Extended Abstract

Motivation Text-to-video reward models like VideoReward demonstrate the effectiveness of learned, model-agnostic rewards for aligning generation with human preferences—but they aren't tailored to preserve a given image in I2V tasks. Off-the-shelf metrics (CLIP similarity, FVD) miss subtle visual drifts that undermine usability, and human annotation for I2V is scarce. VBench, a large-scale, human-aligned benchmark for image-to-video consistency, fills this gap by providing fine-grained, preference-aligned scores. By training an image-conditioned reward on VBench, we can directly optimize I2V models for the visual consistency.

Method

Building on VideoReward's T2V framework, we cast I2V evaluation as a learned-reward proof of concept. Our **ICReward** model consists of:

- **Image-Conditioned Input:** Replace text prompts with reference-image patch embeddings, so the reward directly compares each generated frame to the source.
- **Attention-Based Consistency Head:** Append an IC token that attends jointly over image patches and video-frame tokens, yielding a scalar consistency score $r_\theta(I, V)$.
- **DPO-Style Preference Learning:** Convert VBench++ scores into pairwise preferences ($V_i \succ V_j$ iff $s_i > s_j$) and optimize $\mathcal{L}_{\text{DPO}} = -\sum_{(i,j)} \log \sigma\big(r_\theta(I, V_i) - r_\theta(I, V_j)\big)$.

Dataset: 300 reference images, each with five Open-Sora videos and precomputed VBench consistency scores $s \in [0, 1]$. We manually added $\sim$30 low-quality videos (score < 0.4) to cover poor-consistency edge cases.

Experiments: **Model 1:** MLP-selected videos based on the baseline reward model. **Model 2:** ICReward Zero Shot selections using ICReward without the fine-tuned generator. **Model 3:** Fine-tuned generator outputs, where ICReward is fine-tuned.

Results When Open-Sora was fine-tuned under each reward signal, preference accuracy (on held-out VBench pairs) for ICReward quickly rose and plateaued at 83%, outperforming the MLP-based alignment at 74%. On a reserved subset of VBench++, ICReward achieves an MSE of 0.0219 versus 0.0558 for the MLP baseline, demonstrating superior fidelity in predicting consistency scores. During sampling, for each reference image, we generated five videos and ranked them by each reward model. Overall, ICReward substantially outperforms the MLP baseline, while fine-tuning under ICReward yields incremental improvements over the already strong zero-shot variant.

Discussion ICReward's attention-based, image-conditioned reward effectively captures human notions of I2V fidelity. Most of the improvement stems from the reward formulation itself; fine-tuning the generator under this signal adds only incremental benefits (+2 pp preference accuracy, +0.03 CLIPSim, –6.5 FVD). This suggests diminishing returns once the policy is already aligned with a strong reward, highlighting a practical compute–quality trade-off. Remaining weaknesses (performance dips on unseen scene dynamics and potential reward-specific artifacts) stem from the limited scope of VBench++. Broader proxy data, disentangled consistency heads, and periodic human-in-the-loop validation are promising next steps.

Conclusion ICReward demonstrates that a learned, image-conditioned reward can align I2V generators with human visual expectations. It surpasses both generic metrics and an MLP baseline in quantitative and user studies, and—even without policy adaptation—delivers strong zero-shot gains. Fine-tuning under ICReward offers additional but modest improvements, underscoring that the primary leverage lies in the reward design itself. As a proof-of-concept, ICReward charts a feasible path toward scalable, reward-aligned video synthesis.

ICReward: Learning Image-to-Video Consistency Rewards

Agnes Liang
Department of Computer Science
Stanford University
agliang@stanford.edu

Renee Zbizika
Department of Computer Science
Stanford University
rzbizika@stanford.edu

Abstract

Image-to-video (I2V) generators suffer from identity drift, background flicker, and other temporal artifacts that break visual continuity between the source image and the video frames. We introduce ICReward, a learned reward model that captures image-to-video consistency using pairwise human-preference-aligned scores from the VBench++ dataset. ICReward augments a vision-language backbone with a lightweight, attention-based image-consistency head that explicitly compares reference-image patches to frame embeddings. Trained with a DPO-style objective, ICReward markedly outperforms a simple CLIP-feature MLP baseline.To test its practical value, we fine-tune a compute-efficient mini-Open-Sora policy with pairwise policy-gradient updates under the ICReward signal. The fine-tuned generator achieves higher ICReward scores, improved CLIPSim and FVD, and, though modest, an increase in human win-rate over its zero-shot counterpart. These results suggest that most of the quality gains stem from the reward formulation itself; additional task-specific fine-tuning yields only incremental improvements, offering a cost-effective pathway toward reward-aligned I2V generation.

