IsoVGRBench: Video Reward Modeling Without Human Feedback

Anonymous authors
Paper under double-blind review

Abstract

Reinforcement learning post-trains video generators against reward models, so a generator improves only along the properties that its reward model perceives. Video reward models are trained and evaluated on human preferences between videos that often differ simultaneously on properties such as prompt alignment, visual fidelity, motion, and generator identity. As a result, aggregate accuracy does not reveal which video properties a reward model can reliably distinguish. We construct preference pairs by *controlled interventions* that change one property of a generated video while holding the rest fixed. This determines both the preferred side and the property under test without per-pair human labels, requiring human validation only once per intervention to confirm perceptibility. We propose IsoVGRBench, a benchmark of $8{,}000$ such pairs across eight interventions, and show that current video reward models separate broad quality differences between generators but miss frame order. The best reward model reaches only 60.8% on reordered frames against 100% for humans, and rankings on existing benchmarks do not predict this failure. Intervention pairs can also serve as training data: our IsoVRM-Q2 outperforms VideoReward, which shares its base model but trains on human preferences, by 3.8 points on VBench-2.0 and provides better supervision for video generator post-training. Adding intervention pairs to VideoReward's human-preference supervision further improves transfer. Because intervention pairs are automatically generated rather than collected, the benchmark and training data can be regenerated as generators evolve.

1 Introduction

Reinforcement learning post-trains video generators against reward models (Liu et al., 2025b; Rafailov et al., 2023; Wallace et al., 2024), which rank candidates at sampling time and define the objective for preference optimization. A generator therefore improves only along the properties its reward model perceives. Yet video reward models are trained and evaluated on human preferences between independently generated videos, which differ simultaneously in visual quality, content, motion, and generator identity. Neither the preference labels nor a model's accuracy on these comparisons reveals which properties the model perceives, even when annotators are instructed to judge a particular dimension. Existing benchmarks also rely on outputs from fixed generator sets and per-sample human labels (Liu et al., 2025b; Li et al., 2024; Tong et al., 2025; Zheng et al., 2025), which limits their scale and leaves them stale as generators and their failure modes evolve.

We construct comparisons by controlled intervention instead, and use them to evaluate reward models, train them, and post-train the video generators they supervise (Figure 1). Starting from an anchor video, an intervention changes one property while holding the rest fixed, so both the preferred side and the property under test follow from the construction. **Temporal Shuffle**, for example, preserves every frame of the anchor but changes their order, directly testing whether a model reads order at all, while **Frame Hold** creates stuttering motion from the anchor's own frames. No pair needs a human label, and human validation is needed only once per intervention to confirm that the intended change is perceptible. Because the same interventions apply to any generator's outputs, the pairs can be regenerated as generators evolve. We use this procedure to construct IsoVGRBench, our benchmark of $8{,}000$ such *intervention-isolated* pairs from eight intervention families that target motion, text alignment, visual quality, and overall quality.

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.

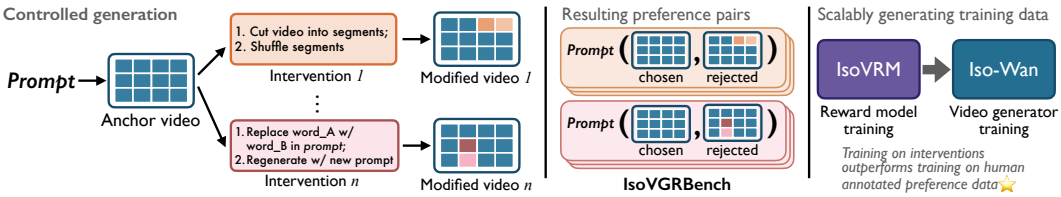

Figure 1: **Controlled intervention data generation procedure.** A prompt is rendered once to create an anchor video. A controlled intervention (e.g., shuffling segments of a video) automatically produces a modified counterpart with a predetermined preference label. The resulting pairs serve reward model evaluation (IsoVGRBench), reward model training, and video generator post-training.

Intervention isolation exposes reward model failures that aggregate evaluation obscures. When we evaluate nine modern video reward models and two general-purpose vision-language judges on IsoVGRBench, we find that models are substantially more accurate on broad differences between generators than on temporal changes that preserve visual content. On temporally shuffled videos, the best reward model reaches only 60.8% against 100% for humans, and no model exceeds 66.1% at any frame count up to 48. Rankings on five existing reward and temporal-understanding benchmarks do not predict sensitivity to frame order.

The same construction yields a training corpus of $160{,}000$ pairs, which we show can serve as effective training supervision. IsoVRM-Q2, trained from the same base model as VideoReward but only on intervention pairs, outperforms VideoReward by 3.8 points on VBench-2.0. When used as the reward for post-training a video generator, IsoVRM-Q2 yields the generator with the highest human win frequency, 55.4%, in a three-way comparison against generators post-trained under IsoVRM-VR and VideoReward. IsoVRM-VR, initialized from human-supervised VideoReward and further trained on intervention pairs, outperforms VideoReward on three of four transfer benchmarks, including 63.5% versus 56.3% on VBench-2.0 and 69.0% versus 66.4% on MJ-Bench-Video. The generator post-trained with its reward, IsoWan-VR, reaches the highest VBench-2.0 score, 51.5 against 49.0 under VideoReward, with a human win frequency of 51.2%. Thus, intervention-derived supervision can either replace human preference supervision or augment it, improving both reward-model transfer and the generators trained against those rewards. We make three contributions.

- **Controlled interventions for scalable reward-model evaluation**, where changing one property at a time determines both the preference label and the property under test without human labeling.
- **IsoVGRBench**, an intervention-isolated benchmark on which current video reward models and general-purpose judges perceive frame content but not frame order, a failure that rankings on existing benchmarks do not predict.
- **Intervention-derived supervision** that trains competitive reward models without human preference labels or in addition to them, and improves downstream video-generator post-training.

2 Constructing IsoVGRBench

2.1 Intervention conditions

Let x be a video generated from prompt p under sampling configuration θ. An intervention $\mathcal{I}_D$ produces a counterpart x' that differs from x along a set of dimensions D, either by transforming x itself or by rendering a new video after changing p or θ. Each intervention targets a set of dimensions D for which the construction specifies the preferred side; throughout this work, the anchor is preferred, such that $x \succ_d x'$ for each $d \in D$. For an intervention to provide a controlled evaluation of a particular property, we consider three conditions.

1. **Perceptibility.** The induced change should be detectable to human viewers, who should prefer the intended side on the targeted dimension.
2. **Confinement.** The intervention should change *only* the targeted dimension, while holding other dimensions fixed. This enables differences in reward to be attributed to the target property.
3. **Plausibility.** The resulting counterpart must resemble a plausible generator output rather than an easily exploitable artifact unrelated to the targeted property.

Table 1: **Interventions in IsoVGRBench, with their descriptions and evaluation dimensions.** IsoVGRBench draws 1,000 pairs from each intervention, resulting in 8,000 total samples.

Intervention	Description	Dimension
Rearranging the anchor's frames		
Temporal Shuffle	Anchor cut into six segments and randomly reordered	Motion
Frame Hold	Frames of the anchor held, giving stuttering motion	Motion
Regenerating one part of the clip		
Motion Re-gen	Regenerated motion between real keyframes	Motion
Attribute Edit	Single-attribute change in input text prompt	Text
Reduced Steps	8 denoising steps (vs. 50 for the anchor) with fixed prompt	Visual
Substituting the generator		
Aligned Weak Gen	Wan2.2-TI2V-5B continuing the anchor's first frame	Overall
Weak Gen (5B)	Rendered by Wan2.2-TI2V-5B from fixed prompt	Overall
Weakest Gen (1.3B)	Rendered by Wan2.1-T2V-1.3B from fixed prompt	Overall

Section 2.3 validates perceptibility through human judgments, while Appendix B provides additional validation of confinement for Motion Re-gen.

2.2 Interventions

Starting from an *anchor*, a clip rendered from a prompt using 50 denoising steps, each intervention produces a counterpart that is worse in a known respect, and the resulting pair is labeled only on the dimensions that change. For the interventions that do not constrain the anchor, we render anchors using Wan2.2-T2V-A14B, Wan2.2-TI2V-5B, and Wan2.1-T2V-1.3B (Wan Team, 2025), in proportions of roughly 60%, 30%, and 10%. Our interventions differ in how much of the original anchor they retain. Table 1 and Figure 2, which shows one example of each intervention, order the interventions by decreasing control over the anchor.

