Verifier Matters: Enhancing Inference-Time Scaling for Video Diffusion Models

Lorenzo Baraldi[*1]
lorenzo.baraldi@phd.unipi.it
Davide Bucciarelli[1,2]
davide.bucciarelli@unimore.it
Zifan Zeng[3,4]
zifan.zeng@huawei.com
Chongzhe Zhang[3,5]
chongzhe.zhang@huawei.com
Qunli Zhang[3]
zhangqunli1@huawei.com
Marcella Cornia[2]
marcella.cornia@unimore.it
Lorenzo Baraldi[2]
lorenzo.baraldi@unimore.it
Feng Liu[3]
feng.liu1@huawei.com
Zheng Hu[3]
hu.zheng@huawei.com
Rita Cucchiara[2]
rita.cucchiara@unimore.it

[1] University of Pisa
Pisa, Italy
[2] University of Modena and Reggio Emilia
Modena, Italy
[3] Huawei Technologies Duesseldorf GmbH, RAMS Lab
Munich, Germany
[4] Technical University of Munich
Munich, Germany
[5] Technical University of Berlin
Berlin, Germany

Abstract

Inference-time scaling has recently gained attention as an effective strategy for improving the performance of generative models without requiring additional training. Although this paradigm has been successfully applied in text and image generation tasks, its extension to video diffusion models remains relatively underexplored. Indeed, video generation presents unique challenges due to its spatiotemporal complexity, particularly in evaluating intermediate generated samples, a procedure that is required by inference-time scaling algorithms. In this work, we systematically investigate the role of the verifier: the scoring mechanism used to guide sampling. We show that current verifiers, when applied at early diffusion steps, face significant reliability challenges due to noisy samples. We further demonstrate that fine-tuning verifiers on partially denoised samples significantly improves early-stage evaluation and leads to gains in generation quality across multiple inference-time scaling algorithms, including Greedy Search, Beam Search, and a new Successive Halving baseline, which we adapt for the inference-time scaling setting.

[*] Work done during internship at Huawei Munich Research Center, Germany.

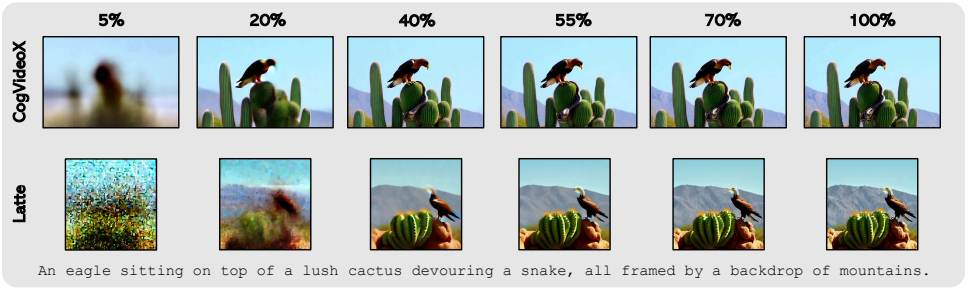

Figure 1: Visualization of intermediate videos during the generation on Latte [23] and CogVideoX [43]. These videos are extracted at intermediate inference timesteps expressed in percentages over a total of 50 steps.

1 Introduction

The introduction of diffusion models [33] has markedly enhanced video synthesis, yielding samples of greater visual fidelity and diversity relative to earlier GAN-based [34, 37] frameworks. Nevertheless, generation failures persist, manifesting as degraded video quality, implausible motion dynamics, or inadequate adherence to the user's prompt that, when taking into account the high computation cost required for sampling, can cause damage to the reliability of the system. While AI-scaling laws [11] posit that increasing dataset size and model capacity may mitigate these shortcomings, different solutions should be investigated. For instance, recent developments in NLP research demonstrate that smaller models can surpass larger counterparts without additional fine-tuning [32], by reallocating computational resources at inference time and employing a verifier (or reward model) to guide generation. This paradigm, called inference-time scaling [32], has been recently adapted to diffusion models [22] in both image [22, 39] and video generation [24]. Indeed, due to the stochasticity of the diffusion process, it is possible to generate multiple intermediate samples (*i.e.*, particles) through the generation trajectory. These particles can then be evaluated by a verifier, typically a pre-trained model designed to assess the quality of generated outputs. This enables the selection of the most promising sample, supporting informed choice among candidates and contributing to the improvement of the overall quality of the final output.