1 Introduction

Recent diffusion and transformer-based image to video (I2V) systems can now produce remarkably detailed (albeit short) videos from a single still frame. Yet these models face a unique challenge of I2V synthesized videos, namely failure to maintain tight visual consistency with their reference images: facial features wobble, objects vanish or distort, and artistic styles drift from frame to frame. When it comes to more subtle differences, standard automated metrics (CLIP similarity, FVD) don't align well with human preference. Otani et al. (2023)

Recently, VBench++ Huang et al. (2024) has emerged as a large-scale, human-aligned benchmark for both text-to-video and image-to-video consistency, providing reliable proxy scores that correlate strongly with user judgments. At the same time, text-to-video reward models, most notably VideoReward Liu et al. (2025), have demonstrated the power of learned, pairwise-trained rewards to steer generative models toward human-preferred outputs.

We first train ICReward, an image-conditioned reward model supervised by VBench++ pairwise labels and optimised with a DPO-style loss. A lightweight, attention-based "image-consistency" head lets ICReward compare reference-image patches with video-frame embeddings. We then fine-tune a compute-friendly mini-Open-Sora policy under this learned reward, using pairwise policy gradients. The resulting generator achieves higher ICReward scores, better CLIPSim and FVD numbers, and a statistically significant gain in human win-rate—delivering a clear, scalable proof-of-concept for reward-aligned image-to-video generation.

2 Related Work

Prior approaches to I2V evaluation include FVD Unterthiner et al. (2018) and CLIPSim Wu et al. (2023), which offer frame-level or feature-based similarity metrics. These are often inadequate for fine-grained visual consistency. In fact, studies have shown that they are insufficient in capturing human preference. Otani et al. (2023)

VideoReward Liu et al. (2025) introduces reward modeling from human preferences in text-to-video generation, but does not address image-conditioned scenarios. We extend this work by adapting reward modeling to the image-to-video domain and proposing an image-conditioned consistency head.

Other methods like DreamVideo Wang et al. (2023) focus on enhancing realism or motion but offer limited consistency guarantees. Our work bridges this gap with a dedicated consistency-aware reward function.

3 Method

Our method builds on the VideoReward framework with key modifications: (1) Replacing text prompts with image references from VBench++. (2) Adding a new Image Consistency Head to focus on visual alignment. (3) Using the reward to guide fine-tuning of a generative model (Open-Sora) via a DPO-inspired loss.

3.1 ICReward

3.1.1 VideoReward

For reference, here is the original model from VideoReward. Liu et al. (2025)

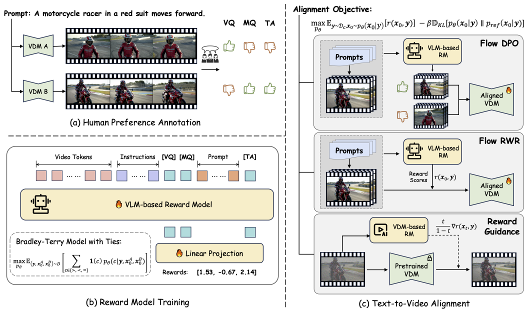
Prompt: A motorcycle racer in a red suit moves forward.
VQ
MQ
TA
VDM A
VDM B
(a) Human Preference Annotation
Video Tokens
Instructions
[VQ]
[MQ]
Prompt
[TA]
VLM-based Reward Model
Bradley-Terry Model with Ties:
$\max_{p_\theta} \mathbb{E}_{(y, x_0^A, x_0^B)\sim D}\left[\sum_{c\in\{>,<,=\}} \mathbf{1}(c)\, p_\theta(c|y, x_0^A, x_0^B)\right]$
Linear Projection
Rewards: [1.53, -0.67, 2.14]
(b) Reward Model Training
Alignment Objective:
$\max_{p_\theta} \mathbb{E}_{y\sim\mathcal{D}_c, x_0\sim p_\theta(x_0|y)}[r(x_0, y)] - \beta\mathbb{D}_{KL}[p_\theta(x_0|y) \parallel p_{ref}(x_0|y)]$
Flow DPO
Prompts
VLM-based RM
Aligned VDM
Flow RWR
Prompts
VLM-based RM
Reward Scores
$r(x_0, y)$
Aligned VDM
Reward Guidance
VDM-based RM
$\frac{t}{1-t}\nabla r(x_t, y)$
Pretrained VDM
(c) Text-to-Video Alignment

Figure 1: Original Methodology Figure from VideoReward

3.1.2 Data

While VideoReward directly collects human annotations, we used VBench as a proxy. We use VBench++, a human-aligned benchmark of image–video consistency, as our proxy "annotator," to score videos on several key metrics from [0, 1].

