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

Baraldi L. [1], Bucciarelli D. [1,2], Zeng Z. [3], Zhang C. [3], Zhang Q. [3], Cornia M. [2], Baraldi L. [2], Liu F. [3], Hu Z. [3], Cucchiara R. [2]
[1] University of Pisa, Italy [2] University of Modena and Reggio Emilia, Italy
[3] Huawei Technologies Duesseldorf GmbH RAMS Lab, Germany

Motivations

Video diffusion models produce impressive results, but generation remains costly, and results can be unstable. Inference-time scaling [1] improves quality without retraining, yet its success hinges on the verifier, the model that scores generated videos.

Our Goal

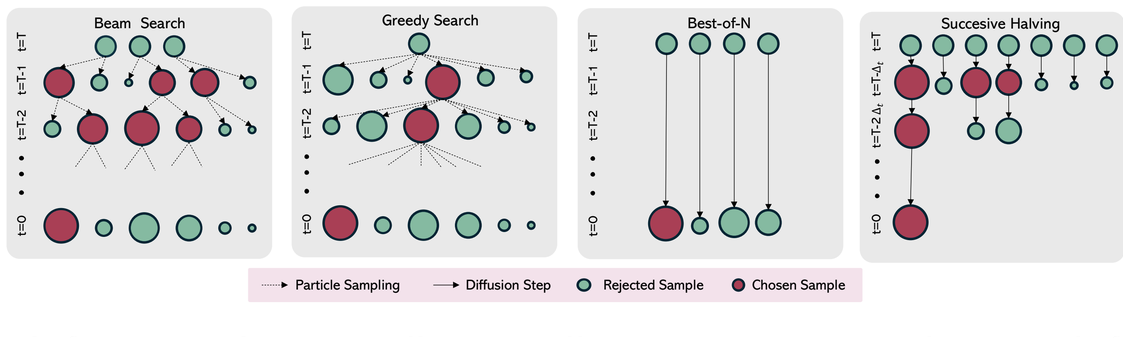
Understand and improve the role of verifiers in video diffusion. We: **(i)** Analyze how verifier scores evolve over timesteps. **(ii)** Fine-tune verifier on partially denoised samples. **(iii)** Show that tuned verifiers boost performance.

Verifiers Temporal Consistency

		VideoReward (Latte)	VisionReward (Latte)	VBench Ens. (Latte)	VideoReward (CogVideoX)	VisionReward (CogVideoX)	VBench Ens. (CogVideoX)
Pearson	0.05	-0.02	-0.01	-0.02	0.16	-0.00	-0.03
	0.1	0.14	0.03	0.03	0.55	0.39	0.43
	0.15	0.23	0.09	0.16	0.67	0.53	0.60
	0.2	0.43	0.25	0.27	0.78	0.63	0.76
	0.25	0.57	0.39	0.38	0.85	0.69	0.83
	0.3	0.70	0.48	0.61	0.90	0.76	0.89
	0.4	0.83	0.60	0.83	0.94	0.85	0.95
	0.45	0.86	0.64	0.87	0.95	0.88	0.96
	0.5	0.91	0.74	0.93	0.96	0.90	0.97
	0.55	0.93	0.80	0.95	0.97	0.92	0.98
	0.6	0.95	0.86	0.97	0.97	0.94	0.99
	0.65	0.96	0.88	0.98	0.97	0.95	0.99
	0.7	0.97	0.92	0.99	0.98	0.96	0.99
	0.8	0.98	0.95	0.99	0.99	0.97	1.00
	0.9	0.99	0.97	1.00	0.99	0.98	1.00
Spearman	0.05	-0.03	0.07	-0.03	0.16	0.02	-0.05
	0.1	0.08	0.03	0.02	0.49	0.39	0.40
	0.15	0.14	0.07	0.12	0.62	0.53	0.59
	0.2	0.33	0.20	0.21	0.76	0.65	0.75
	0.25	0.48	0.33	0.33	0.83	0.70	0.82
	0.3	0.64	0.46	0.56	0.88	0.76	0.88
	0.4	0.80	0.59	0.79	0.93	0.85	0.94
	0.45	0.84	0.63	0.84	0.94	0.88	0.95
	0.5	0.89	0.74	0.91	0.95	0.90	0.97
	0.55	0.92	0.79	0.93	0.96	0.92	0.97
	0.6	0.95	0.85	0.96	0.96	0.94	0.98
	0.65	0.96	0.88	0.97	0.97	0.95	0.99
	0.7	0.97	0.91	0.98	0.97	0.96	0.99
	0.8	0.98	0.95	0.99	0.99	0.97	0.99
	0.9	0.98	0.97	1.00	0.99	0.98	1.00
Kendall	0.05	-0.02	0.05	-0.02	0.10	0.01	-0.04
	0.1	0.06	0.02	0.01	0.33	0.27	0.28
	0.15	0.09	0.05	0.08	0.45	0.38	0.42
	0.2	0.23	0.14	0.14	0.57	0.46	0.56
	0.25	0.34	0.23	0.23	0.64	0.51	0.63
	0.3	0.46	0.32	0.40	0.71	0.57	0.71
	0.4	0.61	0.41	0.60	0.77	0.66	0.80
	0.45	0.65	0.46	0.66	0.80	0.70	0.82
	0.5	0.73	0.55	0.75	0.82	0.75	0.85
	0.55	0.76	0.61	0.79	0.83	0.77	0.87
	0.6	0.80	0.68	0.84	0.85	0.81	0.89
	0.65	0.82	0.73	0.86	0.86	0.83	0.90
	0.7	0.85	0.78	0.89	0.87	0.86	0.92
	0.8	0.88	0.84	0.93	0.90	0.88	0.94
	0.9	0.89	0.89	0.96	0.93	0.91	0.97

