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

NeurIPS 2026

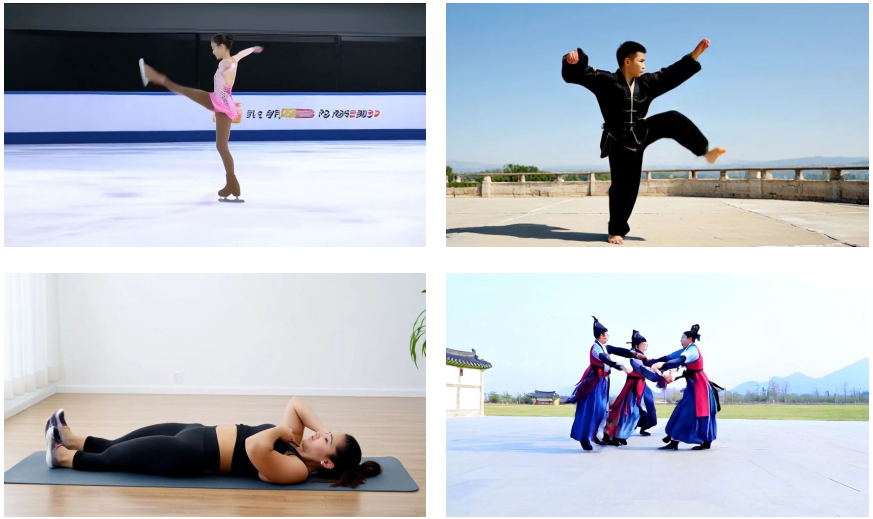
Yidong Huang

December 2026

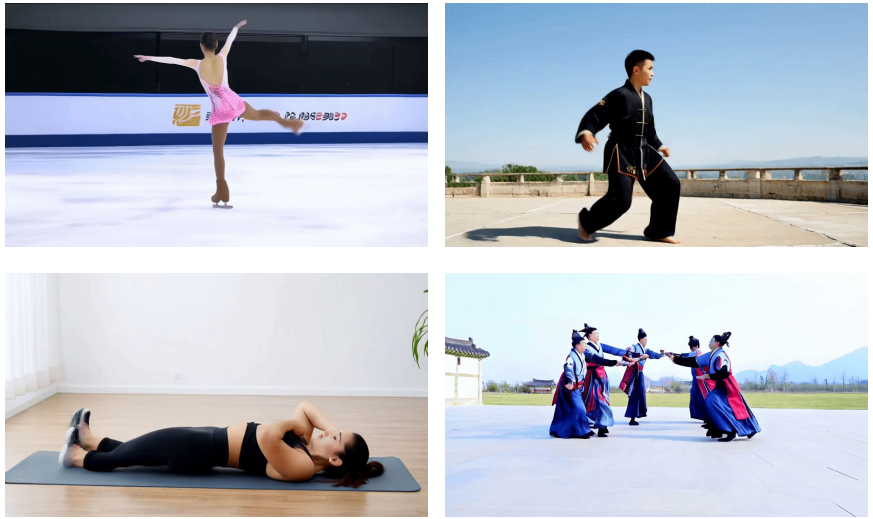
Visual Realism Does Not Guarantee Motion Realism

Human observers are highly sensitive to unnatural motion, even when individual frames appear realistic.

- Detect even subtle motion inconsistencies
- Difficult to articulate exactly what feels unnatural



Photorealistic Appearances



Unnatural Motions

Can Existing Video Metrics Detect Unnatural Motion?

Existing perceptual and VLM-based rewards can assign high scores to videos with physically implausible human motion.

- Visual realism ≠ physical correctness
- Miss force / balance violations



VBench: no human anomaly



VideoAlign: good motion quality

We Can Measure Physical Realism in 3D

Many unnatural human motions violate measurable physical constraints

- **3D motion** reveals errors hidden in 2D appearance
- **Kinematic constraints** capture unnatural body articulation
- **Contact and balance** identify floating, sliding, and instability
- **Physical dynamics** reveal implausible forces and torques



