

# Appendix

## Video Generation Models are Good Latent Reward Models

### A.1. Overview.

We provide detailed experimental settings in Sec. A.2, more experimental analysis about our reward model in Sec. A.3, and more experiments about our post-training algorithm in ??.

### A.2. More Details in Experimental Settings

#### A.2.1. Analysis Experiments

We selected the I2V task and the resolution of 720P for our analysis experiments, with the same dataset of our PAVRM training. The linear probe layer dimension matches the token dimension of the VGM (5120 for Wan2.1-I2V-14B).

For Fig. 3(a), we performed standard 40-step inference on dataset samples, denoise each intermediate timestep noisy latent to clean latent by one step, and decoding clean latent of each intermediate timestep to RGB space for storage. We computed the average score of dataset samples of each timestep using VideoAlign-MQ, a state-of-the-art VLM-based Video Reward Model. The results reveal substantial divergence between high-noise regions and clean videos, demonstrating that RGB-based video reward models fail to directly generalize as latent reward models.

For Fig. 3(b), we fixed the timestep at  $t = 0.2$  and analyzed the impact of varying DiT block counts on VGM performance as a reward model. To isolate the influence of VGM features on video quality assessment, we employed mean pooling for feature aggregation. The linear probe was trained identically to our PAVRM using BCE loss. For accuracy calculation, we applied a threshold of 0 to the linear probe output: predictions  $p_r \geq 0$  were classified as good videos, otherwise as bad videos. The VideoAlign test accuracy of 78.83% was obtained by setting a threshold on VideoAlign MQ reward scores—videos with scores above the threshold were labeled as good, otherwise bad—and computing accuracy against ground-truth labels. This threshold was selected to maximize test set accuracy.

For Fig. 3(c), “Fixed  $t$  (MLP-only)” refers to training only the linear probe at fixed timesteps ( $t = 0.2/0.4/0.6/0.8$ ), resulting in four separate models with test accuracy computed using the same way as (b), based on 8 DiT blocks. “Random  $t$  (MLP-only)” involves training a single model where timesteps are randomly sampled during training using UniPCMultistepScheduler with 1000 training steps. “Random  $t$  (Full fine-tuning)” fine-tunes both DiT blocks and the linear probe (excluding text/image encoders and VAE), with all other settings identical to “Random  $t$  (MLP-only)”.

#### A.2.2. Open Source Test Set

We incorporated the existing open-source benchmark VBench [8] and VBench2 [21] to ensure a fair comparison with state-of-the-art methods. In this paper, we validate the effectiveness of our method in terms of motion quality, which requires the video generation process to be free of distortions, exhibit smooth motion, and comply with physical laws. For the text-to-video (T2V) task, we selected the subject consistency subset from VBench, a total of 72 prompts. Additionally, we employed the human anatomy subset from VBench2, which includes the human anatomy metric with 120 prompts specifically enhanced for the Wan model. For the image-to-video (I2V) task, we selected I2V Subject subset from VBench-I2V (in VBench++ [9]), a total of 246 prompts.

#### A.2.3. Inner Data Collection and Annotations

**Data Generation Pipeline.** Our inner dataset construction begins with an internal collection of 31000 high-quality human portrait videos. Using the first frame and corresponding text prompt from each video, we generated synthetic videos using the Wan2.1-14B-I2V model. Due to computational constraints, we generated one 720P video per input condition, with each video requiring approximately 30 minutes of inference time on a single GPU.

**Annotation Protocol.** The annotation process consists of two stages: automatic filtering and manual quality assessment. *Stage 1: Coarse Filtering.* We first removed videos exhibiting obvious defects, including black screens, or visible watermarks. *Stage 2: Manual Quality Assessment.* The remaining videos were manually annotated by professional annotators across two key dimensions: **Physical Plausibility** and **Subject Deformity**. Each dimension was rated using a three-level scale: qualified, partially qualified, and unqualified. The specific criteria for each rating level are detailed in Table A1. For Physical Plausibility:

- *Qualified:* Motion appears smooth and natural, following real-world physics with realistic acceleration, deceleration, and interactions.
- *Partially Qualified:* Minor physical inconsistencies exist but do not severely impact overall believability.
- *Unqualified:* Significant violations of physical laws, such as objects defying gravity, unnatural motion trajectories, or implausible interactions.