We evaluate **temporal consistency**, how well intermediate verifier scores predict final video quality, by measuring their correlation across diffusion steps. Correlations are weak at early stages, where verifiers fail to assess noisy samples, **limiting reliable decision-making during the initial part of generation.**

Background



- Inference-time scaling reallocates computation at inference, generating multiple candidates through an **algorithm** and using a **verifier** to pick the best.
- Algorithms such as **Best-of-N** [2], **Greedy Search** [3], and **Beam Search** [3] differ in how they explore and select samples within the generation process and may require intermediate verifier evaluation.
- The verifier scores each intermediate video, ideally correlating early predictions with final quality.

However, verifiers trained only on fully generated videos may fail on noisy, partial generations.

Verifiers Fine-Tuning

		Pearson (VideoReward Zero-Shot)	Spearman (VideoReward Zero-Shot)	Kendall (VideoReward Zero-Shot)	Pearson (VideoReward Fine-Tuned)	Spearman (VideoReward Fine-Tuned)	Kendall (VideoReward Fine-Tuned)
	0.05	-0.12	-0.14	-0.095	0.46	0.44	0.31
	0.1	0.022	-0.024	-0.018	0.51	0.47	0.34
	0.15	0.091	0.054	0.035	0.5	0.44	0.31
	0.2	0.24	0.13	0.094	0.52	0.46	0.33
	0.25	0.32	0.27	0.18	0.5	0.45	0.31
	0.3	0.63	0.59	0.42	0.59	0.52	0.38

		Pearson (VideoReward Zero-Shot)	Spearman (VideoReward Zero-Shot)	Kendall (VideoReward Zero-Shot)	Pearson (VideoReward Fine-Tuned)	Spearman (VideoReward Fine-Tuned)	Kendall (VideoReward Fine-Tuned)
	0.05	0.18	0.17	0.11	0.47	0.46	0.31
	0.1	0.62	0.5	0.35	0.69	0.58	0.4

- We fine-tune VideoReward [4] with LoRA on 1,000 pairs of partially denoised videos, using final scores as pseudo-labels. Training is conducted on both CogVideoX [5] and Latte [6] generators.
- After fine-tuning, early-step correlations rise on both CogVideoX and Latte, improving guidance for inference-time scaling in initial steps.

