NoisEasier: Test-Time Noise Optimization for Text-to-Video Generation

Yujiang Pu[1] and Yu Kong[1]

Michigan State University, East Lansing, MI 48824, USA
{puyujian,yukong}@msu.edu
https://yujiangpu20.github.io/noiseasier/

Abstract. Diffusion models have recently advanced text-to-video (T2V) generation, yet they still struggle with fine-grained compositional alignment, such as attribute binding, spatial relations, and object interactions. While reward-based fine-tuning improves alignment, it is susceptible to reward hacking and adapts poorly to new prompt distributions. In this work, we propose **NoisEasier**, a test-time scaling framework that improves T2V generation through differentiable reward-guided noise optimization without modifying the underlying model. By combining efficient short-step generators with a multi-objective reward formulation, NoisEasier enables stable and practical test-time optimization under realistic inference budgets. Our key insight is that jointly optimizing the entire stochastic trajectory accelerates reward convergence and improves compositional alignment over optimizing only the initial latent, with negligible additional computational and time cost. Experiments on VBench and T2V-CompBench demonstrate consistent improvements across multiple backbones, achieving over 10% average gains on challenging dimensions such as attribute binding, object interaction, and numeracy. Overall, NoisEasier serves as both a flexible alternative and a complementary enhancement to reward-based fine-tuning, establishing test-time scaling as an effective paradigm for controllable text-to-video generation.

Keywords: Video generation · Test-time scaling · Noise optimization

1 Introduction

Text-to-video (T2V) diffusion models [3, 5, 16, 18, 40, 51, 54, 68] have recently achieved remarkable progress, producing videos with increasingly realistic appearance, diverse content, and coherent motion. Despite these advances, today's systems still struggle with fine-grained compositional control that faithfully reflects user intent [44, 47]. Subtle details such as attribute bindings, object interactions, and motion dynamics often become unreliable when prompts require precise coordination among multiple semantic concepts. These limitations reveal a broader challenge: how can we align video generation more closely with user-specified semantics while preserving visual fidelity?

Fig. 1: NoisEasier significantly boosts semantic alignment while maintaining the visual fidelity of generated videos. The results are obtained with 25-step noise optimization.

Most existing works draw inspiration from Reinforcement Learning from Human Feedback (RLHF) [9, 35] and adapt similar paradigms to T2V generation. Early methods [38, 64] optimize diffusion models using differentiable rewards [22, 57, 60], while subsequent works rely on human-preference datasets for direct preference optimization [8, 11, 32] or reward-based fine-tuning [31, 58]. While effective in improving global alignment, these methods inherit two intrinsic limitations. First, the learned preferences are baked into the model parameters, making continual adaptation to new reward functions or prompt distributions expensive. Second, these methods remain susceptible to reward hacking, where optimizing imperfect reward functions can degrade visual quality or motion realism. This motivates a natural question: *can we improve compositional alignment at test time without modifying the underlying video generator?*

Recent advances in test-time scaling have shown that increasing inference-time computation can substantially improve model capability without retraining. For diffusion models, latent noise has emerged as an effective optimization variable [13, 29, 39, 61, 69], leading to successful applications in image [6, 17], motion [21], and music generation [33]. However, extending noise optimization to T2V remains largely unexplored due to two fundamental challenges. First, modern T2V models [23,51] require long denoising chains and high memory consumption, making iterative gradient-based refinement prohibitively expensive. Second, video reward signals are considerably more fragile than image rewards [38, 64], making optimization prone to instability and reward hacking.

In this work, we propose **NoisEasier**, a practical test-time scaling framework for T2V generation. To overcome the prohibitive cost of iterative optimization, we build upon recent Video Consistency Models (VCMs) [26, 27, 52],

which distill long diffusion sampling into only a few denoising steps, making end-to-end gradient-based refinement practical. While VCMs address the computational bottleneck, stable optimization still requires reliable reward supervision. We therefore formulate test-time refinement as a robust multi-objective optimization problem that combines complementary semantic, aesthetic, and motion-aware rewards to mitigate reward hacking. To further strengthen compositional supervision, we introduce a lightweight negative-aware reward calibration strategy that calibrates absolute reward scores using semantically hard negative prompts, yielding more discriminative feedback.

Interestingly, the intermediate stochastic perturbations exposed by VCMs provide additional optimization variables beyond the initial latent noise. Rather than optimizing only the initialization [13,17,45,61], we find that jointly refining all available stochastic variables consistently yields faster reward convergence and stronger compositional alignment, while introducing negligible additional runtime or memory overhead. This observation suggests that optimizing the entire denoising process provides finer control over video generation than optimizing only its initialization, offering a simple yet effective test-time scaling strategy for modern T2V models.

Extensive experiments on VBench [20] and T2V-CompBench [43] demonstrate that NoisEasier significantly improves the performance of VCM baselines [26, 52] with only modest additional inference overhead. Compared with optimizing the initial latent alone, jointly optimizing the intermediate perturbations yields more consistent improvements on compositional dimensions, including attribute binding, spatial relations, and object interactions. Furthermore, NoisEasier continues to improve reward-finetuned models, suggesting that offline fine-tuning does not fully amortize reward signals into the model parameters. Consequently, NoisEasier serves not only as a flexible alternative to reward fine-tuning, but also as a complementary test-time scaling mechanism that unlocks additional gains beyond offline alignment.

In summary, our main contributions are as follows:

- We propose **NoisEasier**, a test-time scaling framework for text-to-video generation that enables efficient and robust reward-guided optimization without model fine-tuning.
- We develop an efficient optimization pipeline based on short-step Video Consistency Models together with a robust reward formulation for stable compositional optimization.
- We show that jointly optimizing the intermediate perturbations consistently accelerates reward convergence and improves compositional alignment over initial-noise optimization with negligible additional computational cost.

2 Related Work

Video Generation with Human Feedback. Reinforcement Learning from Human Feedback (RLHF) [9, 35] has shown great success in aligning large language models (LLMs) with human intent, motivating its extension to image

and video generation. Existing work typically follows two paradigms. One is *preference-driven optimization*, where models are trained directly on large-scale human preference datasets. This includes policy gradient methods like DDPO [2] and DPOK [15], as well as direct preference optimization methods [30,32,49,62] that bypass explicit reward models. While effective, these approaches require large-scale preference annotations and cannot easily incorporate structured human knowledge. The other is *reward-driven optimization*, where pretrained reward models provide differentiable feedback for fine-tuning generative models [38,64] or guide distillation for faster inference [25–27]. Although these methods reduce training cost and improve stability, they are generally one-shot systems that lack mechanisms for iterative refinement. In contrast, we directly optimize latent noise via differentiable rewards, enabling preference-controllable text-to-video generation with efficient test-time refinement.

