Comparative Analysis of Euler Ancestral and Res Multistep Samplers in NVIDIA Cosmos for Text-to-Video Generation

Florence Li , Fernando Jia * , Jade Zheng , Yuteng Fu

Posted Date: 19 March 2025

doi: 10.20944/preprints202503.1432.v1

Keywords: Diffusion Models; Text-to-Video Generation; Euler Ancestral Sampler; Res Multistep Sampler; NVIDIA Cosmos; Generative Models; Video Synthesis; Structural Fidelity; Temporal Coherence; Denoising Process; Deep Learning; Image Quality Metrics; PSNR; SSIM; VMAF; Sampling Strategies; Classifier-Free Guidance; Diffusion Steps; Video Resolution; Frame Rate; Benchmarking; Parameter Optimization; Artifact Reduction; High-Dimensional Latent Spaces; Real-Time Generation

 doi:10.20944/preprints202503.1432.v1

Article

Comparative Analysis of Euler Ancestral and Res Multistep Samplers in NVIDIA Cosmos for Text-to-Video Generation

Florence Li [1,2], Fernando Jia [2,3,*], Jade Zheng [2] and Yuteng Fu [4]

[1] Computer Science Department,Stanford University; florence.li@stanford.edu
[2] Intelligence Cubed; jade.zheng@intelligencecubed.com
[3] UC Berkeley RDI
[4] Pratt School of Engineering, Duke University; yuteng.fu@duke.edu
* Correspondence: fernando.jia@berkeley.edu

Abstract: This report presents a comparative study of two samplers—Euler ancestral and Res multistep—within the NVIDIA Cosmos diffusion model for text-to-video generation. Under fixed generation conditions (constant diffusion steps, classifier-free guidance scale, resolution, video length, and frame rate) and using identical positive/negative prompts, an equal number of videos are generated per sampler. Quality is assessed using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Video Multi-method Assessment Fusion (VMAF). The results reveal that although both samplers yield low pixel-level fidelity and nearly zero perceptual quality scores, the Res multistep sampler achieves markedly higher structural similarity, indicating superior structural fidelity. These findings are discussed in the context of prior work on diffusion models and NVIDIA Cosmos [1,2], and it is concluded that Res multistep is preferable when structural coherence is critical.

Keywords: Diffusion Models; Text-to-Video Generation; Euler Ancestral Sampler; Res Multistep Sampler; NVIDIA Cosmos; Generative Models; Video Synthesis; Structural Fidelity; Temporal Coherence; Denoising Process; Deep Learning; Image Quality Metrics; PSNR; SSIM; VMAF; Sampling Strategies; Classifier-Free Guidance; Diffusion Steps; Video Resolution; Frame Rate; Benchmarking; Parameter Optimization; Artifact Reduction; High-Dimensional Latent Spaces; Real-Time Generation

1. Introduction

In recent years, diffusion-based generative models have advanced significantly, incorporating sophisticated architectures, high-dimensional latent spaces, and intricate conditioning mechanisms [3–5]. Concurrently, the proliferation of sampling strategies has led to uncertainty regarding the optimal choice for achieving stable and high-quality outputs, particularly in video generation where temporal coherence is critical [2,6]. Among the available samplers, Euler ancestral has been valued for its simplicity and historical reliability, yet it is often criticized for its variability, while Res multistep employs a multi-step refinement process that promises enhanced structural fidelity at the cost of increased computational demand [3–5]. This debate has left practitioners without clear guidance on which sampler is best suited for modern diffusion models such as NVIDIA Cosmos, designed for real-time text-to-video generation [1]. To address this challenge, we perform a systematic comparison of Euler ancestral and Res multistep under controlled conditions, evaluating their performance using established metrics including Peak Signal-to-Noise Ratio (PSNR) [7], Structural Similarity Index (SSIM) [8], and Video Multi-method Assessment Fusion (VMAF) [9]. Our study aims to resolve the ongoing debate by determining which sampler yields more stable and visually coherent outputs, thereby providing critical insights for researchers and practitioners alike.

In this study, we compare two samplers—Euler ancestral and Res multistep—under strictly controlled conditions. Both samplers use the same input prompts and constant generation parameters.

 doi:10.20944/preprints202503.1432.v1

We evaluate the outputs with PSNR, SSIM, and VMAF to provide a rigorous assessment of which method produces more structurally coherent and perceptually consistent results.

2. Related Work

The evolution of diffusion-based generative models began with the pioneering work on deep unsupervised learning using nonequilibrium thermodynamics [3]. Subsequent research refined these models, leading to significant improvements in denoising diffusion probabilistic models [4,5] and high-resolution image synthesis using latent diffusion models [2]. Moreover, recent work has explored the design space of diffusion-based generative models to enhance performance [6].