Rearranging the anchor's frames. The most controlled interventions modify only the temporal presentation of frames from the anchor. Temporal Shuffle divides the anchor into six segments and permutes them so that no segment remains in place, changing the order of events while preserving every frame. Because temporal order is informative only when there is sufficient motion, we use the 1,000 anchors with the most motion (more details in Appendix C). Frame Hold retains 8 to 12 anchor frames and repeats each until the next, preserving the clip length while turning smooth motion into visible jumps. Because both interventions construct the counterpart entirely from anchor frames, they strongly control for frame-level appearance and are labeled on motion.

Regenerating part of the clip. The next three interventions introduce newly generated content by changing a single input to generation. Motion Re-gen retains 8 or 12 anchor frames as keyframes and regenerates each interval using the anchor's own generator, starting from one keyframe and returning to the next. This allows motion to deviate from the anchor's trajectory between fixed keyframes while preserving the surrounding scene. Random keyframe spacing varies the duration over which this deviation can occur. We exclude Wan2.1-T2V-1.3B anchors from this intervention. Attribute Edit edits or removes one attribute of the prompt and re-renders using the same generator and seed; the resulting video is scored under the original prompt, making the edited attribute the targeted source of misalignment. Reduced Steps re-renders the same prompt and seed using 8 denoising steps instead of 50. The three interventions target motion, text alignment, and visual quality, respectively.

Substituting the generator. The last three interventions replace the anchor's generator with a weaker generator while preserving the prompt. Aligned Weak Gen renders with Wan2.2-TI2V-5B conditioned on the anchor's first frame, while Weak Gen (5B) and Weakest Gen (1.3B) render the prompt with Wan2.2-TI2V-5B and Wan2.1-T2V-1.3B, respectively. Because changing the generator can simultaneously affect visual quality, motion and text alignment, these pairs receive an overall label and deliberately give up confinement. They resemble the independently generated comparisons used in existing preference benchmarks, allowing Section 4.2 to compare published rankings

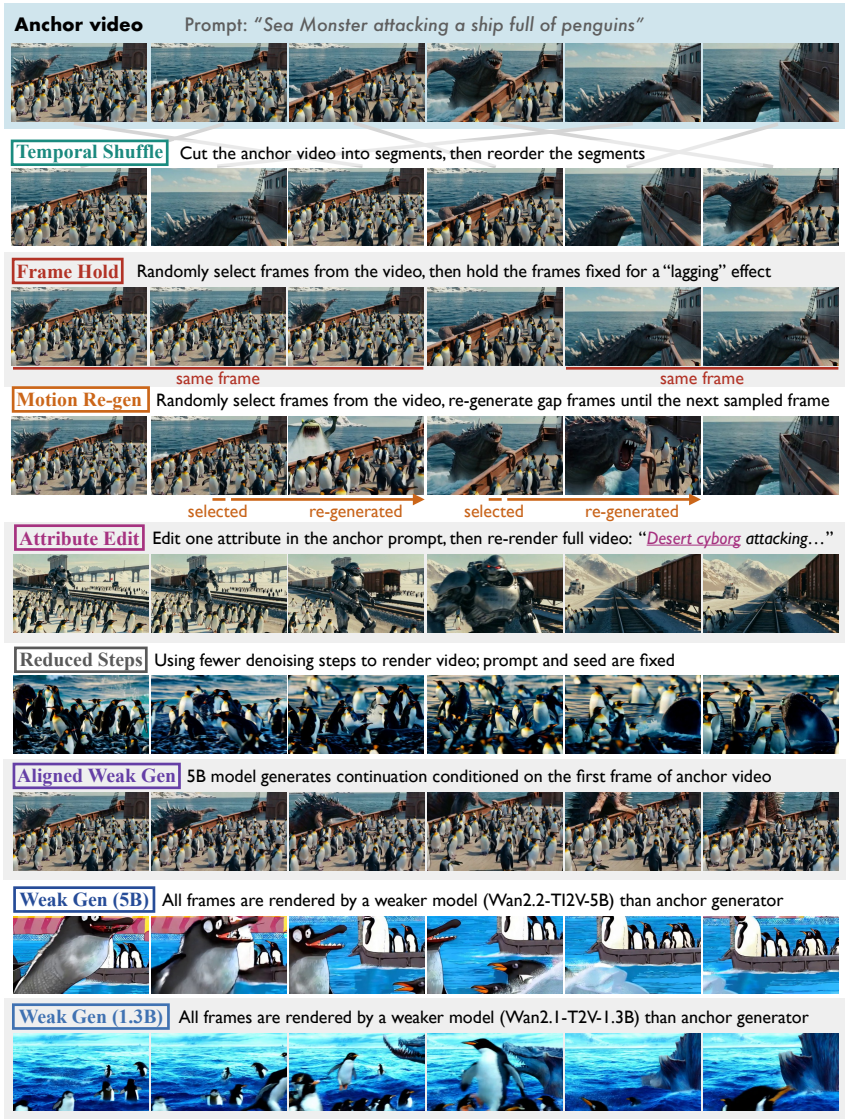

Figure 2: **ISOVGRBENCH interventions on a representative prompt.** The top row is the anchor prompt and every row below is a corresponding clip generated from each intervention. See Table 1 for intervention descriptions.

with performance on these interventions. All three use Wan2.2-T2V-A14B anchors because human judgments did not yield a consistent performance ordering between the two smaller models.

2.3 HUMAN VALIDATION OF PERCEPTIBILITY

We validate each intervention family with human judgments of perceptibility. For each family, raters compared 45 intervention pairs and 8 control pairs that show the same clip on both sides, with presentation order randomized and ties allowed. Each pair was judged separately on visual quality, motion, and text alignment, and the **Human** row of Table 2 reports agreement with the intervention label on its targeted dimension. For the three generator-substitution families, which carry an overall label, we define overall preference as the sign of the sum of the three dimension-level preferences, paralleling VideoReward's overall score as the sum of its three reward heads (Liu et al., 2025b); a sum of zero is a tie. Raters correctly identify all 64 control pairs as ties. Human agreement with the intervention label ranges from 98% to 100% on six of the eight families, with 93% agreement on **Frame Hold** and 84% on **Reduced Steps**. Thus, human judgments strongly agree with the constructed preference labels across all eight intervention families.

Table 2: **Per-intervention accuracy (%) on IsoVGRBench, sorted in descending order by average performance.** Ties are counted as incorrect; LiFT, VideoScore2, and VisionReward tie on 61%, 27%, and 21% of pairs, and PickScore and VideoReward rarely tie (Appendix D). Both UnifiedReward models also return no judgment on 78 pairs, which count as wrong. Accuracy over decided items is in Appendix A.

		Motion			Text	Visual	Overall		
Model	**Avg.**	Temporal Shuffle	Frame Hold	Motion Re-gen	Attribute Edit	Reduced Steps	Aligned Weak Gen	Weak Gen (5B)	Weakest Gen (1.3B)
VideoReward	73.3	**60.8**	**71.6**	**84.7**	74.1	78.5	73.6	77.9	64.9
PickScore	69.7	45.0	54.2	68.3	78.0	68.0	75.0	**89.6**	79.4
UnifiedReward-Flex-9B	68.4	47.7	49.5	54.3	**78.9**	77.4	69.2	88.2	**81.6**
UnifiedReward-2.0-8B	68.1	52.6	54.3	56.8	71.7	76.8	66.7	86.7	79.4
VideoScore v1	66.2	49.5	39.8	75.2	49.6	**80.5**	74.4	81.5	78.8
VideoScore v1.1	64.5	54.2	34.9	64.7	60.1	70.2	**76.3**	82.4	72.9
VisionReward	58.0	27.7	31.7	45.0	66.2	80.0	54.8	86.3	72.6
VideoScore2	47.9	38.2	35.9	42.2	62.1	47.6	47.0	61.0	49.4
LiFT	19.5	11.9	8.9	18.0	26.0	22.8	21.9	21.5	24.8
general vision-language models prompted as judges									
GPT-4.1	70.1	58.9	64.7	57.5	71.4	77.1	69.6	85.5	76.0
Qwen3-VL-8B	69.3	57.1	60.8	64.8	74.2	73.5	68.9	82.6	72.8
Human	96.4	100	93	100	98	84	98	98	100

2.4 Corpus and Splits

IsoVGRBench contains 1,000 pairs per intervention, with the preferred clip appearing first in half of the pairs. In total, the benchmark contains 8,000 pairs spanning 15,047 clips and 7,129 prompts, with chance accuracy at 50%. For the intervention-strength analysis in Appendix B.2, we additionally construct a **Reduced Steps** split using 16 rather than 8 denoising steps. The training corpus contains 164,000 pairs spanning 234,075 clips and 70,075 prompts, with 160,000 pairs for training and 4,000 for validation. Prompts are disjoint across splits because multiple interventions can reuse the same anchor clip, making a pair-level split susceptible to video leakage. Benchmark prompts are also disjoint from the training corpus.