Test-time Noise Optimization. Recent studies show that the initial latent noise strongly influences diffusion generation quality, with certain *golden noise* consistently yielding outputs that are more faithful and aesthetically pleasing. Building on this observation, test-time methods that operate on noise can be broadly categorized into two directions: *selection-based noise search* and *optimization-based noise refinement*. Selection-based methods [29, 61, 69] treat sampling as a search problem over initial noises or trajectories and select the best sample under a fixed backbone. In contrast, optimization-based methods directly refine the noisy latent via gradient-based updates during sampling. DOODL [50] pioneered end-to-end differentiable latent optimization, and subsequent work extended this paradigm to images [6, 17, 45, 69] and other modalities, including music [34] and motion [21]. While effective, optimization-based approaches can be prohibitively slow at inference time; for example, DOODL [50] and D-Flow [1] may take tens of minutes per sample. ReNO [13] partially mitigates this bottleneck by applying optimization to one-step diffusion models, reducing runtime to tens of seconds. In this work, we revisit noise optimization for text-to-video generation, which remains underexplored due to the high cost of video sampling and the scarcity of suitable spatiotemporal reward signals.

3 Method

3.1 Problem Formulation

Let $\mathcal{G}_\theta$ denote a pretrained text-to-video (T2V) diffusion model. Given a text prompt c, the generation process starts from an initial latent noise $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$ and progressively denoises it into a video sequence. Existing test-time optimization methods formulate noise refinement by directly optimizing the initial latent:

$$\mathbf{z}_T^* = \arg\max_{\mathbf{z}_T} \mathcal{R}(\mathcal{G}_\theta(\mathbf{z}_T, c)), \tag{1}$$

where $\mathcal{R}(\cdot)$ denotes a differentiable reward measuring semantic alignment or perceptual quality. While conceptually simple, directly optimizing $\mathbf{z}_T$ for T2V generation is computationally expensive, as gradients must propagate through

dozens of denoising steps. Previous works [10, 37, 38] alleviate this cost through gradient truncation or partial backpropagation, sacrificing end-to-end optimization while remaining computationally demanding. In contrast, we build upon recent Video Consistency Models (VCMs) [26, 52], which distill long diffusion sampling into only a few denoising steps.

Specifically, a consistency model learns a mapping that directly predicts the clean sample along the probability-flow ODE trajectory: $f : (\mathbf{z}_t, t, c) \mapsto \mathbf{z}_\kappa, t \in [\kappa, T]$, where κ is a small positive constant. The model is trained with a self-consistency objective that enforces identical predictions for different noisy states originating from the same trajectory:

$$f(\mathbf{z}_t, t, c) = f(\mathbf{z}_{t'}, t', c), \quad \forall\, t, t' \in [\kappa, T], \tag{2}$$

allowing the original diffusion trajectory to be approximated using only a small number of denoising steps. This substantially reduces the cost of gradient propagation and enables practical end-to-end optimization for video generation. During inference, VCM alternates between predicting the clean sample $\hat{\mathbf{z}}_0 = f(\mathbf{z}_t, t, c)$ and injecting Gaussian perturbations $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \mathbf{I})$ to obtain the next latent

$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, \hat{\mathbf{z}}_0 + \sqrt{1 - \bar{\alpha}_{t-1}}\, \boldsymbol{\varepsilon}_t. \tag{3}$$

Unlike conventional diffusion sampling, the stochastic formulation of VCM naturally exposes the injected perturbations as additional optimization variables. Rather than optimizing only the initial latent as in Eq. 1, we jointly optimize all available stochastic variables throughout the denoising process:

$$(\mathbf{z}_T^*, \{\boldsymbol{\varepsilon}_t^*\}) = \arg \max_{\mathbf{z}_T, \{\boldsymbol{\varepsilon}_t\}} \mathcal{R}(\mathcal{G}_\theta(\mathbf{z}_T, c; \{\boldsymbol{\varepsilon}_t\}))\,. \tag{4}$$

This formulation provides finer control over the denoising trajectory and consistently accelerates reward convergence compared with optimizing only the initial latent, while introducing negligible additional computational overhead.

3.2 Reward Models for Video Generation

While human-preference reward models are well studied for image generation [22, 56, 60], reward modeling for video remains underexplored. Recent works such as VideoRM [58], VideoScore [19], and VisionReward [59] leverage multimodal large language models (MLLMs) to assess generated videos. However, directly using large MLLM-based reward models inside gradient-based noise refinement can be computationally prohibitive, since the reward must be evaluated repeatedly during optimization. Therefore, we focus on lightweight, differentiable rewards built on pretrained vision-language models (VLMs), which provide strong semantic alignment signals while remaining efficient for test-time optimization. Specifically, we consider three types of reward objectives:

Image-Text Alignment Reward. We employ HPSv2 [56] and ImageReward [60], both trained to model human preferences for image-text relevance. HPSv2 uses a

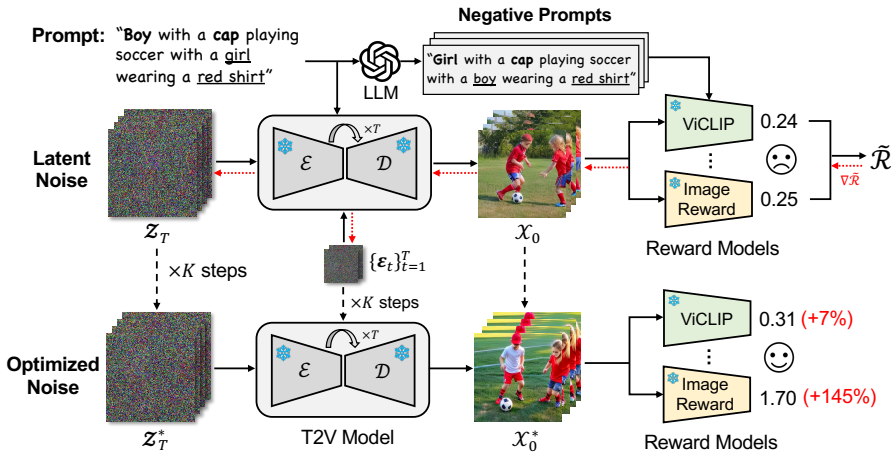


Fig. 2: Overview of the proposed NoisEasier framework. Given a text prompt, both the initial latent noise and intermediate perturbations are iteratively optimized using gradients from multiple reward models. Negative prompts generated by an LLM further calibrate the reward signal for fine-grained compositional feedback.

CLIP-based [41] architecture with a fine-tuned reward head, while ImageReward adopts BLIP [28] as its backbone, fine-tuned with supervised human feedback to improve perceptual alignment. Both models produce scalar scores that reflect the semantic alignment between the generated image and the text prompt. In practice, we compute the reward by evaluating each frame-text pair and take the mean score across all frames as the final image-text alignment reward.