Input video

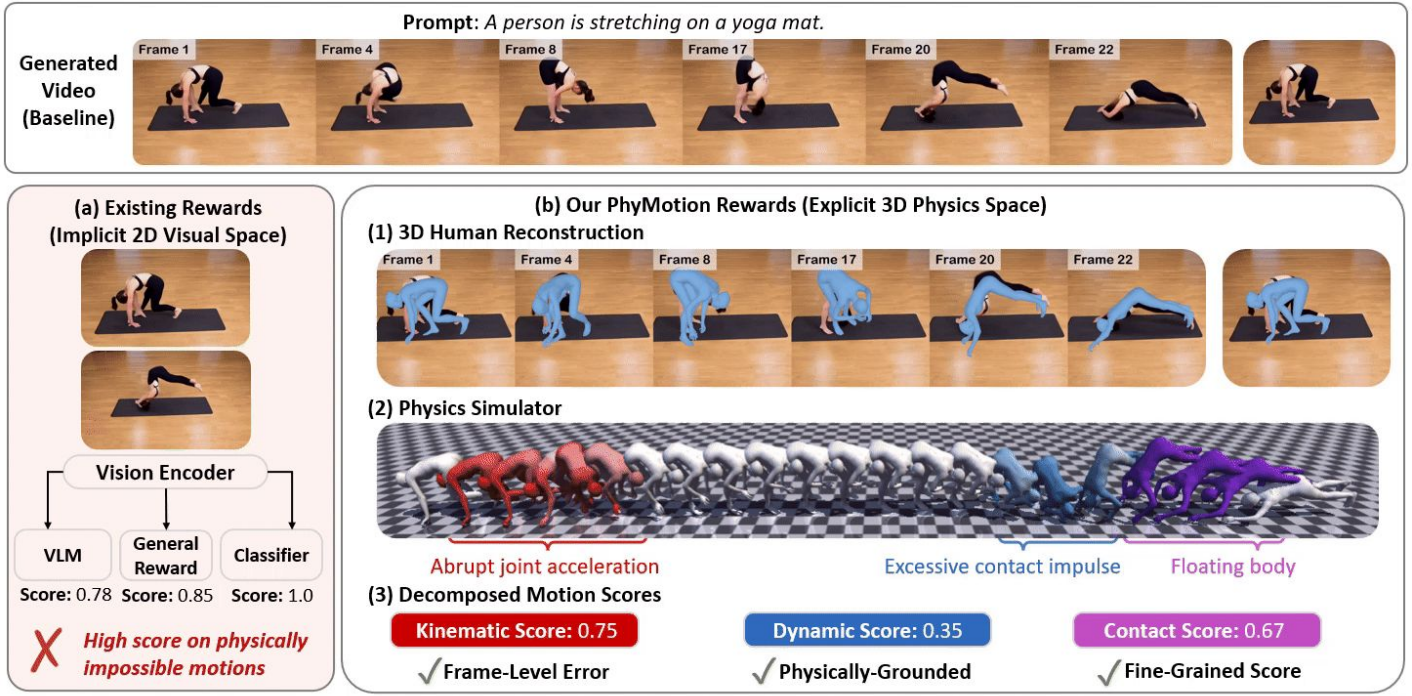


Reconstructed human

PhyMotion: Physics-Grounded Motion Reward

PhyMotion evaluates reconstructed 3D human motion using three physically grounded rewards.

- Kinematic: joint angular velocity
- Dynamic: force feasibility
- Contact: floating / sliding

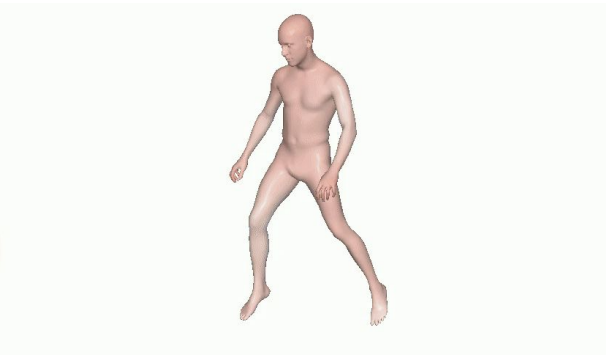


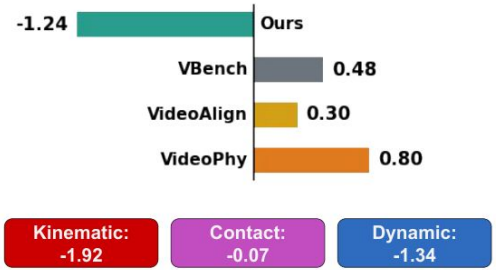
Kinematic Feasibility

Does the reconstructed body move smoothly and remain anatomically valid?