3 Benchmarking Reward Models on IsoVGRBench

We evaluate nine reward models: VideoReward (Liu et al., 2025b), VisionReward (Xu et al., 2024), PickScore (Kirstain et al., 2023), UnifiedReward-Flex-9B and UnifiedReward-2.0-8B (Wang et al., 2025), VideoScore v1 and v1.1 (He et al., 2024), VideoScore2 (He et al., 2025), and LiFT (Wang et al., 2024). For models with property-specific reward heads, we use the head corresponding to the labeled property; for single-score models, we use their overall score. To see whether weaknesses are specific to reward-model training, we also evaluate two general-purpose vision-language judges that are not trained as video reward models: Qwen3-VL-8B-Instruct (Bai et al., 2025) and GPT-4.1 (OpenAI, 2025). Each judge scores a clip independently from its prompt and eight uniformly sampled frames, using the expected value over the ten digit tokens, which keeps the tie rate below 1% (Appendix E).

Reward models perceive frame content but miss frame order. The best reward model on each separate-render intervention scores between 76.3 and 89.6, while VideoReward, the best on both frame-reuse interventions, scores only 60.8 on **Temporal Shuffle** and 71.6 on **Frame Hold**. Every model performs worse on both frame-reuse interventions than on its strongest separate-render intervention, and six of nine reward models are at or below chance on **Temporal Shuffle**. Among decided **Temporal Shuffle** pairs, four models even assign higher reward more often to the degraded clip. People separate both frame-reuse interventions, at 100 on **Temporal Shuffle** and 93 on **Frame Hold**, and agree with the constructed preference on 347 of 360 pairs pooled across all interventions, above

the best reward model on seven of eight, with Reduced Steps (84 against 80.5) the exception (Appendix O). Section 4.1 tests whether the number of frames a model sees explains the gap.

Aggregate scores hide capability trade-offs between releases. VideoScore v1.1 improves over v1 by 4.7 points on Temporal Shuffle and 10.5 on Attribute Edit, while dropping by 10.5 on Motion Re-gen and 10.3 on Reduced Steps. Similarly, UnifiedReward-2.0-8B underperforms UnifiedReward-Flex-9B on five of eight interventions, by as much as 7.2 points, but it improves on both frame-reuse interventions by 4.9 and 4.8 points. These opposing changes largely cancel when averaged, making successive releases appear similar in aggregate despite substantial differences in their capability profiles.

General-purpose vision-language judges exhibit the same weaknesses. GPT-4.1 and Qwen3-VL-8B achieve average accuracies of 70.1 and 69.3, respectively, despite neither being trained specifically as a video reward model. Their performance nevertheless exhibits the same contrast, with 85.5 and 82.6 on Weak Gen (5B), but only 58.9 and 57.1 on Temporal Shuffle. The two judges differ by 1.8 points on Temporal Shuffle, within sampling error for 1,000 pairs, so GPT-4.1 reads frame order no better than the 8B open Qwen3-VL given the same eight frames. Using their continuous scores improves their aggregate ranking under our tie-handling rule but does not alter the frame-order result (Appendix E).

4 Analysis

Temporal Shuffle and Frame Hold are constructed to be the easiest temporal distinctions to perceive, since every frame is the anchor's own and people separate them at 100% and 93%. A reward model that cannot make them has no access to frame order, which judging whether events occur in the right sequence requires. Current generators, commercial ones included, render two prompted actions in the specified order less than a third of the time on VBench-2.0 (Zheng et al., 2025). Two questions follow, whether the failure comes from how few frames the models see, and whether existing benchmarks already reveal it.

4.1 Frame budget does not close the gap on Temporal Shuffle

The evaluated models in Table 2 sample between six and forty-eight frames, and frame count does not order them on Frame Hold, where the two models that sample 48 frames rank fifth and seventh while PickScore, which samples only six, ranks third (Appendix F). We therefore re-evaluate each checkpoint on the same 2,000 pairs with 4, 8, 16, 32, and 48 frames. Each line in Figure 3 varies only the frame count. Appendix L explains the frame counts some models cannot reach.

More frames help on Frame Hold, while Temporal Shuffle remains capped at 66.1. VideoReward, GPT-4.1, and Qwen3-VL-8B gain, respectively, 24.6, 18.9, and 16.2 points on Frame Hold between four frames and their best budget, with VideoReward reaching 82.3. On Temporal Shuffle, however, VideoReward improves by only 16.9 points to 66.1, while GPT-4.1 and Qwen3-VL-8B decline by 6.6 and 6.1 points, respectively, from four to forty-eight frames. No model exceeds 66.1 at any frame budget, compared with 100 for humans. VisionReward and LiFT remain below chance on both interventions at every budget.

For VideoReward, the gap between the two interventions grows from 8.5 points at four frames to 16.2 at thirty-two frames, where performance on both peaks. This is consistent with denser sampling making repeated frames in Frame Hold increasingly apparent, whereas detecting Temporal Shuffle requires recovering relationships between temporally distant frames.

4.2 No published benchmark predicts frame order sensitivity

We compare model rankings on Temporal Shuffle against two classes of benchmark, video generation reward benchmarks that compare independently generated videos (VideoGen-RewardBench (Liu et al., 2025b), GenAI-Bench (Li et al., 2024), and VBench-2.0 (Zheng et al., 2025)) and temporal video-language understanding benchmarks (TemporalBench (Cai et al., 2025) and TempCompass (Liu et al., 2024b)), which present a single video and evaluate its relationship to text.

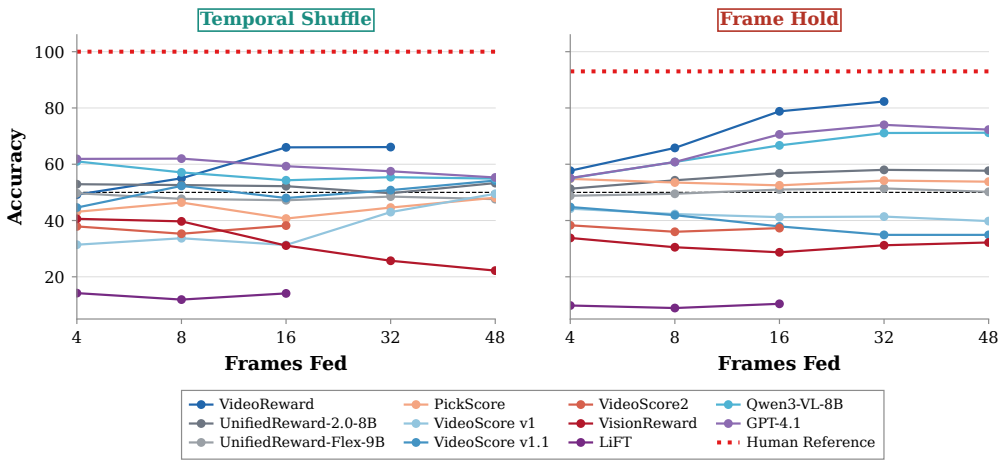

Figure 3: **Accuracy (%) against the frame count on two interventions that reuse the anchor's frames.** Absent points mark frame counts a model cannot be scored at (Appendix L). Seven of the eleven sit at or below chance at every budget on at least one of the two interventions, and VisionReward falls on Temporal Shuffle as it is given more frames.

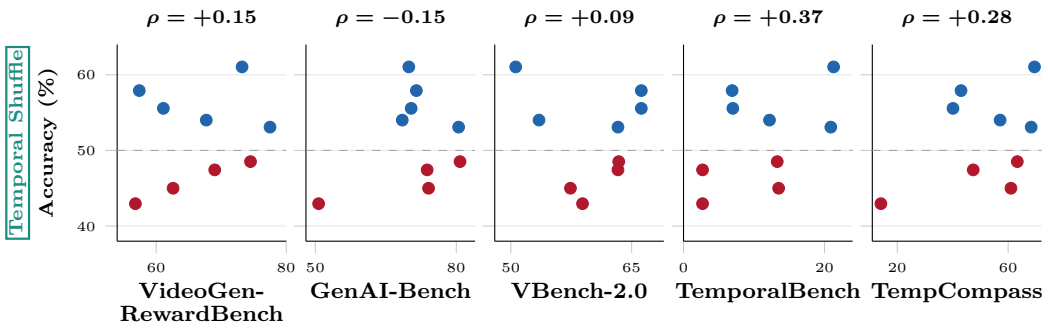

Figure 4: **Accuracy on each published benchmark against IsoVGRBench Temporal Shuffle accuracy.** Each reward model is a dot, and each horizontal axis uses the benchmark's reported metric. Blue is above chance on Temporal Shuffle, red is below, and the dashed line is at chance. ρ is the Spearman rank correlation down each panel. Table 14 gives underlying numbers and Table 15 repeats the correlation for every intervention.

