GEAR: GEOMETRY-AWARE REWARDS FOR HUMAN MOTION IN TEXT-TO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

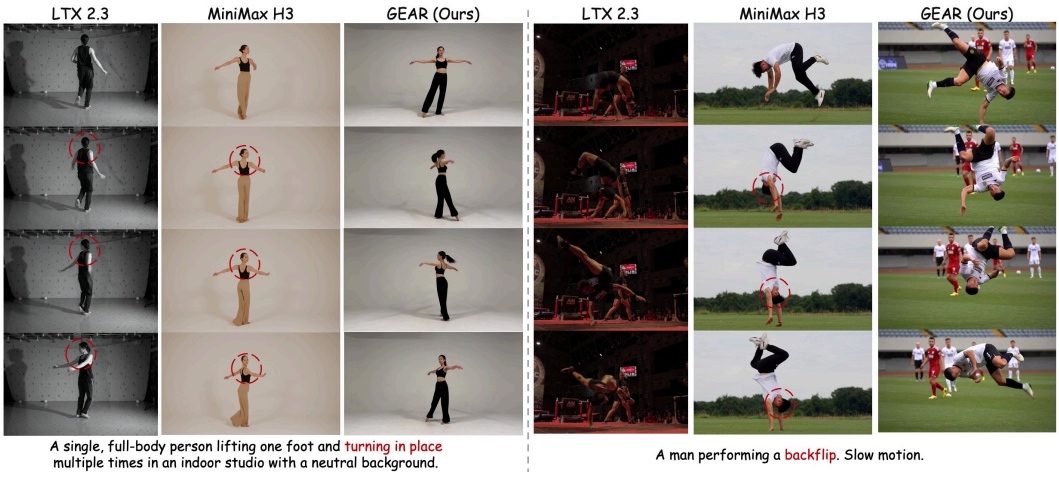

Figure 1: **Qualitative comparison of human-motion generation.** GEAR reduces structural artifacts while retaining the prompted motion execution across challenging examples.

ABSTRACT

Reinforcement learning (RL) has shown strong potential for improving text-to-video generation, yet generated human motion remains prone to structural deformations and temporal inconsistencies. Existing post-training rewards leave two gaps in this setting. First, structural credit is unreliable: holistic learned rewards entangle body correctness with other visual factors, while existing human-centric rewards rely on coarse detection cues or parametric body representations alone, leaving visible structural failures insufficiently captured. Second, optimizing structural consistency alone can favor low-motion shortcuts, compromising prompted motion execution. To address these gaps, we propose **GEAR** (**GE**ometry-**A**ware **R**ewards), an RL post-training framework that improves human-body structural correctness while preserving prompted motion execution. Specifically, it comprises a structural reward that retains image-space evidence from 2D keypoints and complements it with depth- and orientation-aware evidence from reconstructed 3D bodies. Moreover, a requirement-level semantic regularizer evaluates decomposed motion requirements independently without a fixed phase-to-time correspondence, discouraging near-static and incomplete-motion solutions. Experiments show that GEAR substantially reduces human-body deformation while preserving prompted motion execution.

1 INTRODUCTION

Text-to-video (T2V) models (Kong et al., 2024; Chen et al., 2024; Yang et al., 2025; Wan et al., 2025; HaCohen et al., 2026) have made rapid progress in visual quality and temporal coherence. Reinforcement learning (RL) (Xue et al., 2025b; Li et al., 2025b; Liu et al., 2025a; He et al., 2026; Mi et al., 2026) has further emerged as an effective approach for improving T2V generation. However, human motion remains a persistent challenge: generated videos still exhibit structural deformations and temporal inconsistencies, particularly in complex motions.

A central challenge is improving human-body structure while preserving the requested motion. Holistic learned rewards (He et al., 2024; Liu et al., 2025b; Xu et al., 2026; Tong et al., 2025)

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

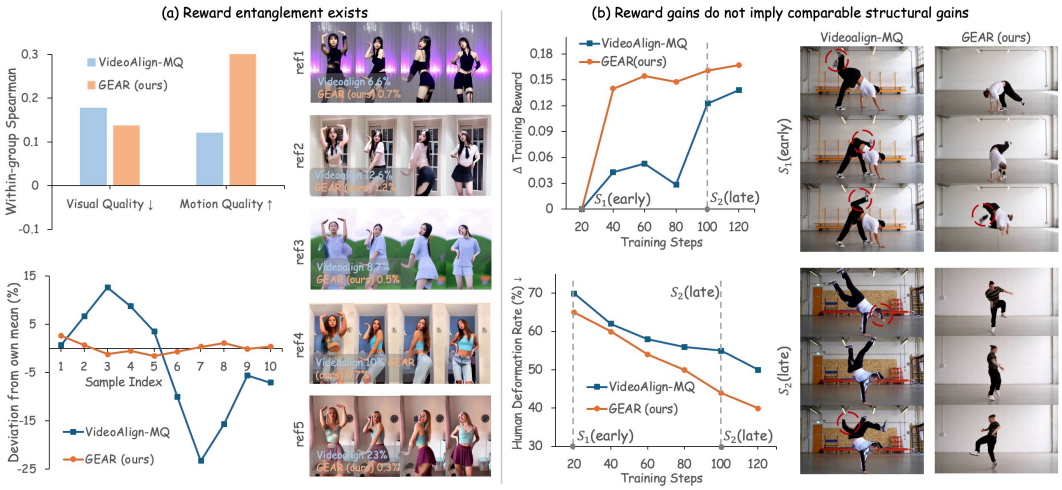

Figure 2: **Motivation for geometry-aware reward supervision. (a)** Across 10 groups of six same-prompt videos, VideoAlign-MQ correlates more with visual quality, while our geometry-based score better matches human motion-quality judgments and varies less under appearance changes with approximately matched motion. **(b)** On the same backbone, VideoAlign-MQ rises from S_1 to S_2, but its aligned policy reduces deformation less than GEAR.

can entangle structural correctness with other visual factors, obscuring structural credit. We examine this ambiguity using VideoAlign-MQ (Liu et al., 2025b) as a representative holistic motion reward. As shown in Figure 2(a), VideoAlign-MQ correlates more strongly with visual quality and is more sensitive to appearance changes, whereas our geometry-based score better matches human motion-quality judgments and remains more stable. During post-training (Figure 2(b)), VideoAlign-MQ increases, yet GEAR achieves a larger reduction in deformation. These observations show that improvements in a holistic motion reward need not translate into comparable gains in body structure.

Human-centric reward methods provide more targeted supervision (Xue et al., 2025a; Cheng et al., 2025; Jiang et al., 2025), but still leave blind spots. HuDA (Ashutosh et al., 2026) combines human detection confidence with phase-based motion supervision, yet its structural signal remains coarse, while its motion preservation relies on fixed temporal correspondence. PhyMotion (Huang et al., 2026) provides physics-based supervision, but may miss visible structural abnormalities and does not explicitly ensure prompted motion execution. These limitations call for more reliable structural supervision while preserving prompted motion execution.

We propose **GEAR** (**GE**ometry-**A**ware **R**ewards), an RL post-training framework that improves human-body structural correctness while preserving prompted motion execution. Its design follows two principles. First, *structural feedback should not rely on a single geometric representation*. GEAR derives cue-level feedback from multiple geometric measurements: 2D keypoints retain image-space evidence that may be attenuated by parametric fitting, while reconstructed 3D bodies provide complementary depth- and orientation-aware evidence. Second, *structural consistency alone does not ensure prompted motion execution*. Phase-based supervision such as HuDA (Ashutosh et al., 2026) assigns motion phases to fixed temporal positions, which may misalign supervision when motion timing varies. GEAR instead decomposes the prompt into motion requirements and evaluates each requirement independently across the video without fixed temporal assignment, discouraging near-static or incomplete-motion solutions.

We summarize our contributions as follows:

- We identify two limitations in human-motion post-training: ambiguous or attenuated structural supervision, and compromised motion execution under structural optimization.
- We propose GEAR, which combines complementary geometric evidence with requirement-level semantic regularizer without fixed temporal assignment.
- We show that GEAR improves human-body structure across multiple T2V backbones while preserving prompted motion execution.