For Subject Deformity:

- *Qualified:* Subjects maintain consistent structure and identity throughout the video with no visible artifacts.
- *Partially Qualified:* Minor temporal inconsistencies or subtle structural artifacts that do not fundamentally distort the subject.
- *Unqualified:* Severe anatomical distortions, identity shifts, or temporal artifacts such as melting, flickering, or morphing.

**Label Construction.** To enhance data distinctiveness and reduce annotation ambiguity, we applied the following labeling strategy: videos rated as *qualified* on both dimensions were labeled as **good videos**, while those rated as *unqualified* on both dimensions were labeled as **bad videos**. Videos with mixed ratings (e.g., qualified on one dimension but unqualified on another) or those marked as *partially qualified* on either dimension were excluded from the final dataset to ensure clear decision boundaries.

**Validation and Test Set Construction.** We randomly sampled 500 videos for evaluation purposes: 100 for validation and 400 for testing. To ensure annotation reliability, each sample in both the validation and test sets was independently annotated by at least three professional annotators, with final labels determined by majority voting.

**Final Dataset Statistics.** After filtering and annotation, our final dataset comprises 24000 video pairs (real and generated), with the generated videos used for reward model training and the real videos used as supervised fine-tuning (SFT) data for video generation. The dataset distribution is

as follows: approximately 23500 samples for training, 100 for validation, and 400 for testing the reward model.

#### A.2.4. Baseline Settings

For reward models, we select VideoAlign-MQ [13] and VideoPhy-PC [1], two state-of-the-art VLM-based reward models that excel particularly in assessing motion quality. Both models are employed in a zero-shot manner. The accuracy (Acc) metric is computed by establishing a threshold on the reward scores: videos with scores at or above the threshold are classified as “good”, while those below are classified as “bad”. The accuracy is then calculated against ground-truth labels, where the reported threshold is selected to maximize accuracy on the test set.

For post-training, we utilize approximately the same number of training samples across all methods, performing one epoch over the text-video pairs with a sequence parallel size of 4, with same learning rate of 5e-6 and a global batch size of 30 (i.e. batch size of 6 with gradient accumulation number of 5).

**Supervised Fine-Tuning (SFT).** SFT [14] is a widely adopted and effective post-training technique that offers high computational efficiency. From a reinforcement learning perspective, it can be viewed as an offline, off-policy algorithm, optimizing the loss function defined in Equation ??.

**Reward Weighted Regression (RWR).** Reward weighted regression (RWR) [13] is a prevalent and effective offline, off-policy RL method that has demonstrated success across traditional RL tasks [16], image generation [6], and video generation [13]. RWR directly learns from pre-sampled training data treated as experience samples, where a reward model scores each sample to determine its weight in the training loss. The loss function is given by:

$$\mathcal{L}_{\text{RWR}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}(0,1), \mathbf{x}_0 \sim q(\mathbf{x}_0), \mathbf{x}_1 \sim p(\mathbf{x}_1)} \left[ \exp(r_\phi(\text{video}, y)) \|\mathbf{v}_\theta(\mathbf{x}_t, t) - \mathbf{v}\|^2 \right], \quad (\text{A.2.1})$$

Following the VideoAlign framework, we utilize VideoAlign-MQ to provide reward signals with varying weight configurations.

**RGB ReFL.** For RGB ReFL, we adopt the implementation from ContentV [12], which performs VAE decoding only on the first frame and employs the image reward model PickScore [10]. The loss function is as follows:

$$\mathcal{L}_{\text{RGB ReFL}} = -\lambda \mathbb{E}_{\mathbf{x}_0 \sim \text{VGM}_\theta} [r_\phi(\mathcal{D}(\mathbf{x}'_0))] + \mathcal{L}_{\text{FM}}(\theta), \quad (\text{A.2.2})$$

where  $\mathbf{x}'_0$  denotes the latent feature corresponding to the first frame.