- **Joint angular velocity:** Are joint rotations unrealistically fast?
- **Self-penetration:** Do non-adjacent body parts intersect?
- **Joint limits:** Do poses exceed valid articulation ranges?







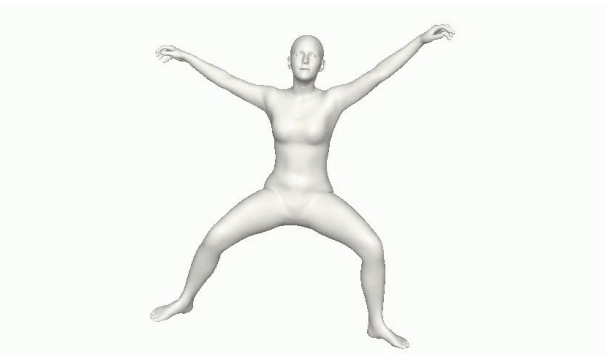
Example: sledgehammer swing → hand–hip self-penetration (PhyMotion Fig. 2a)

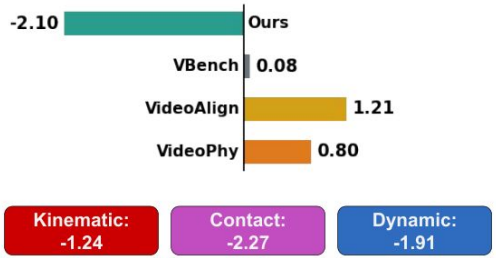
Contact & Balance Feasibility

Does the recovered body maintain physically plausible ground interactions?

- **Foot sliding:** Does a planted foot drift?
- **Ground penetration:** Does the foot sink below the floor?
- **Foot floating:** Is foot motion inconsistent with the root trajectory?
- **Balance:** Does the projected Center of Mass leave the support polygon?







Example: flying kick → implausible support / floating (PhyMotion Fig. 2b)

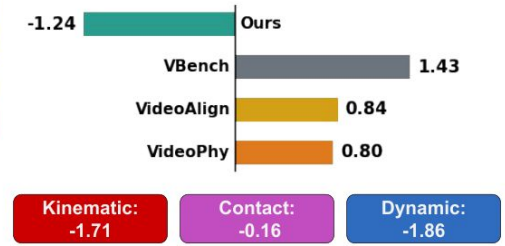
Dynamic Feasibility

Could a physically plausible human body generate the observed motion?

- **Joint torque:** Would tracking require excessive joint torque?
- **Ground reaction force:** Would contact forces exceed plausible limits?
- **Mechanical effort:** Is the summed torque–velocity effort unusually large?





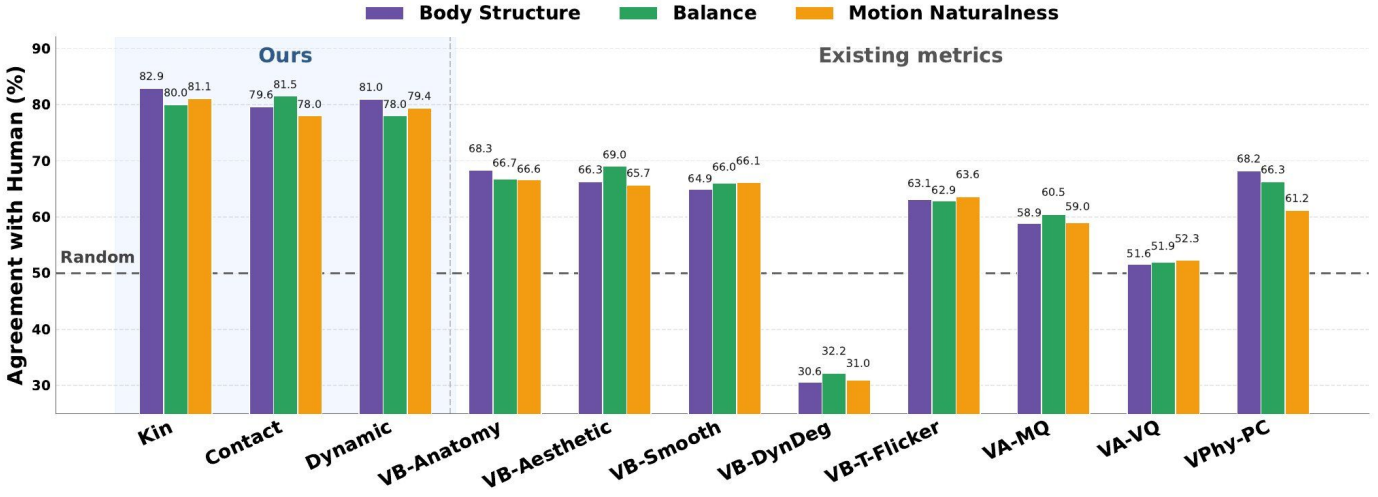


Example: baseball pitch → excessive joint torque (PhyMotion Fig. 2c)