Video-Text Alignment Reward. Image-level reward models provide strong semantic cues but lack temporal perception, making them ineffective at capturing motion consistency in video generation. To address this, we incorporate ViCLIP [55], a video-language model trained on the large-scale InternVid dataset. ViCLIP jointly encodes short video clips and text descriptions to assess temporal and semantic consistency. In practice, we sample two temporally offset clips from the generated video, each consisting of 8 frames, and compute their similarity to the text prompt. The final reward is the average of the two clip-text similarity scores, which provides video-level alignment feedback.

Motion-aware Reward. Although VLM-based rewards help improve semantic alignment with the prompt, they often push the model to favor static content to maximize alignment scores. As a result, the generated videos may lack meaningful motion or introduce temporal artifacts. To mitigate this, we introduce a motion-aware reward that explicitly encourages localized dynamics while regularizing temporal coherence. Specifically, we use RAFT [46] to estimate optical flow $\{\mathbf{F}_t\}_{t=1}^{T-1}$, where $\mathbf{F}_t \in \mathbb{R}^{2\times H\times W}$ captures motion from frame t to $t+1$. We define the spatial mean flow vector $\bar{\mathbf{f}}_t = \frac{1}{HW}\sum_{h,w}\mathbf{F}_t[:,h,w]$ and subtract it to

suppress global camera motion. The motion reward is computed as:

$$\mathcal{R}_{\text{mo}} = \tanh\Big(\tfrac{1}{T-1}\sum_{t=1}^{T-1}\|\mathbf{F}_t - \bar{\mathbf{f}}_t\|_1\Big) + \lambda\Big(1 - \tanh\Big(\tfrac{1}{T-2}\sum_{t=1}^{T-2}\|\mathbf{F}_{t+1} - \mathbf{F}_t\|_1\Big)\Big), \quad (5)$$

where the first term encourages localized motion beyond global translation, and the second term penalizes abrupt changes in flow magnitude to promote temporally coherent dynamics. In practice, this objective effectively mitigates the static bias induced by image-based rewards, leading to more expressive and dynamic video generation.

3.3 Negative-Aware Reward Calibration

Most existing reward models [22, 56, 57, 60] are trained on human preferences using pairwise ranking, providing only weak supervision for fine-grained compositional reasoning. Prior studies [12, 48, 65] further show that VLM-based evaluators often behave like *bag-of-words* classifiers. Consequently, reward signals can be overly coarse and insensitive to subtle semantic errors. To address this limitation, we introduce *negative-aware* reward calibration, which contrasts the target prompt against semantically perturbed distractors rather than relying solely on absolute reward scores.

Given a prompt c^+, we utilize a large language model (LLM) to construct a set of hard negative prompts $\{c_j^-\}_{j=1}^N$ that are syntactically similar but semantically incorrect, as illustrated in Figure 2. These negatives encourage the reward to capture fine-grained semantic distinctions rather than superficial lexical similarity. Given a generated video x and a reward model $\mathcal{R}(\cdot)$ for a video–text pair, the calibrated reward is defined as

$$\tilde{\mathcal{R}} = \mathcal{R}(x, c^+) - \tau \log \sum_{j=1}^{N} \exp\left(\frac{\mathcal{R}(x, c_j^-) - \mathcal{R}(x, c^+) + m}{\tau}\right), \quad (6)$$

where m is a margin enforcing separation between the target prompt and its negatives, and τ is a temperature parameter controlling the sharpness of the contrastive penalty. Intuitively, the calibrated reward encourages the generated video to achieve a high score for the target prompt while suppressing semantically similar distractors, making the optimization process more sensitive to compositional errors such as attribute binding, spatial relations, and numeracy.

Finally, we optimize both the initial latent noise $\mathbf{z}_T$ and the stochastic perturbations $\{\boldsymbol{\varepsilon}_t\}_{t=1}^T$ through gradient ascent on a weighted combination of the calibrated rewards. For brevity, we denote all optimization variables by $\boldsymbol{\epsilon} = \{\mathbf{z}_T, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_T\}$, leading to the update rule

$$\boldsymbol{\epsilon}^{k+1} = \boldsymbol{\epsilon}^k + \eta \cdot \nabla_{\boldsymbol{\epsilon}^k}\left[\sum_i \alpha_i \tilde{\mathcal{R}}_i(\mathcal{G}_\theta(\boldsymbol{\epsilon}^k, c), c)\right], \quad (7)$$

where η denotes the learning rate and α_i are the weights of individual reward components. Together, the proposed reward formulation and joint stochastic optimization enable stable and effective test-time refinement.

Table 1: Quantitative evaluation on VBench (%). The best results among open-source models are in **bold** and the second best is underlined. * denotes our reproduction.

Models	Overall Consist.	Multiple Objects	Color	Spatial Relation.	Motion Smooth.	Dynamic Degree
(Closed-source)						
Gen-3 [42]	26.69	53.64	80.90	65.09	99.23	60.14
Sora [3]	26.26	70.85	80.11	74.29	98.74	79.91
Veo 3 [16]	27.88	82.20	82.48	84.26	99.16	72.43
(Open-source)						
ModelScope [53]	25.67	38.98	81.72	33.68	95.79	66.39
VideoCrafter2 [5]	28.23	40.66	92.92	35.86	97.73	42.50
CogVideoX-2B [63]	27.33	66.59	83.01	**74.27**	97.51	64.79
Vchitect-2.0 [14]	27.57	68.84	87.04	57.55	**98.98**	63.89
AnimateLCM* [52]	25.74	30.78	81.84	39.94	98.19	29.17
+ **NoisEasier (Ours)**	28.79 ↑3.05	44.04 ↑13.26	88.22 ↑6.38	46.70 ↑6.76	98.57 ↑0.38	30.56 ↑1.39
T2V-Turbo (MS) [26]	27.51	58.63	89.67	45.74	95.64	66.39
T2V-Turbo (MS)*	27.40	54.50	89.90	46.62	95.51	70.83
+ **NoisEasier (Ours)**	**32.05** ↑4.65	76.62 ↑22.12	**94.38** ↑4.48	48.93 ↑2.31	95.29 ↓0.22	**71.39** ↑0.56
T2V-Turbo (VC2) [26]	28.16	54.65	89.90	38.67	97.34	49.17
T2V-Turbo (VC2)*	28.12	54.33	90.59	39.95	96.29	51.67
+ **NoisEasier (Ours)**	31.95 ↑3.83	**79.18** ↑24.85	93.63 ↑3.04	52.61 ↑12.66	96.10 ↓0.19	59.17 ↑7.50

