.1, we convert each generated video into a physically interpretable 3D motion representation. Given a video $\{I_t\}_{t=1}^{T}$ at frame rate f, we recover an SMPL-X [Pavlakos et al., 2019] trajectory with GVHMR [Shen et al., 2024]. We denote the recovered pose by $\mathbf{q}_t$, the 3D body joints by $\mathbf{X}_t = \{\mathbf{x}_{t,j}\}_{j=1}^{J}$, and the joint angular velocity by $\mathbf{v}_{t,j} = f(\mathbf{q}_{t+1,j} - \mathbf{q}_{t,j})$. We retarget the motion to a MuJoCo [Todorov et al., 2012] human model with explicit mass, inertia, joint limits, and contact geometry, and run inverse dynamics [Winter, 2009] to estimate joint torques $\boldsymbol{\tau}_t$ and ground reaction forces $\mathbf{F}_t^{\mathrm{GRF}}$. We then score each video along three complementary axes. This conversion makes the failure modes in Sec. 3.1 directly measurable. As shown in Fig. 2, the recovered 3D trajectory lets us identify whether a video fails due to kinematic inconsistency, implausible contact and balance, or dynamically infeasible motion.

Kinematic feasibility. Kinematic feasibility measures whether the recovered body motion is smooth and anatomically valid. We combine three normalized violations: angular-velocity violation v_{vel}, computed by thresholding $\|\mathbf{v}_{t,j}\|_2$ against a clean-motion tolerance; self-penetration violation v_{spen}, computed as the fraction of frames with intersecting non-adjacent mesh triangles; and joint-limit violation v_{lim}, computed as the fraction of joints whose angles fall outside the valid MuJoCo range. The final score is $\mathcal{F}_{\mathrm{kin}} = 1 - \frac{1}{3}(v_{\mathrm{vel}} + v_{\mathrm{spen}} + v_{\mathrm{lim}})$, where higher values indicate smoother motion with fewer anatomical violations.

Contact feasibility. Contact feasibility measures whether the body interacts with the ground plausibly. We infer binary foot–ground contacts $c_{t,k} \in \{0, 1\}$ for each foot $k \in \{L, R\}$ from foot height and velocity, where L and R denote the left and right foot. We compute four normalized violations: foot sliding v_{slip}, which measures displacement of a contacted foot; ground penetration v_{gpen}, which measures how far the foot moves below the floor; foot floating v_{float}, which flags frames where neither foot is in contact while the body does not follow a plausible ballistic trajectory; and balance violation v_{bal} is the fraction of frames where the projected center of mass falls outside the support polygon of contacting feet. The score is $\mathcal{F}_{\mathrm{con}} = 1 - \frac{1}{4}(v_{\mathrm{slip}} + v_{\mathrm{gpen}} + v_{\mathrm{float}} + v_{\mathrm{bal}})$.

Dynamic feasibility. Dynamic feasibility measures whether the recovered motion can be replayed by a physically plausible human body. Using inverse dynamics in MuJoCo, we estimate the forces required to reproduce the trajectory and compute three scores: $s_{\boldsymbol{\tau}}$ penalizes unrealistically large joint torques, s_{GRF} penalizes excessive ground contact forces whose magnitude exceeds a maximum plausible threshold $F_{\max}$, and s_{met} penalizes motions with unusually high joint effort, measured by the torque–velocity work proxy $C_{\mathrm{met}} = \sum_{t,j} |\tau_{t,j}\mathbf{v}_{t,j}|$. The final score is $\mathcal{F}_{\mathrm{dyn}} = \frac{1}{3}(s_{\boldsymbol{\tau}} + s_{\mathrm{GRF}} + s_{\mathrm{met}})$, with higher values indicating more physically realizable motion.

Together, these axes cover articulation quality, environment interaction, and physical realizability, producing continuous and interpretable metrics for both evaluation and reward-based post-training. We provide the detailed reward definitions and implementation specifications in Appendix J.

4 RL Post-Training with PhyMotion

The evaluation metrics in Sec. 3.2 provide decomposed signals for different human motion failure modes. To use them as an optimization signal, we aggregate the three PhyMotion scores into a single motion reward:

$$R_{\text{motion}}(v) = \frac{1}{3}\left(\mathcal{F}_{\text{kin}}(v) + \mathcal{F}_{\text{con}}(v) + \mathcal{F}_{\text{dyn}}(v)\right). \tag{1}$$

This reward is then used as the optimization target in the policy-learning objective below.

Problem formulation. We frame the video generation model as a policy π_θ that maps a text prompt c to a generated video v. Our goal is to fine-tune π_θ to maximize the expected reward $R(v)$ while staying close to the base reference policy π_{ref} via a KL penalty:

$$\max_\theta \ \mathbb{E}_{c\sim\mathcal{D},\, v\sim\pi_\theta(\cdot|c)}\big[R(v)\big] - \lambda\,\mathrm{KL}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big), \tag{2}$$

where λ controls the strength of the KL regularization to prevent mode collapse and preserve general video generation capability. The prompt distribution $\mathcal{D}$ is a curated set of human-motion-specific prompts covering diverse actions and scenarios (details in Sec. 5.2).

Policy optimization. We adopt the forward-process RL formulation of DiffusionNFT [Zheng et al., 2026], following Astrolabe [Zhang et al., 2026]. Given a generated video $v_0 \sim \pi_\theta(\cdot \mid c)$ with normalized reward $\tilde{r} \in [0,1]$, a noisy version v^t is constructed at a randomly sampled timestep t. Using the current velocity predictor v_θ and the old predictor $v_{\theta_{\text{old}}}$, implicit positive and negative policies are defined via interpolation:

$$v^+ = (1-\beta)\, v_{\theta_{\text{old}}} + \beta\, v_\theta, \quad v^- = (1+\beta)\, v_{\theta_{\text{old}}} - \beta\, v_\theta, \tag{3}$$

where β controls the interpolation strength. The policy loss contrasts these implicit policies against the target forward velocity v_{target}:

$$\mathcal{L}_{\text{policy}} = \tilde{r}\,\|v^+ - v_{\text{target}}\|_2^2 + (1-\tilde{r})\,\|v^- - v_{\text{target}}\|_2^2. \tag{4}$$

This trajectory-free formulation requires only clean generated samples and avoids backpropagating through the reverse sampling chain, enabling memory-efficient and solver-agnostic training.

Reward integration. The reward $R_{\text{motion}}(v)$ in Equation (2) follows the equally weighted aggregation defined in Equation (1). Because the reward is decomposed into three independent axes, we can diagnose per-dimension improvement during training and detect reward hacking early. For example, if one axis improves disproportionately at the expense of others, the decomposition makes this immediately visible.

5 Experiments

We organize our experiments around two central questions. First, we evaluate whether our physics-grounded 3D metrics better align with human judgments of synthetic human video quality than existing video metrics (Sec. 5.1). Second, we use the same metrics as rewards for RL post-training and evaluate whether they improve human motion video generation (Sec. 5.2).

5.1 Human Alignment Results of PhyMotion

Experiment setup. We compare against existing evaluators, including VBench [Huang et al., 2024], VBench-2.0 [Zheng et al., 2025], VideoAlign [Liu et al., 2025b], and VideoPhy [Bansal et al., 2025]. To construct the annotation set, we sample prompts from the Motion-X dataset [Lin et al., 2023] and generate videos under the same prompt using six baseline video models: Wan-2.1 1.3B, Wan-2.2 5B, Wan-2.2 14B [Team Wan et al., 2025], Causal Forcing-1.3B [Zhu et al., 2026], EchoMotion-5B [Yang et al., 2026], and FastWan [Zhang et al., 2025]. For each sample, we form video pairs from different models based on the same prompt, so that annotators compare motion quality under identical text conditions rather than across different conditions. From these candidate pairs, we randomly select 1,200 comparisons for annotation. Each annotation presents two anonymized videos side by side with randomized left/right order, together with the shared text prompt. We recruit six annotators and ask them to judge which video is better along three criteria (body structure, balance, and motion

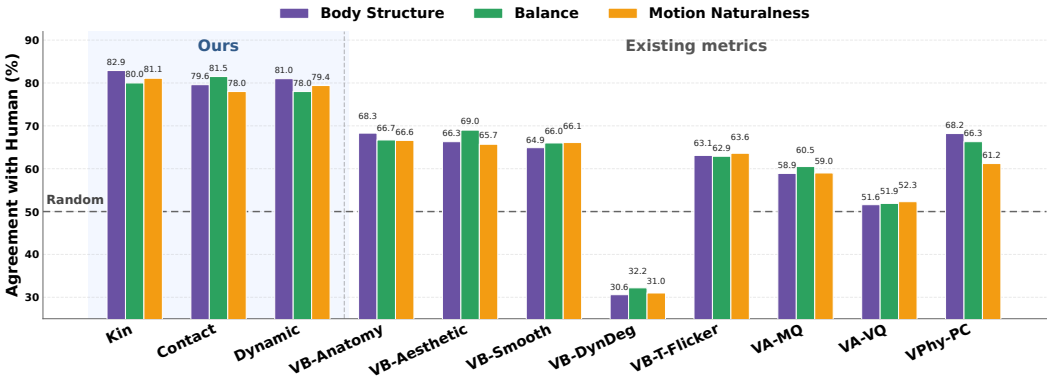


Figure 3: **Agreement with human judgments for motion quality.** Our metrics achieve the highest agreement compared with perceptual (VBench / VBench2) and learned (VideoAlign, VideoPhy) video metrics across three human-evaluation questions: body structure, balance, and motion naturalness.

Table 1: **Correlation with human judgments for motion quality.** We report Spearman's rank correlation (ρ) between automatic metric scores and human judgments across three questions. Best results are in **bold**, and second-best results are underlined.

Question	*VBench/VBench2*					*VideoAlign*		*VideoPhy*	*PhyMotion (Ours)*			
	Aes. Q.	Mot. Sm.	Temp. Fl.	Dyn. Deg.	Hum. An.	VA-MQ	VA-VQ	VP-PC	Kin.	Con.	Dyn.	Overall
Body Structure	+0.248	+0.236	+0.175	−0.135	+0.273	+0.155	+0.045	+0.168	+0.369	+0.290	+0.367	**+0.391**
Balance	+0.238	+0.246	+0.171	−0.118	+0.235	+0.175	+0.053	+0.142	+0.314	**+0.337**	+0.316	+0.333
Motion Naturalness	+0.260	+0.274	+0.208	−0.149	+0.278	+0.156	+0.067	+0.108	+0.375	+0.281	+0.389	**+0.402**
All Questions	+0.248	+0.252	+0.185	−0.135	+0.262	+0.161	+0.055	+0.138	+0.353	+0.290	+0.358	**+0.376**

naturalness) corresponding to the failure modes in Sec. 3.1. For each criterion, annotators choose among "Video A", "Video B", or "Tie", allowing ambiguous cases to be excluded from decisive metric-alignment comparisons. Additional annotation details are provided in Appendix I.