PhyMotion Rewards Better Match Human Judgments

Our three physics rewards achieve around 80% agreement with human judgments, outperforming existing video metrics.

- **Study:** 1,200 same-prompt video pairs from six generators
- **Annotation:** six raters; blind A/B/Tie comparisons
- **Criteria:** body structure, balance, and motion naturalness
- **Baselines:** VBench (perceptual), VideoAlign, and VideoPhy (VLM)



RL Post-Training & Automatic Evaluation

We optimize PhyMotion through RL and evaluate beyond the training reward.

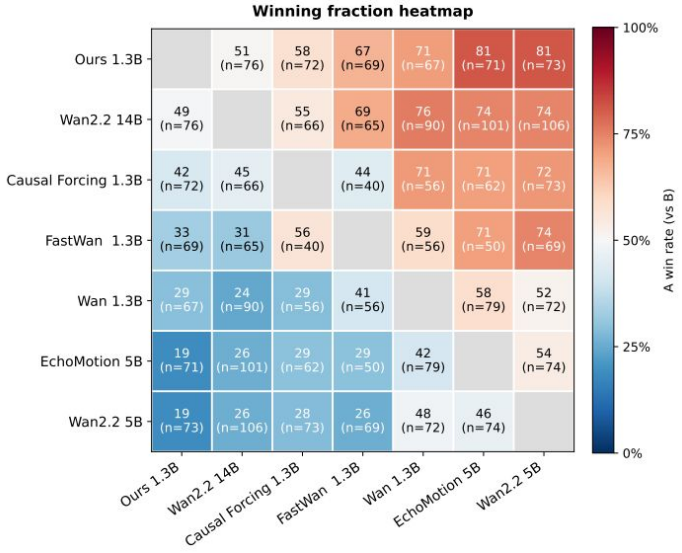
- **Backbones:** 4-step Causal Forcing (AR) and FastWan (bidirectional), 1.3B each
- **Training:**
  - **20K Motion-X prompts:** Curated diverse human-motion descriptions for training.
  - **DiffusionNFT:** Generate 24 videos per prompt and optimize using PhyMotion rewards.

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%)

RL Human Evaluation

Blind human comparisons test whether RL improvements are perceptually meaningful.

- **Pairwise study:** 1,487 same-prompt comparisons with hidden model identities
- **Three questions:** body structure, balance, and motion naturalness
- **Human votes:** Video A / Video B / Tie

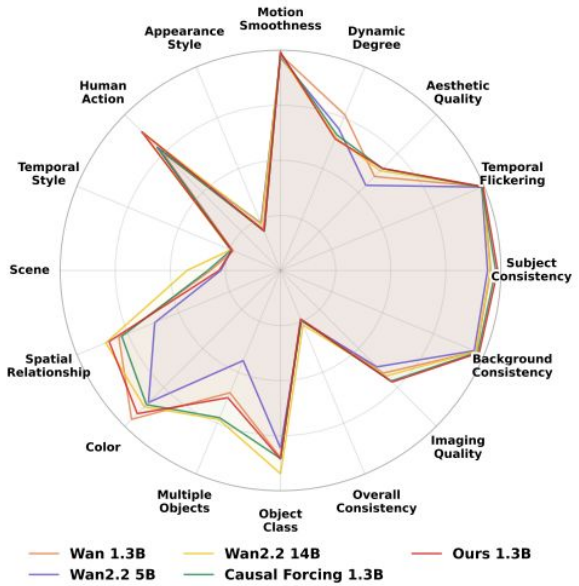


Model	Body	Bal.	Motion	All Q
Wan2.2 5B	1376	1388	1384	1383
EchoMotion 5B	1386	1403	1374	1387
Wan 1.3B	1429	1440	1411	1427
FastWan 1.3B	1526	1521	1528	1525
Causal 1.3B	1562	1546	1553	1553
Wan2.2 14B	1600	1593	1618	1604
Ours 1.3B	**1620**	**1610**	**1632**	**1621**