4 Experiment

Experimental Setup. We evaluate NoisEasier on two VCM backbones: AnimateLCM [52] and T2V-Turbo [26]. For T2V-Turbo, we consider two distilled variants: T2V-Turbo (MS) and T2V-Turbo (VC2), which are distilled from ModelScope [53] and VideoCrafter2 [5], respectively. T2V-Turbo (MS) generates videos at 256×256 resolution, while T2V-Turbo (VC2) produces 512×320 videos. AnimateLCM generates videos at 512×512 resolution. All three backbones support fast sampling with 4–8 denoising steps. To balance generation quality and efficiency, we adopt a 4-step sampler for fast inference. For noise refinement, we optimize the latent variables for 25 iterations using AdamW with gradient clipping and a learning rate of 0.01. Unless otherwise specified, all experiments use the same optimization configuration across backbones. To make noise optimization feasible, we employ gradient checkpointing [7] together with mixed-precision inference to reduce memory overhead. All the experiments are conducted on two RTX 6000 Ada GPUs. More implementation details can be found in the supplementary material.

Benchmarks and Metrics. We evaluate NoisEasier on two widely used benchmarks: VBench [20] and T2V-CompBench [44]. VBench provides a comprehensive evaluation across 16 dimensions covering visual quality and semantic alignment. Following common practice, we report results on four semantic dimensions (*overall consistency*, *multiple objects*, *color*, and *spatial relations*) and two motion-related dimensions, i.e., *motion smoothness* and *dynamic degree*. T2V-

Table 2: Quantitative evaluation on T2V-CompBench. The best results among open-source models are in **bold**, and the second best are underlined. * denotes reproduction.

Models	Consist Attr.	Dynamic Attr.	Spatial	Motion	Action	Interaction	Numeracy
(Closed-source)							
Gen-3 [42]	0.5980	0.0687	0.5194	0.2754	0.5233	0.5906	0.2306
Dreamina [4]	0.6913	0.0051	0.5773	0.2361	0.5924	0.6824	0.4380
PixVerse [36]	0.7060	0.0624	0.5979	0.2867	0.8722	0.8309	0.6066
Kling [24]	0.6931	0.0098	0.5690	0.2562	0.5787	0.7128	0.4413
(Open-source)							
ModelScope [53]	0.5148	0.0161	0.4118	0.2408	0.3639	0.4613	0.1986
Show-1 [66]	0.5670	0.0115	0.4544	0.2291	0.3881	0.6244	0.3086
VideoTetris [47]	0.6211	0.0104	0.4832	0.2249	0.4939	0.6578	0.3467
VideoCrafter2 [5]	0.6182	0.0103	0.4838	0.2259	0.5030	0.6365	0.3330
Open-Sora 1.2 [67]	0.5639	0.0189	0.5063	0.2468	0.4833	0.5039	0.3719
CogVideoX-5B [63]	0.6164	**0.0219**	0.5172	**0.2658**	0.5333	0.6069	0.3706
AnimateLCM* [52]	0.6746	0.0096	0.4651	0.2208	0.3425	0.4504	0.2494
+ **NoisEasier (Ours)**	0.7807 ↑0.1061	0.0183 ↑0.0087	0.5113 ↑0.0462	0.2281 ↑0.0073	0.4362 ↑0.0937	0.5365 ↑0.0861	0.2567 ↑0.0073
T2V-Turbo (MS)* [26]	0.7313	0.0124	0.4620	0.2306	0.4923	0.5822	0.3030
+ **NoisEasier (Ours)**	0.8375 ↑0.1062	0.0147 ↑0.0023	0.5259 ↑0.0639	0.2502 ↑0.0196	0.6477 ↑0.1554	0.6743 ↑0.0921	**0.3983** ↑0.0953
T2V-Turbo (VC2)* [26]	0.7852	0.0113	0.5079	0.2222	0.6241	0.7017	0.2900
+ **NoisEasier (Ours)**	**0.8737** ↑0.0885	0.0097 ↓0.0016	**0.5660** ↑0.0581	0.2343 ↑0.0121	**0.7524** ↑0.1283	**0.7915** ↑0.0898	0.3753 ↑0.0853

CompBench focuses on compositional reasoning in video generation and evaluates seven categories, including *consistent/dynamic attribute binding*, *spatial relationships*, *action/motion binding object interactions*, and *numeracy*.

4.1 Main Results

We first report quantitative results on VBench, comparing against both closed-source models and open-source baselines. As shown in Table 1, NoisEasier consistently improves VCM-based generators across different backbones. On AnimateLCM, NoisEasier increases *Overall Consistency* from 25.74% to 28.79%, with notable gains in compositional dimensions such as *Multiple Objects* and *Spatial Relation*. Similar trends are observed on T2V-Turbo, where the improvements are even more pronounced. For example, *Multiple Objects* increases from 54.50% to 76.62% on T2V-Turbo (MS) and from 54.33% to 79.18% on T2V-Turbo (VC2), while *Spatial Relation* improves by 12.66% on the VC2 variant. Although *Motion Smoothness* decreases slightly for the T2V-Turbo variants, the consistent increase in *Dynamic Degree* suggests that NoisEasier produces richer motion dynamics while largely preserving temporal coherence.

NoisEasier also yields consistent gains on T2V-CompBench across diverse compositional dimensions, as shown in Table 2. On AnimateLCM, it substantially improves *Attribute/Action Binding* while also boosting *Spatial relationship* and *Interaction*. Similar trends hold for T2V-Turbo, where *Consist. Attribute Binding* increases to 0.8375 (MS) and 0.8737 (VC2), and *Action binding* improves notably on both variants. Importantly, Numeracy also sees sizable gains (e.g., 0.0953 on MS), indicating improved robustness in reasoning about discrete

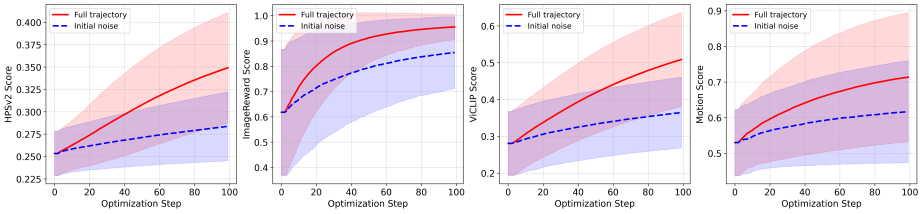


Fig. 3: Reward curve during noise optimization. The red solid line represents full noise optimization, while the blue dotted line denotes initial noise optimization. All scores are scale-normalized.

object counts. These results further support the effectiveness of our negative-aware reward calibration in providing more precise compositional feedback.