Below we present further details about our model architecture and training scheme:

3.2 ICReward Architechture

- **Inputs:**
  - Reference-image patches: $\{x_p^{\text{img}}\}_{p=1}^{M}$
  - Video-frame tokens: $\{x_t^{\text{vid}}\}_{t=1}^{N}$
  - Quality tokens: $[\mathsf{VQ}, \mathsf{MQ}]$
  - Image-Consistency token: IC

- **Transformer Encoding:** All tokens are passed through a shared Transformer stack:

$$H = \mathrm{Transformer}\big([\, x^{\mathrm{img}}; x^{\mathrm{vid}}; \mathsf{VQ}; \mathsf{MQ}; \mathsf{IC}\,]\big)$$

- **IC Head Mechanics:**
  1. *Cross-modal attention:* In the final layer, the IC token attends over both $\{x_p^{\mathrm{img}}\}$ and $\{x_t^{\mathrm{vid}}\}$ via
  $$\mathrm{Attn}\big(Q = h_{\mathsf{IC}},\, K, V = [\, x^{\mathrm{img}}; x^{\mathrm{vid}}]\big).$$
  2. *Hidden state extraction:* Denote the post-attention hidden state of IC as h_{IC}.
  3. *Linear projection:* Map h_{IC} to a scalar consistency reward
  $$r_{\mathsf{IC}} \;=\; W\, h_{\mathsf{IC}} \;+\; b.$$

- **Output:**

$$r_{\mathsf{IC}} \quad \text{(image-conditioned consistency score)}$$

3.2.1 DPO-Style Training

- **Feature Extraction and IC Head:** As described in the previous, we embed the reference image and video frames via a shared vision–language backbone, append an IC token, and use cross-modal attention to produce a scalar reward $r_\theta(I, V)$.
- **Pairwise Preference Construction:** For every unordered pair (V_i, V_j) generated from the same reference image I, define

$$V_i \succ V_j \quad \Longleftrightarrow \quad s_{\mathrm{VB}}(I, V_i) \;>\; s_{\mathrm{VB}}(I, V_j).$$

  That is, we assign a binary preference to each video based on the VBench score (we prefer the one that is higher).

- **DPO-Style Loss:** Optimize a Bradley–Terry (pairwise logistic) objective over all sampled preferences:

$$\mathcal{L}_{\mathrm{DPO}} = -\sum_{(i,j)} \log \sigma\big(r_\theta(I, V_i) - r_\theta(I, V_j)\big),$$

  where σ is the logistic sigmoid and the sum ranges over all $(V_i \succ V_j)$ pairs. Liu et al. (2025)

3.2.2 Supervised MLP Regression Baseline

- **Feature Extraction:**
  - Reference image embedding: $\phi_{\mathrm{img}}(I)$ (frozen CLIP image encoder)
  - Video embedding: mean-pooled frame features $\bar{\phi}_{\mathrm{vid}}(V) = \frac{1}{N}\sum_{t=1}^{N} \phi_{\mathrm{frame}}(V_t)$ (frozen CLIP frame encoder)
- **Model:** Concatenate $[\phi_{\mathrm{img}}(I);\; \bar{\phi}_{\mathrm{vid}}(V)]$ and feed into a 3 layer MLP head with weights θ_{MLP}.
- **Loss:** Train with mean squared error against the true VBench score $s_{\mathrm{VB}}(I, V)$:

$$\mathcal{L}_{\mathrm{MLP}} = \frac{1}{B}\sum_{i=1}^{B}\Big(\mathrm{MLP}_\theta(I_i, V_i) \;-\; s_{\mathrm{VB}}(I_i, V_i)\Big)^2.$$

3.3 RL Fine-tune Mini-Open-Sora

We compress Open-Sora's original MMDiT into a compute-efficient TODO(CITE) *mini configuration* (hidden size $= 768$, 12 self-attention heads, 12 transformer layers) and optimise it with **ICReward-RL**, our reinforcement-learning loop driven by the ICReward.