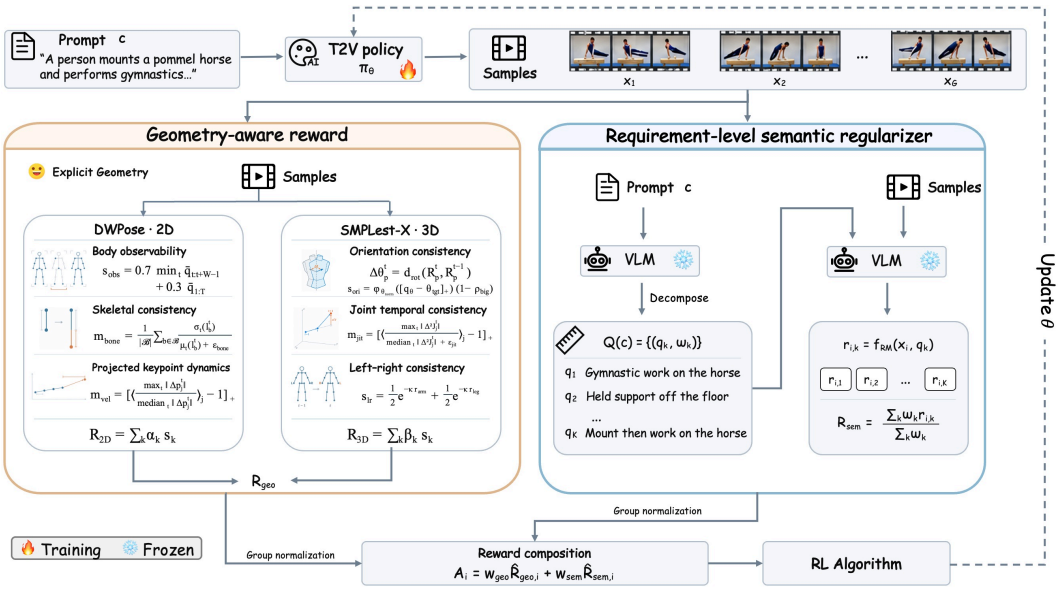

Figure 3: **Overview of GEAR.** GEAR combines structural feedback with motion preservation. **Left:** A multi-cue geometry reward converts complementary evidence into explicit structural feedback. **Right:** Requirement-level semantic regularizer supports prompted motion execution. The two signals are normalized separately and combined for RL optimization.

2 Related Work

Reinforcement learning for text-to-video generation. Recent T2V models achieve strong generation quality with diffusion and flow-based architectures (Ho et al., 2022; Blattmann et al., 2023; Brooks et al., 2024; HaCohen et al., 2024; Kong et al., 2024; Wu et al., 2025; HaCohen et al., 2026). RL and preference optimization further improve generation quality and prompt following (Black et al., 2024; Liu et al., 2025b; Xue et al., 2025b). However, for human motion generation, existing holistic rewards may entangle structural correctness with other visual factors. Our work instead provides explicit geometry-aware structural feedback for human-motion post-training.

Human-centric video generation and evaluation. Human-centric generation methods incorporate structural information through pose guidance (Hu, 2024; Zhu et al., 2024; Zhang et al., 2024; Li et al., 2025a) or joint appearance–motion modeling (Chefer et al., 2025; Yang et al., 2026; Wang et al., 2026). Recent benchmarks increasingly evaluate human fidelity and motion quality in generated videos (Zheng et al., 2025; Fang et al., 2026; Zhang et al., 2026; Thomas et al., 2025). For post-training, data-driven methods such as RealDPO (Cheng et al., 2025) and HuViDPO (Jiang et al., 2025) improve human motion through preference construction. HuDA (Ashutosh et al., 2026) combines detection confidence with phase-based motion supervision; its structural signal is coarse, while fixed temporal correspondence may misalign supervision when motion timing varies. PhyMotion (Huang et al., 2026) evaluates physical feasibility from parametric body reconstructions, where fitting may attenuate visible abnormalities and physical optimization alone does not explicitly ensure prompted motion execution. GEAR instead combines complementary geometric evidence with requirement-level motion supervision without fixed temporal assignment.

3 Method

3.1 Overview

GEAR is organized around two coupled roles in human-motion post-training: *structural credit assignment* and *motion preservation*. The first requires feedback that identifies which measurable aspects of body structure are inconsistent; the second prevents structural optimization from reduc-

Table 1: **Measurements for structural credit.**

Cue	Measured evidence	Space
Body observability	Detection coverage and confidence	2D
Skeletal consistency	Projected bone-length variation	2D
Projected keypoint dynamics	Peak-to-median 2D displacement	2D
Orientation consistency	Relative rotation changes	3D
Joint temporal consistency	Relative 3D second-difference peaks	3D
Left–right assignment consistency	Temporal arm/leg assignment stability	3D

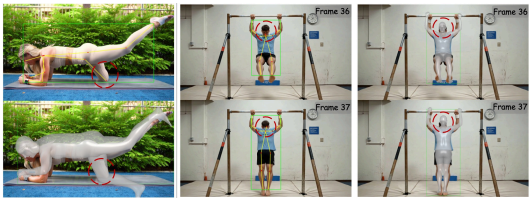

Figure 4: **Complementary geometric evidence.**

ing or omitting the prompted motion. We assign these roles to separate reward branches so that each signal has a clear optimization target.

Figure 3 illustrates the GEAR pipeline. Given a text prompt c, the T2V policy π_θ samples a group of G videos $\mathcal{X} = \{x_i\}_{i=1}^{G}$ under the same prompt. Each video is evaluated by two signals. The *geometry-aware structural reward* $R_{\text{geo},i}$ (Section 3.2) derives structural feedback from multiple interpretable geometric measurements, capturing complementary evidence across geometric representations. The *requirement-level semantic regularizer* (Section 3.3), implemented as an auxiliary reward $R_{\text{sem},i}$, evaluates decomposed motion requirements without fixed temporal assignment, unlike phase-based supervision (Ashutosh et al., 2026). The two rewards are normalized independently within each sampling group and combined into an advantage A_i for RL optimization. This design improves human-body structural correctness while explicitly preserving prompted motion execution.

3.2 Geometry-Aware Structural Reward

Design principle. Structural feedback should not rely on a single geometric representation. We construct R_{geo} from multiple interpretable geometric measurements (Table 1), making structural credit explicit. DWPose (Yang et al., 2023) provides 2D keypoints, while SMPLest-X (Yin et al., 2025) provides reconstructed body orientations and joint trajectories. Figure 4 illustrates their complementary evidence. In the left example, the subject's right leg is visibly deformed; parametric fitting nevertheless recovers a complete body, while 2D keypoints retain the abnormal image-space structure. In the right example, an orientation flip produces similar projected keypoint layouts across frames, whereas reconstructed body orientations make the change explicit. Accordingly, GEAR captures complementary structural and temporal cues across geometric representations. Additional quantitative and qualitative evidence for this complementarity is provided in Appendix A.2.

From measurements to scores. Each cue produces a score $s_k \in [0, 1]$, with larger values indicating greater consistency under that cue. For the scalar bone and trajectory violations, we use

$$s_k = \phi_{\sigma_k}(m_k), \qquad \phi_\sigma(m) = \exp\left(-\frac{1}{2}\left(\frac{m}{\sigma}\right)^2\right), \quad k \in \{\text{bone}, \text{vel}, \text{jit}\}. \tag{1}$$

Here $m_k \geq 0$ measures violation severity, and $\sigma_k > 0$ controls the decay scale. Gaussian calibration provides graded, bounded penalties and, unlike hard clipping, preserves variation for within-group ranking. Observability is scored directly; orientation uses the specialized calibration in Appendix A.3, and left–right consistency uses the mapping below. More details (*e.g.,* thresholds and aggregation weights) are provided in Appendix A.3.