4.2 Ablation Study

Contribution of each reward model. As shown in Table 3, different reward models emphasize different aspects of generation. ImageReward and ViCLIP contribute most to semantic alignment, but optimizing semantic rewards alone tends to yield overly static videos, reflected by reduced *Dynamic Degree*. In contrast, the motion reward alone increases temporal dynamics and motion coherence, yet degrades semantic metrics, highlighting the need to balance complementary objectives rather than relying on a single signal. Figure 4a further explains the dynamic term in our motion reward: by subtracting *global motion* in Eq. 5, the resulting *localized motion* suppresses camera shifts while emphasizing subject movement. Combining all rewards yields complementary gains across semantic and motion dimensions, and negative-aware calibration further strengthens compositional alignment, with only a slight decrease in motion-related scores since it primarily reshapes VLM-based objectives.

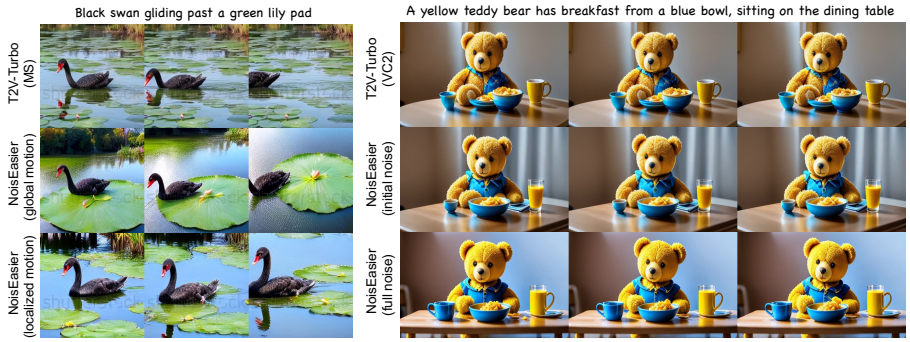
Table 3: The effect of reward models for NoisEasier on VBench. The best score among individual reward models is marked in green, while the second-best is in yellow.

Models	Overall Consist.	Multiple Objects	Color	Spatial Relation.	Motion Smooth.	Dynamic Degree
T2V-Turbo (MS)*	27.40	54.50	89.90	46.62	95.51	70.83
+ HPSv2	27.85	64.66	90.20	46.21	95.19	54.72
+ ImageReward	28.24	73.78	90.27	47.16	95.36	61.11
+ ViCLIP	36.21	68.05	91.19	41.92	94.98	63.89
+ Motion	26.05	47.68	87.32	42.07	95.59	85.56
All w/o negatives	31.74	75.82	93.29	47.95	95.38	73.33
All (NoisEasier)	32.05	76.62	94.38	48.93	95.29	71.39

Initial noise *vs.* Full trajectory optimization. To quantify the benefit of full noise optimization, we compare against a variant that updates only the initial latent noise. As shown in Table 4, full-trajectory optimization provides additional control beyond initial-noise-only refinement, with the largest gains on compositional metrics such as color and spatial relations. Interestingly, motion-related dimensions show smaller gains or

Table 4: Initial *vs.* full noise optimization.

Models	Overall Consist.	Multiple Objects	Color	Spatial Relation.	Motion Smooth.	Dynamic Degree
AnimateLCM+Init.	26.76	43.28	85.61	44.52	98.36	30.83
AnimateLCM+Full.	28.79	44.04	88.22	46.70	98.57	30.56
T2V-Turbo (MS)+Init.	29.67	72.58	90.03	44.10	95.58	66.94
T2V-Turbo (MS)+Full.	32.05	76.62	94.38	48.93	95.29	71.39
T2V-Turbo (VC2)+Init.	29.97	75.41	92.98	45.15	96.16	60.83
T2V-Turbo (VC2)+Full.	31.95	79.18	93.63	52.61	96.10	59.17



(a) Global *vs.* localized motion. **(b)** Initial noise optimization *vs.* Full noise optimization.

Fig. 4: The effect of (a) motion reward; and (b) noise optimization formulation.

occasional advantages for initial noise optimization, suggesting lower sensitivity to intermediate-noise refinement. Figure 3 further demonstrates that jointly optimizing the full noise trajectory accelerates reward convergence. We also provide a qualitative example in Figure 4b: the initial noise mainly controls global appearance and coarse layout, whereas intermediate perturbations refine finer structures and stylistic details, resulting in sharper contrast and more visually pleasing outputs. Overall, these results validate the effectiveness of full trajectory optimization for improving both quantitative performance and visual quality.

Table 5: Effect of different negative prompt construction.

Models	Overall Consist.	Multiple Objects	Color	Spatial Relation.	Motion Smooth.	Dynamic Degree
NoisEaiser w/o neg.	31.74	75.82	93.29	47.95	95.38	73.33
Random sample	31.83	75.41	93.14	46.39	95.46	71.11
Gemini-2.5 Flash	32.00	76.55	93.81	48.72	95.53	74.44
GPT-4o (Ours)	32.05	76.62	94.38	48.93	95.29	71.39

Comparison of different negative prompts. We compare LLM-generated negatives against randomly sampled negatives based on part-of-speech (POS) perturbations. Specifically, we use spaCy[1] to identify key primitives (*verbs*, *nouns*, *adjectives*, *adpositions*, and *numerals*) and randomly replace them with alternatives drawn from the test prompt suite. As shown in Table 5, incorporating negative prompts consistently improves performance over optimizing with the positive prompt alone. Overall Consistency increases slightly across all strategies, with the strongest gains achieved by LLM-constructed negatives. Specifically, GPT-4o negatives yield the best *Color* and *Spatial Relation* scores, suggesting that semantically targeted negatives provide more informative contrast than random perturbations. By contrast, POS-based random negatives offer little improvement and even reduce *Spatial Relation*, indicating that arbitrary replacements are insufficient for effective calibration. Motion Smoothness remains essentially unchanged across variants, while Dynamic Degree changes only modestly, consistent with our design that negative-aware calibration primarily sharpens semantic and attribute alignment without degrading motion.

[1] `https://spacy.io/`

Comparison of test-time scaling methods. We further compare NoisEasier with two canonical test-time scaling (TTS) methods: Best-of-n sampling (BoS) and prompt enhancement (PE). Under the same wall-clock time budget, BoS samples multiple seeds and selects the highest-reward video, while PE uses an LLM to iteratively rewrite the prompt (with fixed noise) and chooses the best result. As shown in Table 6, NoisEasier achieves the highest *Overall Consistency* and *Color* scores, indicating that gradient-based noise optimization is effective for semantic alignment and attribute binding. In contrast, BoS attains the best *Dynamic Degree* and *Multiple Objects*, suggesting that strong motion patterns and complex multi-object layouts are largely driven by the global structure induced by the initial noise. PE slightly improves *Motion Smoothness* but substantially reduces *Color* binding, implying that enriched prompts may distract the model from precise attribute alignment while stabilizing cross-frame predictions. Interestingly, Gemini-based PE achieves the best *Spatial Relation* score, suggesting that some LLMs rewrite prompts with clearer spatial expressions. Overall, different TTS strategies favor different aspects of quality, and NoisEasier offers a complementary path for improving semantic and attribute-level alignment.

