Can Test-Time Scaling Improve World Foundation Model?

Anonymous ICCV submission

Paper ID 15

Abstract

*World foundation models, which simulate the physical world by predicting future states from current observations and inputs, have become central to many applications in physical intelligence, including autonomous driving and robotics. However, these models require substantial computational resources for pretraining and are further constrained by available data during post-training. As such, scaling computation at test time emerges as both a critical and practical alternative to traditional model enlargement or re-training. In this work, we introduce **SWIFT**, a test-time scaling framework tailored for WFMs. SWIFT integrates our extensible WFM evaluation toolkit with process-level inference strategies, including fast tokenization, probability-based Top-K pruning, and efficient beam search. Empirical results on the COSMOS model demonstrate that test-time scaling exists even in a compute-optimal way. Our findings reveal that test-time scaling laws hold for WFMs and SWIFT provides a scalable and effective pathway for improving WFM inference without retraining or increasing model size.*

1. Introduction

World foundation models (WFMs) aim to simulate physical dynamics by predicting future states from current observations and inputs. They underpin physical intelligence across domains such as autonomous driving, robotics, and embodied planning by generating synthetic data for scalable simulation, analysis, and agent training.

Despite their potential, WFMs are expensive to train. Pretraining requires massive resources, especially due to video-based inputs. For example, COSMOS [1] was trained on tens of millions of video hours using thousands of GPUs over several months. Even post-training, performance gains from model scaling diminish due to data limits and scaling law plateaus. These constraints motivate test-time computation scaling—enhancing inference performance by increasing compute usage without retraining.

Inspired by test-time scaling successes in language and vision-language models [8, 10], we explore this paradigm for WFMs—marking the first such effort.

However, key challenges arise: ① **Lack of tailored benchmarks.** Existing video generation evaluations focus on aesthetics or semantics, not the physical realism and consistency needed for world modeling. A modular, extensible toolkit is needed to assess WFMs across diverse downstream tasks. ② **Strategy design constraints.** Unlike LLMs, WFMs often use diffusion decoders—slow and ill-suited for intermediate-step checking strategies like chain-of-thought or tree-of-thought.

To address these, we propose **SWIFT**, a test-time scaling framework designed for WFMs (Fig. 1).

Our contributions are threefold:

- **WFM Evaluation Toolkit.** We introduce the first evaluation suite for WFMs—modular, rule-based, and extensible to various tasks.
- **SWIFT Framework.** We present SWIFT, the first test-time scaling framework for WFMs. It leverages fast tokenization, Top-K pruning to reduce overconfidence, and beam search for efficient sample selection.
- **Empirical Study.** We provide the first empirical analysis of test-time scaling in WFMs, using COSMOS as the base model. Our findings show:
  - A test-time scaling law exists—even under fixed compute budgets—where smaller models with scaling outperform larger models.
  - SWIFT further improves performance as sample count increases, efficiently utilizing inference-time resources.

2. Motivation: Why Test-Time Scaling for World Foundation Models

Test-time scaling—increasing compute at inference—has proven effective in unlocking model potential, often outperforming naive model scaling [10].

World Foundation Models (WFMs), which simulate real-world dynamics to generate synthetic data for domains like robotics and autonomous driving, are prime candidates for this approach. We highlight two core motivations:

① **Training large WFMs is prohibitively expensive.** Unlike LLMs trained on text, WFMs process vast video data,

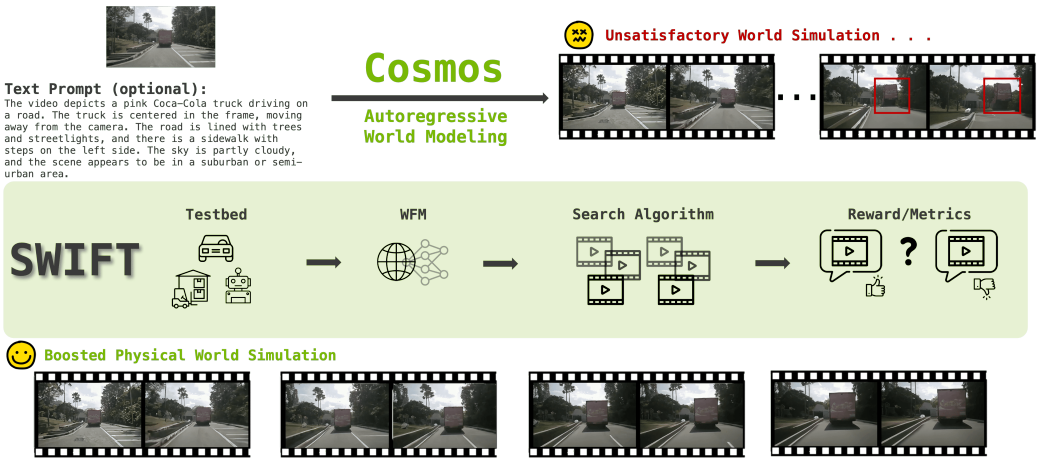

Figure 1. **SWIFT** enables test-time scaling (TTS) for world foundation models (WFMs). Using COSMOS as a base, we show how SWIFT enhances initially unrealistic simulations (top) into more physically plausible ones (bottom).