2D geometric evidence. We extract whole-body keypoints with DWPose (Yang et al., 2023). *Body observability* combines detection coverage and confidence. For each frame t we form a frame-level detection quality

$$q^t = \rho^t \left(0.75\, \bar{c}^t_{\text{body}} + 0.25\, \bar{c}^t_{\text{whole}}\right), \tag{2}$$

where ρ^t is the fraction of body keypoints passing the confidence threshold, and $\bar{c}^t_{\text{body}}$ and $\bar{c}^t_{\text{whole}}$ are the mean confidences of the valid body and whole-body keypoints. The video-level score combines the worst sustained window with the overall mean,

$$s_{\text{obs}} = 0.7 \min_t \left(\frac{1}{W} \sum_{u=t}^{t+W-1} q^u\right) + 0.3\, \frac{1}{T} \sum_{t=1}^{T} q^t. \tag{3}$$

Coverage limits the score while confidence rewards reliable detections, and the minimum-window term prevents a short interval in which the subject becomes undetectable from being averaged away by the rest of the video.

Skeletal consistency measures relative fluctuations in projected bone lengths:

$$m_{\text{bone}} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \frac{\sigma_t(l_b^t)}{\mu_t(l_b^t) + \epsilon_{\text{bone}}}, \tag{4}$$

where l_b^t is the projected length of bone b at frame t, $\mathcal{B}$ is the bone set, μ_t and σ_t are the temporal mean and standard deviation, and $\epsilon_{\text{bone}} > 0$ ensures numerical stability. Normalization by each bone's temporal mean makes relative variation more comparable across bones and subject scales.

Projected keypoint dynamics uses m_{vel}, a peak-to-median measure of 2D keypoint displacement. It measures how far each joint's fastest transition departs from its typical displacement. Normalizing by each joint's own median makes this ratio invariant to uniform displacement rescaling, so high absolute speed alone does not lower the score.

We combine the three cue scores using nonnegative weights α_k:

$$R_{\text{2D}}(x) = \sum_{k \in \mathcal{C}_{\text{2D}}} \alpha_k s_k(x), \quad \mathcal{C}_{\text{2D}} = \{\text{obs}, \text{bone}, \text{vel}\}, \quad \alpha_k \geq 0, \quad \sum_k \alpha_k = 1. \tag{5}$$

3D geometric evidence. To complement image-space measurements with depth and orientation cues (Akhter & Black, 2015), we use SMPLest-X to estimate SMPL-X body geometry (Cai et al., 2023; Yin et al., 2025). Given its parametric body prior, we treat the reconstruction as complementary evidence rather than a replacement for 2D keypoints.

Orientation consistency measures changes in head, torso, and root orientation:

$$\Delta\theta_p^t = d_{\text{rot}}(\mathbf{R}_p^t, \mathbf{R}_p^{t-1}), \qquad t = 2, \ldots, T, \tag{6}$$

where $\mathbf{R}_p^t \in \text{SO}(3)$ is the orientation of part p, and d_{rot} is the geodesic rotation distance. For 3D cues, t indexes frames reconstructed at stride two. Using relative rather than absolute orientation avoids assuming a preferred facing direction. Orientation instability is summarized by the upper tail, extremes, and frequency (Appendix A.3).

Joint temporal consistency uses m_{jit}, the same peak-to-median construction applied to second-order differences of the estimated 3D joint positions $\mathbf{J}_j^t \in \mathbb{R}^3$. Temporal consistency is important for reliable video-based 3D human motion estimation (Kocabas et al., 2020; Choi et al., 2021), and operating in 3D exposes large depth-aware changes that may be weak in the image plane.

Left–right assignment consistency measures temporal instability in estimated limb correspondence:

$$s_{\text{lr}} = \tfrac{1}{2} e^{-\kappa r_{\text{arm}}} + \tfrac{1}{2} e^{-\kappa r_{\text{leg}}}, \qquad \kappa \geq 0. \tag{7}$$

Here r_{arm} and r_{leg} are transition violation rates, and κ controls penalty strength. The exponential decays smoothly with the violation rate and keeps each limb group's contribution bounded, with equal maximum weights for arms and legs.

The three scores are combined using nonnegative weights β_k:

$$R_{\text{3D}}(x) = \sum_{k \in \mathcal{C}_{\text{3D}}} \beta_k s_k(x), \quad \mathcal{C}_{\text{3D}} = \{\text{ori}, \text{jit}, \text{lr}\}, \quad \beta_k \geq 0, \quad \sum_k \beta_k = 1. \tag{8}$$

Structural reward aggregation. We combine the complementary evidence sources as

$$R_{\text{geo}}(x) = \frac{w_{\text{2D}} R_{\text{2D}}(x) + w_{\text{3D}} R_{\text{3D}}(x)}{w_{\text{2D}} + w_{\text{3D}}} \in [0, 1], \tag{9}$$

where $w_{\text{2D}}, w_{\text{3D}} \geq 0$, with a positive sum, balance 2D and 3D structural feedback. All cue thresholds and aggregation weights are summarized in Appendix A.3.

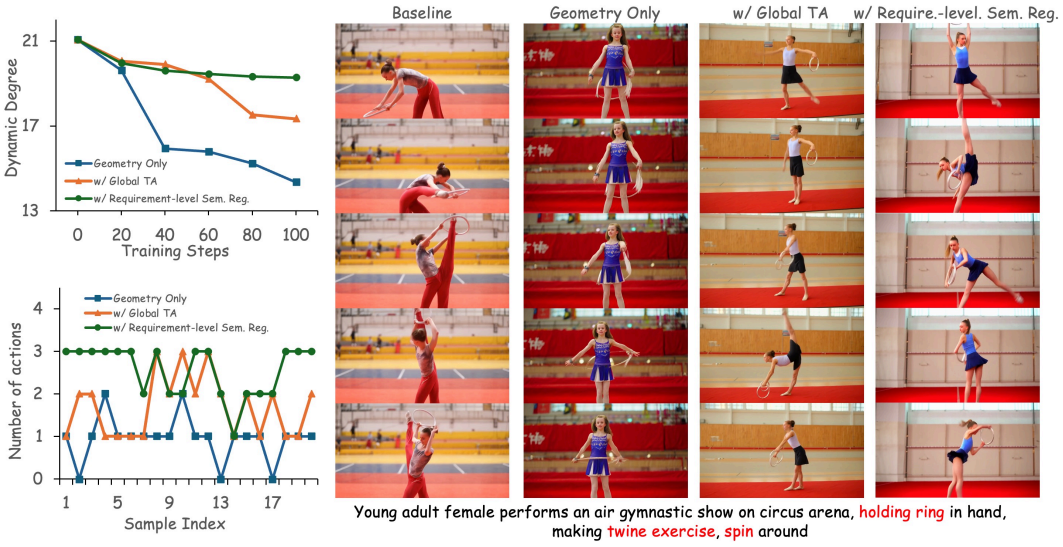

Figure 5: **Motivation for requirement-level semantic regularization.** Compared with geometry-only optimization and global text–video alignment, requirement-level supervision better preserves motion dynamics and motion execution.

3.3 Requirement-Level Semantic Regularizer

Motivation. Geometry-only optimization may favor low-motion and incomplete-motion solutions, compromising prompted motion execution. As shown in Figure 5 (left), the geometry reward alone substantially reduces dynamic degree. A global text–video alignment (TA) reward partially mitigates this drop, but its coarse supervision can reward partial execution while overlooking challenging motions. We therefore evaluate decomposed motion requirements independently across the video, better preserving motion dynamics and prompted motion execution. Figure 5 (right) provides a representative example: the baseline completes all three motions but exhibits body deformation; geometry-only optimization improves structure but omits required motions; and global TA still misses the spin. Our requirement-level supervision preserves structural correctness while completing all three prompted motions.

Semantic regularizer. As illustrated on the right of Figure 3, given a prompt c, we decompose its motion-related semantics, focusing the regularizer on preserving prompted motion execution:

$$\mathcal{Q}(c) = \{(q_k, \omega_k)\}_{k=1}^{K}, \tag{10}$$

where q_k is a motion requirement and $\omega_k \geq 0$ its weight (uniform, $\omega_k = 1/K$, in our implementation). Appearance-only content is discarded, and each requirement specifies *what* motion to execute without fixing *when*.