Table 6: Performance of different test-time scaling methods.

Models	Overall Consist.	Multiple Objects	Color	Spatial Relation.	Motion Smooth.	Dynamic Degree
T2V-Turbo (MS)	27.40	54.50	89.90	46.62	95.51	70.83
+ NoisEasier	32.05	76.62	94.38	48.93	95.29	71.39
+ Best-of-n sampling	29.39	80.06	91.43	48.54	95.52	79.44
+ PE (GPT-4o)	28.70	71.95	87.47	47.58	95.89	70.56
+ PE (Gemini-2.5 Flash)	28.54	72.13	86.54	51.13	95.80	70.83

Comparison with reward fine-tuning. Since NoisEasier optimizes rewards at inference time, we also examine whether incorporating the same rewards directly into the model parameters yields comparable benefits. To this end, we perform reward fine-tuning (RFT) using LoRA on the T2V-Turbo backbone under the same mixed reward objectives. We follow the default LoRA configuration of T2V-Turbo and fine-tune on the full VBench prompt suite (∼1k prompts) for 10k steps, after which further training begins to exhibit overfitting. As shown in Table 7, both approaches improve over the baseline but with different strengths. RFT mainly improves motion-related metrics such as *Motion Smoothness* and *Dynamic Degree*, whereas NoisEasier yields larger gains on compositional dimensions including *Multiple Objects* and *Color*. Notably, applying NoisEasier on top of the RFT model further improves most metrics, indicating that reward signals are not fully amortized into the model parameters during offline tuning. This suggests that noise optimization provides a complementary *test-time scaling* mechanism even when the backbone has already been reward-tuned.

Table 7: Comparison with reward fine-tuning.

Models	Overall Consist.	Multiple Objects	Color	Spatial Relation.	Motion Smooth.	Dynamic Degree
T2V-Turbo (MS)	27.40	54.50	89.90	46.62	95.51	70.83
+ reward fine-tuning (RFT)	28.58	73.49	90.95	51.16	98.65	74.72
+ NoisEasier	32.05	76.62	94.38	48.93	95.29	71.39
+ RFT + NoisEasier	31.91	84.56	94.72	54.71	98.65	77.50

4.3 Efficiency Analysis

Inference Efficiency. Table 8 reports the peak memory and runtime of 25-step optimization under identical settings on a single RTX 6000 Ada GPU. Compared with optimizing only the initial latent, full trajectory optimization incurs *negligible* additional overhead, increasing runtime by less than 1% while leaving

Table 8: Inference Efficiency of 25-step test-time noise optimization.

Model	Peak VRAM	Time	Model	Peak VRAM	Time	Model	Peak VRAM	Time
MS	8.23 GB	1.04s	VC2	8.76 GB	1.41s	Ani-LCM	8.78 GB	2.46s
+Init.	22.93 GB	44.63s	+Init.	33.24 GB	115.19s	+Init.	40.53 GB	176.71s
+Full.	22.95 GB	45.28s	+Full.	33.27 GB	115.63s	+Full.	40.53 GB	177.86s

peak memory virtually unchanged. This is because both methods backpropagate through the same denoising computation graph; our method only introduces additional optimized noise variables without requiring extra forward or backward passes. The dominant computational cost instead comes from multi-step backpropagation through the generator and differentiable reward models. On a single H100 GPU, the runtime can be further reduced to 31.25 s (MS), 72.92 s (VC2), and 105.37 s (AnimateLCM).

Quality–Cost Trade-off. Figure 5 compares the quality–cost trade-off between initial-noise and full-trajectory optimization on the T2V-Turbo (MS) backbone using an RTX 6000 Ada GPU, with the *Overall Consistency* dimension of VBench as the evaluation metric. As the optimization budget increases, both methods improve semantic alignment, but full-trajectory optimization consistently traces a superior Pareto frontier. Notably, the advantage becomes more pronounced with additional optimization steps, indicating that jointly refining all intermediate perturbations utilizes the additional computation more effectively than optimizing only the initial latent. We adopt 25 optimization steps as the default setting, providing a favorable balance between quality and inference cost.

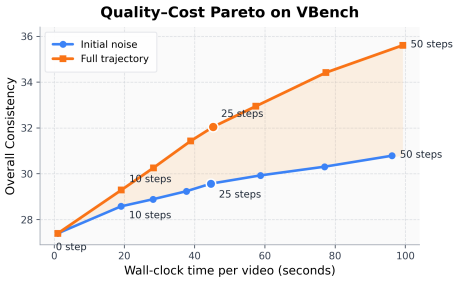


Fig. 5: Quality-Cost Pareto on VBench.

4.4 Human Evaluation.

We conduct a user study to further assess the practical benefits of NoisEasier. We uniformly sample 10 prompts from each of the seven T2V-CompBench dimensions, resulting in 70 test prompts in total. We compare NoisEasier against three baselines: T2V-Turbo (VC2), CogVideoX-2B [63], and Wan2.1-1.3B [51]. For each prompt, we use the same random seed across models. Each comparison is judged by five Amazon Mechanical Turk workers, who select the candidate with better semantic alignment and

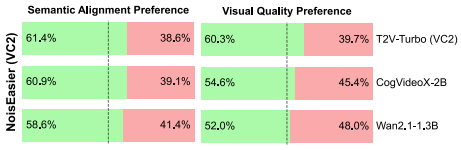


Fig. 6: User study on T2V-CompBench.