The efficacy of this strategy, however, depends critically on the capacity of a verifier to score partially generated content, illustrated in Figure 1, that lies outside its original training distribution of fully synthesized videos. Further, while for image generation a verifier has only to take into account the visual quality of the image, in video generation the verifier must also consider the temporal coherence of the generated content necessitating a more complex understanding of the underlying dynamics of video data, as well as the ability to assess the quality of motion and temporal alignment between frames.

Despite these shortcomings, the role of verifiers in guiding inference-time scaling algorithms remains underexplored. In particular, there is limited understanding of how verifier scores evolve throughout the progress of the sampling trajectory, how these verifiers can be optimized for partially generated content, and whether improvements in verifier quality directly translate to gains in inference-time performance. To this end, we conduct an empirical study on two state-of-the-art video diffusion models: Latte [23] and CogVideoX [43]. Our findings show that verifier scores on intermediate samples correlate weakly with final-output scores during the initial 30% of the timesteps in Latte and 10% in CogVideoX. Furthermore, fine-tuning the verifier on a small corpus of 1,000 intermediate video samples

markedly improves this alignment. For example, when employing a fine-tuned verifier (*e.g.*, VideoReward [20]), correlations increase by 0.41 in Pearson, 0.39 in Spearman, and 0.27 in Kendall correlation scores when applied on samples at 15% of the generated diffusion trajectory compared to the zero-shot verifier. Furthermore, these improvements transfer to inference-time scaling algorithms. For instance, when applying Beam Search [24] with the fine-tuned verifier on early sampling steps, and using Latte under a fixed compute budget of 1,200 function evaluations, we observe significant gains in VideoReward scores: +1.13 on MSR-VTT [42], and +0.59, +0.25, and +0.31 on the DEVIL [19] Static, Medium, and High prompt sets, respectively. These results underscore the potential of verifier fine-tuning for improving the effectiveness of inference-time scaling on video generation.

2 Related Works

Text-to-Video Generative Models. Following the emergence of diffusion models [6, 33], early work primarily targeted text-to-image generation [27, 28, 30]. More recently, significant efforts have aimed to extend the diffusion framework to video generation. For example, Video LDM [2] modifies the standard U-Net [29] by alternating spatial and temporal attention layers, enabling effective cross-frame information exchange during denoising. Similarly, with the introduction of the Diffusion Transformer (DiT) [25], different approaches have been proposed to extend this architecture for video generation.

In this context, Latte [23] explores several strategies for integrating spatial and temporal attention mechanisms into the DiT framework, enabling generation sequences of up to 16 frames. Building on this, CogVideoX [43] introduces a refined latent representation obtained through joint spatial-temporal compression using a 3D-VAE. This compression allows CogVideoX to compute simultaneously spatial and temporal attention, generating sequences containing up to 49 frames. Further refinements in 3D-VAE compression techniques have been investigated in recent works, including OpenSora [26] and Cosmos [1].

Inference-Time Scaling. Inference-time scaling [32] refers to a set of techniques applied during inference that enhance model performance without requiring additional training. These methods operate by increasing computational resources during inference, with the objective of maximizing performance within a fixed computational budget. Inference-time scaling has been widely applied in the context of text generation [9, 32] and, more recently, to image [16, 22, 31, 39] and video [24] generation, by scaling the number of function evaluations (*i.e.*, the number of denoising diffusion operations). All inference-time scaling methods comprehend two fundamental components: the algorithm and the verifier. The former defines how the budget is allocated, while the latter is the scoring mechanism used for the decision-making (*e.g.*, selecting the best output) throughout the process.

The most naive algorithm is the Best-of-N [35, 39], where the model generates N variants of an output, and then selects the best one given a verifier. Differently, diffusion models have explored particle sampling, maintaining a set of candidate samples (*i.e.*, particles) and iteratively propagating high-reward samples while discarding lower-reward ones. In image generation, CoDe [31] performs iterative Best-of-N steps, generating new particles from top-scoring ones, while SVDD [18] applies greedy search by selecting the best particle at each step. For video generation, a recent work [24] introduces a beam search strategy to navigate the particle space. Although recent efforts have primarily focused on algorithmic design, relatively little attention has been paid to the role of the verifier, particularly its reliability and influence during the early, noisier stages of generation.