Quantitative results. As shown in Fig. 3 and Table 1, our physics-grounded metrics achieve the strongest and most consistent alignment with human judgments under both pairwise agreement and Spearman's rank correlation (ρ). In the agreement results, kinematic feasibility performs best on body structure (82.9%), while contact and dynamic feasibility remain consistently high across all criteria (78–81%); in contrast, existing perceptual and learned metrics mostly remain in the 50–66% range, with Dynamic Degree near chance. The same trend holds in correlation: overall feasibility (R_{motion}) achieves the highest aggregate correlation ($+0.376$), outperforming the strongest existing metric, VBench2 Human Anatomy ($+0.262$), as well as VideoAlign-MQ ($+0.161$) and VideoPhy-PC ($+0.138$). The individual components also align with their intended diagnostic roles: kinematic feasibility is strongest for body-structure judgments, contact feasibility is strongest for balance, and dynamic feasibility is highly correlated with motion naturalness, providing non-redundant diagnostic signals that serve as stable, interpretable rewards for RL-based post-training (Sec. 5.2). We further evaluate the robustness of PhyMotion to severe anatomical distortions and 3D reconstruction errors (Appendix F), as well as its extensibility to non-flat terrain (Appendix G).

5.2 RL Post-Training Results of PhyMotion

5.2.1 Experiment Setup

We apply our structured 3D motion reward to two distilled 4-step Wan 1.3B backbones: FastWan [Zhang et al., 2025] (bidirectional) and Causal Forcing [Zhu et al., 2026] (autoregressive). We recover SMPL motion with GVHMR [Shen et al., 2024] and compute physical feedback in MuJoCo [Todorov et al., 2012]. We fine-tune with LoRA adapters [Hu et al., 2022] using RL post-training; full hyperparameters are provided in Appendix K. Training uses 8 NVIDIA A100 80 GB GPUs for 330 steps, taking 2.5 days. We compare against EchoMotion [Yang et al., 2026], Causal Forcing 1.3B [Zhu et al., 2026], FastWan 1.3B [Zhang et al., 2025], and Wan models including Wan 2.1 1.3B, Wan 2.2 5B, and Wan 2.2 14B [Team Wan et al., 2025]. To compare against VLM-based reward optimization, we also

Table 2: **Main results on human-motion video generation.** We compare open-source baselines, a VLM-reward-trained baseline, and our structured 3D reward training. We report VBench perceptual metrics, VideoAlign scores, VideoPhy Physical Commonsense, and our structured 3D reward scores. All metrics are higher-is-better. **Bold**: best; underline: second best.

Method	*VBench/VBench2*					*VideoAlign*		*VideoPhy*	*PhyMotion (Ours)*			
	Mot. Sm.	Dyn. Deg.	Aes. Q.	Hum. An.	Temp. Fl.	VA-MQ	VA-VQ	VP-PC	Kin.	Con.	Dyn.	Overall
Open-source baselines												
Wan2.2 14B [Team Wan et al., 2025]	0.9892	0.393	0.5141	0.943	0.9857	1.133	0.557	3.97	0.881	0.733	0.912	0.842
Wan2.2 5B [Team Wan et al., 2025]	0.9846	0.598	0.4622	0.823	0.9803	1.069	0.514	3.91	0.880	0.710	0.913	0.834
EchoMotion 5B [Yang et al., 2026]	0.9850	0.683	0.4756	0.836	0.9777	1.044	0.565	3.74	0.888	0.702	0.918	0.836
Wan 1.3B [Team Wan et al., 2025]	0.9868	0.700	0.4758	0.834	0.9787	1.213	0.621	3.86	0.862	0.705	0.904	0.824
Causal Forcing 1.3B [Zhu et al., 2026]	0.9895	0.631	0.5391	0.886	0.9819	1.241	0.575	4.06	0.927	0.739	0.948	0.871
FastWan 1.3B [Zhang et al., 2025]	0.9847	**0.928**	0.5035	0.876	0.970	1.307	0.680	3.90	0.908	0.734	0.916	0.853
VLM-reward trained baseline												
VideoAlign-MQ [Liu et al., 2025b]	0.9904	0.570	0.5229	0.910	0.9835	**1.549**	**0.912**	4.12	0.916	0.753	0.946	0.878
PhyMotion (Ours) trained with												
Causal Forcing 1.3B	**0.9956** (+0.6%)	0.411 (-34.9%)	**0.5589** (+3.7%)	**0.951** (+7.3%)	**0.9957** (+1.4%)	1.313 (+5.8%)	0.720 (+25.2%)	4.29 (+5.7%)	**0.982** (+5.9%)	**0.809** (+9.4%)	**0.988** (+4.2%)	0.902 (+3.5%)
FastWan 1.3B	0.9927 (+0.8%)	0.420 (-54.7%)	0.5218 (+3.6%)	0.867 (-1.0%)	0.9928 (+2.4%)	1.318 (+0.8%)	0.687 (+1.0%)	**4.39** (+12.6%)	**0.982** (+8.1%)	0.805 (+9.7%)	0.952 (+3.9%)	**0.913** (+7.0%)

fine-tune the Causal Forcing 1.3B model with the VideoAlign Motion Quality (MQ) reward under the same training setting. We evaluate on two prompt sets: a human-motion prompt set of 22,471 prompts derived from Motion-X [Lin et al., 2023], and the standard VBench and VBench-2.0 prompt sets [Huang et al., 2024, Zheng et al., 2025]. We report three metric groups: VBench/VBench-2.0 metrics, VLM-based rewards (VideoAlign-MQ/VQ [Liu et al., 2025b] and VideoPhy-PC [Bansal et al., 2025]), and our PhyMotion from Sec. 3.2.

5.2.2 Quantitative and Qualitative Results

Quantitative results on automatic metrics. Table 2 shows that our structured 3D reward improves human-motion quality beyond the metrics used for optimization. Compared with the Causal Forcing 1.3B base model, our RL-post-trained model improves all external metrics except VBench Dynamic Degree, including VBench motion smoothness, aesthetic quality, human action, temporal flickering, VideoAlign, and VideoPhy. Notably, it improves VideoAlign video quality by $+25.2\%$ and VideoPhy physical commonsense by $+5.7\%$, and even outperforms larger 5B/14B reference models on most metrics. As expected, these gains are also reflected in our PhyMotion rewards, with improvements of $+3.5\%$ in overall feasibility. The same trend also holds for FastWan 1.3B: after post-training with PhyMotion, the model improves motion smoothness, aesthetic quality, temporal flickering, VideoAlign, VideoPhy, and all PhyMotion scores, including a $+7.0\%$ gain in overall feasibility. The decrease in VBench Dynamic Degree may reflect both reduced motion artifacts and more conservative motion. To investigate this trade-off, we evaluate instruction execution using Gemini and motion pacing through human judgments (Appendix H). Gemini-based evaluation shows that full action execution improves from 13.9% to 21.5% after RL, while human judgments indicate that most videos remain appropriately paced (88% vs. 87%), despite a modest increase in overly slow motions (3% to 8%). Furthermore, compared with the VLM-reward-trained VideoAlign-MQ baseline, our method achieves higher VideoPhy physical commonsense and stronger structured 3D feasibility across all dimensions, suggesting that explicit 3D physical rewards provide a more targeted training signal than generic VLM-based video rewards.

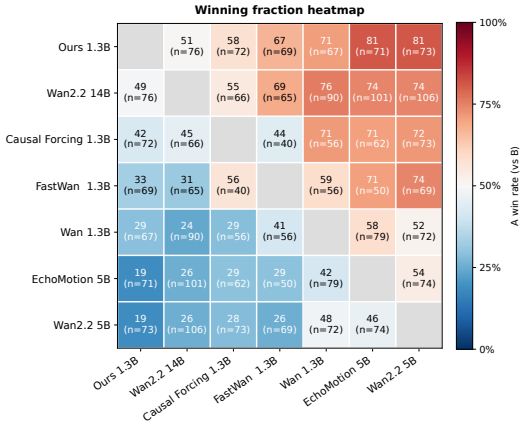


Figure 4: **Winning fraction heatmap of different models.** Each cell shows the win rate between models on the overall human preference evaluation, with ties counted as 0.5.

Quantitative results on human preference evaluation. To verify that improvements in automatic metrics translate to human-perceived motion quality, we further conduct a human preference

Table 3: **Human preference evaluation.** We report Elo ratings computed from pairwise human preferences. Higher is better.

Model	Body	Bal.	Motion	All Q
Wan2.2 5B	1376	1388	1384	1383
EchoMotion 5B	1386	1403	1374	1387
Wan 1.3B	1429	1440	1411	1427
FastWan 1.3B	1526	1521	1528	1525
Causal 1.3B	1562	1546	1553	1553
Wan2.2 14B	1600	1593	1618	1604
Ours 1.3B	**1620**	**1610**	**1632**	**1621**

Table 4: **Ablation results on rewards. VB**, **VA**, and **VP** denote compact summary metrics for VBench average, VideoAlign average, and VideoPhy Physical Commonsense, respectively.

Reward	VB	VA	VP	Kin.	Con.	Dyn.	Ours
Average	**0.782**	**0.979**	**4.12**	0.954	0.763	0.943	**0.883**
Only Kin.	0.767	0.904	3.86	**0.963**	0.730	0.951	0.881
Only Con.	0.773	0.897	3.74	0.911	**0.772**	0.917	0.867
Only Dyn.	0.766	0.915	3.91	0.948	0.734	**0.958**	0.880

evaluation across representative baseline models. We randomly sample 1,487 video pairs and ask human annotators the same preference questions as in Sec. 5.1. As shown in Table 3, our RL-post-trained model achieves the highest overall human preference Elo score, outperforming all baselines, including the much larger Wan2.2 14B model.

Beyond aggregate Elo rankings, we additionally compute pairwise human preference win rates between models. For each model pair, we count a win as 1, a loss as 0, and a tie as 0.5, and report the average win rate between models. As shown in Fig. 4, our PhyMotion optimized model consistently wins against all compared baselines in direct head-to-head comparisons, including the much larger Wan2.2 14B model. This indicates that the human preference improvement of PhyMotion is not driven by a single favorable matchup, but is broadly consistent across diverse baseline comparisons.