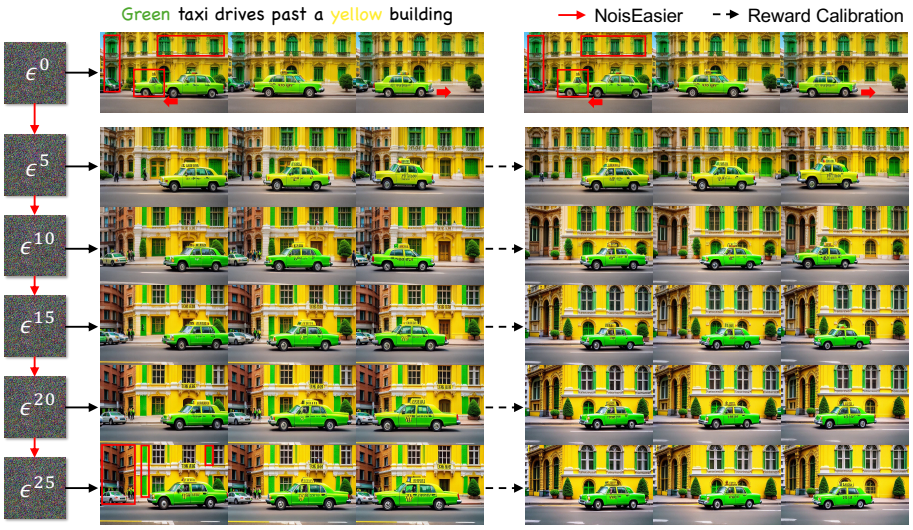


Fig. 7: Effect of negative-aware reward calibration at different noise optimization steps.

visual quality. As shown in Figure 6, NoisEasier achieves a 61.4% win rate over its backbone, indicating that test-time noise optimization improves semantic faithfulness. Moreover, NoisEasier remains competitive against stronger T2V baselines, receiving around 61% preference on alignment and 52% on visual quality compared to CogVideoX-2B and Wan2.1-1.3B, despite operating as a lightweight add-on rather than retraining a larger T2V model.

4.5 Qualitative Analysis

We present qualitative comparisons in Figure 7 to illustrate the effect of our optimization framework. The baseline captures the rough semantics of the prompt but exhibits compositional errors such as incorrect color binding and inconsistent object motion. Noise optimization progressively improves both spatial and temporal coherence: the taxi becomes a coherent object with stable right-to-left motion, and attribute separation becomes clearer, although minor artifacts remain. With negative-aware reward calibration, semantic disentanglement is further enhanced, i.e., the facade becomes consistently yellow while the green taxi is preserved with minimal leakage. Additional examples are provided in Figure 1 and the supplementary material.

While effective, we also observe two common failure modes, as shown in Figure 8. First, although NoisEasier improves visual fidelity and semantic alignment, it still struggles with *temporal state transitions*. In the melting-ice example, the generated video preserves the appearance of the ice cube but fails to capture its gradual transition into water. We attribute this limitation to existing reward models, which primarily assess frame-level semantics and provide only weak supervision for long-term temporal evolution. Second, NoisEasier remains

Fig. 8: Two failure cases of NoisEasier: temporal state transition (left) and object counting (right).

challenged by *fine-grained object counting*. In the example on the right, the model fails to satisfy the exact numeracy constraint despite improving overall semantic consistency. This suggests that current VLM-based rewards are insufficiently sensitive to precise object counting and complex compositional reasoning.

These observations indicate that the performance of test-time optimization is fundamentally bounded by the quality of the underlying reward models. Future work may benefit from incorporating structured reward signals, such as object detection, visual grounding, or temporal-state verification models, to provide more reliable supervision for compositional video generation.

5 Conclusion

In this work, we propose NoisEasier, a test-time scaling framework that improves text-to-video generation through differentiable reward-guided noise optimization without model fine-tuning. By combining efficient optimization with robust reward supervision, NoisEasier consistently enhances compositional alignment while preserving visual fidelity. We further show that jointly optimizing intermediate noises provides more effective control over the denoising trajectory than optimizing only the initial latent, yielding faster reward convergence and stronger compositional alignment with negligible additional computational cost. Moreover, NoisEasier complements offline reward fine-tuning, demonstrating that test-time optimization remains effective even after model alignment. Overall, our findings establish noise optimization as a practical test-time scaling paradigm for improving alignment in video generative models.

Acknowledgements

This work was partially supported by the Office of Naval Research (ONR) grant (N00014-23-1-2417), and the Army Research Office (ARO) grant (W911NF-24-1-0385). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ONR or ARO.

References

1. Ben-Hamu, H., Puny, O., Gat, I., Karrer, B., Singer, U., Lipman, Y.: D-flow: Differentiating through flows for controlled generation. arXiv preprint arXiv:2402.14017 (2024)
2. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
3. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators. Tech. rep., OpenAI (2024), `https://openai.com/research/video-generation-models-as-world-simulators`, technical Report, OpenAI
4. Capcut: Dreamina. `https://dreamina.capcut.com/ai-tool/home` (2024)
5. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7310–7320 (2024)
6. Chen, S.X., Vaxman, Y., Ben Baruch, E., Asulin, D., Moreshet, A., Lien, K.C., Sra, M., Sen, P.: Tino-edit: Timestep and noise optimization for robust diffusion-based image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6337–6346 (2024)
7. Chen, T., Xu, B., Zhang, C., Guestrin, C.: Training deep nets with sublinear memory cost. arXiv preprint arXiv:1604.06174 (2016)
8. Cheng, H., Dong, Q., Peng, L., Sha, Z., Feng, W., Xie, J., Song, Z., Wen, S., He, X., Wu, B.: Discriminator-free direct preference optimization for video diffusion. arXiv preprint arXiv:2504.08542 (2025)
9. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. Advances in neural information processing systems **30** (2017)
10. Clark, K., Vicol, P., Swersky, K., Fleet, D.J.: Directly fine-tuning diffusion models on differentiable rewards. arXiv preprint arXiv:2309.17400 (2023)
11. Ding, X., Zhang, K., Han, J., Hong, L., Xu, H., Li, X.: Pami-vdpo: Mitigating video hallucinations by prompt-aware multi-instance video preference learning. arXiv preprint arXiv:2504.05810 (2025)
12. Doveh, S., Arbelle, A., Harary, S., Herzig, R., Kim, D., Cascante-Bonilla, P., Alfassy, A., Panda, R., Giryes, R., Feris, R., et al.: Dense and aligned captions (dac) promote compositional reasoning in vl models. Advances in Neural Information Processing Systems **36**, 76137–76150 (2023)
13. Eyring, L., Karthik, S., Roth, K., Dosovitskiy, A., Akata, Z.: Reno: Enhancing one-step text-to-image models through reward-based noise optimization. Advances in Neural Information Processing Systems **37**, 125487–125519 (2024)
14. Fan, W., Si, C., Song, J., Yang, Z., He, Y., Zhuo, L., Huang, Z., Dong, Z., He, J., Pan, D., et al.: Vchitect-2.0: Parallel transformer for scaling up video diffusion models. arXiv preprint arXiv:2501.08453 (2025)
15. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. Advances in Neural Information Processing Systems **36**, 79858–79885 (2023)
16. Google: Veo-3 technical report. Tech. rep., Google DeepMind (May 2025)