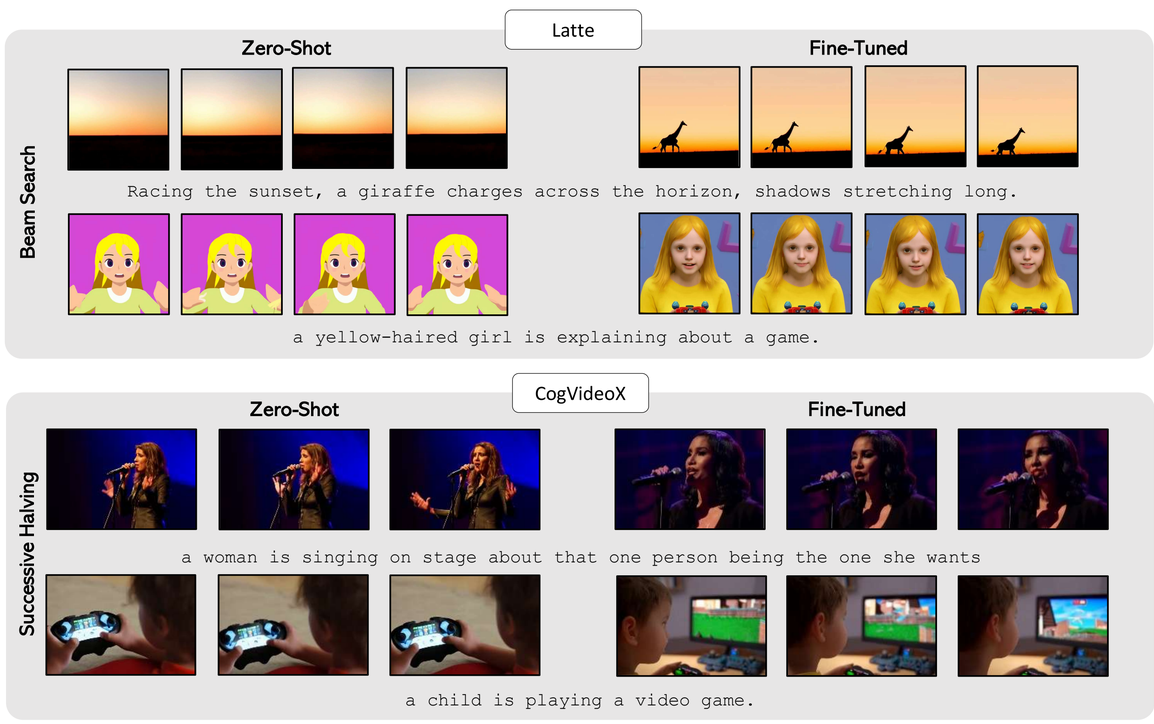
Inference-Time Scaling Results

Method	FT	MSR-VTT 500	MSR-VTT 1200	MSR-VTT 2000	DEVIL Static 500	DEVIL Static 1200	DEVIL Static 2000	DEVIL Medium 500	DEVIL Medium 1200	DEVIL Medium 2000	DEVIL High 500	DEVIL High 1200	DEVIL High 2000
Latte		(-1.91)			(-2.20)			(-1.82)			(-2.68)		
Best-of-N		-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		-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	✓	**-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		-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	✓	**-0.21**	**1.18**	**0.83**	**-0.10**	**0.45**	0.95	-0.66	**-0.09**	**0.28**	**-1.25**	**-0.90**	**-0.72**

Method	FT	MSR-VTT 100	MSR-VTT 200	MSR-VTT 300	DEVIL Static 100	DEVIL Static 200	DEVIL Static 300	DEVIL Medium 100	DEVIL Medium 200	DEVIL Medium 300	DEVIL High 100	DEVIL High 200	DEVIL High 300
CogVideox		(0.69)			(1.47)			(0.79)			(-0.41)		
Best-of-N		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**

Fine-tuned verifiers obtained improved results on video generation benchmarks on both static and dynamic videos.

Qualitative Results



[1] Snell et al. *Scaling llm test-time compute optimally can be more effective than scaling model parameters*. ICLR, 2025.
[2] Xie et al. *Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer*. ArXiv, 2025.
[3] Oshima et al. *Inference-Time Text-to-Video Alignment with Diffusion Latent Beam Search*. ArXiv, 2025.
[4] Liu et al. *Improving Video Generation with Human Feedback*. ArXiv, 2025.
[5] Yang et al. *CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer*. ICLR, 2025.
[6] Ma et al. *Latte: Latent Diffusion Transformer for Video Generation*, TMLR 2025.
[7] Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. ICLR, 2022.