Video Generation Quality Assessment. Given the inherent complexity and richer dimensionality of videos compared to static images, video quality assessment methodologies typically partition evaluation into multiple distinct dimensions. Recently, several approaches have leveraged Multimodal Large Language Models (MLLMs) [4] to facilitate comprehensive video evaluation. For instance, VideoReward [20] employs a fine-tuned variant of Qwen2-VL [38], trained to align the score with human preference data. The final score is a sum of three sub-scores: (i) visual quality, which assesses static frame quality; (ii) motion quality, which captures temporal dynamics such as smoothness and consistency; and (iii) text alignment, which evaluates the semantic consistency of the video with the user's prompt. During training, the model processes a couple of videos and is optimized to predict which video is better on each of the three axes, via a Bradley-Terry loss [3]. Conversely, VisionReward [41] employs an MLLM that evaluates generated videos by answering predefined binary (*i.e.*, Yes/No) questions, and subsequently aggregates these responses via a linear layer to produce an overall evaluation score. Differently, VBench [13] dissects video quality into 16 disentangled evaluation dimensions, and assesses each one independently.

While these models are typically employed for evaluating fully generated videos, the primary challenge in the context of inference-time scaling involves accurately assessing noisy video estimation obtained during the diffusion process. In our analysis, we quantitatively investigate the degree of alignment between evaluation scores assigned to partially generated videos and those corresponding to the completed video generations.

3 Background

This section provides the necessary background for the concepts analyzed in this work. We begin by introducing diffusion models and then review recent inference-time scaling strategies applied to video generation.

Diffusion Models. Denoising Diffusion Probabilistic Models (DDPMs) [6, 33] are generative probabilistic frameworks that approximate a target distribution by learning to invert a T-step Gaussian Markov noising process. For computational reason, recent work on image [28] and video [23, 43] generation, adopt latent diffusion: an autoencoder $E : \mathcal{X} \to \mathcal{Z}$ compresses the visual signal to a lower-dimensional latent space [7, 17], resulting in $z = E(x)$ where x is a video in our case. The diffusion process is solved entirely in $\mathcal{Z}$, and a decoder $D : \mathcal{Z} \to \mathcal{X}$ reconstructs the final output [7, 17].

Further, to fasten the generation process, Denoising Diffusion Implicit Models (DDIMs) [34] make the sampling non-Markovian and parametrized by a variance coefficient $\sigma_t \in [0,1]$:

$$z_{t-1} = \sqrt{\alpha_{t-1}}\,\hat{z}_t^0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\;\varepsilon_\theta(z_t,t) + \sigma_t\,\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(\mathbf{0},\mathbf{I}), \tag{1}$$

where $\hat{z}_t^0$ consists of an approximation to the final result at the current timestep t, and it is defined as $\hat{z}_t^0 = \frac{z_t - \sqrt{1-\alpha_t}\,\varepsilon_\theta(z_t,t)}{\sqrt{\alpha_t}}$, α_t is a noise scheduling parameter, and $\varepsilon_\theta(z_t,t)$ is the noise predicted from the diffusion model. Decoding $\hat{z}_t^0$ provides an early estimate of the final video in pixel space $\hat{x}_t^0$. Inference-time scaling verifiers can exploit this proxy to guide the reverse diffusion trajectory toward higher-quality samples.

Inference-Time Scaling on Video Diffusion Models. Inference-time scaling for diffusion models can be viewed as an optimization-oriented search over the space of noise samples [22]. The framework is centered on two essential components: (i) an algorithm that

proposes candidate noise sequences and (ii) a verifier V that assesses the quality of the corresponding generated samples, where $V : \mathcal{X} \to \mathbb{R}$. On the algorithmic side, recent work adopts particle-generation techniques [16, 22, 24, 31, 39] that, given a budget expressed in number of function evaluations (NFEs), perturb the DDIM sampling (cf. Eq. 1) with stochastic noise ε_t, producing a set of alternative diffusion trajectories that the verifier can subsequently rank.

A simple baseline is the Best-of-N algorithm, which starts from N different generation seeds and scores each sample at $t = 0$, resulting in $\{v_0^i\}_{i=1}^N$, subsequently choosing the max-scoring element. A different solution is to embed Beam Search into the sampling loop [24]. Let b denote the number of beams and k the number of noise perturbations explored per beam. At each time step t, the DDIM sampling is executed k times for every beam, using independent noise vectors $\{\boldsymbol{\varepsilon}_t^{(j,i)}\}$ with $j = 1, \dots, b$ and $i = 1, \dots, k$ to produce bk latent proposals $\{\mathbf{z}_{t-1}^{(j,i)}\}$. Each proposal is deterministically rolled out to $t = 0$ via DDIM sampling to obtain an approximate reconstruction $\hat{\mathbf{x}}_{t-1}^{0(j,i)} = D(\hat{\mathbf{z}}_{t-1}^{0(j,i)})$, which is decoded to a video and evaluated by a verifier $V(\hat{\mathbf{x}}_{t-1}^{0(j,i)})$ producing a score $v_{t-1}^{(i,j)}$. The top-b candidates according to the verifier score are retained as beams for the next time step, and the procedure is repeated until $t = 0$. The highest-scoring video at the final step is returned as the output. When $b = 1$, the algorithm always selects at each timestep the best-scoring sample, resulting in a Greedy Search algorithm. However, it is important to note that both Beam and Greedy Search strategies require the generative model to permit stochasticity via noise injection. Consequently, these algorithms are not directly applicable to models lacking such stochastic mechanisms, including flow-based models [21] and diffusion variants such as CogVideoX [43].