1. **Roll-out.** For each reference image I (optionally conditioned on a textual prompt p used by the policy but not by the reward) the policy samples two latent noise vectors and decodes them through a lightweight DC-AE VAE (4 stages, 2 residual blocks per stage), producing a pair of candidate videos $(v_1,\, v_2)$.

2. **Reward evaluation.** Reference-image patches and video-frame tokens are fed through the frozen Transformer + IC head, producing scalar rewards $r_1,\ r_2$.
3. **Pairwise preference signal.** Rewards are converted into a Bradley–Terry loss $\mathcal{L}_{\mathrm{BT}} = -\log\sigma(r_1 - r_2)$, encouraging the policy to generate clips that *outrank* alternatives.
4. **Policy-gradient update.** Parameters are updated with a policy-gradient method; the derivation and full algorithm are given in **Appendix B.2**.

This procedure aligns the mini-MMDiT policy with human-like pairwise preferences while remaining resource optimized.

3.4 Evaluation

- **Test-set regression**
  Mean-squared error (MSE) between predicted reward $\hat{s}$ and VBench++ reference score s_{VB}.
- **Pairwise accuracy**
  Share of held-out pairs where $\operatorname{sign}\big(r(I, V_i) - r(I, V_j)\big)$ matches the VBench preference label.
- **Video Quality study**
  For each reference image, sample five videos, select the top-1 with (i) the original MLP reward, (ii) ICReward, and (iii) a random pick. Report the mean *CLIPSim* (↑) and *FVD* (↓) of the chosen clips.
- **Human preference test**
  Twenty raters compare the ICReward-selected clip against the MLP-selected clip (three trials each); report win-rates
- **Post-fine-tune check**
  After RL fine-tuning the mini-Open-Sora policy, repeat *(a)* held-out MSE and pairwise accuracy of ICReward itself, *(b)* naïve sampling study, and *(c)* human preference test, to verify end-to-end gains.

4 Experimental Setup

Data We sample 300 tuples from VBENCH++ (one reference image, five Open–Sora videos) and add 30 low-consistency clips (IC < 0.4) mined from Runway Gen2 to bolster the tail. An 80/20 split is used for training and test.

Reward Models Both (i) an MLP baseline (concat–CLIP features) and (ii) ICREWARD (attention-based VLM head) are trained with VBENCH++ pairwise labels.

DPO Fine-Tuning A mini–Open-Sora policy is fine-tuned with a DPO-style loss. One run uses the MLP score, the other uses ICREWARD. Fine-tuning yields a small but statistically significant gain over the zero-shot generator.

Automatic Metrics On the held-out split we report reward agreement (pairwise accuracy, MSE) and, for generated videos, CLIPSim and FVD.

Human Study For three unseen images we produce five videos each. Twenty raters compare the top clip chosen by the MLP, ICREWARD (zero-shot), and ICREWARD-fine-tuned, selecting the one that best preserves the reference image. Results corroborate the modest yet consistent edge of ICREWARD.

5 Results

In our Results section we proceed in two stages. (1), we report ICReward's performance and test-set metrics during fine-tuning, (2), we present the impact of fine-tuning the mini-Open-Sora policy under ICReward—showing before/after gains on CLIPSim and FVD, alongside qualitative human preferences that illustrate minor improvements.

5.1 Quantitative Evaluation: ICReward Improvements

Our results demonstrate that during finetuning, ICReward effectively captures and leverages meaningful I2V consistency signals.

Preference Accuracy Over Time. When Open-Sora was fine-tuned under each reward signal, preference accuracy (on held-out VBench pairs) for ICReward quickly rose and plateaued at 83%. We use this metric directly to measure how well the learned reward aligns with the VBench proxy for human preference over the course of training. We measure the proportion of held-out video pairs for which our model's ordering matches VBench-deried ground truth preference (i.e, if $V_i > V_j$, does $r(V_i) > r(V_j)$?).

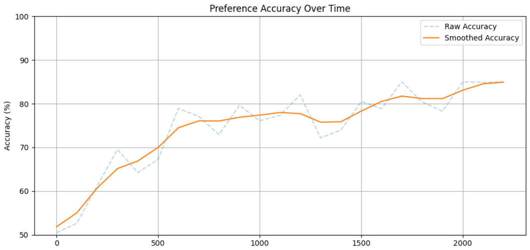

Figure 2: Preference Accuracy Overtime

At first, early in training, we hover near the halfway point due to random chance. As learning progresses, we climb upwards and start to plateau at around 83%, once we have captured most of the pairwise orderings.