Qualitative comparison. Fig. 5 provides qualitative comparisons across diverse human-motion prompts, including (a) martial arts, (b) multi-human dancing, (c) soccer, (d) handstand, (e) figure skating, and (f) floor exercise / side lifting. Compared with baseline models, our model better preserves the intended action structure over time. In challenging motions such as handstand, figure skating, and floor exercises (Fig. 5d–f), baseline models often exhibit unstable body support, floating feet or hands, implausible limb bending, or sudden pose discontinuities. In contrast, our model more consistently maintains physically plausible motions.

5.2.3 Additional Analysis

We further analyze PhyMotion from three perspectives: the contribution of each reward component, whether reward post-training preserves general-domain video generation ability, and the computational overhead of our reward during training.

Effect of individual reward components. We ablate each reward component by training single-reward variants from the same Causal Forcing base model, using identical hyperparameters and 120 training steps for all variants. As shown in Table 4, each single-reward variant primarily improves its corresponding feasibility dimension. However, optimizing any single component alone does not yield the best overall performance. In contrast, the full reward achieves the highest overall feasibility while remaining competitive across all individual dimensions, indicating that jointly optimizing kinematic, contact, and dynamic feasibility produces a better-balanced human-motion prior.

Table 5: **Generalization to general prompts.** Gray rows denote reference models at larger scales.

Model	VBench	VBench-2.0
Wan 2.2 5B	0.607	0.403
Wan 2.2 14B	0.657	0.510
Wan 1.3B	0.630	0.456
Causal Forcing 1.3B	**0.657**	0.458
Ours	0.654	**0.462**

Generalization to broader prompts. To assess whether our human-centric reward compromises general video generation quality, we evaluate on standard VBench and VBench-2.0 prompts. As shown in Table 5, PhyMotion maintains comparable VBench performance to the Causal Forcing 1.3B backbone (0.654 vs. 0.657) while slightly improving VBench-2.0 (0.462 vs. 0.458). These results suggest that optimizing physical plausibility does not substantially sacrifice general generation capabilities. Per-category results are provided in Appendix D.

Reward computation overhead. As shown in Table 6, PhyMotion requires 2.80s per video but introduces only 7% effective training overhead through pipelined reward computation, compared with 35% for HPSv3. Further details are provided in Appendix E.

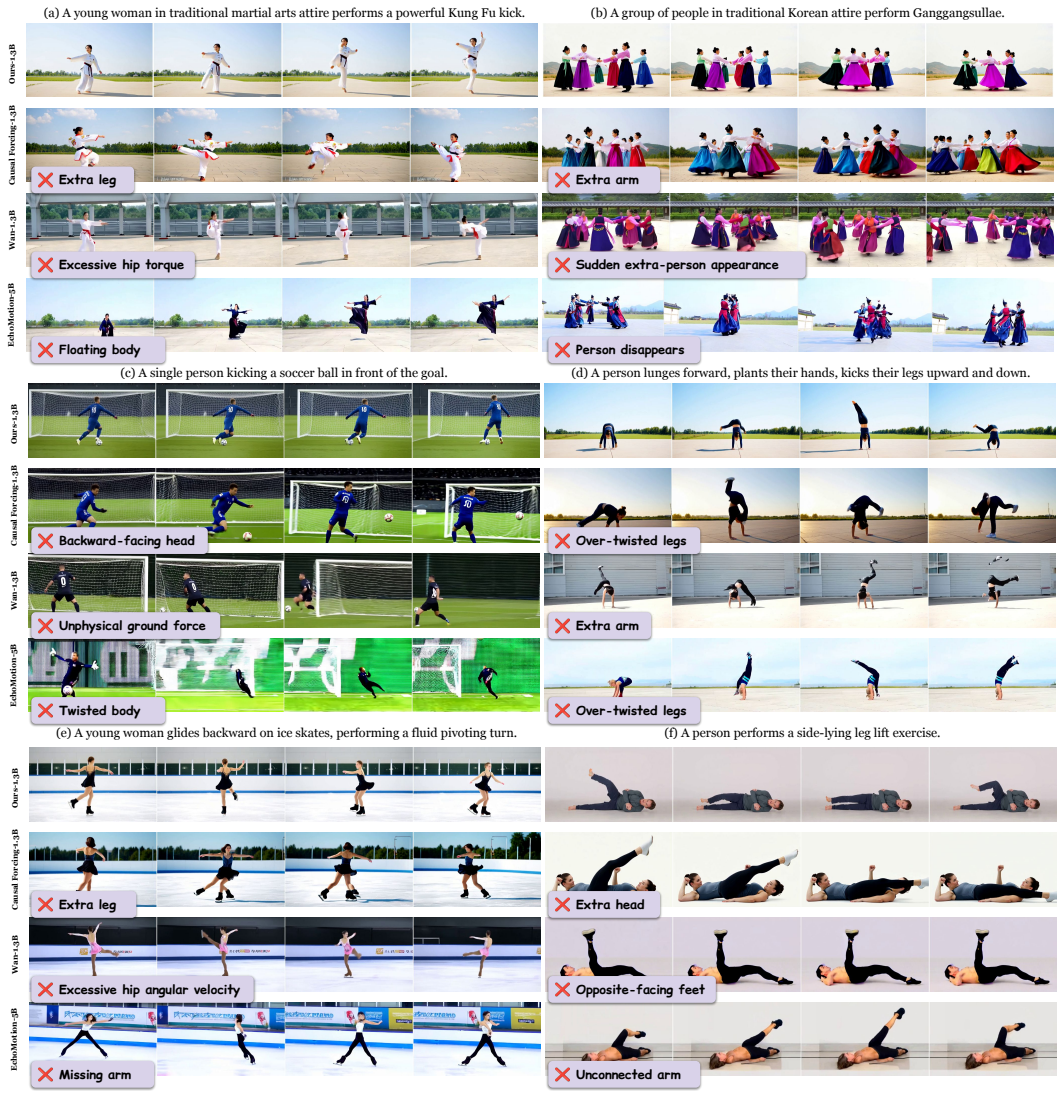


Figure 5: **Qualitative comparison on diverse human-motion prompts.** We compare our model against baseline models across challenging motion categories. Baseline models often produce artifacts, while our model maintains more physically plausible motion.

Table 6: **Reward computation overhead.** We report the per-video reward computation time and the effective training overhead introduced by each reward module.

	HPSv3	VideoAlign-MQ	VideoPhy-PC	PhyMotion (Ours)
Time / video (s)	4.72	0.25	0.20	2.80
Effective overhead	35%	0.2%	0.2%	7%

6 Conclusion

We introduce PhyMotion, a structured, fine-grained motion reward for physics-grounded human video generation. Instead of relying only on perceptual or learned video-level signals, PhyMotion lifts generated videos into 3D human motion, grounds them in a physics simulator, and evaluates motion feasibility through kinematic, contact/balance, and dynamic consistency. Experiments show that PhyMotion aligns better with human judgments than existing video rewards, and our post-training experiment also shows that PhyMotion serves as an effective RL post-training objective, improving automatic metrics performance and human preference scores while preserving general video quality across different video-generation backbones.

Acknowledgments and Disclosure of Funding

This work was supported by ONR Grant N00014-23-1- 2356, ARO Award W911NF2110220, and NSF-AI Engage Institute DRL-2112635. The views contained in this article are those of the authors and not of the funding agency.

References

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, pages 102075–102121, 2025.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. VideoJAM: Joint appearance-motion representations for enhanced motion generation in video models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 7595–7616. PMLR, 2025. URL https://proceedings.mlr.press/v267/chefer25a.html.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations (ICLR)*, 2025.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pages 79858–79885, 2023. doi: 10.52202/075280-3497. URL https://openreview.net/forum?id=8OTPepXzeh.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Haoran He, Yuxiao Ye, Jie Liu, Jiajun Liang, Zhiyong Wang, Ziyang Yuan, Xintao Wang, Hangyu Mao, Pengfei Wan, and Ling Pan. GARDO: Reinforcing diffusion models without reward hacking. *arXiv preprint arXiv:2512.24138*, 2025.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2105–2123. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.127. URL https://aclanthology.org/2024.emnlp-main.127/.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. MotionBench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8450–8460, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8153–8163, 2024.

Yidong Huang, Zun Wang, Han Lin, Dong-Ki Kim, Shayegan Omidshafiei, Jaehong Yoon, Yue Zhang, and Mohit Bansal. Planning with sketch-guided verification for physics-aware video generation. *arXiv preprint arXiv:2511.17450*, 2025. URL https://arxiv.org/abs/2511.17450.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818, 2024.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3268–3285, 2026. doi: 10.1109/TPAMI.2025.3633890.

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

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-X: A large-scale 3D expressive whole-body human motion dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 25268–25280, 2023.

Jing Lin, Ruisi Wang, Junzhe Lu, Ziqi Huang, Guorui Song, Ailing Zeng, Xian Liu, Chen Wei, Wanqi Yin, Qingping Sun, Zhongang Cai, Lei Yang, and Ziwei Liu. The quest for generalizable motion generation: Data, model, and evaluation. In *International Conference on Learning Representations (ICLR)*, 2026.

Xinran Ling, Chen Zhu, Meiqi Wu, Hangyu Li, Xiaokun Feng, Cundian Yang, Aiming Hao, Jiashu Zhu, Jiahong Wu, and Xiangxiang Chu. VMBench: A benchmark for perception-aligned video motion generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13087–13098. IEEE, 2025.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. In *Advances in Neural Information Processing Systems*, volume 38, pages 40783–40818, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3a10c46572628d58cb44fb705f25cbbf-Abstract-Conference.html.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22139–22149, 2024.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. FETV: A benchmark for fine-grained evaluation of open-domain text-to-video generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 34 (6):248:1–248:16, 2015.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, Yujun Shen, and Min Zhang. Reward Forcing: Efficient streaming video generation with rewarded distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 34385–34397, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lu_Reward_Forcing_Efficient_Streaming_Video_Generation_with_Rewarded_Distribution_Matching_CVPR_2026_paper.html.

Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow Your Pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4117–4125, 2024. doi: 10.1609/aaai.v38i5.28206.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15086–15095, 2025.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5442–5451, 2019.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 43781–43806. PMLR, 2025. URL https://proceedings.mlr.press/v267/meng25c.html.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 948–958, 2026.

OpenAI. Sora: Creating video from text. https://openai.com/sora, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. 360-degree human video generation with 4D diffusion transformer. *ACM Transactions on Graphics*, 43(6):1–13, 2024a. doi: 10.1145/3687980.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. doi: 10.1145/3680528.3687565.

Kewei Sui, Anindita Ghosh, Inwoo Hwang, Bing Zhou, Jian Wang, and Chuan Guo. A survey on human interaction motion generation. *International Journal of Computer Vision*, 134(3):113, 2026.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8406–8416, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Sun_T2V-CompBench_A_Comprehensive_Benchmark_for_Compositional_Text-to-video_Generation_CVPR_2025_paper.html.