For each sampled video x_i, a frozen semantic video judge (Qwen3.5-397B-A17B (Qwen Team, 2026) in our implementation) evaluates every requirement independently:

$$r_{i,k} = f_{\mathrm{RM}}(x_i, q_k). \tag{11}$$

We aggregate the requirement-level scores into

$$R_{\mathrm{sem}}(x_i, c) = \frac{\sum_{k=1}^{K} \omega_k r_{i,k}}{\sum_{k=1}^{K} \omega_k}. \tag{12}$$

R_{geo} measures structural consistency, while R_{sem} measures prompted motion execution. Both are normalized within each sampling group and combined for RL optimization. Requirement construction and judging are detailed in Appendix A.4.

Table 2: **Quantitative comparison on public benchmarks.** Reward baselines share the HunyuanVideo 1.5-distilled backbone. Best results are in **bold**, second-best underlined. †Human-motion prompts are manually selected, as VideoJAM-bench has no official human-centric split.

Method	*VBench / VBench-2.0*			*VideoJAM-bench*†	
	Hum. An. ↑	Mot. Sm. ↑	Hum. Act. ↑	Mot. ↑	App. ↑
CogVideoX1.5-5B (Yang et al., 2025)	71.82	98.16	82	89.29	75.16
Wan2.2-5B (Wan et al., 2025)	80.22	98.48	81	88.73	77.94
Wan2.2-14B (Wan et al., 2025)	85.53	97.69	82	91.33	**79.43**
EchoMotion (Yang et al., 2026)	71.46	97.86	64	89.24	75.93
LTX-2.3 (HaCohen et al., 2026)	91.24	98.63	73	92.22	76.13
HunyuanVideo 1.5-distilled (Wu et al., 2025)	91.96	98.44	83	90.75	75.21
+ VideoAlign-MQ (Liu et al., 2025b)	93.09	98.79	82	92.45	76.59
+ HuDA (Ashutosh et al., 2026)	92.25	98.57	83	91.31	75.38
+ PhyMotion (Huang et al., 2026)	93.12	**99.20**	81	92.89	75.27
+ GEAR (Ours)	**94.89**	99.12	**85**	**95.17**	75.33

4 Experiments

4.1 Experimental Setup

Implementation details. We apply **GEAR** to HunyuanVideo 1.5-distilled (Wu et al., 2025) and optimize it with GRPO. For each prompt, we sample eight videos at 480×864 resolution and combine group-normalized geometry and requirement-level semantic rewards for policy optimization.

Data and evaluation. We train on curated single-person prompts featuring diverse motions, their generation procedure is detailed in Appendix A.1. For evaluation, we report Motion Smoothness and Human Action on VBench (Huang et al., 2024), Human Anatomy on VBench-2.0 (Zheng et al., 2025), and Appearance and Motion on selected human-motion subset of VideoJAM-bench (Chefer et al., 2025). For the backbone-portability and ablation studies, we additionally evaluate on the hard subset of HumanScore (Fang et al., 2026) (HumanScore-Hard), applying the same VBench-2.0 Human Anatomy, VBench Motion Smoothness, and VBench Human Action metrics to its prompts. Human evaluation focuses on Human Anatomy and Motion Smoothness on human-motion subset of VideoJAM-bench and on HumanScore-Hard.

Baselines. We compare GEAR with recent T2V models, including CogVideoX1.5 (Yang et al., 2025), Wan 2.2 (Wan et al., 2025), EchoMotion (Yang et al., 2026), and LTX-2.3 (HaCohen et al., 2026). For a controlled reward comparison, we implement VideoAlign-MQ (Liu et al., 2025b), HuDA (Ashutosh et al., 2026), PhyMotion (Huang et al., 2026), and GEAR on the same HunyuanVideo 1.5-distilled backbone with an identical policy-optimization setup. As HuDA has no publicly available implementation, we reproduce it following the paper. This setting reduces backbone confounding and isolates the effects of different reward formulations.

4.2 Comparison with existing methods

Quantitative results. Table 2 compares GEAR with open-source T2V models and same-backbone reward-alignment baselines. On HunyuanVideo 1.5-distilled, GEAR achieves the highest Human Anatomy, Human Action, and VideoJAM motion scores among the compared methods, reaching 94.89, 85, and 95.17, respectively. These correspond to gains of 2.93, 2.00, and 4.42 points over the base model. GEAR also achieves the second-best Motion Smoothness score of 99.12, while maintaining visual quality comparable to the base model (75.33 vs. 75.21 in appearance score). Together, these results support the effectiveness of combining geometric structural feedback with requirement-level semantic regularization to improve anatomical quality while preserving motion execution. The qualitative comparison below further illustrates structural quality and prompted motion execution.

Qualitative results. Figure 6 compares figure-skating and aerial-hoop motions. GEAR exhibits fewer visible limb and body distortions while retaining the prompted movements in these examples.

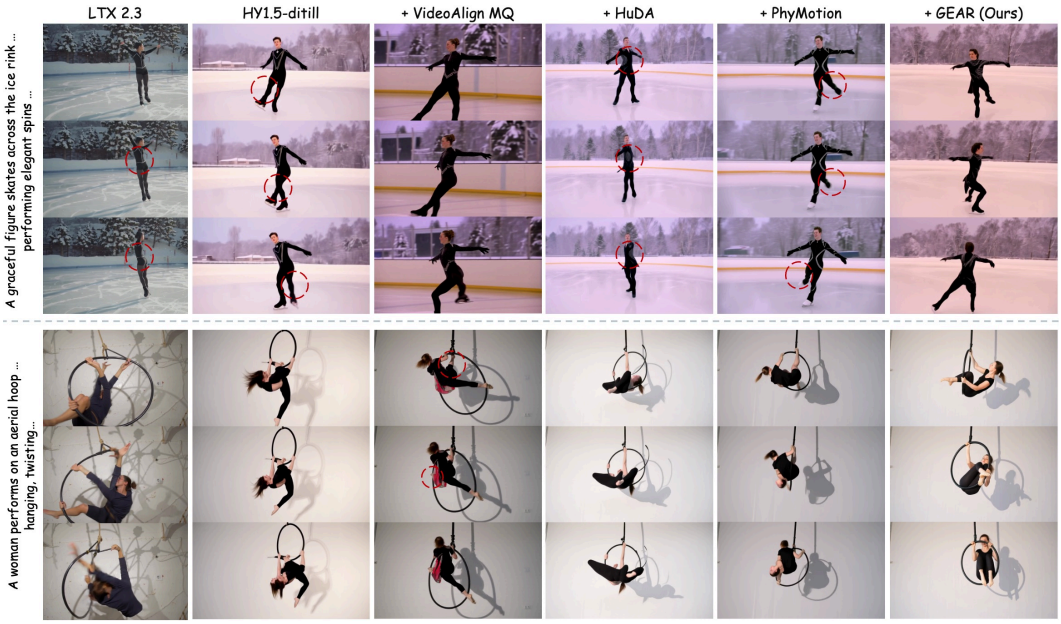

Figure 6: **Qualitative comparison on challenging human motions.** Compared with the base model, reward-based post-training, and T2V baselines, GEAR yields more coherent body structure while preserving the prompted motion.

Table 3: **User study.** Human preferences on VideoJAM-bench and HumanScore-Hard for Human Anatomy and Motion Smoothness.

Model	VideoJAM-bench		HumanScore-Hard	
	Hum. An. ↑	Mot. Sm. ↑	Hum. An. ↑	Mot. Sm. ↑
HunyuanVideo 1.5-distilled	40.50	41.50	19.67	25.17
+ PhyMotion	40.83	42.50	55.67	53.83
+ GEAR (Ours)	**68.67**	**66.00**	**74.67**	**71.00**

Table 4: **Application to alternative backbones.** Evaluation on MiniMax H3 and Wan2.2-5B on HumanScore-Hard.