Concerning the verifier, in [24] an ensemble of embedding metrics tested in VBench [13] is proposed for evaluating video generation. In detail, it integrates the DINO [5], RAFT [36], LAION aesthetic predictor, MUSIQ [15], and ViCLIP [40] metrics, aggregating their outputs using learned coefficients chosen to maximize Pearson's correlation with a GPT/Gemini-based scorer. In the following, we refer to this verifier as the VBench ensemble.

4 Experimental Analysis

To better understand the role of verifiers in inference-time scaling for video diffusion models, we analyze how these components behave across the generation process. Recent works [24] integrate a verifier at every diffusion timestep, scoring the approximate reconstructions, $\hat{x}_t^0$. However, in the early phases of inference, these reconstructions remain highly corrupted by noise. As a result, the verifier is often required to assess inputs that lie far outside its training distribution, potentially compromising the reliability of its guidance. To systematically study this issue, we pose the following research questions:

- **RQ1:** How consistent are the scores of multimodal verifiers across intermediate steps of the diffusion process?
- **RQ2:** How can we train verifiers to more effectively score intermediate video samples?
- **RQ3:** Does improving the verifier enhance inference-time scaling performance?

4.1 Temporal Coherence of Verifier Scores

To address **RQ1**, we investigate the consistency of verifier scores throughout the reverse diffusion process. Specifically, we study how well intermediate scores, computed at early

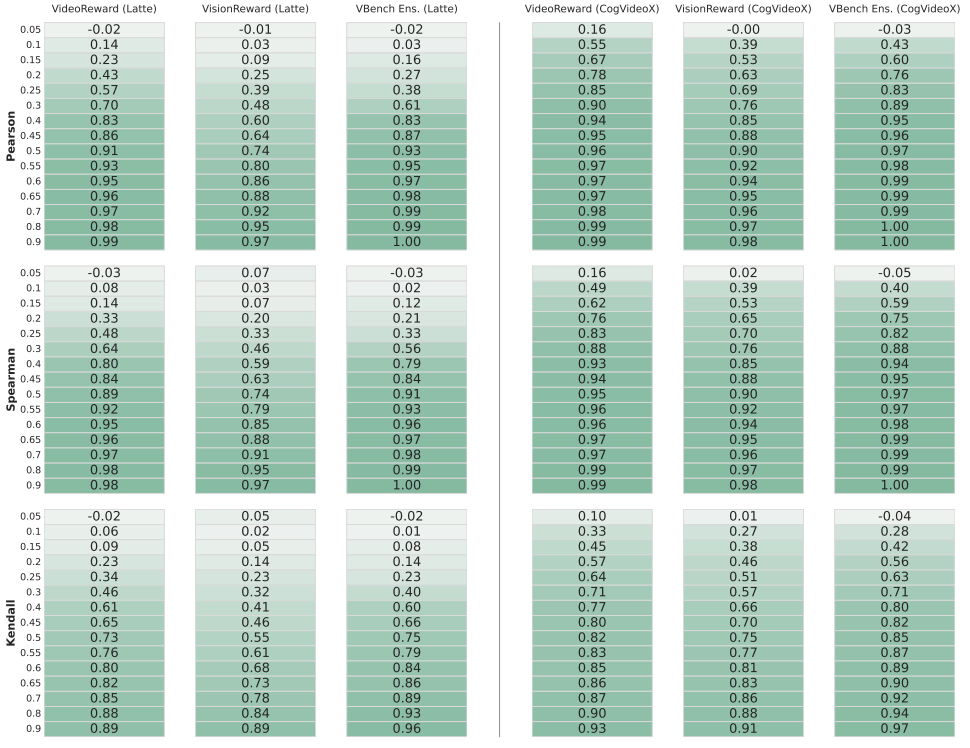

Figure 2: Temporal coherence of the verifiers on Latte (left) and CogVideoX (right). The correlation is measured between the verifier scores at intermediate diffusion steps (indicated as percentages of the total number of steps) and the score on the final output video.

timesteps, correlate with the final score assigned to the generated video. We define this property as the *temporal coherence* of a verifier. Although the raw scores at early diffusion steps are expected to be lower due to high noise levels, our primary concern lies in their association with the final evaluation, not their absolute values. This focus stems from the fact that inference-time scaling algorithms (*e.g.*, Beam Search) operate by identifying local maxima of verifier scores across the sampling trajectory. For such approaches to be effective, verifier outputs must exhibit consistent relative ranking even under partially denoised inputs.