Team Wan, Ang Wang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2019.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8228–8238, 2024.

Haoru Wang, Wentao Zhu, Luyi Miao, Yishu Xu, Feng Gao, Qi Tian, and Yizhou Wang. Aligning human motion generation with human perceptions. In *International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c129741a2451e5fefe447591e39de30e-Abstract-Conference.html.

Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. DisCo: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9326–9336, 2024.

Yiping Wang, Xuehai He, Kuan Wang, Luyao Ma, Jianwei Yang, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Is your world simulator a good story presenter? A consecutive events-based benchmark for future long video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13629–13638, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Is_Your_World_Simulator_a_Good_Story_Presenter_A_Consecutive_CVPR_2025_paper.html.

Zehan Wang, Tengfei Wang, Haiyu Zhang, Xuhui Zuo, Junta Wu, Haoyuan Wang, Wenqiang Sun, Zhenwei Wang, Chenjie Cao, Hengshuang Zhao, Chunchao Guo, and Zhou Zhao. WorldCompass: Reinforcement learning for long-horizon world models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 131140–131155. PMLR, 2026a. URL https://proceedings.mlr.press/v306/wang26iu.html.

Zun Wang, Jaemin Cho, Jialu Li, Han Lin, Jaehong Yoon, Yue Zhang, and Mohit Bansal. EPiC: Efficient video camera control learning with precise anchor-video guidance. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 125551–125586. PMLR, 2026b. URL https://proceedings.mlr.press/v306/wang26af.html.

Zun Wang, Han Lin, Jaehong Yoon, Jaemin Cho, Yue Zhang, and Mohit Bansal. AnchorWeave: World-consistent video generation with retrieved local spatial memories. In *European Conference on Computer Vision (ECCV)*, pages 203–224, 2026c. URL https://arxiv.org/abs/2602.14941.

David A Winter. *Biomechanics and motor control of human movement*. John Wiley & Sons, fourth edition, 2009.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Dan Zhang, Ming Ding, Xiaohan Zhang, Shiyu Huang, Xiaotao Gu, Minlie Huang, Jie Tang, and Yuxiao Dong. VisionReward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 11269–11277, 2026. doi: 10.1609/aaai.v40i13.38107.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. MagicAnimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1481–1490, 2024.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. DanceGRPO: Unleashing GRPO on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Yuxiao Yang, Hualian Sheng, Sijia Cai, Jing Lin, Jiahao Wang, Bing Deng, Junzhe Lu, Haoqian Wang, and Jieping Ye. EchoMotion: Unified human video and motion generation via dual-modality diffusion transformer. In *International Conference on Learning Representations (ICLR)*, 2026.

Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhengzhong Liu, and Hao Zhang. Fast video generation with sliding tile attention. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 74714–74731. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25m.html.

Songchun Zhang, Zeyue Xue, Siming Fu, Jie Huang, Xianghao Kong, Yue Ma, Haoyang Huang, Nan Duan, and Anyi Rao. Astrolabe: Steering forward-process reinforcement learning for distilled autoregressive video models. *arXiv preprint arXiv:2603.17051*, 2026.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations (ICLR)*, 2026.

Jingkai Zhou, Benzhi Wang, Weihua Chen, Jingqi Bai, Dongyang Li, Aixi Zhang, Hao Xu, Mingyang Yang, and Fan Wang. RealisDance: Equip controllable character animation with realistic hands. *arXiv preprint arXiv:2409.06202*, 2024.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal Forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 168030–168049. PMLR, 2026. URL https://proceedings.mlr.press/v306/zhu26bf.html.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3D parametric guidance. In *European Conference on Computer Vision (ECCV)*, pages 145–162, 2024a.

Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2430–2449, 2024b. doi: 10.1109/TPAMI.2023.3330935.

Appendix

A Limitations

This work studies structured reward design for improving human-motion realism in video generation. As with other reward-based post-training methods, our approach may be susceptible to reward hacking. Although our quantitative analysis shows improved action execution and largely preserved motion pacing after RL post-training (Appendix H), isolated failure cases still exhibit motion suppression or simplification of challenging actions (Fig. 6(a)). Moreover, our reward focuses on human-body motion without explicitly modeling human–object interactions, which may lead to physically inconsistent object manipulation (Fig. 6(b)). Finally, our post-training primarily covers motions on flat surfaces, limiting generalization to out-of-distribution scenarios such as stair climbing (Fig. 6(c)). Although our evaluation framework can incorporate non-flat terrain geometry, such scenarios remain underrepresented during training. Future work could address these limitations through more diverse training data and reward signals that explicitly account for action completion, object interactions, and complex environments.

B Broader Impact

While PhyMotion aims to improve the physical realism of human motion in video generation, it is important to consider its broader ethical, legal, and societal implications.

Potential benefits. Improving the physical realism of human motion in generated videos can benefit applications that rely on reliable human-centric simulation, including education, accessibility, sports and exercise instruction, entertainment, virtual communication, and embodied-agent training. In addition, because PhyMotion decomposes motion quality into kinematic, contact, and dynamic feasibility, it provides more interpretable feedback than a single holistic video score. This can help researchers and practitioners diagnose specific motion failures, develop safer generation systems, and better understand when synthetic human videos are physically plausible or unreliable.

Deceptive misuse. More realistic human motion may make generated videos appear more convincing, which could increase the risk of misuse in misinformation, impersonation, or fraudulent media. Although our work focuses on evaluating and improving physical plausibility rather than identity preservation or face realism, the resulting models may still contribute to more realistic synthetic human videos. We therefore encourage responsible release practices, clear usage guidelines, watermarking or provenance mechanisms when appropriate, and safeguards against deceptive or non-consensual use.

Consent, privacy, and human likeness. Human-centric video generation raises concerns around consent and privacy, especially when generated videos resemble real individuals or depict sensitive activities. Our experiments use text prompts and generated videos for evaluation, rather than collecting private human video data for model training. Nevertheless, future applications of human-motion video generation should respect individual consent, avoid unauthorized likeness generation, and follow applicable privacy and publicity-rights regulations.

Bias and fairness. PhyMotion builds on existing video generation backbones and 3D human reconstruction tools, which may inherit biases from their training data and model design. These biases may affect motion quality across different body types, clothing styles, activities, camera viewpoints, or cultural contexts. Although our reward is designed to evaluate physical feasibility in 3D, it does not by itself guarantee fairness across all populations or motion categories. Careful evaluation across diverse human appearances, body shapes, and actions is needed before deployment.

C Additional Qualitative Results

C.1 Qualitative Analysis of Individual Reward Components

We further provide qualitative examples to illustrate the diagnostic role of each PhyMotion submetric. As shown in Fig. 7, each example is selected by filtering for videos that score poorly on one target

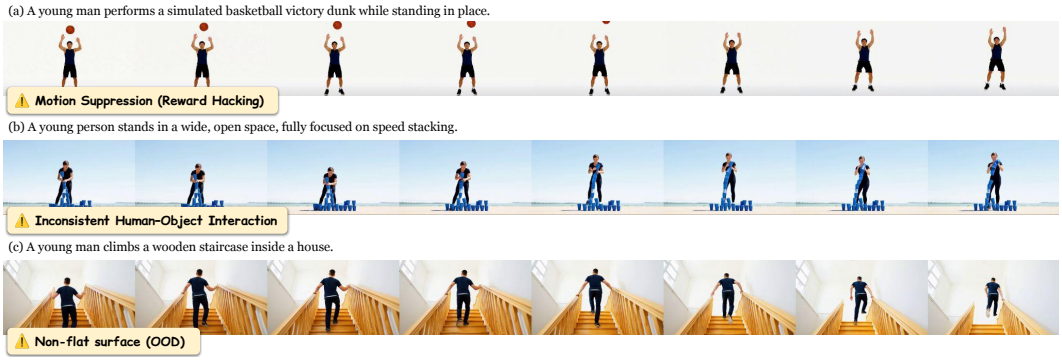


Figure 6: **Qualitative limitations of PhyMotion.** Representative failure cases involving (a) motion suppression due to reward hacking, (b) physically inconsistent human–object interactions, and (c) out-of-distribution non-flat terrain.

submetric while scoring relatively well on the other two. This isolates failure cases primarily captured by a single component.

The examples show that the three submetrics capture complementary artifacts. Kinematic feasibility Fig. 7(a) detects articulated-body errors such as self-penetration, extra limbs and impossible joint angles. Contact/balance feasibility Fig. 7(b) captures unstable support, balance violations, and foot sliding. Dynamic feasibility Fig. 7(c) identifies motions requiring implausible forces and torques, such as excessive horizontal ground force. These results support the need for a decomposed reward: each component provides an interpretable signal for a distinct class of physical failures.

D Per-category Comparison on General Video Benchmarks.

To assess whether our human-centric reward over-specializes the generator, we evaluate on the standard VBench and VBench-2.0 prompt sets. We use Causal Forcing 1.3B as the primary comparison, since it is the backbone of our post-trained model, and include larger Wan2.2 models as additional references.

As shown in Table 5, our model maintains comparable average VBench performance to the base model (0.654 vs. 0.657) and slightly improves VBench-2.0 (0.462 vs. 0.458). These results suggest that optimizing human-motion feasibility does not substantially compromise general video generation quality.

In Fig. 8, the category-level results show that our model remains competitive across broad perceptual, temporal, and compositional dimensions, while retaining strong performance on human- and motion-related categories such as subject consistency, motion smoothness, temporal flickering, human action, and human anatomy.

E Reward Computation Overhead

We compare the standalone reward-evaluation latency and effective training overhead of PhyMotion against HPSv3 [Ma et al., 2025], VideoAlign-MQ [Liu et al., 2025b], and VideoPhy-PC [Bansal et al., 2025]. These baselines evaluate human preference, motion quality, and physical commonsense, respectively.

Evaluation protocol. We generate a fixed batch of $B{=}8$ FastWan videos at $480{\times}832$ resolution. Each reward is executed once for warm-up and five times for timing, and we report the steady-state per-video average.