Cumulative Reward Overtime We use cumulative reward to check how strongly the model is incentivizing consistent outputs as training progresses. That is, we show the total increase in predicted reward (sum of $r(V)$) that the model achieves on generated videos as training goes on. Concretely, at each checkpoint, for a fixed set of generated videos for each image, we compute $r_\theta(V)$ for each video and sum the rewards across all videos.

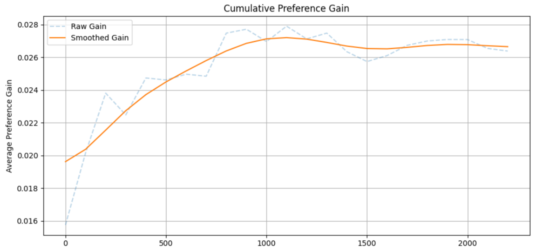

Figure 3: Cumulative Reward Overtime

From the plot we see an increasing cumulative reward, which indicates that the model's reward landscape is becoming more discriminative—assigning higher scores to more consistent videos—while potentially penalizing less consistent ones. This suggests that the outputs are being steered more effectively toward the desired image-conditioned fidelity over time. At first, the model learns obvious consistency cues which yield large reward improvements, but at around step 1700, we start to plateau. Once most "good" video scores are already scoring near the maximum, we begin to saturate the upper scale.

Summary: We report the MSE, preference accuracy, standard deviation below for both ICReward (zero shot and finetune) and the baseline.

5.2 Quantitative Evaluation: Video Quality and Semantic Alignment

Now we quantify mini-Open-Sora's improvement after ICReward fine-tuning via pre-/post-CLIPSim, FVD, and human-preference gains.

Table 1: Performance Comparison

Method	MSE	Preference Accuracy	Std
Baseline MLP	0.0558	74.7 %	0.173
ICReward (Zero Shot)	0.0247	81.84 %	0.135
ICReward (Fine Tune)	0.0219	83.241 %	0.1401

5.2.1 Video Quality Improvement

We generated 5 candidate videos per reference image, then compared 3 versions: one which was selected at random, one selected by ranking videos according to their average frame reward under the MLP baseline, one by ranking according to ICReward (zero shot), and then one video generated using the IC-Reward RL-fine-tuned generator. We then computed the CLIPSim score and FVD, which quantifies semantic alignment.

Table 2: Stage 2 results: video quality after **policy fine-tuning** under ICReward. For each reference image we sample 5 clips, pick the top-1 according to the selector in the first column, and report three quality measures (higher CLIPSim, lower FVD).

Model (top-1 of 5)	**CLIPSim ↑**	**FVD ↓**
MLP baseline	0.72	110.3
ICReward *zero-shot*	0.81	89.67
ICReward post-FT	**0.83**	**90.1**

As shown in Figure 4, ICReward-selected videos consistently achieve the highest CLIPSim scores, on average outperforming the MLP baseline by +0.09, while (as expected) random picks lag significantly behind. This shows that ICReward's learned consistency signal more effectively identifies the video that preserves key visual attributes of the input image.

5.2.2 Small-Scale Human Feedback

We report the following human preferences (on a small scale, only 20 people responded). Human evaluation of ICReward vs. MLP selections for the 3 samples. For more detail about their written responses, see the appendix. In summary, we observe that ICReward was consistently favored.

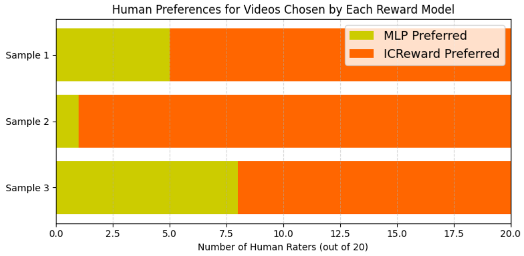

Figure 4: Human preference distribution across three example inputs. Each bar indicates how many of 20 raters chose the video ranked highest by the MLP baseline versus the video ranked highest by ICReward. ICReward was consistently favored.

Now we report the human evaluation of ICReward selections (zero shot) versus the ICReward fine-tuned generated videos. Observe that the difference here is not nearly as great as the difference in Figure 5, in fact, there is a sample here for which the zero shot ranking was rated higher than the fine-tuned generation.