Table A1. Definitions of evaluation dimensions and assessment criteria used in our human annotation framework.

Evaluation Dimension	Definition and Assessment Criteria
Physical Plausibility	<p>Evaluates whether video dynamics adhere to <b>real-world physical principles</b>.</p> <ul style="list-style-type: none"> <li>- <b>Motion Dynamics:</b> Assesses whether object motion exhibits realistic acceleration, deceleration, and inertia consistent with natural physics.</li> <li>- <b>Interaction Realism:</b> Evaluates the plausibility of physical interactions, including gravitational effects (e.g., falling objects), collision dynamics, and force propagation (e.g., splashing water).</li> <li>- <b>Material Behavior:</b> Examines the realistic deformation and dynamics of complex materials, including fluid motion (water, smoke) and soft body dynamics (cloth, skin).</li> </ul>
Subject Deformity	<p>Assesses <b>structural integrity and temporal consistency</b> of subjects (humans, animals, objects).</p> <ul style="list-style-type: none"> <li>- <b>Structural Integrity:</b> Evaluates anatomical correctness and structural coherence, penalizing severe distortions, unnatural proportions, or implausible body parts (e.g., malformed faces, extra limbs).</li> <li>- <b>Temporal Consistency:</b> Measures the stability of subject identity and form across frames, penalizing artifacts such as shape morphing, flickering, melting effects, or sudden appearance changes.</li> </ul>

### A.2.5. Evaluation

**Experimental Configuration.** Following the standardized protocol recommended by Wan2.1, we generate evaluation videos at 720P/480P resolution. During the inference phase of video generation models, we maintain a classifier-free guidance (CFG) weight of 5.5. The sampling process employs the UniPCMultistepScheduler [20] over 40 iterative steps. The early, middle, and late stages of the denoising process correspond to steps 1-13, 14-26, and 27-40, respectively.

**Automatic Evaluation Metrics.** To assess the performance of our reward model, we implement a stratified sampling approach across the temporal dimension. The timestep  $t$  is partitioned into five uniform intervals:  $[0, 0.2]$ ,  $(0.2, 0.4]$ ,  $(0.4, 0.6]$ ,  $(0.6, 0.8]$ , and  $(0.8, 1.0]$ . Within each interval, test samples undergo random sampling exactly once, and the reward accuracy metric is derived by averaging the accuracy across all intervals.

For text-to-video generation tasks, we adopt multiple evaluation dimensions inspired by the VBench framework, encompassing dynamic degree, motion smoothness, subject consistency, and human anatomy accuracy. In image-to-video generation scenarios, we additionally incorporate the image-video subject consistency metric. Furthermore, we utilize the PAVRM score to quantify the proportion of qualified samples across the entire test set.

**Motion Smoothness.** Following VBench, we evaluate motion fluidity using frame interpolation priors. Given a video

sequence  $[f_0, f_1, \dots, f_{2n}]$ , we remove odd-indexed frames to create  $[f_0, f_2, \dots, f_{2n}]$ , then reconstruct the missing frames  $[\hat{f}_1, \hat{f}_3, \dots, \hat{f}_{2n-1}]$  via interpolation. The normalized MAE between reconstructed and original frames yields a score in  $[0, 1]$ , with higher values indicating smoother motion.

**Dynamic Degree.** To measure generation dynamism, we adopt VBench’s approach using RAFT [17] to estimate inter-frame optical flow. We compute the mean of the top 5% flow magnitudes as a static/dynamic threshold, with the final score representing the proportion of non-static videos generated.