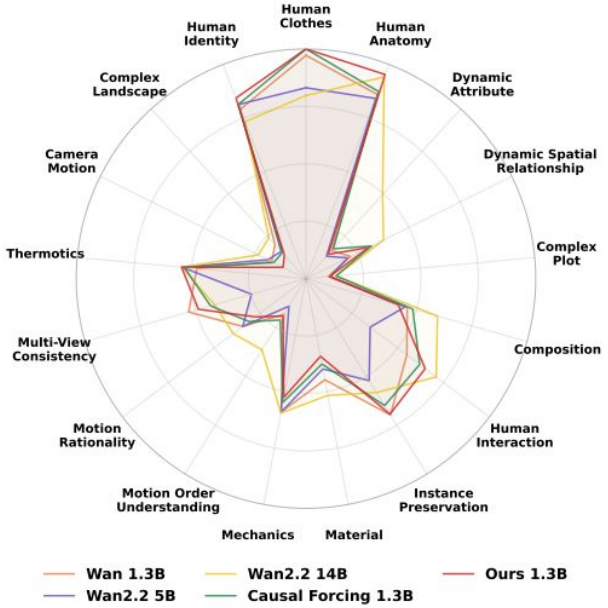
Does Physics-Grounded RL Hurt General Video Quality?

We evaluate whether human-motion-focused RL sacrifices performance on broader video prompts.

- **Evaluation:** standard VBench and VBench-2.0 evaluation suites
- **Matched comparison:** Causal Forcing 1.3B before and after RL



(a) VBench per-dimension comparison.



(b) VBench-2.0 per-dimension comparison.

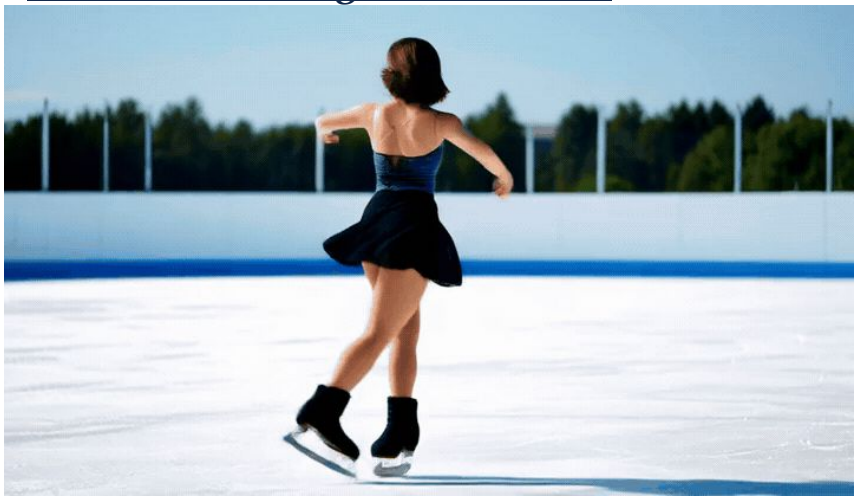
Qualitative Examples: Ice Skating

Prompt: A young woman glides backward on ice skates, pivoting with sweeping arm movements and alternating supporting legs.

PhyMotion RL (1.3B)



Causal Forcing Base (1.3B)



Source: PhyMotion project page · matched text prompt · 2.8-second loops

Qualitative Examples: Martial Arts

Prompt: A person performs a series of fluid martial arts strikes and a powerful jumping kick.

PhyMotion RL (1.3B)



Causal Forcing Base (1.3B)



Source: PhyMotion project page · matched text prompt · 2.8-second loops

Qualitative Examples: Handstand

Prompt: A young person lunges forward, plants both hands, inverts into a handstand, then lands back on their feet.

PhyMotion RL (1.3B)



Causal Forcing Base (1.3B)



Source: PhyMotion project page · matched text prompt · 2.8-second loops

Conclusion

From photorealistic appearances to physically plausible human motion.

- **Perceptual gap:** Visual realism does not guarantee physical correctness.
- **Structured reward:** Kinematics, contact, and dynamics provide interpretable 3D feedback.
- **RL improvement:** More natural motion without substantial loss of general video quality.