Experimental Setup. We generate intermediate video reconstructions $\hat{x}_t^0$ at selected timesteps $t \in T_{\text{chosen}}$ during inference, and collect the corresponding verifier scores v_t. The degree of alignment between these intermediate scores and the final score v_0 is quantified using three rank and value-based correlation metrics: Pearson's ρ, Spearman's ρ_s, and Kendall's τ. We analyze three scoring models: VideoReward [20], VisionReward [41], and the VBench ensemble [13] following prior work [24]. Generation is performed using two state-of-the-art text-to-video diffusion models: Latte [23] and CogVideoX [43], with inference steps fixed at $T = 50$. Prompts are sampled from the VideoFeedback dataset [10], and we compute scores at ∼5-90% of the diffusion steps[1].

Results. Figure 2 summarizes the results. Across both generators, correlation coefficients rise monotonically with diffusion progress, mirroring the gradual denoising and refinement

[1]Specifically, we compute scores at $T_{\text{chosen}} = \{2, 5, 7, 10, 12, 15, 20, 22, 25, 27, 30, 32, 35, 40, 45\}$.

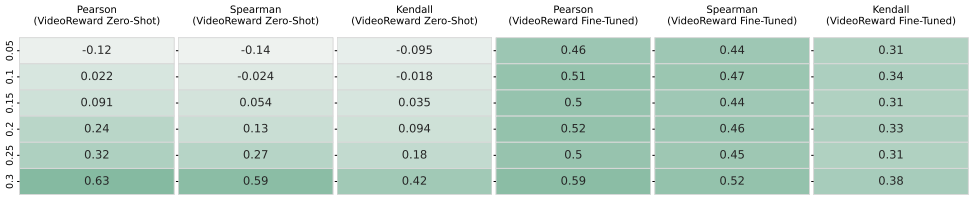

Figure 3: Temporal coherence of VideoReward zero-shot (left) and VideoReward after fine-tuning (right) on Latte.

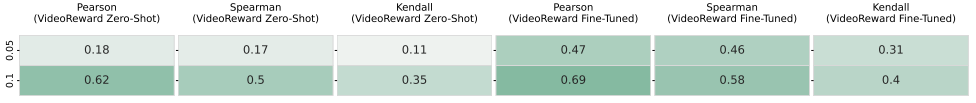

Figure 4: Temporal coherence of VideoReward zero-shot (left) and VideoReward after fine-tuning (right) on CogVideoX.

of the generated sequences. Additionally, for the Latte generator, intermediate scores become strongly predictive of final quality scores beyond the 50% mark of the generation trajectory. For example, VideoReward achieves correlation coefficients of 0.91 (Pearson), 0.89 (Spearman), and 0.73 (Kendall) at 50%, indicating a high degree of consistency with the final output evaluations. Comparable trends are observed with the VBench ensemble and VisionReward metrics, further proving the predictive reliability of intermediate scores in the second half of the generation process. For CogVideoX, predictive correlations emerge even earlier. High correlation values are achieved from $\sim$30% of the generation process, indicating that denoising advances more rapidly than in Latte.

For both generators, VideoReward consistently shows stronger early-stage correlations than the other verifiers. For example, at 20% of Latte's sampling, VideoReward achieves a Pearson's correlation of 0.43, compared to 0.25 and 0.27 for VisionReward and VBench. A similar trend holds for CogVideoX, where VideoReward leads by margins of 0.12 (Pearson) at 10% of the diffusion trajectory, highlighting its effectiveness in the early stages.

Discussion. These results indicate that verifier scores become more informative as reverse diffusion progresses, and that the choice of verifier strongly affects early-stage coherence. While VideoReward offers the strongest early signals, all verifiers show limited reliability within the first 30% of steps for Latte and the first 10% for CogVideoX, potentially limiting their effectiveness in inference-time scaling. Motivated by this limitation, we next examine whether targeted fine-tuning can improve verifier performance in early-stage sampling.

4.2 Improving Verifier Training for Intermediate Representations

To address **RQ2**, we investigate whether verifiers can be trained to more effectively evaluate intermediate video samples. Unlike existing video quality metrics that rely on human annotations, our study aims to enhance the scorer by relying solely on automated supervision. To this end, we treat the predicted scores assigned to generated videos (at final timesteps) as proxy ground-truth labels and perform fine-tuning on intermediate video representations.