demanding orders of magnitude more compute. Training COSMOS on 20M hours of video required 10K H100 GPUs for 3 months—just for a 13B model. Larger models also need more data, but high-quality video is scarce, creating a *chicken-and-egg dilemma*: strong WFMs require synthetic data to train, but good synthetic data only emerges after strong WFMs exist.

② **Inference for large WFMs is as costly as multiple runs of smaller ones.** Due to autoregressive decoding, generating long, high-resolution videos is slow and memory-intensive. For instance, a 12B model costs 3× more FLOPs than a 4B model, meaning we could run the 4B model multiple times instead—enabling exploration strategies or sample selection at test time.

Goal: Improve WFM performance at inference without retraining or enlarging the base model.

3. WFM Evaluation Toolkit

World Foundation Models (WFMs) increasingly rely on video generation to create digital twins—synthetic representations of physical environments—enabling simulation, analysis, and training in domains such as autonomous driving and robotics.

Yet, no standard evaluation toolkit exists for WFMs. ① General video benchmarks (e.g., VBench [7], VideoScore [5]) prioritize aesthetic or semantic fidelity, misaligned with the physical realism and consistency WFMs demand. ② Task-specific benchmarks (e.g., ACT-Bench [2]) focus on downstream control but overlook generative video quality and coherence.

To bridge this gap, we propose the first general-purpose evaluation toolkit tailored for WFMs. Our toolkit is modular and extensible, enabling domain-specific analysis while remaining broadly applicable. **Included Metrics:**

- **3D Consistency:** Assesses geometric coherence using CUT3R [11], which reconstructs 3D from videos feed-forward.
- **Temporal Consistency:** Measures frame-to-frame smoothness and object permanence using CLIP and DINO similarity.
- **Spatial Relationship Awareness:** Evaluates whether spatial layouts—especially human-environment interactions—are physically plausible, e.g., left-right/top-bottom relations in factory scenes.
- **Perceptual Quality:** Uses LAION's aesthetic predictor to evaluate visual fidelity. Natural noise or blur is not penalized, as it may reflect real sensor outputs.
- **Text-to-Video Alignment:** Assesses prompt-video coherence via CLIPScore (frame-level) and X-CLIPScore (video-level).

While WFM evaluation remains complex due to the models' generality, our toolkit is designed to grow with the field. It will be open-sourced to support reproducibility and community adoption.

We adopt *autonomous driving* as our primary testbed, aligning with COSMOS's application domain and reflecting the high importance—and difficulty—of generating diverse, high-fidelity videos in this space.

4. SWIFT

We propose **SWIFT**, the first test-time scaling framework for WFMs, to address two core questions:

Q① Can test-time scaling improve WFM quality under a fixed compute budget? Can a smaller model, augmented with test-time search, rival or outperform a larger one?

Q② What strategy best fits WFM's unique video generation needs? Unlike LLMs, WFMs rely on autoregressive decoding and diffusion-based video generation, which is expensive and poorly suited for intermediate-step verification.

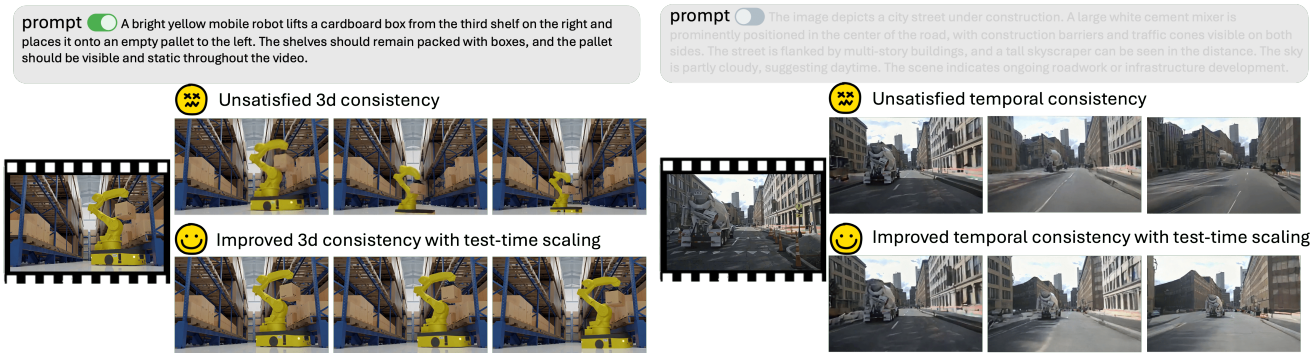

Figure 2. Generated videos without (top) and with (bottom) test-time scaling. TTS improves 3D (left) and temporal (right) consistency.