Method	**Hum. An. ↑**	**Mot. Sm. ↑**
MiniMax H3	90.33	98.11
MiniMax H3 + GEAR (Ours)	**92.25**	**98.35**
Wan2.2-5B	78.59	97.38
Wan2.2-5B + GEAR (Ours)	**82.73**	**97.67**

User study. GEAR achieves the highest human preference scores on both benchmarks (Table 3). Compared with PhyMotion, it substantially improves both Human Anatomy and Motion Smoothness on VideoJAM-bench and HumanScore-Hard. See Appendix A.1 for the user study protocol.

4.3 Backbone Portability

We further apply GEAR to MiniMax H3 (MiniMax, 2026) and Wan2.2-5B (Wan et al., 2025) using the same reward formulation and optimization procedure, and evaluate them on the challenging HumanScore-Hard benchmark. As shown in Table 4, GEAR improves Human Anatomy from 90.33 to 92.25 on MiniMax H3 and from 78.59 to 82.73 on Wan2.2-5B, with modest gains in Motion Smoothness on both. These results support the portability of GEAR across different T2V backbones.

4.4 Structural Reward Validation

We evaluate reward–human agreement on 300 annotated video pairs from a separate diagnostic set, labeled for human-body structural and temporal stability. We report *coverage*, the fraction of pairs with a non-tied reward preference, and *conditional accuracy*, agreement with human preferences on these decided pairs. As shown in Table 5, on the same pairs GEAR's structural reward reaches 81.00% accuracy at full coverage, versus 60.33% for

Table 5: **Structural reward agreement with human judgments.**

Reward	**Coverage (%)**	**Cond. Acc. (%)**
VideoAlign-MQ	100.0	60.33
HuDA	97.3	57.88
PhyMotion	100.0	62.33
GEAR (R_{geo})	100.0	**81.00**

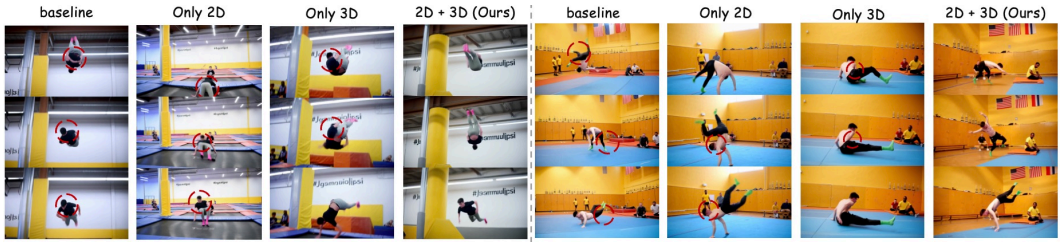

Figure 7: **Ablation of structural evidence.** The 2D branch mainly reduces image-space deformation but provides limited temporal gains, while the 3D branch improves temporal consistency but is less effective on visible image-space abnormalities. Their combination provides complementary improvements in both aspects.

Table 6: **Ablation studies of GEAR on HumanScore-Hard. Left:** geometric evidence ablation. **Right:** semantic supervision ablation. Best results are in **bold**.

Structural Evidence	Hum. An. ↑	Mot. Sm. ↑
Baseline	86.75	98.74
2D Cues	93.15	98.86
3D Cues	92.99	98.93
2D + 3D Cues	**95.63**	**99.09**

Optimization	Hum. Act. ↑	Hum. An. ↑
Baseline	**84**	86.75
Geometry only	79	**95.63**
Geometry + Global TA	82	91.36
Geometry + Req.-level Sem. (GEAR)	**84**	94.37

VideoAlign-MQ, 62.33% for PhyMotion, and 57.88% for HuDA at 97.3% coverage. Appendix A.2 provides individual-cue analysis and qualitative examples.

4.5 Ablation Studies

Complementarity of geometric evidence. We evaluate 2D-only, 3D-only, and combined structural evidence on HumanScore-Hard, with semantic regularization disabled for all variants (Table 6, left, and Figure 7). Either branch improves structural quality, while their combination performs best. Qualitatively, without 2D evidence, visible image-space abnormalities may remain despite plausible body reconstruction; without 3D evidence, depth- or orientation-related temporal inconsistencies are harder to capture. These results support the use of complementary geometric evidence.

Need for requirement-level semantic supervision. Figure 5 shows that geometry-only optimization can reduce motion dynamics and omit prompted motions, while requirement-level supervision better preserves motion execution. We quantify this effect on HumanScore-Hard in Table 6 (right). Geometry-only optimization improves Human Anatomy from 86.75 to 95.63, but reduces Human Action from 84 to 79. Adding global text–video alignment partially restores Human Action to 82, while sacrificing part of the structural gain (91.36). In contrast, GEAR restores Human Action to the baseline level of 84 while retaining a Human Anatomy score of 94.37, showing that requirement-level supervision preserves prompted motion execution while retaining most of the structural gain.

5 Conclusion

In this paper, we identify two limitations in human-motion RL post-training: structural feedback can be ambiguous in holistic rewards or attenuated by parametric body fitting, while geometry-only optimization may favor low-motion or incomplete-motion solutions. Accordingly, we present GEAR, which combines complementary geometric evidence with requirement-level semantic supervision without fixed temporal assignment. Experiments show consistent improvements in human-body structure across different T2V backbones, while preserving motion dynamics and prompted motion execution on the primary backbone. An offline analysis of 300 human-annotated video pairs further supports the agreement between the structural reward and human judgments.

Limitations. The structural reward may be unreliable for fine-grained hand or facial motion. Incorporating dedicated hand and face estimators with finer-grained geometric cues is a promising direction for future work.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript writing and refinement, LaTeX editing, portions of the code implementation, and the generation of a portion of the data. The authors reviewed the AI-assisted text, code, and generated data, verified the experimental results, claims, and citations, and take responsibility for the final manuscript.

REFERENCES

Ijaz Akhter and Michael J Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1446–1455, 2015.

Kumar Ashutosh, XuDong Wang, Xi Yin, Kristen Grauman, Adam Polyak, Ishan Misra, and Rohit Girdhar. Human detectors are surprisingly powerful reward models. *arXiv preprint arXiv:2601.14037*, 2026.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468, 2023.

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. *arXiv preprint arXiv:2502.02492*, 2025.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320. IEEE, 2024.

Guo Cheng, Danni Yang, Ziqi Huang, Jianlou Si, Chenyang Si, and Ziwei Liu. Realdpo: Real or not real, that is the preference. *arXiv preprint arXiv:2510.14955*, 2025.

Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1964–1973. IEEE, 2021.

Yusu Fang, Tiange Xiang, Tian Tan, Narayan Schuetz, Scott Delp, Li Fei-Fei, and Ehsan Adeli. Humanscore: Benchmarking human motions in generated videos. *arXiv preprint arXiv:2604.20157*, 2026.

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

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

Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. In *International Conference on Learning Representations*, volume 2026, pp. 132869–132895, 2026.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. 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*, pp. 2105–2123, 2024.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

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*, pp. 8153–8163, 2024.

Yidong Huang, Zun Wang, Han Lin, Dong-Ki Kim, Shayegan Omidshafiei, Jaehong Yoon, Jaemin Cho, Yue Zhang, and Mohit Bansal. Phymotion: Structured 3d motion reward for physics-grounded human video generation. *arXiv preprint arXiv:2605.14269*, 2026.

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

Lifan Jiang, Boxi Wu, Jiahui Zhang, Xiaotong Guan, and Shuang Chen. Huvidpo: Enhancing video generation through direct preference optimization for human-centric alignment. *arXiv preprint arXiv:2502.01690*, 2025.

Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5252–5262. IEEE, 2020.

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.

Hongxiang Li, Yaowei Li, Yuhang Yang, Junjie Cao, Zhihong Zhu, Xuxin Cheng, and Long Chen. Dispose: Disentangling pose guidance for controllable human image animation. In *International Conference on Learning Representations*, volume 2025, pp. 72213–72231, 2025a.

Junzhe Li, Yutao Cui, Tao Huang, Weijie Kong, Yiming Cheng, Chuxuan Zeng, Yinping Ma, Chun Fan, Miles Yang, Zhao Zhong, et al. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025b.