Experimental Setup. Given the superior performance of VideoReward, we adopt it as the base model for fine-tuning. In light of the Bradley-Terry loss [3] employed by VideoReward, we construct pairs of videos generated by the same model (*i.e.*, Latte or CogVideoX) using identical textual prompts but different random seeds. Specifically, our experimental setup

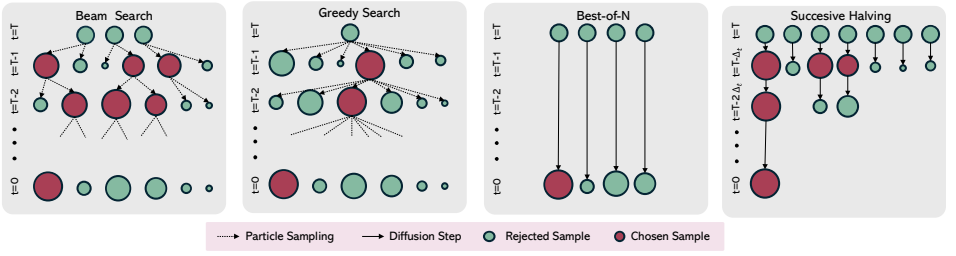

Figure 5: Schematic illustration of inference-time scaling algorithms. Circle size corresponds to the verifier score, with larger circles indicating higher scores.

comprises 1,000 video pairs, each composed of their respective early-stage generations. For Latte, early generations are sampled at $\{5\%, 10\%, 15\%, 20\%, 25\%, 30\%\}$ of the full generation process, whereas for CogVideoX, we consider $\{5\%, 10\%\}$.

For each evaluation dimension defined by VideoReward (*i.e.*, visual quality, motion quality, and text alignment), we automatically select the preferred video in each pair based on predicted scores for the final outputs. The model is then fine-tuned on the corresponding intermediate video pairs, with the goal of aligning the intermediate predictions with the preferences inferred from the final outputs. We adopt LoRA [12] for efficient fine-tuning on this small data collection. After fine-tuning, we calculate the correlations of the zero-shot VideoReward model compared to the fine-tuned one on the validation set (composed of 120 samples). Additional experimental details are provided in the supplementary materials.

Results. Figure 3 and Figure 4 compare the fine-tuned verifier against the original (zero-shot) model, respectively for the Latte and CogVideoX generators. Across both settings, we observe a consistent improvement in correlation coefficients when comparing the fine-tuned verifier to its zero-shot counterpart, proving the effectiveness of the proposed training strategy. For example, on Latte, at 15% of the generation, we obtain a 0.41 increase in Pearson, 0.39 in Spearman, and 0.27 in Kendall with the fine-tuned model, compared to the zero-shot one. Similarly, when comparing results on CogVideoX there is a marked improvement across all correlation metrics. Indeed, at 5% of the generation process we obtain an increase 0.29, 0.29 and 0.20 on respectively Pearson, Spearman, and Kendall correlations.

Discussion. These results confirm that a verifier can be fine-tuned to better capture quality signals in early-stage generations. The resulting improvement in temporal coherence suggests that the verifier can now provide more reliable guidance earlier in the diffusion process. Therefore, we next investigate whether these improvements in intermediate prediction quality translate into performance gains when integrated into inference-time scaling.

4.3 Do Better Verifiers Yield Better Inference-Time Scaling?

To address **RQ3**, we assess whether improvements in intermediate verifier predictions can translate into better performance for inference-time scaling in video generation. Building on prior work [24], we evaluate the impact of fine-tuned verifiers across different inference-time scaling algorithms applied to both the Latte and CogVideoX.

Inference-Time Scaling Algorithms. We consider three established algorithms: Best-of-N sampling [39], Greedy Search, and Beam Search [24]. In addition, we introduce a new baseline, Successive Halving, adapted from its use in hyperparameter optimization [8, 14].