For the ranking comparison in Figure 4, we use each published benchmark's reported metric and decided-item accuracy on Temporal Shuffle from Table 5, and Table 14 reports the underlying scores.

No published ranking tracks frame order. None of the five published benchmark rankings predicts performance on Temporal Shuffle. Rank correlations range from -0.15 to $+0.37$ (Figure 4), and none differs significantly from zero under an exact permutation test; the smallest p-value is 0.33 for TemporalBench, followed by 0.46 for TempCompass.

Under the decided-item protocol, UnifiedReward-2.0-8B and UnifiedReward-Flex lead both VideoGen-RewardBench and GenAI-Bench and score only 53.1 and 48.5 on Temporal Shuffle. On the same protocol, VideoReward leads the reward models on Temporal Shuffle by more than three points and Frame Hold by fifteen, despite ranking third, seventh, and last on the three video reward benchmarks.

Rankings transfer between comparisons built alike. Across all forty benchmark–intervention comparisons, only five correlations are significant (Table 15), and three of them pair a benchmark with the intervention whose comparison resembles its own. GenAI-Bench orders the nine models exactly as Weak Gen (5B) does ($\rho = +1.00$) and nearly as Weakest Gen (1.3B) does ($\rho = +0.87$); both interventions substitute a weaker generator, resembling the independently gener-

Table 3: Transfer accuracy (%) on benchmarks built from human labels. IsoVRM-VR, trained with synthetic data on top of human labels, generally performs the best. The VBench-2.0 column covers all eighteen dimensions the benchmark defines, and the MJ-Bench-Video estimate uses 116 decided pairs.

Model	VGRB	GenAI-Bench	VBench-2.0	MJ-Bench-Video
Benchmark Size	20,997	1,270	11,029	116
IsoVRM-Q2	71.5	67.9	60.1	61.2
IsoVRM-VR	**73.3**	68.2	**63.5**	**69.0**
VideoReward	73.2	**69.9**	56.3	66.4

Table 4: VBench-2.0 score and three-way human win frequency (%) of Wan2.2-TI2V-5B after 300 GRPO steps under each reward model.

GRPO Reward	VBench-2.0	Human
Pretrained	47.3	—
VideoReward	49.0	43.3
IsoVRM-Q2	49.8	**55.4**
IsoVRM-VR	**51.5**	51.2

ated comparisons used by video reward benchmarks. TempCompass correlates with **Attribute Edit** at $\rho = +0.80$, and both evaluate whether video content matches text. The remaining two are VideoGen-RewardBench and TempCompass on **Frame Hold**, at $+0.73$ and $+0.72$, and both are driven by LiFT, without which each falls to $+0.64$ and is not significant, while VBench-2.0 runs the other way on **Frame Hold**, at -0.61, strengthening to -0.80 without LiFT. TemporalBench correlates significantly with none of the eight interventions, with its largest correlation reaching $\rho = +0.67$ on **Attribute Edit** ($p = 0.055$).

A benchmark ranking therefore measures the distinctions its own comparisons contain, and none of the five contains a comparison that differs only in frame order. Appendix K finds the same distinction when comparing against IsoVGRBench's own overall generator-substitution pairs.

5 Leveraging intervention preferences as training data

5.1 Training reward models

We train two reward models on preference labels constructed by our intervention protocol. Both employ a Qwen2-VL-2B backbone and follow the VideoReward architecture (Liu et al., 2025b), with a scalar reward head at per-dimension marker tokens and a Bradley-Terry objective with ties (Bradley & Terry, 1952), and Appendix G gives the training configuration.

We first ask whether intervention supervision can *replace* human preference supervision by training IsoVRM-Q2 from a Qwen2-VL-2B initialization without prior reward-model training. We next ask whether intervention supervision can *augment* human preference supervision, and train IsoVRM-VR by initializing from released VideoReward weights, which have already been trained on human preference data.

Our intervention training data contains 20,000 pairs per intervention family, or 160,000 in total, with another 500 prompt-disjoint pairs per intervention family for validation. All IsoVGRBench prompts are excluded, although the interventions remain in distribution (Appendix H). We benchmark our reward models against VideoReward, which shares the same backbone, model size, and reward architecture but receives solely human preference supervision during reward-model training.

As Table 3 shows, IsoVRM-Q2, trained without human preference labels, outperforms VideoReward on VBench-2.0 by 3.8 points, 60.1 against 56.3, and trails it on VideoGen-RewardBench (71.5 against 73.2), GenAI-Bench (67.9 against 69.9), and MJ-Bench-Video (61.2 against 66.4). VBench-2.0 includes two generators absent from VideoReward's training data, and VideoReward rates both below their human win rates and the two it trained on above them (Appendix J). This is the generator-identity shortcut of Section 6, since in preference benchmarks the two generators' identities alone predict 67.3% to 77.4% of the labels (Appendix I). Intervention pairs carry little of this signal, because five of the eight families render both sides with the same generator, and those five supply most of the transfer. Added to VideoReward's supervision, the intervention pairs raise VBench-2.0 from 56.3 to 63.5 in IsoVRM-VR, which also scores 73.3 against 73.2 on VideoGen-RewardBench and 69.0 against 66.4 on MJ-Bench-Video (Tong et al., 2025), and trails on GenAI-Bench, scoring 68.2 against 69.9.

5.2 Training a video generator with intervention-supervised reward models

We fine-tune Wan2.2-TI2V-5B with diffusion GRPO (Liu et al., 2025a) using each reward model in Table 3. Appendix R shows more training details.

People prefer the generator post-trained under IsoVRM-Q2, whose reward model saw no human preference labels. Across every comparison in the three-way human study, it wins 266 of its 480 appearances, the IsoVRM-VR generator wins 246, and the VideoReward-supervised generator wins 208, win frequencies of 55.4%, 51.2%, and 43.3% (Appendix P reports the pairwise rates and agreement). On VBench-2.0 all runs improve on the pretrained generator (Table 4), and the generators post-trained under IsoVRM-VR and IsoVRM-Q2 reach 51.5 and 49.8, compared with 49.0 under VideoReward, the same order as their reward models' transfer accuracy on VBench-2.0 (Appendix Q gives every dimension). Both intervention-trained reward models therefore yield generators that improve on the VideoReward-supervised one under human judgments and under VBench-2.0.

6 Related work

Video reward models and their evaluation. Image and video reward models learn scores from human choices between independently generated outputs, through scalar or multidimensional heads, decomposed questions, reasoning-based scoring, or pairwise judgments (Kirstain et al., 2023; Xu et al., 2023; Wu et al., 2023; Liu et al., 2025b; He et al., 2024; Xu et al., 2024; He et al., 2025; Wang et al., 2024; 2025), and general vision-language models are also prompted as judges (Zheng et al., 2023; Chen et al., 2024). Reward models are evaluated by agreement with human preferences between generated videos in VideoGen-RewardBench, GenAI-Bench, and MJ-Bench-Video (Liu et al., 2025b; Li et al., 2024; Tong et al., 2025), and generators by per-dimension protocols in VBench, VBench-2.0, and EvalCrafter (Huang et al., 2024; Zheng et al., 2025; Liu et al., 2024a). Even when an annotation names the dimension to judge, the two videos differ in other dimensions and in generator identity, so a model can succeed on correlated differences instead of the requested property, a shortcut documented in other domains (Gururangan et al., 2018; Geirhos et al., 2020). Our pairs differ in one property by construction.

Constructed supervision. Synthetic data transforms scarce source data into forms a model learns from more readily (Yang et al., 2024). Image-quality assessment applies distortions of known severity but still collects human quality labels (Ponomarenko et al., 2015), attribute-specific constructions support question generation and latent preference models (Li et al., 2025a;b), and comparisons between stronger and weaker generators give preferences without human labeling but only for overall quality (Geng et al., 2025). Our generator substitutions reuse that signal, and the other interventions extend construction to one named property per video pair, with no human labeling.