**Subject Consistency.** We adopt VBench’s DINO-based [3] metric to assess subject identity preservation across frames. The consistency score is:

$$S_{\text{subject}} = \frac{1}{T-1} \sum_{t=2}^T \frac{1}{2} (\langle d_1, d_t \rangle + \langle d_{t-1}, d_t \rangle), \quad (\text{A.2.3})$$

where  $d_i$  is the normalized DINO feature of frame  $i$ , and  $\langle \cdot, \cdot \rangle$  computes cosine similarity. This jointly measures consistency with the first frame and temporal continuity.

**Human Anatomy.** We use the VBench metric that they train three ViT-based [18] anomaly detectors for human torso, hands, and faces. Training data includes  $\sim 1\text{K}$  real videos (YOLO-World [4] extracted patches as positives) and  $\sim 1\text{K}$  synthetic videos from CogVideo [7, 19] and HunyuanVideo [11], plus HumanRefiner [5] negatives, totaling  $\sim 150\text{K}$  annotated frames. The score is the percentage of frames without detected anomalies.

*I2V Subject Consistency.* We use the VBench++ [9] metric to evaluate input image-to-video subject correspondence. DINOv1 [3] features are extracted from the input image and video frames. The final score combines weighted similarities between the input image and each frame, plus inter-frame similarities, addressing variations in how models handle input images. *PAVRM Score.* To estimate the qualified sample ratio across the test set, we adopt a randomized evaluation protocol. For each test sample, we randomly sample a timestep  $t$  from the interval  $[0, 1.0]$  and feed it to the PAVRM model, which produces a binary prediction: 0 for unqualified and 1 for qualified. The overall qualified ratio serves as the PAVRM score metric.

**Complete Prompt in Fig.??** *Case 1.* A woman in a flowing white dress is dancing gracefully in a modern dance studio. Her movements are fluid and expressive, with arms sweeping widely and legs moving in elegant, rhythmic patterns. She has long wavy hair that flows freely with each movement, catching the soft lighting from above. The background is a minimalist setup with black walls and a few abstract paintings hanging on them. The camera follows her from a medium shot, capturing her full body as she dances, then moves to a close-up of her face, highlighting her joyful expression and the sparkle in her eyes. The video has smooth transitions and dynamic camera movements, including tracking shots and slow-motion sequences to emphasize her graceful movements.

*Case 2.* Four people are seated on an ornate rug in a room with exposed brick walls. One man holds an acoustic guitar. Instruments including a saxophone and harmonica rest on the floor near them. The individuals have varying hair colors and styles; one woman has long dreadlocks. They wear casual clothing like t-shirts, jeans, and sneakers.

Table A2. Ablation study on training objectives. The models are trained on Wan2.1 generated videos and evaluated on the held-out test set. The metric is classification accuracy (%). **Bold** indicates the best performance.

Loss	[0, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.6, 0.8]	(0.8, 1.0]	Avg
BT	73.50	77.25	78.25	<b>82.50</b>	<b>87.75</b>	79.85
BCE	<b>77.00</b>	<b>78.50</b>	<b>79.75</b>	82.00	83.00	<b>80.05</b>

**User Study.** We ask each evaluator: For each question, two options represent videos generated from the title text using two different models. Select the option with the