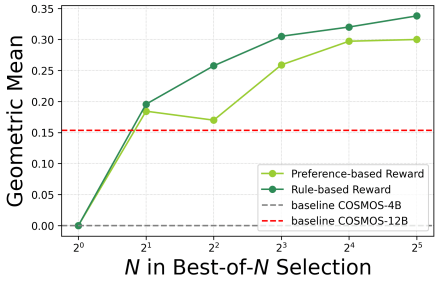

Figure 3. Rule-based vs. preference-based rewards.

4.1. Framework

We cast autoregressive video generation as a Markov Decision Process (MDP). A pretrained WFM p_Θ generates a video sequence $\mathcal{V} = \{v_1, ..., v_N\}$ given prompt c and initial video chunk v_0:

$$v_i = p_\Theta(v_i \mid c, v_0, v_1, ..., v_{i-1})$$

Each state $s \in \mathcal{S}$ is a partially generated video. The reward function $\mathcal{R}(s, a)$ (verifier) scores outputs; actions $\mathcal{A}$ correspond to sampling strategies.

4.2. Verifier Design

We compare rule-based and preference-based verifiers (Figure 3). Rule-based rewards (e.g., aesthetic quality, object permanence) are objective, robust, and extendable. Preference-based ones (e.g., VideoScore [5]) require fine-tuning and large-scale human feedback.

Our experiments (best-of-N sampling under COSMOS-4B) show that rule-based rewards outperform preference-based ones in both alignment and stability—echoing findings in DeepSeek R1 [4]. Therefore, SWIFT adopts rule-based metrics for reliable and scalable test-time verification.

4.3. Action Design

Best-of-N Sampling. Our baseline strategy samples N candidate videos and selects the best via verifier scores. We observe:

- *Test-time scaling exists:* More samples lead to better quality (Table 1).

- *Compute-optimality:* 2–4 passes from a 4B model match 1 pass from 12B–a cost-efficient alternative to scaling.

Toward Efficient Search. While best-of-N improves output, it wastes compute. We propose an efficient beam-style search tailored for WFMs with three key designs (Figure 4):

- **Fast Tokenizer Proxy:** To avoid costly diffusion decoding (137s), we use the model's discrete-token decoder as a proxy (0.015s) for verifier scoring. Figure 5 shows strong correlation in scores, enabling cheap early pruning.
- **Probabilistic Top-K Selection:** At each step, we sample N continuations and compute rewards $\{r_i\}$. We apply softmax selection with temperature τ to avoid overconfidence and encourages diversity (Figure 6).

$$p_i = \frac{\exp(r_i/\tau)}{\sum_j \exp(r_j/\tau)}$$

- **Beam Search with Pruning:** We maintain K partial trajectories. Each spawns M continuations; we score all and retain top-K. This keeps growth linear in sequence length and prevents compute explosion.

5. Experiments

Setup. We evaluate test-time scaling on WFMs using the COSMOS family, focusing on the autonomous driving domain. Two input modalities are used: image-to-video and image+text-to-video. Each model receives 9 input frames to capture key motion priors (position, velocity, acceleration) as motivated in [3, 6].

Datasets. We use 900 videos sampled from the test splits of nuScenes (150 scenes) and Waymo (750 scenes), ensuring zero training-set leakage. Text prompts are generated using a COSMOS-trained prompt model fine-tuned from [9].