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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 40783–40818. Curran Associates, Inc., 2025a. doi: 10.52202/085713-1362. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3a10c46572628d58cb44fb705f25cbbf-Paper-Conference.pdf.

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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 82155–82192. Curran Associates, Inc., 2025b. doi: 10.52202/085713-2750. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/76227feb18ea0ee40bd15cf02c33e18e-Paper-Conference.pdf.

Xiaoyue Mi, Wenqing Yu, Jiesong Lian, Shibo Jie, Ruizhe Zhong, Zijun Liu, Guozhen Zhang, Zixiang Zhou, Zhiyong Xu, Yuan Zhou, et al. Video generation models are good latent reward models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4719–4728, 2026.

MiniMax. MiniMax H3. https://huggingface.co/MiniMaxAI/MiniMax-H3, 2026. Hugging Face model card, accessed September 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. URL https://qwen.ai/blog?id=qwen3.5.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Xavier Thomas, Youngsun Lim, Ananya Srinivasan, Audrey Zheng, and Deepti Ghadiyaram. Generative action tell-tales: Assessing human motion in synthesized videos. *arXiv preprint arXiv:2512.01803*, 2025.

Haibo Tong, Zhaoyang Wang, Zhaorun Chen, Haonian Ji, Shi Qiu, Siwei Han, Kexin Geng, Zhongkai Xue, Yiyang Zhou, Peng Xia, Mingyu Ding, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Mj-video: Benchmarking and rewarding video generation with fine-grained video preference. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 79142–79173. Curran Associates, Inc., 2025. doi: 10.52202/085713-2647. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/71ad539a57b1fd49b19e5c80070cb8b9-Paper-Conference.pdf.

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

Haoyu Wang, Hao Tang, Donglin Di, Zhilu Zhang, Wangmeng Zuo, Feng Gao, Siwei Ma, and Shiliang Zhang. Mosa: motion-coherent human video generation via structure-appearance decoupling. In *International Conference on Learning Representations*, volume 2026, pp. 24875–24905, 2026.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11269–11277, 2026.

Haotian Xue, Qi Chen, Zhonghao Wang, Xun Huang, Eli Shechtman, Jinrong Xie, and Yongxin Chen. Mogan: Improving motion quality in video diffusion via few-step motion adversarial post-training. *arXiv preprint arXiv:2511.21592*, 2025a.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025b.

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*, volume 2026, pp. 102259–102285, 2026.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4212–4222. IEEE, 2023.

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

Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, et al. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Bingzi Zhang, Kaisi Guan, and Ruihua Song. Hum-eval: A coarse-to-fine framework for human-centric video evaluation. *arXiv preprint arXiv:2604.25361*, 2026.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024.

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.

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*, pp. 145–162. Springer, 2024.

A Appendix

A.1 Implementation Details

We instantiate GEAR for RL-based post-training of a pretrained text-to-video diffusion (flow-matching) model and optimize it using GRPO (Shao et al., 2024; Liu et al., 2025a). Unless otherwise specified, experiments on the primary HunyuanVideo 1.5-distilled backbone use the same training configuration. We provide the complete hyperparameters in Table 7.

Group-relative reward composition. Because the geometry and semantic branches can have different numerical scales, we normalize them independently within each sampling group:

$$\widehat{R}_{m,i} = \frac{R_{m,i} - \mu_g(R_m)}{\max(\sigma_g(R_m),\, \epsilon_{\mathrm{grp}})}, \qquad m \in \{\mathrm{geo}, \mathrm{sem}\}, \tag{13}$$

where μ_g and σ_g denote the mean and standard deviation over the current sampling group $\{x_i\}_{i=1}^{G}$, and $\epsilon_{\mathrm{grp}} > 0$ lower-bounds the denominator for numerical stability when a group is nearly degenerate. The GRPO advantage is then

$$A_i = w_{\mathrm{geo}} \widehat{R}_{\mathrm{geo},i} + w_{\mathrm{sem}} \widehat{R}_{\mathrm{sem},i}, \tag{14}$$

where $w_{\mathrm{geo}}, w_{\mathrm{sem}} \geq 0$ with a positive sum, so the weights balance the normalized geometry and semantic signals.

User study protocol. We conducted an online, blinded ranking study with 20 tasks: 10 prompts from VideoJAM-bench and 10 from HumanScore-Hard. Each task presented three videos generated for the same prompt by HunyuanVideo 1.5-distilled, PhyMotion, and GEAR. Model identities were hidden, and video positions were shuffled for each participant. Participants ranked the videos separately for human anatomy and motion smoothness, with ties allowed.

We converted each ranking into within-prompt pairwise outcomes. For each benchmark and criterion, a method's preference score was

$$S = 100\, \frac{W + 0.5T}{W + T + L},$$

where W, T, and L denote its numbers of wins, ties, and losses. The main results use the 15 participants who completed all 20 tasks.

Table 7: **Training configuration of GEAR on the primary backbone.** Rollouts use classifier-free guidance 1.0 because the backbone is distilled.

Configuration	Value
Backbone	HunyuanVideo 1.5-distilled
Resolution	480×864, 121 frames, 24 fps
Sampling group size G	8
Sampling steps	12
Noise level η (SDE sampler)	0.5
Optimizer	Muon
Learning rate	5×10^{-6} (constant, no warmup)
Weight decay	0.01
Gradient accumulation	1
Training steps	120
KL coefficient	1×10^{-3}
Policy-ratio clipping	1×10^{-4}
Advantage clipping	5.0
Group-normalization constant ϵ_{grp}	1×10^{-6}
Geometry reward weight w_{geo}	0.5
Semantic reward weight w_{sem}	0.5
Precision / sharding	bf16, hybrid FSDP, sequence parallel 4

Training Data. We synthetically generate 1,600 training prompts with Gemini 3.1 Flash-Lite Preview, covering artistic gymnastics, trampoline, figure skating, and freestyle skateboarding. Each prompt describes one or two complex technical skills under a single-person, full-body, static-camera setting. More details are shown in Figure 8.

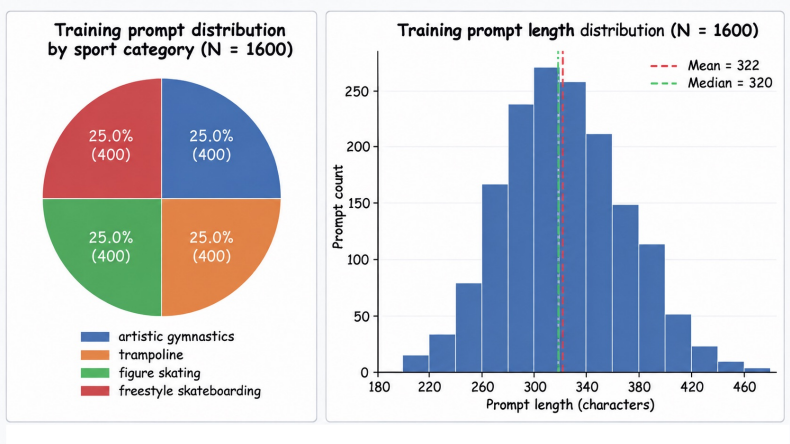

Figure 8: **Training prompt statistics.** Left: balanced distribution across four sport categories (400 prompts each). Right: prompt-length distribution of the 1600 synthetic training prompts.

Table 8: **Offline agreement of geometric reward components on 300 human preference pairs.** *Coverage* denotes the fraction of pairs receiving different scores, and accuracy is computed over these decided pairs. The full reward aggregates all cues.

Reward Variant	**Space**	**Coverage (%)**	**Cond. Acc. (%)**
Body observability	2D	53.0	62.89
Skeletal consistency	2D	99.7	59.53
Projected keypoint dynamics	2D	99.7	71.57
Orientation consistency	3D	90.0	72.59
Joint temporal consistency	3D	95.0	73.68
Left–right assignment consistency	3D	83.3	71.20
Full R_{geo}	**2D+3D**	**100.0**	**81.00**

A.2 Offline Human Agreement of the Structural Reward

Quantitative Results. This section extends the offline reward validation in Section 4.4 to individual geometric cues. We construct **300 human-annotated preference pairs** from a separate internal human-motion diagnostic set, labeled by *human-body structural and temporal stability*. For each pair (x^+, x^-), a reward is correct if it assigns a higher score to the human-preferred video x^+.