Effective training overhead. Our training pipeline overlaps video sampling with reward computation: while the model generates the next batch, the previous batch is evaluated in parallel. Therefore, only the non-overlapped reward computation contributes to the effective training overhead, which we measure relative to a constant-reward baseline:

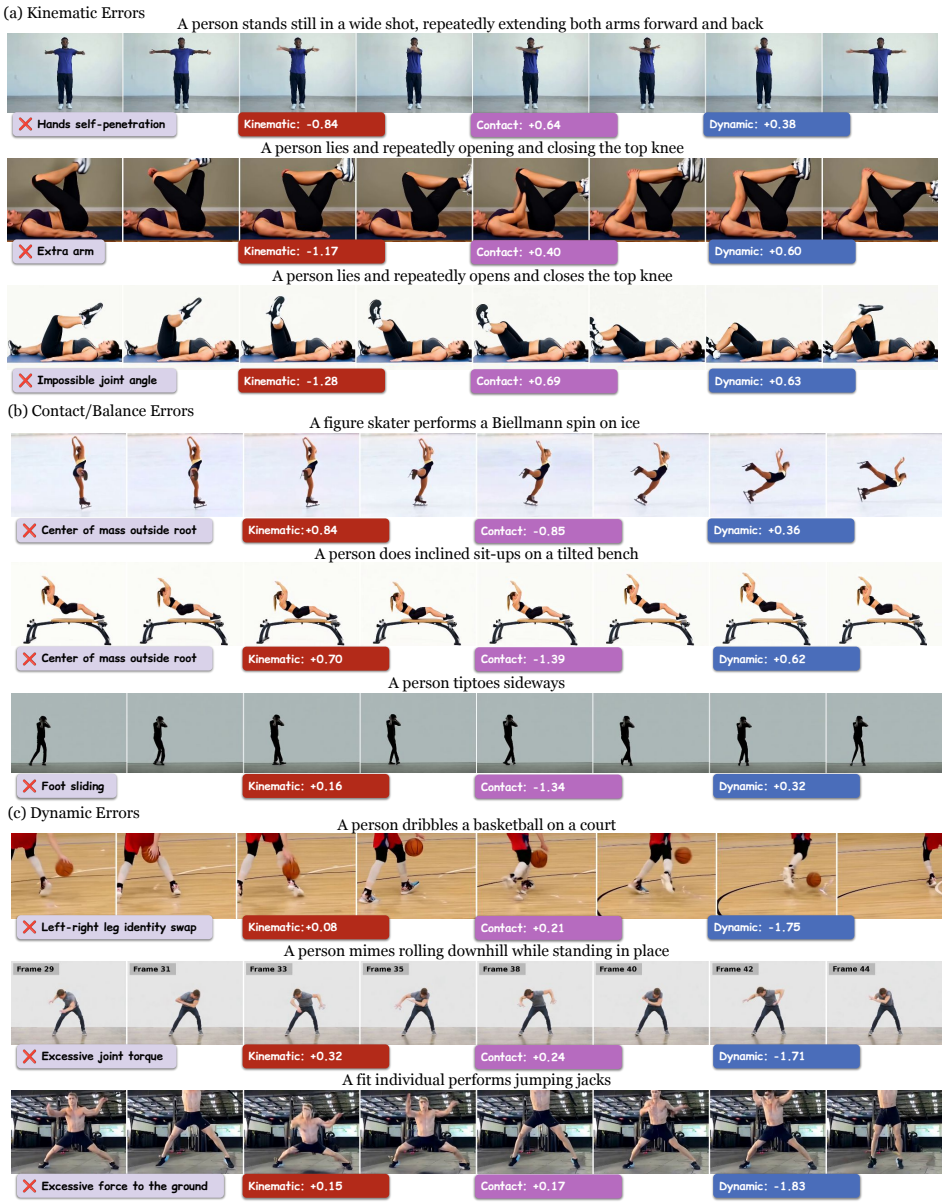
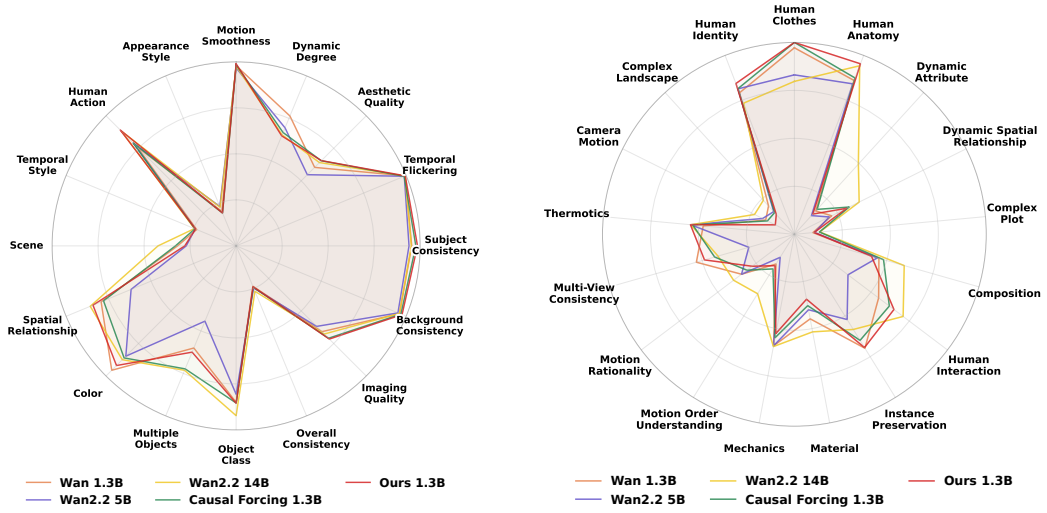


Figure 7: **Qualitative examples selected by individual PhyMotion submetrics.** Each row shows a generated video that scores poorly on one target submetric while scoring relatively well on the other two. The examples show complementary failures captured by kinematic feasibility, contact/balance feasibility, and dynamic feasibility. All scores are reported in z-scores after normalization.

$$\text{Overhead} = \frac{T_{\text{epoch}}^{\text{reward}} - T_{\text{epoch}}^{\text{const}}}{T_{\text{epoch}}^{\text{const}}}. \tag{5}$$

As reported in Table 6, PhyMotion requires 2.80s per video for standalone reward computation but introduces only 7% effective training overhead due to pipelining. In contrast, HPSv3 introduces 35% overhead despite its comparable standalone latency. These results demonstrate that structured 3D motion evaluation can be efficiently integrated into RL post-training.



(a) VBench per-dimension comparison.

(b) VBench-2.0 per-dimension comparison.

Figure 8: **Comparison on general video benchmarks.** We show normalized radar charts for VBench and VBench-2.0 dimensions. Our model remains competitive across general perceptual, temporal, and compositional categories, while showing strong performance on human- and motion-related dimensions. These results support that PhyMotion improves human-centered physical plausibility without substantially sacrificing general video generation quality.

F Robustness to Human Motion Reconstruction Errors

F.1 Detection of Severe Anatomical Distortions

A possible limitation of a reconstruction-based reward is that the body reconstructor may map visibly distorted human anatomy to a plausible SMPL-X pose, concealing the visual artifact. To investigate this issue, we first analyze 3,000 randomly sampled generated videos using Gemini-2.5-Pro; 2.7% are identified as containing severe anatomical distortions. We then evaluate a balanced set of 100 videos from four generators, including 50 severely distorted and 50 clean clips, with distortion labels agreed upon by three human raters without access to automatic metric scores.

We measure how reliably each evaluation signal distinguishes the distorted from the clean group using area under the receiver operating characteristic curve (AUROC). As shown in Table 7, the overall PhyMotion reward achieves an AUROC of 0.913, exceeding VBench Human Anatomy (0.875), VideoAlign-MQ (0.732), and VideoAlign-VQ (0.587). Each individual PhyMotion axis also discriminates between the two groups. These findings suggest that severe visual artifacts frequently leave detectable evidence in the recovered trajectories, such as implausible joint motion and physical violations. Nevertheless, reconstruction can still fail to preserve some image-space defects; our physics-grounded reward complements, rather than replaces, visual-quality evaluation.

F.2 Consistency Under Video-to-3D Reconstruction

To isolate the effect of reconstruction errors on the reward, we conduct a controlled motion–video–motion consistency experiment. We apply temporally smooth perturbations to ground-truth AMASS [Mahmood et al., 2019] human motions, producing a range of nearly valid to moderately distorted trajectories. For each of 240 perturbed motion pairs, we compute PhyMotion through two paths: (i) scoring the perturbed 3D motion directly, and (ii) rendering the same motion into video, reconstructing the human trajectory using GVHMR, and then scoring the reconstruction. The two paths share the same underlying motion, allowing us to quantify how the intermediate video reconstruction changes reward rankings.

Table 8 reports the Spearman correlation between direct and reconstructed scores. Correlations reach 0.854 for kinematics, 0.790 for contact, 0.899 for dynamics, and 0.874 for the overall reward. This

Table 7: Detection of severe anatomical distortion on a balanced set of 50 distorted and 50 clean generated videos. Higher AUROC is better.

Metric	AUROC ↑
VBench Human Anatomy	0.875
VideoAlign-MQ	0.732
VideoAlign-VQ	0.587
PhyMotion Kinematic	0.848
PhyMotion Contact	0.869
PhyMotion Dynamic	0.880
PhyMotion Overall	**0.913**

Table 8: Spearman rank correlation between PhyMotion scores from direct 3D motions and scores after video rendering and GVHMR reconstruction.

PhyMotion component	Spearman ρ ↑
Kinematic	0.854
Contact	0.790
Dynamic	0.899
Overall	**0.874**

agreement indicates that the relative physical assessments are largely preserved under the tested reconstruction perturbations. It does not imply perfect 3D reconstruction or guarantee robustness to severe tracking failures, unobserved anatomy, or arbitrary scene conditions.

F.3 Challenging Reconstruction Conditions

We also inspect representative generated videos involving multiple people, substantial articulation and self-occlusion, partial-body visibility, and moving cameras. For multi-person videos, we track and reconstruct each person separately and aggregate the resulting per-person rewards using a weighted average. Occluded joints are estimated using temporal evidence and the SMPL-X body prior; this can provide a coherent trajectory under moderate occlusion, but also creates a risk of concealing visual defects. Changes in camera pose further require a consistent reference frame to distinguish body motion from viewpoint changes. Full temporal examples include group dancing, martial-arts actions, skating, inversion, and floor exercises. These qualitative demonstrations illustrate the scope of the pipeline but do not establish reliability under complete occlusions or lost tracks.

G Extending PhyMotion to Non-Flat Terrain

Our default contact and balance evaluation assumes a flat ground plane, which may incorrectly penalize valid contacts on non-flat terrain. To extend PhyMotion, we incorporate terrain geometry into the MuJoCo simulator and evaluate foot contact, penetration, and balance relative to local support surfaces rather than a single global plane. This requires only modifying the support geometry and contact evaluation, while preserving the overall reward formulation.

We validate this extension on six ground-truth stair-climbing motions from AMASS [Mahmood et al., 2019]. Incorporating local staircase geometry improves the average contact score from 0.75 to 0.90. In a controlled synthetic ascent, valid-contact recognition increases from 0.25 to 0.94. Furthermore, a sim-to-real-to-sim experiment shows that GVHMR recovers all five staircase levels and 0.73 m of vertical displacement from a photorealistic video, compared with 0.81 m in the original motion. These results demonstrate that PhyMotion can accommodate non-flat terrain with minimal modifications.