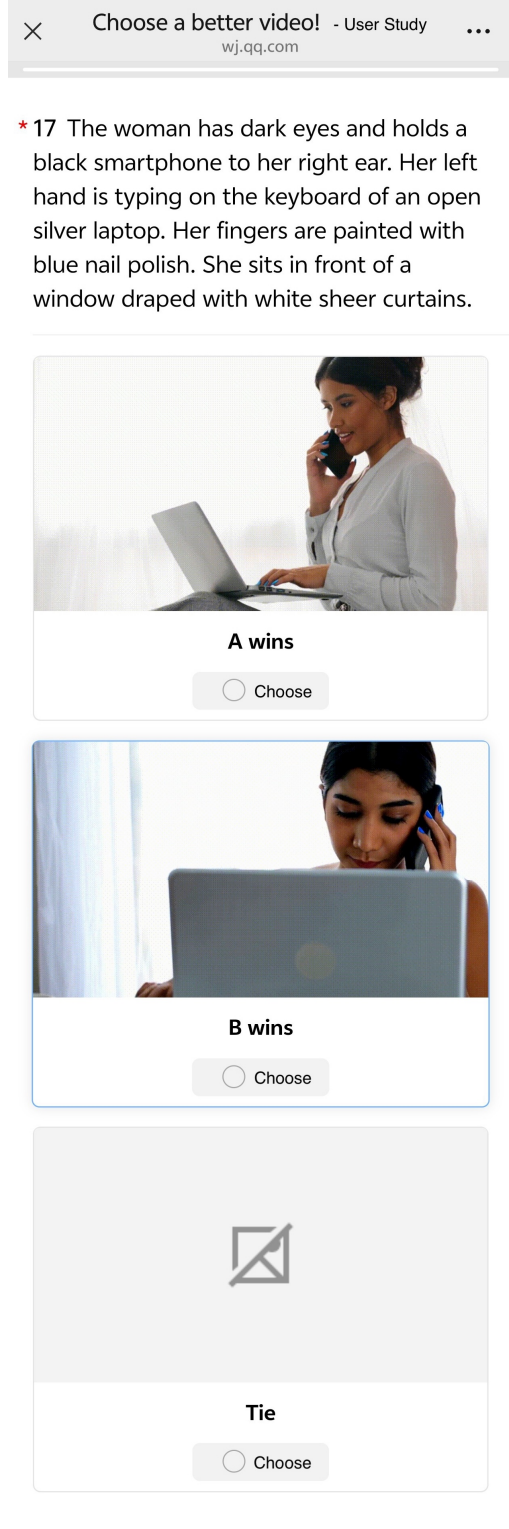


Figure A1. A case of user study page.

higher overall quality (greater text-video consistency, more natural motion, and no human deformities or physically implausible elements). There are 100 questions in total.



Table A3. Ablation study on the number of trainable DiT blocks. The metric is classification accuracy (%). **Bold** indicates the best performance.

Layer	[0, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.6, 0.8]	(0.8, 1.0]	Avg
8	83.42	84.95	84.69	84.44	83.42	84.18
16	<b>84.95</b>	<b>85.46</b>	<b>85.71</b>	<b>86.73</b>	84.69	<b>85.51</b>
24	83.93	85.20	<b>85.71</b>	86.22	<b>85.20</b>	85.25
32	80.36	83.67	84.95	85.97	<b>85.20</b>	84.03
40 (Full)	79.85	81.12	84.95	85.46	84.95	83.27

The page is shown in Fig. A1.

### A.3. More Experiments on Process-Aware Video Reward Models

#### A.3.1. The Influence of Training Loss

To assess the robustness of our proposed method against different optimization objectives, we compare the standard binary cross-entropy (BCE) loss with the pairwise Bradley-Terry (BT) loss [2]. Specifically, we construct preference pairs  $(x_{\text{win}}, x_{\text{lose}})$  by randomly sampling a positive sample (label 1) as  $x_{\text{win}}$  and a negative sample (label 0) as  $x_{\text{lose}}$ . As shown in Table A2, the average accuracy gap between the two objectives is marginal (0.2%). Interestingly, we observe a trade-off across timesteps: BT loss performs better in high-noise regions ( $t > 0.6$ ), while BCE demonstrates superior precision in the structure-forming and detailing stages ( $t \leq 0.6$ ). Given that BCE achieves a slightly higher overall average accuracy and eliminates the computational overhead of pair construction, we adopt BCE as our default training objective.

#### A.3.2. The Influence of the Number of DiT Blocks

In our feasibility analysis, we observed that fixed DiT features are effective. Here, we investigate the impact of model depth when the DiT blocks are *fully fine-tuned*. We vary the number of trainable DiT blocks (from the first 8 to the full 40 blocks) to determine the optimal capacity for the reward task. The results in Table A3 reveal a non-monotonic trend, contradicting a simple scaling law. The performance peaks at 16 blocks (**85.51%**) and subsequently degrades as more layers are added, with the full 40-block model performing worse than the 8-block baseline. This suggests that the critical semantic information for assessing motion quality is concentrated in the early-to-middle layers of the network. Using the full generation backbone for the reward task is not only computationally expensive but potentially leads to optimization difficulties or overfitting to high-frequency generation details rather than high-level quality. Consequently, using the first 8 or 16 blocks offers the best trade-off between efficiency and accuracy.