In the domain of video generation, maintaining temporal and spatial consistency remains challenging. NVIDIA Cosmos [1] addresses these challenges through parameter optimization. The choice of sampler has been identified as a key variable affecting generation quality. The Euler ancestral sampler, an earlier approach, has been associated with high variability, while the Res multistep sampler, with its multi-step refinement process, tends to preserve structural details more effectively. Video quality is typically measured using metrics such as PSNR [7], SSIM [8], and VMAF [9].

3. Dataset and Features

In this project, videos are generated using a single fixed positive prompt and a single fixed negative prompt. This controlled input ensures that any variations in output quality are attributable solely to the sampling strategy rather than to differences in textual content. The generation parameters—including resolution, video length, frame rate, diffusion steps, and classifier-free guidance scale—are kept constant across both samplers.

For each sampler (Euler ancestral and Res multistep), an identical set of videos is generated. The outputs are evaluated against a common reference using PSNR, SSIM, and VMAF, ensuring consistency in the evaluation process.

4. Methods

4.1. Video Generation with NVIDIA Cosmos

NVIDIA Cosmos leverages diffusion-based methods to generate videos from textual descriptions by iteratively denoising a latent representation until a complete video is formed [1]. Two sampling strategies are compared in this study:

- Euler ancestral: Implements an ancestral approach that can introduce variability between frames [3,4].
- Res multistep: Refines the denoising process by dividing it into multiple sub-steps, thereby enhancing structural stability [6].

4.2. Quality Metrics

4.2.1. Peak Signal-to-Noise Ratio (PSNR)

PSNR quantifies the fidelity of a generated frame relative to a reference frame:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right), \tag{1}$$

where MAX is the maximum pixel value and MSE is the mean squared error between the generated frame and the reference [7].

4.2.2. Structural Similarity Index (SSIM)

SSIM measures the similarity in luminance, contrast, and structure between two images x and y:

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{2}$$

 doi:10.20944/preprints202503.1432.v1

where μ_x and μ_y are local means, σ_x^2 and σ_y^2 are variances, σ_{xy} is the covariance, and C_1, C_2 are constants for numerical stability [8].

4.2.3. Video Multi-method Assessment Fusion (VMAF)

VMAF is a perceptual metric that integrates several quality measures such as Visual Information Fidelity, detail loss, and blockiness:

$$\text{VMAF} = f(\text{VIF}, \text{Detail Loss}, \text{Blockiness}, \ldots), \tag{3}$$

where f is a learned function that produces a score correlating with human subjective quality [9].

5. Experiments, Results, and Discussion

5.1. Experimental Setup

- **Prompts**: One fixed positive and one fixed negative prompt.
- **Generation Parameters**:
  - Diffusion Steps: 20
  - CFG Scale: 7.5
  - Resolution: 1280×704
  - Video Length: 121 frames
  - Frame Rate: 12 fps
- **Samplers**: Euler ancestral vs. Res multistep
- **Number of Videos**: 50 per sampler
- **Evaluation Metrics**: PSNR, SSIM, VMAF

5.2. Euler Ancestral Sampler Results

- **PSNR**: Approximately 14.0 dB
- **SSIM**: Ranges from 0.04 to 0.07
- **VMAF**: Between 0.000 and 0.009

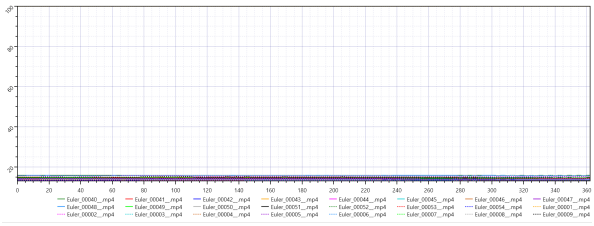

Figure 1. Euler PSNR.

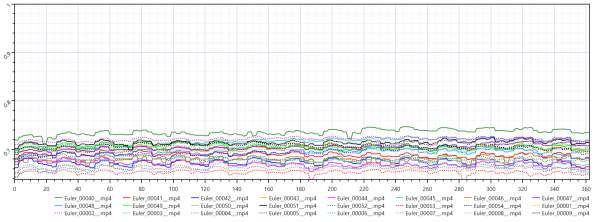

Figure 2. Euler SSIM.

 doi:10.20944/preprints202503.1432.v1

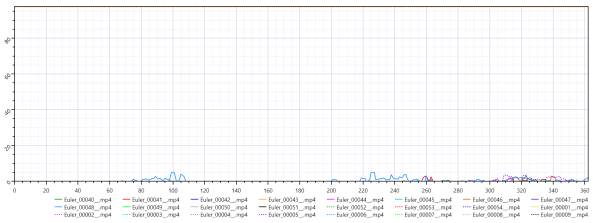

Figure 3. Euler VMAF.