Table 9: Prompt-conditioned action execution judged by Gemini (%).

Generator	Full ↑	Partial	None ↓	Full+Partial ↑
Causal Forcing 1.3B	13.9	40.4	45.7	54.3
PhyMotion RL (1.3B)	**21.5**	36.3	**42.3**	**57.8**
Wan2.2 5B	11.4	43.0	45.6	54.4
Wan2.2 14B	34.4	36.3	29.3	70.7

Table 10: Human judgments of motion pacing on 500 videos, with three raters.

Generator	Overly fast (%)	Normal (%)	Overly slow (%)
Causal Forcing 1.3B	10	88	3
PhyMotion RL (1.3B)	4	87	8

H Analysis of Action Execution and Motion Pacing

H.1 Instruction-Conditioned Action Execution

The decrease in VBench Dynamic Degree following RL post-training raises the question of whether the generator avoids physically difficult actions in order to earn higher feasibility rewards. To test this, we use Gemini to evaluate anonymized generated videos conditioned on their text prompts, classifying the requested action as fully executed, partially executed, or not executed. Table 9 reports the results, with larger Wan models included as references rather than controlled before/after comparisons.

Relative to the same Causal Forcing 1.3B backbone before RL, PhyMotion post-training increases the full-execution rate from 13.9% to 21.5% (+7.6 percentage points). Full-or-partial execution increases from 54.3% to 57.8%, while non-execution decreases from 45.7% to 42.3%. This provides evidence against broad action avoidance as the explanation for improved rewards, while leaving open the possibility of more conservative motion timing.

H.2 Human Evaluation of Motion Pacing

We separately evaluate perceived pacing on 500 generated videos, equally split between the Causal Forcing baseline and the PhyMotion-finetuned model (250 per model). Three human annotators judge whether each video's motion is overly fast or jittery, appropriately paced for the requested action, or overly slow or insufficiently dynamic. As reported in Table 10, the share of normally paced clips is nearly unchanged after RL (88% to 87%), and overly fast or jittery motion falls from 10% to 4%. However, the overly slow category rises from 3% to 8%. Thus, the results do not indicate widespread collapse to static videos, but they do reveal a modest slowdown trade-off. Combining physical feasibility rewards with complementary action-completion or motion-amplitude objectives is a possible direction for addressing this limitation.

I Human Study Detail

This appendix describes the human preference studies used to evaluate whether PhyMotion aligns with human judgments of generated human-motion quality. We conduct two pairwise-comparison studies: one for validating our automatic feasibility metrics against human preferences, and one for ranking post-trained and baseline video generation models. Both studies were conducted with non-author participants, including students and researchers in related areas. Participants compared short generated videos under the same text prompt and judged them along three axes: body structure, balance, and motion naturalness.

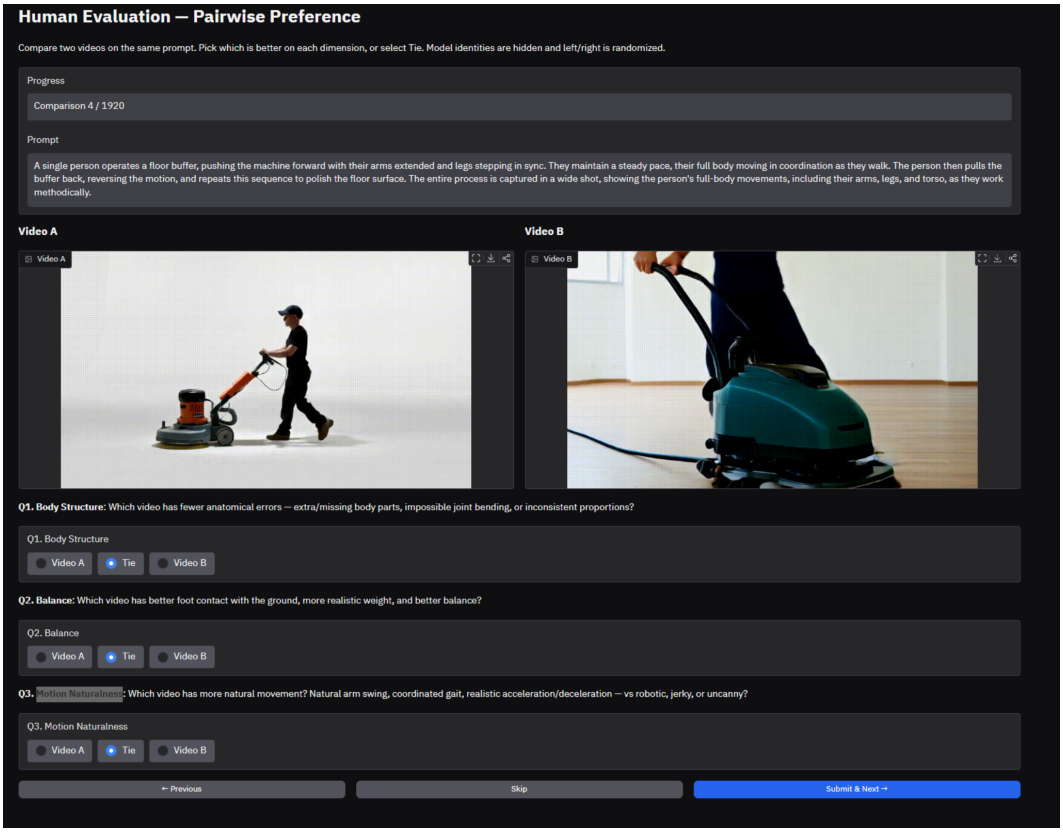


Figure 9: **Human preference annotation interface.** Annotators compare two anonymized videos generated from the same prompt and answer three pairwise preference questions on body structure, balance, and motion naturalness.

I.1 Task Design

Each annotation example consists of two short videos generated from the same text prompt by two different models. The videos are displayed side by side and played in sync. Model identities are hidden from annotators, and the left/right order is randomized independently for each pair.

For each video pair, annotators answer three preference questions, corresponding to the main perceptual dimensions studied in this paper:

- **Body structure:** Which video shows a more anatomically correct human body? Annotators are asked to consider limb consistency, body-part distortions, missing or extra limbs, and self-intersections.
- **Balance:** Which video shows more physically plausible body support? Annotators are asked to consider whether the person appears stably supported by the contacting feet, hands, or other body parts.
- **Motion naturalness:** Which video shows motion that looks more like a real human action? Annotators are asked to consider temporal smoothness, realistic timing, and the absence of teleporting, flickering, or floating artifacts.

For each question, annotators choose one of three options: Video A, Video B, or Tie. The default option is Tie, so unanswered questions do not introduce a preference toward either side. The same instruction text is shown to every annotator, and a screenshot of the annotation interface is shown in Fig. 9.

Table 11: **Inter-annotator agreement on the human pilot study.** For each question, we report the mean pairwise agreement between human annotators on overlapping video pairs.

Question	Body Structure	Balance	Motion Naturalness
Mean Pairwise Agreement	95.9%	91.4%	96.9%

Table 12: **Correlation with human judgments and statistical significance.** We report Spearman's rank correlation (ρ) between automatic metric scores and human judgments across three questions, together with significance levels. Best results are in **bold**, and second-best results are underlined.

Metric Group	Metric	Body Structure	Balance	Motion Naturalness	All Questions
VBench / VBench2	Aesthetic	$+0.248 \pm 0.032$	$+0.238 \pm 0.033$	$+0.260 \pm 0.032$	$+0.248 \pm 0.018$
	Motion	$+0.236 \pm 0.035$	$+0.246 \pm 0.034$	$+0.274 \pm 0.034$	$+0.252 \pm 0.019$
	Temporal	$+0.175 \pm 0.032$	$+0.171 \pm 0.035$	$+0.208 \pm 0.034$	$+0.185 \pm 0.020$
	Dynamic	-0.135 ± 0.033	-0.118 ± 0.035	-0.149 ± 0.034	-0.135 ± 0.019
	Anatomy	$+0.273 \pm 0.032$	$+0.235 \pm 0.033$	$+0.278 \pm 0.033$	$+0.262 \pm 0.018$
VideoAlign	Motion	$+0.155 \pm 0.035$	$+0.175 \pm 0.036$	$+0.156 \pm 0.034$	$+0.161 \pm 0.020$
	Visual	$+0.045 \pm 0.035$	$+0.053 \pm 0.035$	$+0.067 \pm 0.034$	$+0.055 \pm 0.020$
VideoPhy	Physics	$+0.168 \pm 0.033$	$+0.142 \pm 0.035$	$+0.108 \pm 0.035$	$+0.138 \pm 0.020$
PhyMotion(Ours)	Kinematic	$\underline{+0.369 \pm 0.031}$	$+0.314 \pm 0.033$	$+0.375 \pm 0.031$	$+0.353 \pm 0.018$
	Contact	$+0.290 \pm 0.032$	$\mathbf{+0.337 \pm 0.033}$	$+0.281 \pm 0.032$	$+0.290 \pm 0.018$
	Dynamic	$+0.367 \pm 0.030$	$+0.316 \pm 0.032$	$\underline{+0.389 \pm 0.030}$	$\underline{+0.358 \pm 0.017}$
	Overall	$\mathbf{+0.391 \pm 0.030}$	$\underline{+0.333 \pm 0.032}$	$\mathbf{+0.402 \pm 0.030}$	$\mathbf{+0.376 \pm 0.017}$

I.2 Study Composition

We form the dataset using a wide variety of video models including Causal Forcing 1.3B [Zhu et al., 2026], FastWan 1.3B [Zhang et al., 2025], Wan 2.1 1.3B, Wan2.2 5B, Wan2.2 14B [Team Wan et al., 2025], and EchoMotion 5B [Yang et al., 2026]. Each pair contains two videos generated from the same prompt by different models. We filter out unsafe videos, nearly identical pairs, and pairs where both videos already have very high feasibility scores, since these provide little discriminative signal. After filtering incomplete sessions and low-coverage variants, the final Elo analysis uses $1{,}487$ vote rows per question, or $4{,}461$ axis-level annotations in total. Ties are counted as 0.5, and Elo ratings are computed with base rating 1500 and $K = 32$.

I.3 Inter-Annotator Agreement

We also compute inter-annotator agreement on overlapping video pairs in the pilot study. Since some comparisons are ambiguous, we treat Tie as compatible with either Video A or Video B; a disagreement is counted only when two annotators make hard opposite choices, i.e., one selects Video A and the other selects Video B. As shown in Table 11, annotators achieve high agreement across all three questions, suggesting that the criteria are consistently understood.