#### A.3.3. Cross-Model Generalization

To evaluate the generalization capability of PAVRM, we extend our evaluation beyond the source domain. While our reward model is initialized and trained solely on data generated by Wan2.1, we test its performance on samples from two other state-of-the-art video generation models: HunyuanVideo [11] and Veo3. The test sets for these models share the same annotation format but specifically focus on human structural deformities (e.g., limb distortions), a common challenge in video generation.

The results in Tab. A4 reveal two key insights regarding transferability and timestep sensitivity:

**Feasibility of Cross-Model Evaluation.** First, PAVRM demonstrates strong zero-shot transferability. Despite being trained exclusively on Wan2.1 latents, it effectively identifies quality degradation in HunyuanVideo and Veo3. This suggests that the spatiotemporal features learned by the backbone VGM are not strictly model-specific but encode universal representations of motion and structure validity.

**Inverted Generalization across Timesteps.** We observe a distinct behavior in performance distribution across diffusion timesteps ( $t$ ) between in-domain and out-of-domain (OOD) settings:

- **In-Domain (Wan2.1):** The model achieves higher performance in the middle and last trajectory ( $t \in [0.2, 1]$ ). This aligns with the intuition that intermediate states balance signal and noise, containing the most critical information for motion formation.
- **Out-of-Domain (Hunyuan/Veo3):** Surprisingly, generalization is stronger in high-noise regions ( $t \rightarrow 1$ ) compared to the near-data regions ( $t \rightarrow 0$ ).

**Analysis.** We attribute this phenomenon to the nature of the denoising process. In the late stages of generation ( $t \rightarrow 0$ ), the latents are dominated by model-specific high-frequency details and “fingerprints” (unique texture patterns or artifact types inherent to the specific generator architecture). A reward model trained on Wan2.1 overfits to these specific patterns, leading to poor transfer when evaluating clean latents from other models. Conversely, at high noise levels ( $t \rightarrow 1$ ), the latent representation is dominated by Gaussian noise and low-frequency structural layouts. The “fingerprints” of the specific generative model are less pronounced, while fundamental structural errors (such as severe human deformities) remain detectable as gross geometric inconsistencies. Consequently, the reward model relies on these universal structural cues rather than model-specific textures, resulting in superior generalization in high-noise regimes.

Table A4. Dataset analysis on PAVRMs, the model is based on Wan2.1. The metric is average accuracy. Task is I2V and resolution is 720P.

Train Set	Test Set	[0, 0.2]	(0.2, 0.4]	(0.4, 0.6]	(0.6, 0.8]	(0.8, 1.0]	Avg
Veo3&HunyuanVideo	Veo3&HunyuanVideo	<b>77.00%</b>	<b>86.00%</b>	<b>89.00%</b>	<b>89.00%</b>	<b>87.00%</b>	<b>85.60%</b>
Wan2.1	Veo3&HunyuanVideo	70.00%	72.00%	73.00%	76.00%	81.00%	74.40%

Table A5. Full VBench evaluation results (480P T2V).

Models	Background Consistency	Temporal Flickering	Aesthetic Quality	Imaging Quality	Object Class	Multiple Objects	Color
Wan2.1	<b>97.79</b>	98.85	<b>68.95</b>	<b>64.87</b>	80.38	63.19	76.93
+PRFL	97.64	<b>99.01</b>	68.60	64.55	<b>97.15</b>	<b>84.38</b>	<b>91.32</b>
Models	Spatial Relationship	Human Action	Scene	Appearance Style	Temporal Style	Overall Consistency	Temporal Style
Wan2.1	61.28	93.00	32.78	21.83	22.20	25.47	22.20
+PRFL	<b>82.53</b>	<b>96.00</b>	<b>50.36</b>	<b>21.21</b>	<b>23.32</b>	<b>25.77</b>	<b>23.32</b>

### A.3.4. Comprehensive Evaluation on VBench and VBench 2.0

Tables A5 and A6 present full per-dimension results on VBench and VBench 2.0 respectively. Beyond motion and artifact improvements, PRFL achieves notable gains in semantic understanding (Scene +17.58) and spatial reasoning (Spatial Relationship +21.25) with minimal aesthetic degradation.