Temporal order. Frame order has long served as self-supervision, for recognizing forward playback, detecting shuffled frames, and sorting clips (Pickup et al., 2014; Wei et al., 2018; Misra et al., 2016; Lee et al., 2017; Xu et al., 2019). Video-language benchmarks find models weak on temporal understanding beyond static content (Cai et al., 2025; Liu et al., 2024b), and TempCompass, the closest construction to ours, builds conflicting videos that share static content and differ in one temporal property, by reversal or by spatial or temporal concatenation (Table 12). These benchmarks test a single video against text, and **Temporal Shuffle** poses the same contrast as a preference pair.

7 Conclusion

Controlled interventions make a reward model's perception testable one property at a time, and they show that current video reward models see frame content but not frame order, a failure no existing benchmark reveals. Because the construction fixes the preferred side, the resulting pairs need no human labels, and human validation is required only once per intervention. Used as supervision, these pairs train reward models that transfer to human-labeled benchmarks and yield generators people prefer, with or without human preference data. With our proposed model-agnostic intervention procedure, any property that can be written as an intervention can therefore be measured in a reward model and taught to it, without collecting new human preferences.

AI USE STATEMENT

We used generative AI tools for manuscript drafting and language editing, code drafting and debugging, and figure prototyping. The authors reviewed and revised all AI-assisted text, tested AI-assisted code, and checked numerical claims against stored experiment outputs. Generative AI did not provide human annotation or determine evaluation labels.

ETHICS STATEMENT

ISOVGRBENCH contains model-generated videos and does not require collecting videos of research participants. Human participation is limited to perceptibility validation and preference judgments on generated clips. Generated content and learned reward models may inherit biases, unsafe depictions, or representational gaps from the source prompts and base generators. These artifacts are intended for research on video reward modeling and should not be used as standalone measures of safety, factuality, or human values.

REPRODUCIBILITY STATEMENT

Sections 2 and 3 specify the intervention construction and evaluation protocol. Sections 2 and 5 report data splits, Section 4 reports frame budgets, and the appendix gives training hyperparameters and per-dimension generator results.

AUTHOR CONTRIBUTIONS

All authors contributed to conceptualization, methodology, experimental design, result interpretation, and manuscript revision. The first author led implementation, data construction, model training, evaluation, and initial manuscript preparation. All authors reviewed and approved the final manuscript.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. URL https://doi.org/10.2307/2334029.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. In *ICLR*, 2025. URL https://arxiv.org/abs/2410.10818.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *ICML*, 2024. URL https://arxiv.org/abs/2402.04788.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. URL https://arxiv.org/abs/2004.07780.

Scott Geng, Hamish Ivison, Chun-Liang Li, Maarten Sap, Jerry Li, Ranjay Krishna, and Pang Wei Koh. The delta learning hypothesis: Preference tuning on weak data can yield strong gains. In *COLM*, 2025. URL https://arxiv.org/abs/2507.06187.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *NAACL*, 2018. URL https://arxiv.org/abs/1803.02324.

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. *arXiv preprint arXiv:2406.15252*, 2024. URL https://arxiv.org/abs/2406.15252.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, Keming Wu, Benjamin Schneider, Quy Duc Do, Zhuofeng Li, Yiming Jia, Yuxuan Zhang, Guo Cheng, Haozhe Wang, Wangchunshu Zhou, Qunshu Lin, Yuanxing Zhang, Ge Zhang, Wenhao Huang, and Wenhu Chen. Videoscore2: Think before you score in generative video evaluation, 2025. URL https://arxiv.org/abs/2509.22799.

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 *CVPR*, 2024. URL https://arxiv.org/abs/2311.17982.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2305.01569.

Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. URL https://arxiv.org/abs/1708.01246.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation. In *CVPR Workshops*, 2024. URL https://arxiv.org/abs/2406.13743.

Shuyue Stella Li, Jimin Mun, Faeze Brahman, Pedram Hosseini, Bryceton G. Thomas, Jessica M. Sin, Bing Ren, Jonathan S. Ilgen, Yulia Tsvetkov, and Maarten Sap. ALFA: Aligning LLMs to ask good questions, a case study in clinical reasoning. *arXiv preprint arXiv:2502.14860*, 2025a. URL https://arxiv.org/abs/2502.14860.

Shuyue Stella Li, Melanie Sclar, Hunter Lang, Ansong Ni, Jacqueline He, Puxin Xu, Andrew Cohen, Chan Young Park, Yulia Tsvetkov, and Asli Celikyilmaz. PrefPalette: Personalized preference modeling with latent attributes. *arXiv preprint arXiv:2507.13541*, 2025b. URL https://arxiv.org/abs/2507.13541.

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. *arXiv preprint arXiv:2505.05470*, 2025a. URL https://arxiv.org/abs/2505.05470.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025b. URL https://arxiv.org/abs/2501.13918.

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*, 2024a. URL https://arxiv.org/abs/2310.11440.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics*, 2024b. URL https://arxiv.org/abs/2403.00476.

Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, 2016. URL https://arxiv.org/abs/1603.08561.

OpenAI. Introducing gpt-4.1 in the api, April 2025. URL https://openai.com/index/gpt-4-1/.

Lyndsey C Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Schölkopf, and William T Freeman. Seeing the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014. URL https://doi.org/10.1109/cvpr.2014.262.

Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C-C Jay Kuo. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30: 57–77, 2015. URL https://doi.org/10.1016/j.image.2014.10.009.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

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: Fine-grained benchmarking and rewarding video preferences in video generation. *arXiv preprint arXiv:2502.01719*, 2025. URL https://arxiv.org/abs/2502.01719.

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*, 2024. URL https://arxiv.org/abs/2311.12908.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*, 2024. URL https://arxiv.org/abs/2412.04814.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025. URL https://arxiv.org/abs/2503.05236.

Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. URL https://doi.org/10.1109/cvpr.2018.00840.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. URL https://arxiv.org/abs/2306.09341.

Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://doi.org/10.1109/cvpr.2019.01058.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2304.05977.

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. *arXiv preprint arXiv:2412.21059*, 2024. URL https://arxiv.org/abs/2412.21059.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, and Tatsunori Hashimoto. Synthetic continued pretraining, 2024. URL https://arxiv.org/abs/2409.07431.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. URL https://arxiv.org/abs/2503.21755.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NeurIPS Datasets and Benchmarks*, 2023. URL https://arxiv.org/abs/2306.05685.

A ACCURACY OVER DECIDED ITEMS

Published reward-model benchmarks usually leave ties out of the denominator. Table 5 reports accuracy over decided pairs with each model's tie rate, and Table 2 counts ties as wrong. Only VideoReward, VideoScore2, and LiFT keep their rank under both rules. VisionReward moves from seventh to second, and PickScore moves from second to fifth.

Model	Avg.	tie %	Temporal Shuffle	Frame Hold	Motion Re-gen	Attribute Edit	Reduced Steps	Aligned Weak Gen	Weak Gen (5B)	Weakest Gen (1.3B)
VideoReward	73.6	0	61.0	71.9	84.9	74.6	78.7	74.0	78.2	65.2
VisionReward	71.3	21	47.4	56.9	67.0	70.0	83.4	76.8	89.2	80.1
VideoScore v1	70.3	7	55.6	48.8	81.2	51.9	82.7	78.5	83.1	80.3
UnifiedReward-Flex-9B	69.8	2	48.5	51.8	55.2	79.1	77.6	71.9	91.1	83.5
PickScore	69.7	0	45.0	54.2	68.3	78.0	68.0	75.0	89.6	79.4
UnifiedReward-2.0-8B	69.5	2	53.1	54.6	57.4	72.4	77.4	68.6	90.5	81.9
VideoScore v1.1	67.5	5	57.9	39.9	69.7	62.4	71.8	79.5	83.9	74.6
VideoScore2	64.8	27	54.0	54.3	59.4	72.5	67.0	67.1	76.0	68.2
LiFT	50.7	61	43.0	43.4	67.2	52.3	48.0	57.0	40.4	54.0

Table 5: Accuracy over decided items only, with the share of pairs on which each model expressed no preference. Compare Table 2, which counts those pairs as wrong.

B ADDITIONAL VALIDATION

B.1 CONFINEMENT HAS TO BE JUDGED FRAME BY FRAME

Judging visual quality on whole videos mixes appearance with motion. Human raters preferred the Motion Re-gen anchor on 17 of 20 decided video pairs but on none of 17 corresponding frame pairs. We therefore test confinement on single frames.