Method	FT	MSR-VTT			DEVIL Static			DEVIL Medium			DEVIL High		
		500	1200	2000	500	1200	2000	500	1200	2000	500	1200	2000
Latte [23]		(-1.91)			(-2.20)			(-1.82)			(-2.68)		
Best-of-N [39]		-1.14	-0.88	-0.66	0.21	-0.09	0.02	-1.16	-0.85	-0.71	-1.54	-1.33	-1.24
Successive Halving		**-1.66**	-1.54	-1.46	**-1.18**	**-0.68**	**-0.5**	**-1.54**	-1.58	-1.22	-2.26	-2.04	-2.02
Successive Halving	✓	-1.67	**-1.25**	**-1.21**	-1.54	-0.88	-0.68	**-1.54**	**-1.18**	**-1.28**	**-2.17**	**-1.95**	**-1.94**
Greedy Search [24]		-0.86	-0.35	-0.32	-0.42	0.06	0.31	-1.14	-0.83	-0.52	-1.71	-1.43	-1.58
Greedy Search [24]	✓	**-0.37**	**-0.33**	**0.13**	**-0.14**	**0.43**	**0.51**	**-0.59**	**-0.61**	**-0.50**	**-1.63**	**-1.33**	**-1.19**
Beam Search [24]		-0.38	0.15	0.42	-0.29	-0.14	**1.12**	**-0.61**	-0.34	0.15	-1.50	-1.21	-0.87
Beam Search [24]	✓	**-0.21**	**1.18**	**0.83**	**-0.10**	**0.45**	0.95	-0.66	**-0.09**	**0.28**	**-1.25**	**-0.90**	**-0.72**

Table 1: Inference-time scaling algorithms applied to the Latte [23] generator, using VideoReward (zero-shot or our fine-tuned version) as verifier. Latte baseline runs at 50 NFEs.

Method	FT	MSR-VTT			DEVIL Static			DEVIL Medium			DEVIL High		
		100	200	300	100	200	300	100	200	300	100	200	300
CogVideox [43]		(0.69)			(1.47)			(0.79)			(-0.41)		
Best-of-N [39]		1.27	1.65	1.77	2.05	2.43	2.80	1.15	1.56	1.71	0.07	0.42	0.82
Successive Halving		1.19	1.73	1.81	1.93	2.3	**2.77**	1.35	1.71	1.99	**0.13**	0.23	0.29
Successive Halving	✓	**1.52**	**2.03**	**2.23**	**2.17**	**2.49**	2.69	**1.37**	**1.94**	**2.15**	0.12	**0.31**	**0.53**

Table 2: Inference-time scaling algorithms applied to the CogVideoX [43] generator, using VideoReward (zero-shot or our fine-tuned version) as verifier. CogVideoX baseline runs at 50 NFEs.

In our adaptation, we start with N independently sampled noisy latents. At regular intervals of Δ_t diffusion steps (every 5% in our implementation), each sample is evaluated by the verifier, and only the top 50% based on scores are retained until a single sample remains. The selected trajectory is then completed to the final timestep, and its corresponding score is reported as the final output. An overview of the employed algorithms is shown in Figure 5.

Experimental Setup. Following the evaluation protocol in [24], we assess model performance using prompts drawn from two benchmark datasets: MSR-VTT [42] and DEVIL [19], which include subsets of varying motion complexity (*i.e.*, Static, Medium, High). For Latte, we consider budgets of 500, 1,200, and 2,000 NFEs, while for CogVideoX, we use 100, 200, and 300 NFEs. These budget choices reflect the higher computational cost associated with each diffusion step in the CogVideoX model. When using the fine-tuned verifier, we restrict its application to the early stages of generation, consistent with its training. Specifically, the fine-tuned verifier is applied during the initial 30% of diffusion steps for Latte and 10% for CogVideoX. The original (zero-shot) verifier is used for the remaining steps.

Results. Table 1 reports results for Latte across the evaluated inference-time scaling algorithms. As expected, Best-of-N sampling (2000 NFEs) significantly outperforms naive generation ($N = 1$), with VideoReward gains of 1.25, 2.22, 1.11, and 1.44 on MSR-VTT, DEVIL Static, Medium, and High, respectively. Both Greedy Search and Beam Search surpass Best-of-N across all budgets and datasets. For example, at 1,200 NFEs, Greedy Search yields improvements of 0.55, 0.52, and 0.24 on MSR-VTT, DEVIL Static, and Medium. In contrast, Successive Halving does not outperform Best-of-N in this setting, due to low alignment of VideoReward scores during the early 50% of the generation process (cf. Sec. 4.1).

Incorporating the fine-tuned verifier consistently improves performance. When used with Beam Search (budget 1,200), it yields VideoReward gains of 1.13, 0.59, 0.25, and 0.31 across MSR-VTT and the three DEVIL subsets. These gains suggest that fine-tuning the verifier improves the reliability of early-step evaluations, thereby enabling more effective trajectory selection during sampling. This trend is further confirmed by similar improvements observed with Greedy Search. For instance, with budget 2,000 we obtain 0.45, 0.20, 0.02, and 0.39 increases on VideoReward metrics on MSR-VTT and the three DEVIL subsets.