A closer inspection reveals a consistent pattern: improvements are most pronounced in dimensions that require high-level semantic or relational understanding, while low-level perceptual metrics remain largely stable. On VBench, PRFL yields substantial gains in Object Class (+16.77), Multiple Objects (+21.19), Color (+14.39), and Human Action (+3.00), whereas Background Consistency (−0.15) and Aesthetic Quality (−0.35) show only marginal changes, indicating that the reward signal steers the model toward semantic fidelity without sacrificing visual naturalness.

On VBench 2.0, the gains are even more pronounced in motion-related and compositional dimensions. Camera Motion improves by +38.89, Motion Rationality by +25.86, and Motion Order Understanding by +27.27, demonstrating that process-level supervision effectively enhances the model’s ability to reason about physical plausibility and temporal causality. Composition (+17.08), Human Interaction (+20.00), and Complex Plot (+9.28) further confirm that PRFL strengthens the model’s understanding of structured, multi-element scenarios. Human Clothes achieves a perfect score of 100.00. The only dimension that exhibits a notable decline is Multi-view Consistency (−3.28), suggesting a mild trade-off between dynamic expressiveness and cross-view geometric coherence, which we leave for future investigation.

### A.3.5. Generalization to a Different Backbone: Wan2.2

To verify that PRFL is not limited to a single architecture, we apply it to Wan2.2, which adopts a Mixture-of-Experts (MoE) design different from Wan2.1. As shown in Table A7, PRFL consistently improves performance across all evaluated dimensions, demonstrating the backbone-agnostic nature of our framework.

Concretely, the average score increases from 85.34 to 90.96 (+5.62). The most significant gains are observed in Dynamic Degree, which improves from 39.00 to 64.00 (+25.00) on the inner test set and from 76.39 to 84.72 (+8.33) on VBench, echoing the trend observed on Wan2.1. Human Action similarly benefits, improving from 90.36 to 96.47 (+6.11) on the inner test set and from 76.13 to 89.81 (+13.68) on VBench 2.0. These results indicate that the process reward signal captures generalizable quality criteria that transfer across fundamentally different architectural designs, and that the benefits of PRFL are not an artifact of any specific model inductive bias.

### A.3.6. Combination with Inference-Time Methods

We further examine whether PRFL is complementary to inference-time alignment methods. Table A8 shows results when combining PRFL with Diffusion Latent Beam Search (DLBS) [15]. The two approaches improve different aspects: PRFL primarily boosts dynamic degree and subject consistency during training, while DLBS contributes further subject consistency gains at inference time, and their combination achieves competitive overall performance.

Looking at the numbers more closely, PRFL alone achieves the highest average score (92.52), outperforming both DLBS alone (89.50) and the combined setting (91.44). Notably, DLBS alone reduces Dynamic Degree below the

Table A6. Full VBench 2.0 evaluation results (480P T2V).

Models	Human Clothes	Human Identity	Composition	Diversity	Mechanics	Material	Thermotics	Multi-view Consistency	Dynamic Spatial Relationship
Wan2.1	98.44	<b>66.82</b>	43.74	<b>48.27</b>	60.98	68.75	55.00	<b>32.90</b>	23.19
+PRFL	<b>100.00</b>	66.65	<b>60.82</b>	48.23	<b>74.47</b>	<b>73.81</b>	<b>60.82</b>	29.62	<b>39.13</b>
Models	Dynamic Attribute	Motion Order Understanding	Human Interaction	Complex Landscape	Complex Plot	Camera Motion	Motion Rationality	Instance Preservation	
Wan2.1	37.36	18.18	68.00	12.00	8.00	19.44	29.31	80.70	
+PRFL	<b>43.96</b>	<b>45.45</b>	<b>88.00</b>	<b>23.33</b>	<b>17.28</b>	<b>58.33</b>	<b>55.17</b>	<b>89.47</b>	

Table A7. Comparison based on Wan2.2 (480P T2V).