For each of 45 clips, raters compare the anchor and Motion Re-gen frames at the index where the two differ most, excluding the keyframes, which are identical. Raters judged 41 pairs equal, and the other four split three to one ($p = 0.63$, sign test). All eight identical controls were judged equal. The regenerated intervals use 30 denoising steps against the anchor's 50, and the frame comparison shows no loss of per-frame quality from it.

B.2 INTERVENTION STRENGTH SETS DIFFICULTY

We render Reduced Steps at 8 and 16 denoising steps with the prompt, seed, anchor, and every other sampling parameter fixed, and score all nine reward models on both over decided pairs (Figure 5).

Eight models gain 7.4 to 14.2 points at the stronger degradation. VideoReward passes UnifiedReward-Flex-9B, and VideoScore v1.1 passes PickScore. PickScore, which scores frames independently, gains least. LiFT falls from 53.8 to 48.0 and is the only model below chance at 8 steps. The spread across models widens from 53.8–69.2 at 16 steps to 48.0–83.4 at 8. We release the 16-step setting as an additional split outside the 8,000 benchmark pairs.

On **Reduced Steps**, the strength of the intervention therefore sets its difficulty for reward models.

C TEMPORAL SHUFFLE ANCHOR SELECTION

Reordering is visible only in a clip that moves, so **Temporal Shuffle** draws on the anchors with the most motion. We measure motion as the mean absolute difference between sampled frames in 8-bit pixel values, averaged over pixels and color channels. For the benchmark, motion is computed on every third frame, and the 1,000 anchors with the most motion range from 18.3 to 56.1, against about 12 for a typical clip. The training corpus instead keeps every anchor above 25, computed on every fourth frame, which admits 23% of them.

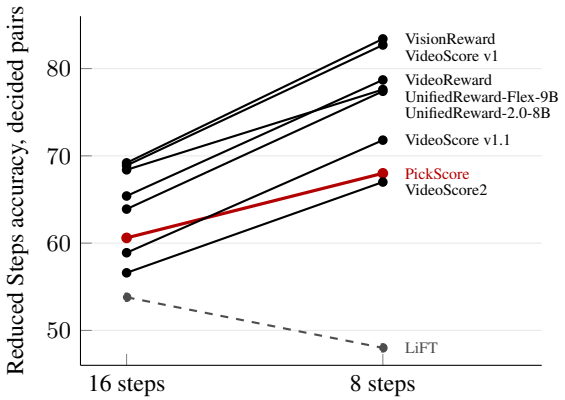

Figure 5: Accuracy on decided pairs against degradation strength on Reduced Steps, with prompt, seed, anchor and every other sampling parameter held fixed. Eight of the nine models gain as the budget falls from 16 to 8 denoising steps, by 7.4 to 14.2 points. PickScore, in red, gains least. LiFT, dashed, falls by 5.8. Tie rates are in Table 5.

D TIE HANDLING

Tie rates range from zero to 61% and follow how finely each model scores. PickScore and VideoReward each produce about 15,000 distinct scores and rarely tie. LiFT produces ten, assigns 0.9 to 40% of clips, and ties on 61% of pairs. VisionReward and VideoScore2 tie on 21% and 27%.

Paired clips share a prompt and most of their content, which raises the tie rate further for coarse scores. Independent draws from LiFT's score distribution would tie on 25.8% of pairs, against the observed 61.3%. We count a tie as wrong because it gives no ordering for a difference people perceive.

E PROMPTED JUDGE PROTOCOL

Each judge answers with a digit from 0 to 9, and we rescale the expected digit under its token probabilities to 0–100. For Qwen3-VL the probabilities come from its logits. For GPT-4.1 they come from the ten most likely tokens the API returns, which hold nearly all the probability of a single-digit answer.

Taking the decoded digit instead produces many ties. On samples of 200 pairs for Qwen3-VL and 60 for GPT-4.1, Qwen3-VL ties on 47%–54% of pairs and GPT-4.1 on 17%–20%. The expected digit brings both below 1%. Asking Qwen3-VL for two decimals between 0 and 1 still yields only nine distinct values and a 54% tie rate. We use the expected digit throughout, which changes the judges' average rank under the tie rule but not their frame-order results, so we compare judges on the frame-order columns.

F FRAMES SAMPLED PER MODEL

Table 6 lists how many frames each model samples from a 5.1-second clip, next to its accuracy on the two frame-reuse interventions.

Each model has its own architecture, so this comparison cannot separate frame count from model, and Section 4.1 varies the count within each model instead.

Model	Frames	Temporal Shuffle	Frame Hold
VideoReward	10	**60.8**	**71.6**
UnifiedReward-2.0-8B	8	52.6	54.3
PickScore	6	45.0	54.2
UnifiedReward-Flex-9B	8	47.7	49.5
VideoScore v1	48	49.5	39.8
VideoScore2	10	38.2	35.9
VideoScore v1.1	48	54.2	34.9
VisionReward	24	27.7	31.7
LiFT	8	11.9	8.9

Table 6: Frames sampled per clip against accuracy on the two interventions that reuse the anchor's frames. Clips are 81 frames at 16 fps. Underlined entries fall below chance. Frame count orders neither column. Accuracies count a tie as wrong, as in Table 2, and so run lower than in Table 14, which drops ties as VideoGen-RewardBench does.

G Reward model training

Both released models use the same training configuration and differ only in initialization. They train on 160,000 pairs and validate on 4,000, drawn equally from the eight interventions. The training, validation, and benchmark prompt sets are disjoint.

Setting	Value
Backbone	Qwen2-VL-2B-Instruct
Initialization	VideoReward, or the bare backbone
Reward head	scalar, read at three per-dimension marker tokens
Objective	Bradley-Terry with ties
LoRA	rank 64, α 128, dropout 0.05
LoRA exclusions	`lm_head`, `rm_head`, `embed_tokens`, `visual`
Vision tower	trained, LoRA not applied
Learning rate	2×10^{-6}
Schedule	constant with warmup, warmup ratio 0.05
Weight decay	0
Epochs	3
Batch size	1 per device, no gradient accumulation
Hardware	8 GPUs on one node
Precision	bfloat16
Seed	42
Frames	2 fps, at most 200,704 pixels per frame

Table 7: Training configuration for both released reward models.

H The trained reward models on IsoVGRBench

Table 8 reports the two trained reward models on IsoVGRBench, with ties counted as wrong. Training excludes every benchmark prompt but uses the same eight interventions, so this tests new prompts and not new interventions.

The two models average 23.0 and 23.4 points above VideoReward. Both reach at least 99.0% on six intervention families, with the remaining errors concentrated on Temporal Shuffle and Attribute Edit. VideoReward initialization raises Temporal Shuffle accuracy from 84.9% to 89.5%, while the Qwen2-VL initialization is slightly better on Attribute Edit at 86.8% versus 85.5%. Their averages differ by 0.4 points. Transfer to benchmarks built from human labels is in Table 3.

Model	Avg.	motion			text	visual	overall		
		Temporal Shuffle	Frame Hold	Motion Re-gen	Attribute Edit	Reduced Steps	Aligned Weak Gen	Weak Gen (5B)	Weakest Gen (1.3B)
VideoReward (2B), human preferences	73.3	60.8	71.6	84.7	74.1	78.5	73.6	77.9	64.9
IsoVRM-Q2	96.3	84.9	99.8	100.0	86.8	100.0	99.9	100.0	99.3
IsoVRM-VR	96.7	89.5	100.0	100.0	85.5	99.7	99.9	99.7	99.0

Table 8: The two released reward models on IsoVGRBench, ties counted as wrong, with VideoReward from Table 2 for reference.

I Generator identity in preference benchmarks

Preference benchmarks pair videos drawn from a fixed set of generators. We measure how many labels can be predicted from which two generators made the pair. The predictor fits pairwise generator win rates and is evaluated by five-fold cross-validation. It never sees a video.

Benchmark	Generators	Generator only
VideoGen-RewardBench	12	77.4
GenAI-Bench	10	74.1
VBench-2.0	4	67.3
MJ-Bench-Video	*not measurable, the release does not name the generators*	

Table 9: Accuracy of a predictor shown only which two generators produced a pair, and never a video, by five-fold cross-validation. Chance is 50.

It predicts 67.3% to 77.4% of the labels (Table 9), so a model can gain accuracy on these benchmarks by learning their generator ranking. MJ-Bench-Video does not record which generator made each clip, so the test cannot be run there. On VideoGen-RewardBench, generator win rates range from 16.8% for easyanimatev4 to 80.9% for kling1.5.