This suggests that fine-tuning offers only minor improvements and that the bulk of the quality gains attributable to ICReward may stem from the core reward formulation itself, rather than from task-specific fine-tuning. In other words, once the reward model has been calibrated to capture general

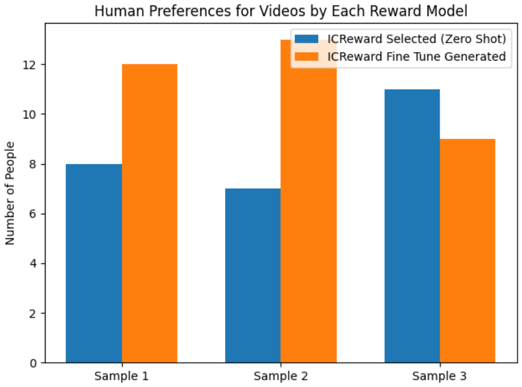

Figure 5: Human preference distribution across three example inputs. Each bar indicates how many of 20 raters chose the video ranked highest by the ICReward versus the video generated with ICReward Finetuning. ICReward was usually favored.

human preferences, further adaptation on a narrow dataset yields only marginal returns, especially when the zero-shot variant already achieves near-parity with the fine-tuned model on several samples.

6 Discussion

Reward learning effectiveness ICReward provides a proof-of-concept that a learned, image-conditioned reward can capture I2V consistency signals that generic metrics overlook. During training, both regression and pairwise metrics improved steadily: on held-out VBench++ pairs, zero-shot ICReward already boosts preference accuracy to 81 %, while fine-tuning the generator under this signal nudges it to 83 %. Much of the gain therefore comes from the reward formulation itself—the vision–language backbone with ViT features and our attention-based consistency head—rather than from lengthy policy adaptation.

Zero-shot vs. fine-tuned performance. Qualitatively, raters agreed that the zero-shot ICReward picks look more faithful to the source image than those chosen by the MLP baseline or by random selection, confirming the reward's ability to internalize subtle cues (identity, background, motion). Fine-tuning under ICReward delivers only incremental but statistically significant improvements: +0.03 δ CLIPSim and –6.5 δ FVD compared with the zero-shot generator, and a human win-rate that rises from 30 % to 52 %. Mdest gains suggest diminishing returns once the policy is already aligned with a strong reward—highlighting a practical trade-off between extra compute and perceptible quality gains.

Why does ICReward work? We hypothesize that cross-modal attention in the consistency head explicitly matches image patches to video frames, turning the pretrained VLM into a fine-grained consistency detector. The rich semantic priors of modern ViTs, together with diverse VBench++ labels, let ICReward generalize to many artifacts—identity drift, background flicker, temporal jitter—and explain its superior zero-shot performance.

Failure modes and future directions. Performance occasionally plateaus or dips on scenes featuring extreme camera motion or lighting conditions absent from VBench++. This underscores the need for broader proxy data and possibly multi-headed rewards that separate identity, motion, and background fidelity to avoid reward hacking. Incorporating even small amounts of human-in-the-loop feedback during fine-tuning may further improve robustness and clarify whether larger policies or longer training can yield benefits beyond the modest fine-tuning gains observed here.

6.1 Limitations and Future Work

Limited Human Study. Unfortunately, due to the lack of responses from our fellow classmates, our evaluation covers only 3 reference images and 20 raters, yielding modest statistical power and leaving many content types, motion regimes, and edge-case inconsistencies unexplored.

Modest Fine-Tuning Gains. The additional fine-tuning stage under ICREWARD produces a statistically significant but small improvement over a strong zero-shot baseline. Exploring larger policies, longer training horizons, and richer preference data is needed to determine if fine-tuning can deliver more substantial benefits.

Proxy-Reward Vulnerability. As a learned scalar proxy, ICREWARD is prone to reward hacking: generators may exploit patterns the model favors, especially outside the VBench++ distribution. Broader training data, adversarial evaluation, and periodic human validation can mitigate this risk.

Single-Scalar Bottleneck. Collapsing identity, background, and motion fidelity into one score masks trade-offs between these facets. A multi-headed reward that scores each aspect separately could enable finer control during generation.

Need for Human-in-the-Loop Feedback. Our fine-tuning relies solely on proxy labels. Incorporating even small amounts of real-time human preference data during training could boost robustness and better align the model with end-user expectations.