Similar trends are observed for CogVideoX, as presented in Table 2. In particular, Successive Halving combined with the fine-tuned verifier consistently improves performance across nearly all settings. At 200 NFEs, we observe gains of 0.30, 0.19, 0.23, and 0.08 on MSR-VTT and the three DEVIL subsets compared to the zero-shot verifier. Moreover, due to the generally stronger temporal coherence exhibited by CogVideoX, Successive Halving with a fine-tuned verifier outperforms Best-of-N in nearly all tested configurations. These results highlight its potential as a competitive and computationally efficient baseline for inference-time scaling in video generation.

5 Conclusions

In this study, we have explored the role of verifiers in inference-time scaling for video generation, with a focus on their ability to evaluate partially denoised samples during the diffusion process. We have shown that fine-tuning the verifier on intermediate representations significantly enhances the alignment between intermediate predicted scores and the final ones. Finally, our results demonstrate that integrating a fine-tuned verifier into inference-time scaling algorithms leads to consistent improvements in overall generation performance, resulting in higher-quality video outputs across multiple settings and datasets.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources. This work has been supported by the EU Horizon projects "ELSA" (GA No. 101070617) and "ELLIOT" (GA No. 101214398), and by the EuroHPC JU project "MINERVA" (GA No. 101182737).

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos World Foundation Model Platform for Physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023.

[3] 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.

[4] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The Revolution of Multimodal Large Language Models: A Survey. In *ACL Findings*, 2024.

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *ICCV*, 2021.

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.

[8] Matthias Feurer and Frank Hutter. *Hyperparameter optimization*. Springer International Publishing, 2019.

[9] Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*, 2025.

[10] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. VIDEOSCORE: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation. In *EMNLP*, 2024.

[11] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

[13] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, , et al. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *CVPR*, 2024.

[14] Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial intelligence and statistics*. PMLR, 2016.

[15] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale Image Quality Transformer. In *ICCV*, 2021.

[16] Jaihoon Kim, Taehoon Yoon, Jisung Hwang, and Minhyuk Sung. Inference-Time Scaling for Flow Models via Stochastic Generation and Rollover Budget Forcing. *arXiv preprint arXiv:2503.19385*, 2025.

[17] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[18] Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, et al. Derivative-Free Guidance in Continuous and Discrete Diffusion Models with Soft Value-Based Decoding. *arXiv preprint arXiv:2408.08252*, 2024.

[19] Mingxiang Liao, Qixiang Ye, Wangmeng Zuo, Fang Wan, Tianyu Wang, Yuzhong Zhao, Jingdong Wang, Xinyu Zhang, et al. Evaluation of Text-to-Video Generation Models: A Dynamics Perspective. In *NeurIPS*, 2024.

[20] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving Video Generation with Human Feedback. *arXiv preprint arXiv:2501.13918*, 2025.

[21] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *ICLR*, 2023.

[22] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-Time Scaling for Diffusion Models beyond Scaling Denoising Steps. In *CVPR*, 2025.

[23] Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent Diffusion Transformer for Video Generation. *Transactions on Machine Learning Research*, 2025.

[24] Yuta Oshima, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Inference-Time Text-to-Video Alignment with Diffusion Latent Beam Search. *arXiv preprint arXiv:2501.19252*, 2025.

[25] William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *ICCV*, 2023.

[26] Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, et al. Open-Sora 2.0: Training a Commercial-Level Video Generation Model in $200k. *arXiv preprint arXiv:2503.09642*, 2025.

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*, 2022.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Karagol, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*, 2022.

[31] Anuj Singh, Sayak Mukherjee, Ahmad Beirami, and Hadi Jamali-Rad. CoDe: Blockwise Control for Denoising Diffusion Models. *arXiv preprint arXiv:2502.00968*, 2025.

[32] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. In *ICLR*, 2025.

[33] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *ICML*, 2015.

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2021.

[35] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *NeurIPS*, 2020.

[36] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.

[37] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating Videos with Scene Dynamics. In *NeurIPS*, 2016.

[38] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[39] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, et al. SANA 1.5: Efficient Scaling of Training-Time and Inference-Time Compute in Linear Diffusion Transformer. *arXiv preprint arXiv:2501.18427*, 2025.

[40] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. *arXiv preprint arXiv:2109.14084*, 2021.

[41] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. VisionReward: Fine-Grained Multi-Dimensional Human Preference Learning for Image and Video Generation. *arXiv preprint arXiv:2412.21059*, 2024.

[42] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *CVPR*, 2016.

[43] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *ICLR*, 2025.