J Generator ranking in VideoReward and in our reward models

Table 10 compares each VBench-2.0 generator's win rate under human labels with the win rate each model's scores imply. It accounts for the 7.2-point gap between IsoVRM-VR and VideoReward on VBench-2.0 in Table 3.

Generator	In VideoReward's training data	VBench-2.0 people	VideoReward win	VideoReward Δ	IsoVRM-VR win	IsoVRM-VR Δ
Kling	59.3 / 80.9	64.6	70.9	+6.3	81.2	+16.6
Sora	absent	61.9	47.9	−14.0	70.2	+8.3
HunyuanVideo	absent	47.6	41.1	−6.5	42.9	−4.7
CogVideo	44.8	28.1	40.9	+12.8	9.7	−18.4

Table 10: Percentage of its comparisons each generator wins, by human count and as each model's scores imply, with each model's deviation from the human value. The first column is that generator's standing in VideoGen-RewardBench, the data VideoReward was trained on. Kling has two entries in VideoGen-RewardBench and VBench-2.0 uses a third version, so its training-set entry is the least certain of the four.

VideoReward rates both generators present in its training data above their human win rates and both absent generators below them. It assigns CogVideo 40.9, within 4 points of its 44.8 VideoGen-RewardBench win rate and 13 points above its 28.1 VBench-2.0 human win rate. With four generators, this pattern arises by chance one time in eight.

Five of our interventions use the same generator on both sides and so carry no generator ranking. The inherited ranking affects VideoReward on VBench-2.0, whose generators it partly never saw, and not on VideoGen-RewardBench, whose generators it saw in training.

On VideoGen-RewardBench, GenAI-Bench, and MJ-Bench-Video, dropping the generator substitutions costs 1.9, 1.6, and 0.5 points, 69.6, 66.3, and 60.7 against 71.5, 67.9, and 61.2 for IsoVRM-Q2, so the five same-generator interventions supply most of the transfer.

K Rank on overall quality does not predict rank on frame order

The three generator substitutions carry an overall label, like the independently generated pairs of published benchmarks. Table 11 compares each model's mean over them with its **Temporal Shuffle** accuracy on the same prompts, corpus, and scoring rule.

	Overall		Temporal Shuffle	
Model	mean	rank	mean	rank
PickScore	**81.3**	1	45.0	8
UnifiedReward-Flex-9B	79.7	2	47.7	7
VideoScore v1	78.2	3	49.5	6
UnifiedReward-2.0-8B	77.6	4	52.6	5
VideoScore v1.1	77.2	5	54.2	4
GPT-4.1	77.0	6	58.9	2
Qwen3-VL-8B	74.8	7	57.1	3
VideoReward	72.1	8	**60.8**	1
VisionReward	71.2	9	27.7	10
VideoScore2	52.5	10	38.2	9
LiFT	22.7	11	11.9	11

Table 11: Mean accuracy over the three generator-substitution interventions, which carry an overall label because a weaker generator degrades every dimension at once, against accuracy on **Temporal Shuffle**. Underlined entries are below chance. LiFT is the only model to hold its rank, at the bottom of both.

PickScore ranks first of eleven on overall pairs and eighth on **Temporal Shuffle** at 45.0. VideoReward ranks eighth on overall pairs and first on **Temporal Shuffle**. Without LiFT, the Pearson correlation between the overall mean and accuracy is $r = +0.40$ for **Temporal Shuffle** and $r = +0.31$ for **Frame Hold**, and neither differs from zero. With LiFT, which is below chance on all eight interventions, they rise to $+0.79$ and $+0.72$.

The two best models on overall pairs do not exceed chance on frame order.

L Frame counts the sweep cannot reach

Figure 3 omits three sets of points. LiFT stops at 16 frames because its VILA1.5-13B backbone has 4,096 positions, and more frames push the prompt out of context so that every clip gets the top score. VideoScore2 stops at 16 because with more frames it fails to finish on half the clips, repeating a closing tag until its token limit without producing a score. VideoReward stops at 32 because 48 frames exceed the memory of an 80 GB GPU at batch size one.

M Caption ranking on temporal-understanding benchmarks

A reward model gives one score per video and text, so we use the TemporalBench (Cai et al., 2025) and TempCompass (Liu et al., 2024b) settings in which it picks among complete captions for a fixed video. TemporalBench poses binary caption questions on short videos and reports Multiple Binary Accuracy, and TempCompass poses binary caption selection and reports caption-matching accuracy.

We score each caption with the model's text-alignment head, or its only output for a single-score model, and select the unique maximum, counting exact ties as wrong. For UnifiedReward we use the continuous alignment score of its pointwise video prompt.

Table 12 sets the three designs side by side.

Design property	IsoVGRBench	**TemporalBench**	**TempCompass**
Original target	Video-generation reward models	Video text-generation and embedding models	Video and image language models
What the model ranks	One prompt, rank two videos	One video, rank binary captions	One video, rank two captions
What changes	One named video property	A caption detail	The candidate caption
Human role	Validate each intervention family, with no per-item label	Write and refine positive captions, then review generated negatives	Annotate video metadata, then review generated instructions
Frame-order isolation	Same prompt and frames, reordered	None, since the evaluated video is fixed	None, since the evaluated video is fixed

Table 12: Differences among the benchmark designs. The TemporalBench and TempCompass columns describe the settings evaluated here.

TemporalBench counts a video correct only when every one of its binary questions is answered correctly, the group-level score its paper treats as primary, and we report it over 2,179 videos.

TempCompass contains 1,503 binary choices between complete captions. Its other three tasks do not fit a single score, since multiple-choice answers are fragments such as "up" or "forward" that need the question to make sense, yes/no questions need a calibrated threshold, and caption generation needs a language model.

Model	**TemporalBench**	**TempCompass**
VideoReward	**21.3**	69.5
UnifiedReward-Flex-9B	13.3	63.3
PickScore	13.5	61.0
UnifiedReward-2.0-8B	20.9	68.3
VideoScore v1	7.0	40.1
VideoScore v1.1	6.9	43.0
VisionReward	2.7	47.4
VideoScore2	12.2	57.1
LiFT	2.7	14.1
Ours, VideoReward init	18.4	**69.7**
Ours, Qwen2-VL init	18.1	66.9

Table 13: Accuracy on the caption-ranking settings of TemporalBench and TempCompass. All values are percentages. Bold marks the best result in each column.

VideoReward has the highest TemporalBench score, 21.3, and IsoVRM-VR, initialized from it, has the highest TempCompass score, 69.7, with 18.4 on TemporalBench.

N Published benchmarks against IsoVGRBench

Figure 4 plots Temporal Shuffle accuracy against each published benchmark, Table 14 gives the values, and Table 15 gives the rank correlation for every intervention.

O Every intervention against the human ceiling

Figure 6 orders the interventions by median reward-model accuracy, and the two frame-reuse interventions come last.

Model	Published Benchmarks					IsoVGRBench	
	VideoGen-RewardBench	GenAI-Bench	VBench-2.0	TemporalBench	TempCompass	Temporal Shuffle	Frame Hold
UnifiedReward-2.0-8B	**77.5**	80.5	63.3	20.9	68.3	53.1	54.6
UnifiedReward-Flex-9B	74.5	**80.8**	63.4	13.3	63.3	48.5	51.8
VideoReward	73.2	69.9	50.6	**21.3**	**69.5**	61.0	71.9
VisionReward	69.0	73.8	63.3	2.7	47.4	47.4	56.9
VideoScore2	67.7	68.5	53.5	12.2	57.1	54.0	54.3
PickScore	62.6	74.1	57.4	13.5	61.0	45.0	54.2
VideoScore v1	61.1	70.4	**66.2**	7.0	40.1	55.6	48.8
VideoScore v1.1	57.4	71.5	**66.2**	6.9	43.0	57.9	39.9
LiFT	56.8	50.7	58.9	2.7	14.1	43.0	43.4
ρ with Temporal Shuffle	+0.15	−0.15	+0.09	+0.37	+0.28		
ρ with Frame Hold	+0.73	+0.13	−0.61	+0.52	+0.72		

Table 14: Accuracy on five published benchmarks against accuracy on the two IsoVGRBench interventions that reuse the anchor's frames. All columns use each benchmark's reported metric, and the IsoVGRBench columns use the decided-item protocol of Table 5. Underlined entries fall below the corresponding random baseline. The last two rows give the Spearman rank correlation down each benchmark's column. Bold marks the best model in each published-benchmark column.