I.4 Human Study Significance

Correlation with human judgments. In addition to the correlation values reported in the main paper, we report the statistical significance of the metric–human alignment results in Table 12.

Human preference evaluation. To complement the compact human preference results in the main paper, we report the full per-question Elo ratings with bootstrap uncertainty in Table 13.

I.5 Annotator Recruitment, Compensation, and Data Handling.

Annotators were recruited from paid research assistants within our laboratory, and annotation was conducted as part of their regular research responsibilities. After data collection, personal identifying information was removed from the annotation records, and only de-identified annotator IDs (e.g., Annotator 1, Annotator 2) were retained for subsequent analysis.

Table 13: **Detailed human preference Elo ratings.** We provide the full Elo ratings corresponding to the human preference evaluation summarized in the main paper. Ratings are computed from pairwise human preferences on body structure, balance, motion naturalness, and all questions. Higher is better, and uncertainty denotes bootstrap standard deviation.

Model	Body Struct.	Balance	Motion	Overall
Ours	$\mathbf{1620 \pm 34}$	$\mathbf{1610 \pm 34}$	$\mathbf{1632 \pm 39}$	$\mathbf{1621 \pm 11}$
Wan2.2 14B	$\underline{1600 \pm 34}$	$\underline{1593 \pm 33}$	$\underline{1618 \pm 40}$	$\underline{1604 \pm 13}$
Causal-Forcing 1.3B	1562 ± 35	1546 ± 33	1553 ± 39	1553 ± 8
FastWan 1.3B	1526 ± 33	1521 ± 32	1528 ± 35	1525 ± 4
Wan 1.3B	1429 ± 37	1440 ± 33	1411 ± 39	1427 ± 15
EchoMotion 5B	1386 ± 36	1403 ± 35	1374 ± 38	1387 ± 14
Wan2.2 5B	1376 ± 35	1388 ± 34	1384 ± 39	1383 ± 6

J Detailed Definition of PhyMotion Metrics

This appendix provides the detailed computation of the three PhyMotion feasibility scores used in the main paper: kinematic feasibility, contact feasibility, and dynamic feasibility. The goal of these metrics is to convert a generated human video into interpretable physical signals that can be used both for evaluation and as rewards for RL post-training.

Given a generated video $v = \{I_t\}_{t=1}^{T}$ with frame rate f, we first recover an SMPL-X body trajectory using GVHMR. Let q_t denote the recovered body pose at frame t, $X_t = \{x_{t,j}\}_{j=1}^{J}$ denote the 3D body joints, and M_t denote the body mesh. For SMPL-X, we use $J = 55$ joints and a mesh with $V = 10{,}475$ vertices and $F = 20{,}908$ triangular faces. We then retarget the recovered motion to a MuJoCo human model and compute physics-related quantities such as joint torques and ground reaction forces. We write $\Delta t = 1/f$.

For foot-related metrics, we define two foot-sole vertex sets, $\mathcal{V}_L$ and $\mathcal{V}_R$, corresponding to the left and right feet. These sets are precomputed from the SMPL-X template by selecting the lowest vertices on each foot. In our experiments, all metrics are computed from the GVHMR-recovered camera-frame trajectory, with $f = 16$.

J.1 Kinematic Feasibility

Kinematic feasibility measures whether the reconstructed human body is anatomically valid and temporally smooth. It penalizes three common artifacts: joints moving too fast, body parts intersecting each other, and unstable limb geometry. The final kinematic score is high when the motion has smooth joints, little self-penetration, and a stable body structure.

Joint angular velocity. Generated videos sometimes contain abrupt limb snapping, where a joint rotates too quickly between adjacent frames. We approximate the angular velocity of joint j at frame t by finite differences:

$$\omega_{t,j} = f\, d_{\mathcal{Q}}(q_{t+1,j}, q_{t,j}),$$

where $d_{\mathcal{Q}}(\cdot,\cdot)$ measures the distance between two joint rotations, such as the geodesic distance on $SO(3)$. We then compute the fraction of joint-frame pairs whose angular velocity exceeds a plausible joint-specific threshold:

$$v_{\text{vel}} = \frac{1}{(T-1)J} \sum_{t=1}^{T-1} \sum_{j=1}^{J} \mathbf{1}\left[\omega_{t,j} > \omega_j^{\max}\right].$$

Self-penetration. Another common failure is body-part penetration, such as a hand passing through the torso or a leg intersecting the body. For each frame, we detect intersections between non-identical mesh triangles using a BVH-based triangle-intersection test. Let $\mathcal{F}$ be the set of mesh triangles. We compute the percentage of intersecting triangle pairs:

$$\text{spen}_t = 100 \cdot \frac{\left|\left\{(f_1, f_2) \in \mathcal{F}^2 : f_1 \cap f_2 \neq \emptyset,\ f_1 \neq f_2\right\}\right|}{F}.$$

We then average this quantity over time:

$$\mathrm{spen} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{spen}_t.$$

Because clean SMPL meshes can have small soft-body overlaps, we treat roughly 2% self-penetration as a normal baseline and regard values above 20% as severe. The normalized self-penetration violation is

$$v_{\mathrm{spen}} = \mathrm{clip}\left(\frac{\mathrm{spen}-2}{18}, 0, 1\right).$$

Joint-limit violation. Generated videos may contain anatomically impossible poses, such as a knee bending backward, a hip twisting too far, or an arm rotating beyond a plausible range. To detect these failures, we compare each reconstructed joint angle with the valid joint range defined by the MuJoCo human model. Let $q_{t,j}$ denote the pose parameter of joint j at frame t, and let $[q_j^{\min}, q_j^{\max}]$ denote its valid range. We define the joint-limit violation as the fraction of joint-frame pairs that fall outside this range:

$$v_{\mathrm{lim}} = \frac{1}{TJ}\sum_{t=1}^{T}\sum_{j=1}^{J}\mathbf{1}\left[q_{t,j} < q_j^{\min} \ \vee \ q_{t,j} > q_j^{\max}\right].$$

For multi-dimensional joints, the violation is computed per degree of freedom and then averaged across all joint dimensions. This term penalizes poses that may look locally plausible in RGB frames but require anatomically invalid joint configurations in 3D.

Finally, the kinematic feasibility score is

$$F_{\mathrm{kin}}(v) = 1 - \frac{1}{3}\left(v_{\mathrm{vel}} + v_{\mathrm{spen}} + v_{\mathrm{lim}}\right).$$

Higher F_{kin} indicates smoother and more anatomically plausible body motion.

J.2 Contact Feasibility

Contact feasibility measures whether the body interacts with the ground in a physically plausible way. It penalizes foot sliding, ground penetration, floating feet, and poor balance.

For each foot $k \in \{L, R\}$, let $u_{t,v}$ denote the 3D position of mesh vertex v at frame t. We compute the foot height using the lowest sole vertex:

$$h_{t,f} = \min_{v\in\mathcal{V}_f} u_{t,v}^{(z)}.$$

We also compute the average foot-sole position:

$$p_{t,f} = \frac{1}{|\mathcal{V}_f|}\sum_{v\in\mathcal{V}_f} u_{t,v}.$$

A foot is considered to be in contact with the ground if it is both close to the ground and nearly stationary:

$$c_{t,k} = \mathbf{1}\left[h_{t,f} < 0.02 \ \wedge \ \|\dot{p}_{t,f}\|_2 < 0.05\right].$$

Foot sliding. If a foot is in contact with the ground, it should not slide significantly. We therefore measure the amount of foot displacement while contact is active:

$$\mathrm{slip}_{t,f} = c_{t,k}\cdot\|\dot{p}_{t,f}\|_2\cdot\Delta t.$$

The average foot-sliding violation is

$$v_{\mathrm{slip}} = \frac{1}{2T}\sum_{t=1}^{T}\sum_{k\in\{L,R\}}\mathrm{slip}_{t,f}.$$

Ground penetration. Generated or reconstructed feet can sometimes go below the ground plane. We measure penetration depth as

$$\mathrm{gpen}_{t,f} = \max(0, -h_{t,f}),$$

and average it over both feet and all frames:

$$v_{\mathrm{gpen}} = \frac{1}{2T} \sum_{t=1}^{T} \sum_{k \in \{L,R\}} \mathrm{gpen}_{t,f}.$$

Foot floating. A common artifact is that the root body moves while the feet do not make plausible contact with the ground, producing floating or whipping feet. To capture this, we compare the foot motion relative to the root with the root motion itself:

$$\rho_{t,f} = \frac{\|(\dot{p}_{t,f} - \dot{x}_{t,0})\|_2}{\|\dot{x}_{t,0}\|_2 + \varepsilon},$$

where $x_{t,0}$ denotes the pelvis/root joint. We flag a frame if the foot moves implausibly slowly relative to the root, or implausibly fast:

$$\rho_{t,f} < 0.6 \quad \text{or} \quad \rho_{t,f} > 1.75.$$

The floating violation is the fraction of frames where such implausible foot-root motion is detected:

$$v_{\mathrm{float}} = \frac{1}{2T} \sum_{t=1}^{T} \sum_{k \in \{L,R\}} \mathbf{1}\left[\rho_{t,f} < 0.6 \ \vee \ \rho_{t,f} > 1.75\right].$$

In implementation, we additionally include sequence-level checks for sustained non-contact and whole-body floating, which help detect cases where both feet remain airborne with an irregular trajectory.

Balance. A physically plausible standing or walking person should keep the projected center of mass close to the support region formed by the contacting feet. We approximate the center of mass as a weighted average of body joints, giving the pelvis three times the weight of other joints:

$$C_t = \frac{1}{\sum_j w_j} \sum_{j=0}^{J-1} w_j x_{t,j}, \qquad w_0 = 3, \quad w_{j>0} = 1.$$

Let $\mathcal{P}_t$ denote the support polygon, defined as the convex hull of the contacting ankle positions in the ground plane. We compute the distance from the projected center of mass $C_t^{(xy)}$ to this support polygon:

$$d_t = \begin{cases} 1.0, & \text{if there is no support contact,} \\ \min_{e \in \partial \mathcal{P}_t} \mathrm{dist}_e(C_t^{(xy)}), & \text{otherwise.} \end{cases}$$

The balance violation is the normalized average distance:

$$v_{\mathrm{bal}} = \frac{1}{T} \sum_{t=1}^{T} \frac{\mathrm{clip}(d_t, 0, 0.5)}{0.5}.$$

We use this continuous balance measure instead of a binary stable/unstable indicator because binary thresholds often saturate and provide less useful training signal.