We evaluate each geometric cue independently using the same DWPose and SMPLest-X outputs. Because a cue may assign tied scores, we report *coverage*, the fraction of pairs receiving a non-tied preference, and *conditional accuracy* over these decided pairs. This analysis measures reward–human agreement and is separate from policy-level evaluation.

As shown in Table 8, all cues achieve conditional accuracy above 50%, with different coverage and discriminative strength. Projected keypoint dynamics reaches 71.57% accuracy with nearly full coverage, while body observability covers 53.0% of pairs. Orientation, joint temporal, and left–right consistency achieve 72.59%, 73.68%, and 71.20% accuracy, respectively, with 83.3–95.0% coverage. The full geometric reward reaches **100% coverage** and **81.00% accuracy**. Overall, individual cues provide partial structural evidence, while their combination yields broader coverage and higher agreement.

Qualitative results. Figure 9 illustrates reward preferences on human-labeled video pairs. In the first pair, VideoAlign-MQ favors the human-dispreferred video, whereas GEAR (R_{geo}) agrees with human judgments, consistent with the ambiguity of holistic structural feedback discussed in Section 1. In the second pair, the human-dispreferred video exhibits image-space structural abnormalities. The 2D reward correctly ranks this pair, while the 3D reward fails, illustrating how parametric



Figure 9: **Qualitative reward agreement with human judgments.** GEAR (R_{geo}) correctly ranks all three pairs. Higher scores indicate stronger preference; ✓/× mark agreement/disagreement with human judgments.

reconstruction may attenuate visible artifacts. In the third pair, the 3D reward better reflects the temporal inconsistency and correctly ranks the pair, whereas the 2D reward fails. GEAR combines these complementary signals and agrees with human preferences across all three examples.

A.3 Geometry-Aware Structural Reward Details

Section 3.2 defines the multiple geometric cues and the 2D/3D reward aggregation used by GEAR. To avoid repeating the main text, we only provide the implementation details omitted there: the peak-to-median computation for temporal cues, the specialized orientation calibration, the left–right assignment event-rate formulation, and the thresholds and aggregation weights used in our experiments.

Peak-to-median transition cues. Projected keypoint dynamics and joint temporal consistency share the same construction. Let $\mathcal{J}_k$ denote the set of joints with at least three strictly positive transitions, and let d_j^t be the transition magnitude of joint j at frame t: the first-order displacement $\|\mathbf{p}_j^{t+1} - \mathbf{p}_j^t\|$ of frame-normalized 2D keypoints for $k = \text{vel}$, and the second-order difference $\|\mathbf{J}_j^{t+1} - 2\mathbf{J}_j^t + \mathbf{J}_j^{t-1}\|$ of estimated 3D joint positions for $k = \text{jit}$. We compute

$$m_k = \left[\frac{1}{|\mathcal{J}_k|} \sum_{j \in \mathcal{J}_k} \frac{\max_t d_j^t}{\text{median}_t\, d_j^t + \epsilon_k} - 1 \right]_+, \qquad k \in \{\text{vel}, \text{jit}\}, \tag{15}$$

where the median is taken over strictly positive transitions, $\epsilon_{\text{jit}} = 0.01$ stabilizes the ratio for nearly static joints, and $\epsilon_{\text{vel}} = 0$. For constant positive transition magnitudes, the ratio is at most one and contributes no positive violation. Thus, m_k measures excess transition peaks relative to typical

transition magnitudes. The resulting values are converted to scores using Eq. (1); if no joint has enough valid transitions, the cue returns its maximum score and the estimator reliability gates below determine how much it is trusted. Normalizing each joint by its own median measures relative rather than absolute transition peaks. Because legitimate rapid motion can also produce a large peak-to-median ratio, this cue is treated as one proxy within the aggregated reward rather than as a standalone test of temporal failure.

Orientation aggregation and calibration. The main text defines the frame-wise part rotation changes $\Delta\theta_p^t$ in Eq. (6). For $\mathcal{P} = \{\text{head}, \text{torso}, \text{root}\}$, we collect

$$\Delta\Theta = \{\Delta\theta_p^t \mid p \in \mathcal{P},\ t = 2, \dots, T\} \tag{16}$$

and summarize both persistent and extreme changes as

$$q_\theta = \max\left(\mathrm{P}_{95}(\Delta\Theta), \tfrac{1}{2} \max_{\delta \in \Delta\Theta} \delta\right), \qquad \rho_{\text{big}} = \frac{1}{|\Delta\Theta|} \sum_{\delta \in \Delta\Theta} \mathbf{1}[\delta > \theta_{\text{big}}]. \tag{17}$$

The percentile term reduces sensitivity to isolated estimator noise, whereas the maximum and large-change rate retain sensitivity to abrupt or persistent orientation changes. We use

$$s_{\text{ori}} = \exp\left[-\frac{1}{2}\left(\frac{[q_\theta - \theta_{\text{tgt}}]_+}{\theta_{\text{norm}}}\right)^2\right](1 - \rho_{\text{big}}), \tag{18}$$

where $[z]_+ = \max(z, 0)$. The first factor leaves q_θ values below θ_{tgt} unpenalized and decays for larger values, while ρ_{big} separately penalizes the frequency of changes above θ_{big}.

Left–right assignment events. Equation (7) uses the arm and leg transition violation rates r_{arm} and r_{leg}. For each limb group $g \in \{\text{arm}, \text{leg}\}$, let π_g^t denote the estimator's canonical left–right assignment at frame t, and let $\mathcal{V}_g$ be the set of transitions with valid assignments. We compute

$$r_g = \frac{1}{|\mathcal{V}_g|} \sum_{t \in \mathcal{V}_g} \mathbf{1}[\pi_g^t \neq \pi_g^{t-1}]. \tag{19}$$

This is an estimator-based temporal consistency proxy: it can also respond to ambiguous limb crossings or left–right labeling errors from the body estimator, so we do not interpret it as a direct detector of visible identity swaps.

Estimator reliability gates. Both pose estimators can fail on generated videos, making cue scores unreliable even when the video itself is plausible. We therefore apply two gates that interpolate a score toward a neutral value s_0 whenever the evidence supporting it is weak. Both use the smoothstep $\psi(x) = \hat{x}^2(3 - 2\hat{x})$ with $\hat{x} = \text{clip}(x, 0, 1)$.

The *per-estimator quality gate* asks whether an estimator produced usable output at all. Given its mean detection quality $\bar{q}$ and valid-frame ratio ρ,

$$g = \text{clip}(\max(g_{\min}, \min(\psi(\bar{q}/q^*), \psi(\rho/\rho^*))), 0, 1), \qquad s_k \leftarrow g\, s_k + (1 - g)\, s_0, \tag{20}$$

applied to $\{s_{\text{bone}}, s_{\text{vel}}\}$ for the 2D estimator and to $\{s_{\text{ori}}, s_{\text{jit}}, s_{\text{lr}}\}$ for the 3D estimator, each with its own quality target q^*.

The *cross-estimator gate* moderates low temporal scores using support between first-order 2D keypoint transitions and second-order 3D joint transitions. For $k \in \{\text{vel}, \text{jit}\}$, let k' denote the other temporal cue. We convert the counterpart score into an anomaly level and a support gate,

$$a_{k'} = \max(0,\, 1 - s_{k'}), \qquad g_k = g_{\min}^{\times} + \left(1 - g_{\min}^{\times}\right) \psi(a_{k'}/a^*), \tag{21}$$

and use it to soften penalties only:

$$s_k \leftarrow \begin{cases} s_k, & s_k \geq s_0, \\ g_k\, s_k + (1 - g_k)\, s_0, & s_k < s_0. \end{cases} \tag{22}$$

A low score is retained in full when the counterpart anomaly reaches a^*; otherwise its penalty is softened toward s_0. Scores at or above s_0 are unchanged by this gate. The positive floor $g_{\min}^{\times}$ retains part of a one-sided penalty. The quality gate is applied first, and branch rewards are computed from the resulting cue scores.