Benchmark	Temporal Shuffle	Frame Hold	Motion Re-gen	Attribute Edit	Reduced Steps	Aligned Weak Gen	Weak Gen (5B)	Weakest Gen (1.3B)
VideoGen-RewardBench	+0.15	**+0.73**	−0.50	+0.63	+0.43	−0.20	+0.62	+0.57
GenAI-Bench	−0.15	+0.13	−0.48	+0.50	+0.37	+0.27	**+1.00**	**+0.87**
VBench-2.0	+0.09	−0.61	−0.08	−0.46	+0.34	+0.56	+0.39	+0.60
TemporalBench	+0.37	+0.52	+0.03	+0.67	+0.07	−0.15	+0.34	+0.19
TempCompass	+0.28	**+0.72**	−0.20	**+0.80**	+0.22	−0.17	+0.48	+0.27

Table 15: Spearman rank correlation between accuracy on each published benchmark and accuracy on each IsoVGRBench intervention, over the 9 reward models common to all columns. Each benchmark uses its reported metric. Bold marks a correlation an exact permutation test separates from zero at $p < 0.05$.

P Human preference comparisons for the trained generators

Table 16 reports three annotation passes over the pairwise comparisons among the three generators after 300 training steps. Each pass contains all 240 comparisons, with the same 80 held-out prompts used for each model pair. Item order and side assignment are randomized independently.

Preferred generator	Win Rate	Decisions	Agreement	κ
IsoVRM-Q2 over VideoReward	56.2	240	63.3	0.26
IsoVRM-VR over VideoReward	57.1	240	59.2	0.17
IsoVRM-Q2 over IsoVRM-VR	54.6	240	55.8	0.11

Table 16: Blinded human preference comparisons on 80 prompts. Win rate pools the three annotation passes. Agreement is the mean exact agreement over the three pairs of passes on the same comparison, and the final column gives Fleiss' κ across the three passes.

Pooled over passes, both generators trained under our reward models win more comparisons than the VideoReward-supervised one. Raters agree most on IsoVRM-Q2 against VideoReward, 63.3% with $\kappa = 0.26$, and least on IsoVRM-Q2 against IsoVRM-VR, 55.8% with $\kappa = 0.11$, the comparison closest to even. Over all 240 comparisons, mean pairwise agreement is 59.4% and Fleiss' κ is 0.18. Those two generators also differ by only 1.7 points on VBench-2.0.

Each pass is nearly transitive. A pass sees the same three videos for each prompt, one per generator, and its three choices form a cycle on 5 of the 240 pass-prompt combinations, against 25% for random

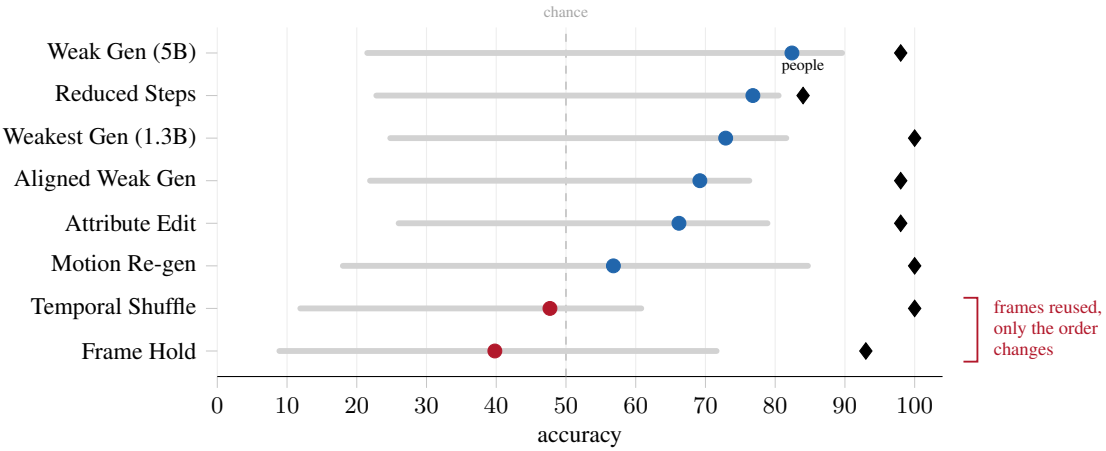

Figure 6: Interventions ordered by median reward-model accuracy. Each bar spans the nine reward models from worst to best, the dot marks their median, blue above chance and red below, and the diamond marks human agreement on the same 45 pairs per intervention, scored the same way, which stays between 84 and 100. On the two frame-reuse interventions the median model reaches 39.8 and 47.7, against 93 and 100 for people.

choices (Table 17). Raters therefore rank the three generators consistently within a pass, and the low agreement comes from different rankings across passes.

	Cyclic Prompts	Prompts	Cyclic (%)
Pass 1	1	80	1.2
Pass 2	4	80	5.0
Pass 3	0	80	0.0
All passes	5	240	2.1
Majority of passes	2	80	2.5

Table 17: Transitivity of the human comparisons. A prompt is cyclic for a pass when its three pairwise choices there form a cycle, such as IsoVRM-Q2 over IsoVRM-VR, IsoVRM-VR over VideoReward, and VideoReward over IsoVRM-Q2. Random choices would give a cycle on 25% of prompts. The last row applies the majority of the three passes to each comparison.

Q VBench-2.0 by dimension for the trained generators

Table 18 breaks the aggregate VBench-2.0 results into all eighteen dimensions. Within each dimension we score every generator on the same number of videos per prompt, keeping the smaller count where two differ, since overlapping evaluation shards and four extra dimensions otherwise give unequal counts. Dynamic_Spatial_Relationship, for example, has 74 and 128 videos over the same 69 prompts. Diversity uses the twenty generations per prompt it requires, which are equal by construction.

Fourteen of eighteen dimensions separate the two trained runs by less than 5 points. Removing Multi-View_Consistency alone, or that dimension and Diversity, reverses the sign of the IsoVRM-VR margin over the VideoReward-supervised run.

The aggregate weights every dimension equally, so Instance_Preservation, at 91.9 for all four generators, and Complex_Landscape, spanning 2.0 points, count as much as Multi-View_Consistency, spanning 39.6, and a dimension scored on 10 videos counts as much as one scored on 264.

Dimension	Videos	Pretrained	VideoReward	IsoVRM-Q2	IsoVRM-VR
Camera_Motion	108	15.7	7.4	**16.7**	10.2
Complex_Landscape	30	14.7	**16.7**	14.7	16.0
Complex_Plot	61	13.5	**14.5**	11.7	11.1
Composition	53	**49.1**	47.2	43.0	48.2
Diversity	10	47.6	64.3	60.9	**81.0**
Dynamic_Attribute	94	**12.8**	11.7	7.4	8.5
Dynamic_Spatial_Relationship	69	27.5	**47.8**	36.2	30.4
Human_Anatomy	264	84.2	86.7	**91.3**	87.7
Human_Clothes	191	91.7	97.8	97.8	**100.0**
Human_Identity	128	72.5	76.9	**77.9**	77.1
Human_Interaction	129	56.6	62.0	55.8	**63.6**
Instance_Preservation	62	**91.9**	**91.9**	**91.9**	**91.9**
Material	50	**58.3**	51.7	54.2	56.0
Mechanics	57	**64.4**	54.8	52.3	55.3
Motion_Order_Understanding	181	**19.9**	17.1	17.7	13.8
Motion_Rationality	58	**44.8**	43.1	37.9	41.4
Multi-View_Consistency	97	30.7	33.4	**70.3**	70.0
Thermotics	50	55.3	56.2	59.1	**65.2**
Mean		47.3	49.0	49.8	51.5

Table 18: Every VBench-2.0 dimension for the four generators, in percent. **Videos** is the count each is scored on, held equal across all four dimension by dimension. Bold marks the best in a row. Against the VideoReward-supervised run, the IsoVRM-VR run leads on 11 dimensions and trails on 6; the median difference is $+0.8$ and a bootstrap over dimensions puts the mean difference in $[-1.9, +7.8]$. Multi-View_Consistency separates by initialization rather than by run.

R GENERATOR TRAINING

Each run draws from 12,303 training prompts and samples eight videos per prompt, with advantages computed within each group. Videos contain 121 frames at 704×1280 and use 30 sampling steps. The reward weights visual and motion quality by 0.5 and text alignment by 1.0. A dual KL penalty keeps the policy near a reference updated every 10 steps. Each run uses 128 H200 GPUs and validates every 50 steps on 512 held-out prompts, with one video per prompt at 50 denoising steps.