Metrics. Following convention, we report FVD and FID per dataset, plus VBench and VideoScore dimensions. To avoid overlap with our rule-based rewards, we exclude metrics like temporal consistency from VBench and use only

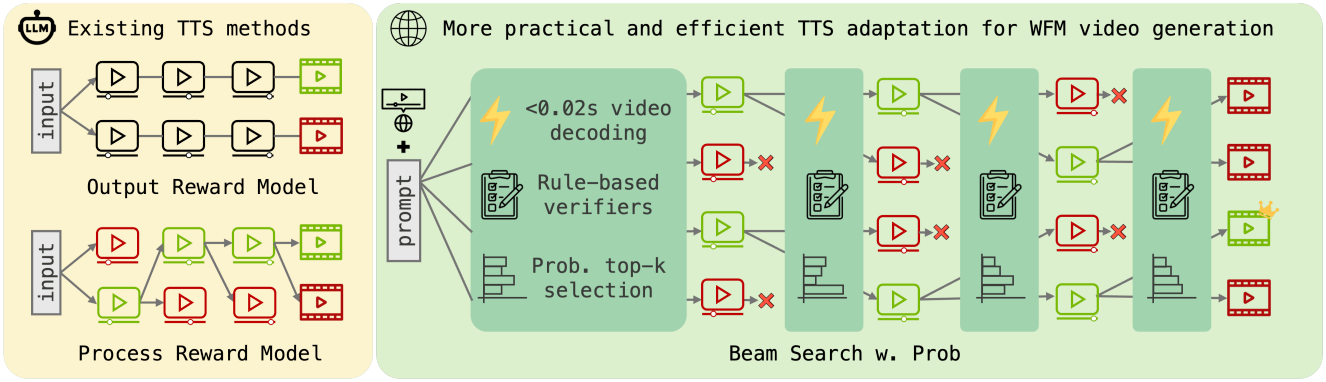

Figure 4. Our proposed beam-style search improves over ORM and PRM by addressing WFM-specific efficiency bottlenecks.

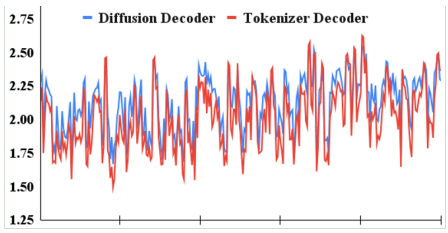

Figure 5. Fast tokenizer and diffusion decoder yield similar score trends.

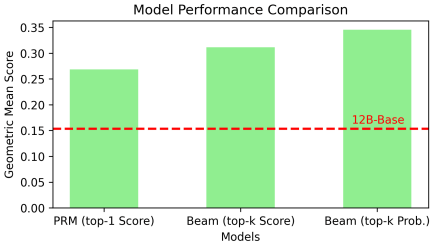

Figure 6. Beam search with probabilistic top-K outperforms others.

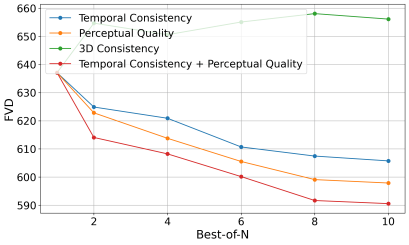

Figure 7. COSMOS-4B ablation on reward design.

motion smoothness (MS) and image quality (IQ) for cross-validation. We aggregate results via geometric mean of normalized scores for summary evaluation.

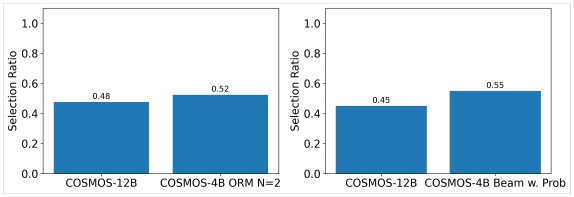

Figure 8. Human preference: 4B+TTS vs. 12B.

5.1. Reward Ablation

We first identify effective reward functions by running best-of-N selection with COSMOS-4B. As shown in Figure 7, *temporal consistency* and *perceptual quality* are the most reliable for autonomous driving. In contrast, 3D consistency is less helpful due to noisy point clouds, and spatial awareness is trivial in road scenes.

5.2. Test-Time Scaling Results

Naive Best-of-N. Table 1 shows consistent improvements across metrics with increasing N. Notably: - Best-of-2 already rivals COSMOS-12B, showcasing compute-efficient scaling. - Performance improves monotonically with N, verifying a test-time scaling law. - Smaller models benefit more, reducing dependence on larger pretrained models.
Probabilistic Beam Search. Table 2 compares our strategy with PRM. We observe: - PRM suffers from greedy trajectory selection and instability. - Our approach maintains a diverse candidate pool and uses soft top-K pruning, yielding stronger performance with minimal extra cost.
Image+Text Modality. Table 3 shows similar gains using our strategy on COSMOS-5B vs. 13B, confirming that test-time scaling generalizes across modalities.

5.3. Human & Qualitative Evaluation

Human Study. We conduct a 2AFC human study with 3,685 responses from 24 participants. Test-time scaled outputs from COSMOS-4B are often preferred over COSMOS-12B, indicating perceptual gains not fully captured by automatic metrics.