17. Guo, X., Liu, J., Cui, M., Li, J., Yang, H., Huang, D.: Initno: Boosting text-to-image diffusion models via initial noise optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9380–9389 (2024)
18. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
19. He, X., Jiang, D., Zhang, G., Ku, M., Soni, A., Siu, S., Chen, H., Chandra, A., Jiang, Z., Arulraj, A., et al.: Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. arXiv preprint arXiv:2406.15252 (2024)
20. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
21. Karunratanakul, K., Preechakul, K., Aksan, E., Beeler, T., Suwajanakorn, S., Tang, S.: Optimizing diffusion noise can serve as universal motion priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1334–1345 (2024)
22. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 36652–36663 (2023)
23. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
24. Kuaishou: Kling. `https://kling.kuaishou.com/` (2024)
25. Li, J., Feng, W., Chen, W., Wang, W.Y.: Reward guided latent consistency distillation. arXiv preprint arXiv:2403.11027 (2024)
26. Li, J., Feng, W., Fu, T.J., Wang, X., Basu, S., Chen, W., Wang, W.Y.: T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. arXiv preprint arXiv:2405.18750 (2024)
27. Li, J., Long, Q., Zheng, J., Gao, X., Piramuthu, R., Chen, W., Wang, W.Y.: T2v-turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design. arXiv preprint arXiv:2410.05677 (2024)
28. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
29. Li, S., Le, H., Xu, J., Salzmann, M.: Enhancing compositional text-to-image generation with reliable random seeds. In: The Thirteenth International Conference on Learning Representations (2024)
30. Liang, Y., He, J., Li, G., Li, P., Klimovskiy, A., Carolan, N., Sun, J., Pont-Tuset, J., Young, S., Yang, F., et al.: Rich human feedback for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19401–19411 (2024)
31. Liu, J., Liu, G., Liang, J., Yuan, Z., Liu, X., Zheng, M., Wu, X., Wang, Q., Qin, W., Xia, M., et al.: Improving video generation with human feedback. arXiv preprint arXiv:2501.13918 (2025)
32. Liu, R., Wu, H., Ziqiang, Z., Wei, C., He, Y., Pi, R., Chen, Q.: Videodpo: Omni-preference alignment for video diffusion generation. arXiv preprint arXiv:2412.14167 (2024)

33. Novack, Z., McAuley, J., Berg-Kirkpatrick, T., Bryan, N.: Ditto-2: Distilled diffusion inference-time t-optimization for music generation. arXiv preprint arXiv:2405.20289 (2024)
34. Novack, Z., McAuley, J., Berg-Kirkpatrick, T., Bryan, N.J.: Ditto: Diffusion inference-time t-optimization for music generation. arXiv preprint arXiv:2401.12179 (2024)
35. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
36. PixVerse: Pixverse. `https://app.pixverse.ai` (2024)
37. Prabhudesai, M., Goyal, A., Pathak, D., Fragkiadaki, K.: Aligning text-to-image diffusion models with reward backpropagation (2023)
38. Prabhudesai, M., Mendonca, R., Qin, Z., Fragkiadaki, K., Pathak, D.: Video diffusion alignment via reward gradients. arXiv preprint arXiv:2407.08737 (2024)
39. Qi, Z., Bai, L., Xiong, H., Xie, Z.: Not all noises are created equally: Diffusion noise selection and optimization. arXiv preprint arXiv:2407.14041 (2024)
40. Qing, Z., Zhang, S., Wang, J., Wang, X., Wei, Y., Zhang, Y., Gao, C., Sang, N.: Hierarchical spatio-temporal decoupling for text-to-video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6635–6645 (2024)
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
42. Runway: Introducing gen-3 alpha: A new frontier for video generation. `https://runwayml.com/blog/introducing-gen-3-alpha/` (2024)
43. Sun, K., Huang, K., Liu, X., Wu, Y., Xu, Z., Li, Z., Liu, X.: T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. arXiv preprint arXiv:2407.14505 (2024)
44. Sun, K., Huang, K., Liu, X., Wu, Y., Xu, Z., Li, Z., Liu, X.: T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8406–8416 (2025)
45. Tang, Z., Peng, J., Tang, J., Hong, M., Wang, F., Chang, T.H.: Inference-time alignment of diffusion models with direct noise optimization. arXiv preprint arXiv:2405.18881 (2024)
46. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 402–419. Springer (2020)
47. Tian, Y., Yang, L., Yang, H., Gao, Y., Deng, Y., Wang, X., Yu, Z., Tao, X., Wan, P., ZHANG, D., et al.: Videotetris: Towards compositional text-to-video generation. Advances in Neural Information Processing Systems **37**, 29489–29513 (2024)
48. Trager, M., Perera, P., Zancato, L., Achille, A., Bhatia, P., Soatto, S.: Linear spaces of meanings: compositional structures in vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15395–15404 (2023)
49. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2024)

50. Wallace, B., Gokul, A., Ermon, S., Naik, N.: End-to-end diffusion latent optimization improves classifier guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7280–7290 (2023)
51. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
52. Wang, F.Y., Huang, Z., Bian, W., Shi, X., Sun, K., Song, G., Liu, Y., Li, H.: Animatelcm: Computation-efficient personalized style video generation without personalized video data. In: SIGGRAPH Asia 2024 Technical Communications, pp. 1–5 (2024)
53. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
54. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. International Journal of Computer Vision **133**(5), 3059–3078 (2025)
55. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv preprint arXiv:2307.06942 (2023)
56. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
57. Wu, X., Sun, K., Zhu, F., Zhao, R., Li, H.: Human preference score: Better aligning text-to-image models with human preference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2096–2105 (2023)
58. Wu, X., Huang, S., Wang, G., Xiong, J., Wei, F.: Boosting text-to-video generative model with mllms feedback. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
59. Xu, J., Huang, Y., Cheng, J., Yang, Y., Xu, J., Wang, Y., Duan, W., Yang, S., Jin, Q., Li, S., et al.: Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. arXiv preprint arXiv:2412.21059 (2024)
60. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 15903–15935 (2023)
61. Xu, K., Zhang, L., Shi, J.: Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3024–3034. IEEE (2025)
62. Yang, K., Tao, J., Lyu, J., Ge, C., Chen, J., Shen, W., Zhu, X., Li, X.: Using human feedback to fine-tune diffusion models without any reward model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8941–8951 (2024)
63. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
64. Yuan, H., Zhang, S., Wang, X., Wei, Y., Feng, T., Pan, Y., Zhang, Y., Liu, Z., Albanie, S., Ni, D.: Instructvideo: Instructing video diffusion models with human feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6463–6474 (2024)
65. Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., Zou, J.: When and why vision-language models behave like bags-of-words, and what to do about it? arXiv preprint arXiv:2210.01936 (2022)

66. Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation. International Journal of Computer Vision pp. 1–15 (2024)
67. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)
68. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)
69. Zhou, Z., Shao, S., Bai, L., Zhang, S., Xu, Z., Han, B., Xie, Z.: Golden noise for diffusion models: A learning framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17688–17697 (2025)