The final contact feasibility score is

$$F_{\mathrm{con}}(v) = 1 - \frac{1}{4}\left(v_{\mathrm{slip}} + v_{\mathrm{gpen}} + v_{\mathrm{float}} + v_{\mathrm{bal}}\right),$$

where v_{slip} and v_{gpen} are normalized to $[0, 1]$ using fixed thresholds. Higher F_{con} indicates more plausible body-ground interaction.

J.3 Dynamic Feasibility

Dynamic feasibility measures whether the recovered motion could be produced by a physically plausible human body. It penalizes motions that require excessive ground reaction forces, excessive joint torques, or unusually large mechanical effort.

Ground reaction force. We estimate the ground reaction force needed to produce the observed center-of-mass acceleration. Let $m = 70\,\mathrm{kg}$ be the body mass and $g = 9.81\,\mathrm{m/s^2}$ be gravity. Using Newton's second law, we estimate

$$F_t^{\mathrm{GRF}} = \begin{bmatrix} m\ddot{C}_t^{(x)} \\ m\ddot{C}_t^{(y)} \\ m(g + \ddot{C}_t^{(z)}) \end{bmatrix}.$$

We penalize vertical forces larger than 3 times body weight:

$$v_{\mathrm{grf}}^{\mathrm{vert}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left[(F_t^{\mathrm{GRF}})^{(z)} > 3mg\right].$$

We also penalize horizontal forces larger than 0.5 times body weight:

$$v_{\mathrm{grf}}^{\mathrm{horiz}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left[\sqrt{(F_t^{\mathrm{GRF}})^{(x)2} + (F_t^{\mathrm{GRF}})^{(y)2}} > 0.5mg\right].$$

The ground-reaction-force score is

$$s_{\mathrm{GRF}} = 1 - \frac{1}{2}\left(v_{\mathrm{grf}}^{\mathrm{vert}} + v_{\mathrm{grf}}^{\mathrm{horiz}}\right).$$

Joint torque. We estimate the torque required to drive each joint using a simple segment-level inertia model:

$$\tau_{t,j} = I_j \|\ddot{x}_{t,j}\|_2,$$

where I_j is the approximate inertia for joint j. In our implementation, we use $I_j = 1.0\,\mathrm{kg\,m^2}$ for all joints.

For each joint type, we compute the fraction of frames where the estimated torque exceeds a plausible limit:

$$v_{\mathrm{torque}}^{(j)} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left[\tau_{t,j} > \tau_{\max}^{(j)}\right].$$

We use torque limits of 200, 300, 400, and $200\,\mathrm{N\,m}$ for the ankle, knee, hip, and spine, respectively. The torque score is

$$s_\tau = 1 - \frac{1}{J}\sum_{j=1}^{J} v_{\mathrm{torque}}^{(j)}.$$

Mechanical effort. Even if a motion does not exceed a single force or torque threshold, it may still require unrealistically large total effort. We therefore compute a mechanical-work proxy by multiplying joint torque with joint velocity:

$$\mathrm{MET} = \sum_{t=1}^{T}\sum_{j=0}^{J-1} \tau_{t,j}\|\dot{x}_{t,j}\|_2 \Delta t.$$

We convert this value into a normalized score:

$$s_{\mathrm{met}} = \max\left(0, 1 - \frac{\mathrm{MET}}{10000}\right).$$

The final dynamic feasibility score is

$$F_{\mathrm{dyn}}(v) = \frac{1}{3}\left(s_\tau + s_{\mathrm{GRF}} + s_{\mathrm{met}}\right).$$

Higher F_{dyn} indicates that the motion requires more plausible forces, torques, and total effort.

J.4 Overall Reward

The final PHYMOTION reward is the average of the three feasibility axes:

$$R_{\text{motion}}(v) = \frac{1}{3}\left(F_{\text{kin}}(v) + F_{\text{con}}(v) + F_{\text{dyn}}(v)\right).$$

All violation terms are clipped or normalized to $[0, 1]$, so higher scores consistently indicate more physically plausible human motion. This decomposition also makes the reward interpretable: kinematic feasibility captures body-shape and joint-motion artifacts, contact feasibility captures body-ground interaction failures, and dynamic feasibility captures physically infeasible forces and motions.

Table 14: Training hyperparameters for PHYMOTION training on Causal Forcing-1.3B and FastWan-1.3B. For the training time of FastWan, part of the GPU was running another job, so the reported time is longer than expected.

Setting	**Causal Forcing-1.3B**	**FastWan-1.3B**
Backbone & reward		
Base model	Causal Forcing-1.3B (chunkwise)	FastWan-1.3B (bidirectional)
Dataset	MotionX (21,348 training prompts)	MotionX (21,348 training prompts)
Resolution / frames	480×832, 45 frames	480×832, 45 frames
LoRA		
Rank r	256	256
Scale α	256	256
Dropout	0	0
Optimization		
Optimizer	AdamW	AdamW
Learning rate	1×10^{-5}	1×10^{-5}
(β_1, β_2)	$(0.9,\ 0.999)$	$(0.9,\ 0.999)$
Adam ϵ	1×10^{-8}	1×10^{-8}
Weight decay	1×10^{-4}	1×10^{-4}
Gradient clipping	$\|g\|_2 \leq 1.0$	$\|g\|_2 \leq 1.0$
Mixed precision	bf16	bf16
NFT objective		
KL interpolation β	0.1	0.1
KL coefficient λ	1×10^{-4}	1×10^{-4}
Advantage normalization	per-prompt, $\|A\| \leq 5$	per-prompt, $\|A\| \leq 5$
Rollout / sampling		
Denoising step list	$\{1000, 750, 500, 250\}$	$\{999, 856, 599\}$
`num_frame_per_block`	3	N/A (bidirectional, full-sequence)
KV-cache	enabled	n/a (bidirectional)
Guidance scale	1.0 (post-distillation)	1.0 (post-distillation)
Samples per prompt	24	24
Rollout batches per epoch	16	16
EMA		
EMA on LoRA params	yes	yes
EMA start step	0	0
Schedule & hardware		
Per-rank micro-batch size	3	3
Gradient accumulation steps	16	16
Reported step	330	330
GPUs	8× A100 80 GB	8× A100 80 GB
Wall-clock	$\sim$ 60.9 hours	$\sim$ 66 hours (crowded GPU)

Table 15: Licenses of external assets, models, datasets, codebases, and evaluation tools used in this work.

Asset / Method	Website / Source	License
Wan series [Team Wan et al., 2025]	https://github.com/Wan-Video/Wan2.2	Apache-2.0
Causal Forcing [Zhu et al., 2026]	https://github.com/thu-ml/Causal-Forcing	Apache-2.0
FastWan [Zhang et al., 2025]	https://github.com/hao-ai-lab/fastvideo	Apache-2.0
EchoMotion [Yang et al., 2026]	https://github.com/D2I-ai/EchoMotion	CC BY-NC 4.0
VideoAlign [Liu et al., 2025b]	https://github.com/KlingAIResearch/VideoAlign	MIT
VideoPhy [Bansal et al., 2025]	https://github.com/Hritikbansal/videophy	MIT
GVHMR [Shen et al., 2024]	https://github.com/zju3dv/GVHMR	Educational, research, and non-profit use only
Motion-X [Lin et al., 2023]	https://github.com/IDEA-Research/Motion-X	Non-commercial scientific research only
VBench / VBench-2.0 [Huang et al., 2024, Zheng et al., 2025]	https://github.com/Vchitect/VBench	Apache-2.0
Astrolabe [Zhang et al., 2026]	https://github.com/franklinz233/Astrolabe	Apache-2.0
SMPL [Loper et al., 2015]	https://smpl.is.tue.mpg.de/	SMPL-Body Creative Commons License / CC BY 4.0
MuJoCo [Todorov et al., 2012]	https://github.com/google-deepmind/mujoco	Apache-2.0

K Training Details

We provide the full training and rollout configuration used for our method in Table 14. Unless otherwise specified, all experiments use the same configuration. We report the checkpoint at training step 330.

L Licenses for Existing Assets

We show the licenses of all assets we use in Table 15

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the scope of PhyMotion: a structured 3D motion reward for human video generation, its agreement with human judgments, and its use for RL post-training of video generators.

   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: We discuss the assumptions, scope, and limitations of the proposed reward in Appendix A.

   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: Not applicable: the paper does not present formal theoretical claims requiring proofs.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: We provide the reward formulation and computation details in Appendix J, and the training setup in Appendix K. The public repository also provides implementation code and instructions for training and inference.

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: We publicly release the reward implementation, RL post-training code, and training/inference instructions at `https://github.com/h6kplus/PhyMotion`. We also release the Causal Forcing 1.3B LoRA adapter at `https://huggingface.co/6kplus/PhyMotion-CausalForcing-1.3B` and the MotionX-derived train/test prompt splits at `https://huggingface.co/datasets/6kplus/PhyMotion-MotionX-Prompts`; the base model and third-party dependencies are obtained separately.

   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: The training and evaluation settings are documented in Appendix K; the public repository additionally provides environment setup, training configurations, and inference commands.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: We report statistical significance tests for the correlation with human judgments in Appendix I.4. We do not report uncertainty estimates from repeated automatic-evaluation samples because of the computational cost of generating additional videos.

   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: Compute resources and training settings are described in Appendix K. The public repository further lists reference GPU hardware and approximate runtimes for reward computation, RL post-training, and inference.

   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: We reviewed the NeurIPS Code of Ethics and describe the human-subjects study, informed consent, potential risks, data handling, and responsible release considerations in the paper and appendix.

   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: We discuss potential benefits and risks, including misuse of synthetic human videos, in Appendix B.

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: [Yes]

Justification: We discuss risks associated with releasing a video-generation LoRA adapter, document the applicable base-model terms, and describe NSFW filtering of generated videos used in the study. These measures do not eliminate the potential for misuse.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: We credit the external codebases, models, datasets, and evaluation tools, and document their source URLs and applicable licenses or terms in Appendix L. The released prompt subsets are derived from MotionX, whose original source and terms are identified in the dataset card.

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: The public code repository documents setup, reward computation, training, and inference; the linked Hugging Face model and dataset cards document the released LoRA adapter and MotionX-derived prompt splits. Each public release includes documentation describing its contents and usage.

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: Appendix I provides the participant instructions, annotation interface, recruitment and compensation details, and data handling procedures. Annotators were compensated through their regular research assistant appointments, and personal identifiers were removed from the collected annotation records.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

    Answer: [N/A]

    Justification: The primary purpose of our human evaluation is to assess the quality of AI-generated videos, rather than to study human participants or their behavior.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: Not applicable to the core methodology: LLMs were used for assistance with code writing, debugging, and language editing, but not as an original or non-standard component of the proposed method. The authors reviewed and verified the code, experiments, and scientific claims.

    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.