Method	Inner Test Set					VBench				VBench2		Avg
	MS	DD	SC	HA	PAVRM	MS	DD	SC	PAVRM	HA	PAVRM	
Wan2.2	98.81	39.00	96.08	90.36	<b>99.00</b>	96.74	76.39	88.69	<b>100.00</b>	76.13	77.50	85.34
+PRFL	<b>98.91</b>	<b>64.00</b>	<b>96.87</b>	<b>96.47</b>	<b>99.00</b>	<b>97.47</b>	<b>84.72</b>	<b>92.43</b>	<b>100.00</b>	<b>89.81</b>	<b>80.83</b>	<b>90.96</b>

Table A8. Combined with DLBS [15] based on Wan2.1 (subset of VBench, 480P T2V).

Method	Motion Smoothness	Dynamic Degree	Subject Consistency	Average
Wan2.1	98.31	80.00	93.37	90.56
+PRFL	98.37	<b>85.00</b>	94.19	<b>92.52</b>
+DLBS	98.36	75.00	95.13	89.50
+PRFL+DLBS	<b>98.46</b>	80.00	<b>95.87</b>	91.44

baseline (75.00 vs. 80.00), whereas PRFL alone pushes it to 85.00; when the two methods are combined, Dynamic Degree returns to the baseline level (80.00), suggesting a degree of tension between DLBS’s beam search objective and the dynamic diversity encouraged by PRFL. On the other hand, Subject Consistency benefits from complementary optimization: PRFL+DLBS attains the best score (95.87), surpassing either method alone. These findings suggest that PRFL and DLBS occupy partially overlapping but distinct regions of the quality landscape, and that their combination is most beneficial when the deployment priority is consistency over dynamism.

## References

- [1] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *Workshop on Video-Language Models @ NeurIPS 2024*, 2025. 2
- [2] 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. 5
- [3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3, 4
- [4] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xingang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16901–16911, 2024. 3
- [5] Guian Fang, Wenbiao Yan, Yuanfan Guo, Jianhua Han, Zutaotao Jiang, Hang Xu, Shengcai Liao, and Xiaodan Liang. Humanrefiner: Benchmarking abnormal human generation and refining with coarse-to-fine pose-reversible guidance. In *European Conference on Computer Vision*, pages 201–217. Springer, 2024. 3
- [6] Hiroki Furuta, Heiga Zen, Dale Schuurmans, Aleksandra Faust, Yutaka Matsuo, Percy Liang, and Sherry Yang. Improving dynamic object interactions in text-to-video generation with ai feedback. *arXiv preprint arXiv:2412.02617*, 2024. 2
- [7] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 3
- [8] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1
- [9] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024. 1, 4
- [10] 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. *Advances in neural information processing systems*, 36:36652–36663, 2023. 2
- [11] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 3, 5
- [12] Wenfeng Lin, Renjie Chen, Boyuan Liu, Shiyue Yan, Ruoyu Feng, Jiangchuan Wei, Yichen Zhang, Yimeng Zhou, Chao Feng, Jiao Ran, et al. Contentv: Efficient training of video generation models with limited compute. *arXiv preprint arXiv:2506.05343*, 2025. 2
- [13] 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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2
- [14] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. 2
- [15] Yuta Oshima, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Inference-time text-to-video alignment with diffusion latent beam search. In *Advances in Neural Information Processing Systems*, 2025. 6, 7
- [16] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019. 2
- [17] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 3
- [18] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022. 3
- [19] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3
- [20] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36:49842–49869, 2023. 3
- [21] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, 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. 1