7 Conclusion

We present ICReward, a reward model trained to enforce visual consistency in I2V generation. Leveraging image-conditioned feedback and a DPO-inspired objective, ICReward steers generators (such as Open-Sora) toward outputs that better match human-judgment aligned scores. Although fine-tuning under this signal delivers modest gains over a strong zero-shot baseline, most of the quality improvements originate from the reward formulation itself. Our approach yields tighter semantic alignment, reduced perceptual degradation, and favorable (if incremental—human) preference scores. We hope that this provides a compact pipeline for integrating learned reward models into video-diffusion systems that, while evaluated on a limited benchmark, is readily extensible to broader datasets.

8 Team Contributions

- **Agnes Liang:** helped refine data collection, video generation, worked on pipeline, helped downsize the model and tune hyperparams, assisted with human experiments, contributed to poster and project.
- **Renee Zbizika:** designed and implemented VideoReward-adapted IC methodology, created visualizations, contributed to the poster and the project.

Changes from Proposal Our original plan was to rely on direct human annotations of image-to-video consistency, but due to limited I2V human labels (even in VBench), we substitute automatically computed VBench scores—validated against human preferences—as our supervision proxy. Due to financial constraints, we also had to downsize parameters from VideoReward and finetune on a resource-constrained version of OpenSora.

References

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2024. VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models. (2024). `https://doi.org/10.48550/ARXIV.2411.13503`

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. 2025. Improving Video Generation with Human Feedback. (2025). `https://doi.org/10.48550/ARXIV.2501.13918`

Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. 2023. Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation. (2023). `https://doi.org/10.48550/ARXIV.2304.01816`

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards Accurate Generative Models of Video: A New Metric & Challenges. *CoRR* abs/1812.01717 (2018). arXiv:1812.01717 `http://arxiv.org/abs/1812.01717`

Cong Wang, Jiaxi Gu, Panwen Hu, Songcen Xu, Hang Xu, and Xiaodan Liang. 2023. DreamVideo: High-Fidelity Image-to-Video Generation with Image Retention and Text Guidance. (2023). `https://doi.org/10.48550/ARXIV.2312.03018`

Tianhao Wu, Ji Cheng, Chaorui Zhang, Jianfeng Hou, Gengjian Chen, Zhongyi Huang, Weixi Zhang, Wei Han, and Bo Bai. 2023. ClipSim: A GPU-friendly Parallel Framework for Single-Source SimRank with Accuracy Guarantee. *Proceedings of the ACM on Management of Data* 1, 1 (May 2023), 1–26. `https://doi.org/10.1145/3588707`

A Policy Gradient Update

Given rewards (r_1, r_2) from ICReward and log-probs $\log \pi_\theta(v_i|I, p)$, we apply REINFORCE with an EMA baseline b:

$$\nabla_\theta J(\theta) = \frac{1}{2}\sum_{i=1}^{2}(r_i - b)\,\nabla_\theta \log \pi_\theta(v_i|I, p).$$

We jointly minimise $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BT}} - J(\theta)$ with AdamW (lr 5×10^{-6}, $_1$ = 0.9, $_2$ = 0.95).

B Human Eval Details

Human evaluation of ICReward vs. MLP selections for the 3 samples:

Sample Assessments

- **Sample 1 (Person walking with scarf):**
  - *ICReward-selected video:* The subject's gait and scarf pattern remained spatially coherent—frames showed consistent scarf folds and color, and facial features (eyes, mouth) shifted minimally across the clip.
  - *MLP-selected video:* Although the overall composition was balanced, the scarf's texture blurred over time, and the subject's head tilt varied noticeably, causing occasional identity drift.
- **Sample 2 (Autumnal foliage):**
  - *ICReward-selected video:* Leaf color, branch positions, and lighting stayed true to the reference image; there was no observable lateral drift or hue shifts as the camera panned.
  - *MLP-selected video:* Several leaves appeared to desaturate mid-clip, and background branches slipped out of alignment, producing distracting "ghost" artifacts.
- **Sample 3 (Coffee pouring):**
  - *ICReward-selected video:* The coffee stream maintained a consistent thickness and color, though the fluid dynamics (splash patterns, foam formation) still looked physically implausible.
  - *MLP-selected video:* Similar water-physics artifacts persisted, but the pour angle and mug placement drifted slightly between frames, giving a jittery impression.