Table 9: **Implementation parameters of the geometry-aware structural reward.** Cue weights are reported after the normalization $\sum_k \alpha_k = \sum_k \beta_k = 1$; the unnormalized 3D weights are 0.85 for orientation, 1.0 for joint temporal consistency, and 0.6 for left–right assignment consistency. σ_{vel} and σ_{jit} share a value but are not directly comparable: the 2D branch uses every frame while the 3D branch uses a stride of two.

Geometric cue	Parameter(s)	Value	Weight
Body observability	keypoint confidence / visible-ratio validity, window W	0.3, 0.25, 6 frames	$\alpha_{\text{obs}} = 0.250$
Skeletal consistency	$\epsilon_{\text{bone}}, \sigma_{\text{bone}}$	10^{-6}, 0.18	$\alpha_{\text{bone}} = 0.375$
Projected keypoint dynamics	σ_{vel}	8.0	$\alpha_{\text{vel}} = 0.375$
Orientation consistency	$\theta_{\text{tgt}}, \theta_{\text{norm}}, \theta_{\text{big}}$	20°, 35°, 90°	$\beta_{\text{ori}} = 0.347$
Joint temporal consistency	$\epsilon_{\text{jit}}, \sigma_{\text{jit}}$	0.01, 8.0	$\beta_{\text{jit}} = 0.408$
Left–right assignment consistency	κ; confidence / motion / margin gates	5.0; 0.5, 0.08, 0.05	$\beta_{\text{lr}} = 0.245$
2D / 3D branch combination	–	–	$w_{\text{2D}} = 0.5,\ w_{\text{3D}} = 0.5$
Estimator quality gate	$s_0, g_{\min}, q^*_{\text{2D}}$ / q^*_{3D}, ρ^*	0.75, 0.15, 0.35 / 0.70, 0.8	–
Cross-estimator gate	$s_0, g^{\times}_{\min}, a^*$	0.75, 0.25, 0.25	–

Table 10: **Implementation details of the requirement-level semantic regularizer.** Requirements are decomposed once per prompt and cached, so all samples of a group share the same list. Motion filtering keeps only motion, temporal-order, completeness, and persistence requirements.

Setting	Value
Requirement construction model	Qwen3.5-397B-A17B (greedy, 8192 max tokens, seed 42)
Number of requirements K	median 3, range 1–7 (mean 3.1)
Requirement weighting ω_k	uniform, $\omega_k = 1/K$
Semantic video judge	Qwen3.5-397B-A17B (frozen, same model)
Video evidence	24 frames: 16 uniform + 8 change-peak, long side ≤ 512
Judge output range	fulfillment and confidence, each an integer in $[0, 100]$
Confidence shrinkage	$r_{i,k} = \frac{1}{2} + \left(\frac{f_{i,k}}{100} - \frac{1}{2}\right)\frac{c_{i,k}}{100}$
Decoding temperature	0.0 (top-p 0.01, thinking disabled)

Implementation parameters. Unless otherwise specified in the ablations, cue thresholds and aggregation weights are fixed across experiments. The parameters omitted from the main text are summarized in Table 9. Here, σ values set the decay scale of the calibration in Eq. (1), α, β, and w_{2D} and w_{3D} are the cue- and branch-level aggregation weights, and the last two rows give the gate constants of Eqs. (20)–(22).

A.4 Requirement-Level Semantic Regularizer Details

Section 3.3 defines the requirement-level semantic reward and its aggregation. Here we provide the construction and judging protocol omitted from the main text for reproducibility.

Motion-requirement construction. For each prompt, we retain only motion-related content and decompose compositional motions into individually judgeable requirements. Static attributes such as appearance, background, and camera are excluded. Requirements specify *what* motion should be executed without assigning fixed temporal positions, while explicit relative relations such as "A followed by B" are preserved. This allows motions with varying durations or execution speeds to be evaluated without fixed phase-to-time correspondence.

Semantic judging protocol. We use the same frozen Qwen3.5 (Qwen Team, 2026) judge for all samples. Requirements are decomposed once per prompt and cached, so all G samples are evaluated with an identical requirement list. Each requirement q_k is scored independently using 24 frames sampled across the video, and the scores are aggregated by Eq. (12). Video-wide sampling allows motions to occur at varying temporal positions without fixed temporal anchors.

Additional implementation settings, including requirement construction, weighting, and judge configuration, are summarized in Table 10.

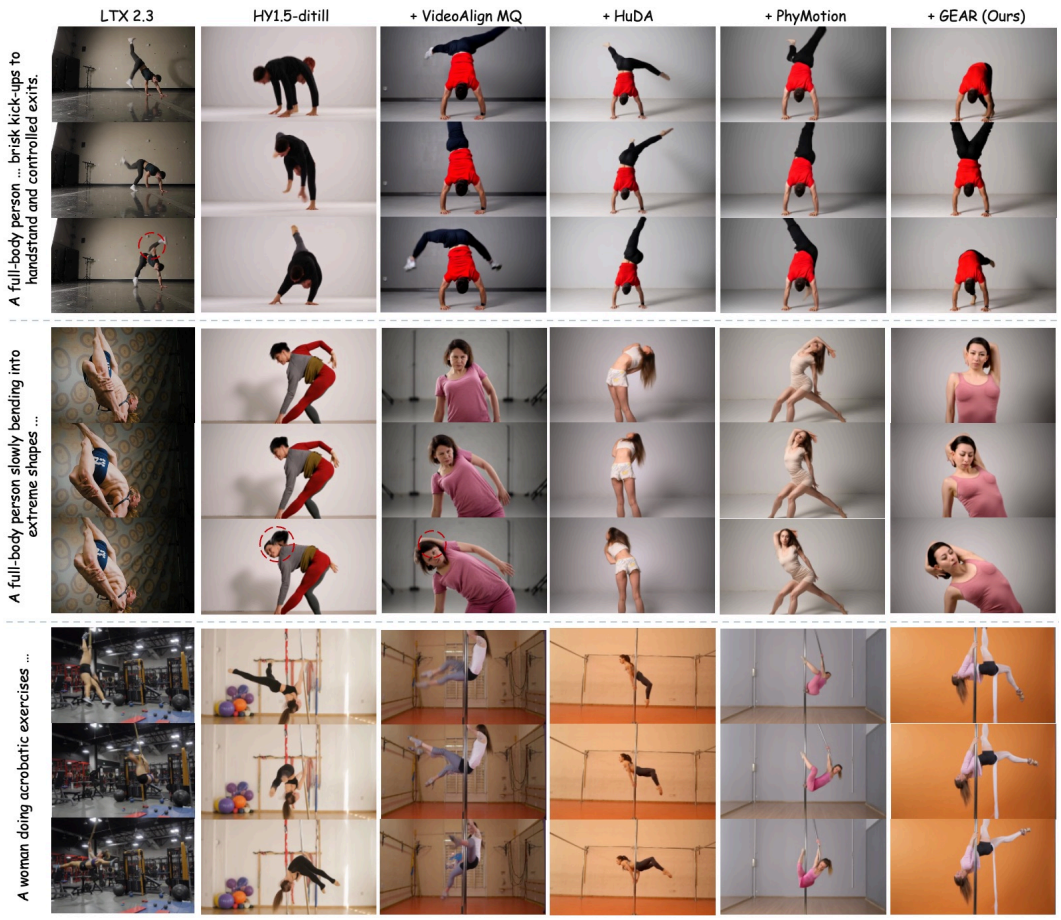

Figure 10: **Additional qualitative comparisons.** GEAR produces more structurally coherent human motion while preserving the prompted motion across challenging examples.

A.5 More Qualitative Results

Figure 10 provides additional qualitative results that further demonstrate the effectiveness of **GEAR**.