Model	N	FVD	FID	IQ	MS	VQ	TC	DD	FC
4B	1	637.1 / 120.3	67.8 / 10.6	63.5	0.982	3.86	3.56	3.86	3.68
	2	622.8 / 120.2	58.9 / 10.3	64.0	0.983	3.86	3.56	3.86	3.68
	4	613.8 / 117.4	52.0 / 10.2	64.3	0.983	3.87	3.57	3.86	3.68
	8	599.1 / 116.1	49.3 / 10.1	64.7	0.984	3.88	3.58	3.87	3.70
	16	599.1 / 120.0	45.8 / 10.1	64.8	0.984	3.90	3.59	3.89	3.73
12B	1	560.9 / 117.2	67.1 / 10.7	63.7	0.981	3.94	3.63	3.93	3.76

Table 1. COSMOS-4B vs. 12B under best-of-N. FVD/FID are split for nuScenes / Waymo.

Model	N	Alg.	FVD	FID	IQ	MS	VQ	TC	DD	FC
4B	1	-	637.08/120.30	67.75/10.58	63.48	0.9822	3.86	3.56	3.86	3.68
	4	PRM	614.00/116.51	52.68/11.48	63.79	0.9836	3.68	3.38	3.68	3.48
		Ours	612.68/114.27	50.39/10.35	64.39	0.9837	3.86	3.56	3.85	3.67
	16	PRM	616.89/121.07	47.29/10.33	64.87	0.9844	3.89	3.58	3.88	3.71
		Ours	590.34/120.48	43.69/10.27	64.98	0.9846	3.92	3.64	3.90	3.74
12B	1	-	560.86/117.23	67.10/10.67	63.73	0.9807	3.94	3.63	3.93	3.76

Table 2. COSMOS-4B and COSMOS-12B under different search algorithms. N is sample number. For Ours, M is set as sqrt(N).

Model	N	Alg.	FVD	FID	IQ	MS	VQ	TC	DD	FC
5B	1	-	728.68/126.63	59.71/10.47	63.48	0.9822	3.80	3.44	3.81	3.63
	2	ORM	659.79/111.96	59.45/10.01	63.90	0.9840	3.89	3.55	3.88	3.71
	4	PRM	641.5/112.15	52.60/10.07	64.39	0.9843	3.89	3.55	3.88	3.71
		Ours	628.01/110.02	50.16/9.98	64.77	0.9848	3.91	3.57	3.92	3.73
13B	1	-	616.92/109.77	59.71/11.48	63.79	0.9834	3.92	3.55	3.94	3.75

Table 3. COSMOS-5B and COSMOS-13B under different search algorithms. N is sample number. For Ours, M is set as sqrt(N).

Figure 9. COSMOS-4B w/o (top), with ORM (middle), and with our method (bottom).

Figure 10. COSMOS-5B w/o (top), with ORM (middle), and with our method (bottom).

Qualitative Samples. Figures 9–10 show visual improvements from TTS: smoother transitions, more stable objects, and reduced visual artifacts.

6. Conclusion

In this work, we presented **SWIFT**, the first test-time scaling framework specifically designed for world foundation models (WFMs). Addressing the high computational cost and data limitations of training and scaling WFMs, SWIFT offers an efficient alternative by reallocating computation during inference. Empirical results on the COSMOS model demonstrate that test-time scaling not only improves output quality, but does so in a compute-optimal manner—allowing smaller models to match or even outperform larger ones under the same compute budget. These findings establish that test-time scaling laws hold for WFMs and open up a practical, scalable pathway for deploying WFMs more efficiently, without retraining or model enlargement.

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. 1

[2] Hidehisa Arai, Keishi Ishihara, Tsubasa Takahashi, and Yu Yamaguchi. Act-bench: Towards action controllable world models for autonomous driving. *arXiv preprint arXiv:2412.05337*, 2024. 2

[3] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024. 3

[4] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3

[5] 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. *arXiv preprint arXiv:2406.15252*, 2024. 2, 3

[6] Xiaotao Hu, Wei Yin, Mingkai Jia, Junyuan Deng, Xiaoyang Guo, Qian Zhang, Xiaoxiao Long, and Ping Tan. Drivingworld: Constructingworld model for autonomous driving via video gpt. *arXiv preprint arXiv:2412.19505*, 2024. 3

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 2

[8] 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. *arXiv preprint arXiv:2501.09732*, 2025. 1

[9] Mistral and NVIDIA. Mistral-nemo-12b-instruct: A 12b parameter large language model. https://mistral.ai/news/mistral-nemo. 2024. 3

[10] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. 1

[11] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. *arXiv preprint arXiv:2501.12387*, 2025. 2