These metrics indicate that outputs generated by the Euler ancestral sampler have significant pixel-level discrepancies (low PSNR) and very limited structural similarity (extremely low SSIM). The near-zero VMAF suggests minimal perceptual alignment with the reference.

5.3. Res Multistep Sampler Results

- **PSNR**: Ranges from 14.1 to 14.6 dB
- **SSIM**: Ranges from 0.64 to 0.65
- **VMAF**: Between 0.0050 and 0.0054

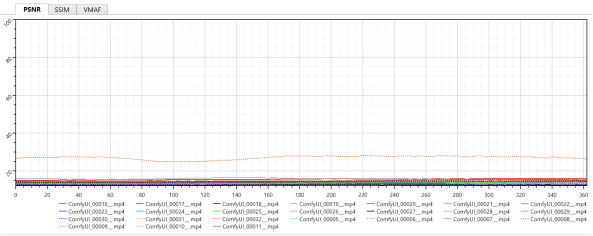

Figure 4. Res multistep PSNR.

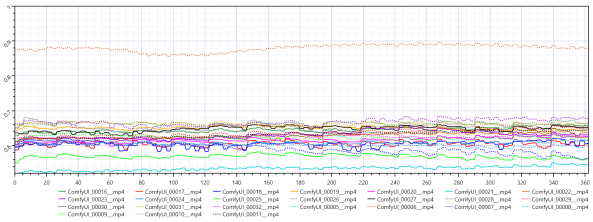

Figure 5. Res multistep SSIM.

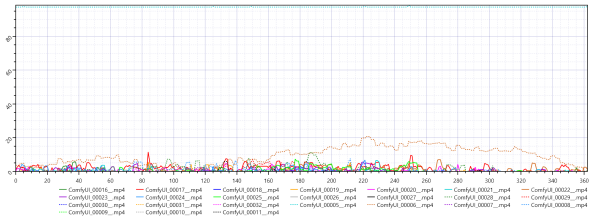

Figure 6. Res multistep VMAF.

While the PSNR values remain similar between the two samplers, Res multistep exhibits significantly higher SSIM, indicating a marked improvement in preserving structural details. The slightly higher VMAF values, although still near zero, further support its relative advantage.

5.4. Comparative Analysis

Key Observations:

- *Structural Similarity*: Res multistep achieves SSIM values roughly 10 times higher than Euler ancestral, indicating vastly better preservation of structure [8].

 doi:10.20944/preprints202503.1432.v1

- *Pixel-Level Fidelity*: Both samplers show similar PSNR values, suggesting comparable pixel-wise reconstruction errors [7].
- *Perceptual Quality*: Despite near-zero VMAF scores for both, the slight improvement with Res multistep indicates marginally better perceptual alignment.

Table 1. Comparison of Euler ancestral and Res multistep Samplers.

ine **Sampler**	**PSNR (dB)**	**SSIM**	**VMAF**
ine **Euler ancestral**	∼14.0	0.04 – 0.07	0.000 – 0.009
Res multistep	14.1 – 14.6	0.64 – 0.65	0.0050 – 0.0054
ine			

Reasons for Res multistep's Superiority:

- **Multi-Step Refinement**: Res multistep refines the denoising process in multiple sub-steps, leading to enhanced structural stability [6].
- **Artifact Reduction**: This iterative approach helps reduce high-frequency artifacts, resulting in a significantly improved SSIM.
- **Temporal Consistency**: Improved frame-to-frame coherence is achieved, although further work is needed to boost VMAF scores.

6. Conclusion and Future Work

6.1. Final Summary

Under identical generation conditions using NVIDIA Cosmos, the Res multistep sampler outperforms the Euler ancestral sampler in terms of structural similarity (SSIM) and shows a slight advantage in PSNR. Although both methods yield near-zero VMAF scores, the substantial improvement in SSIM suggests that Res multistep produces more coherent and structurally faithful videos. Consequently, for applications where structural fidelity is paramount, Res multistep is the recommended sampling method.

6.2. Future Work

To further advance this research, future work will focus on:

- **Parameter Exploration**: Varying the number of diffusion steps and CFG scale to examine broader performance trade-offs.
- **Additional Metrics**: Incorporating video FID and temporal flow consistency to better capture perceptual quality.
- **Reference Alignment**: Employing more suitable or domain-specific reference videos to yield more meaningful VMAF evaluations.
- **Real-Time Enhancement**: Investigating algorithmic and hardware optimizations to achieve near-real-time text-to-video generation.

Current key parts of the code are available in the GitHub repository: https://github.com/JadeeeZh/Text-to-Video-Generation.git.

References

1. Li, F. Efficient Adaptive Parameter Tuning for Real-Time Text-to-Video Generation using NVIDIA Cosmos Diffusion Models. In Proceedings of the Milestone Report, 2023, pp. 93–95.
2. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.
3. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International conference on machine learning. pmlr, 2015, pp. 2256–2265.

 doi:10.20944/preprints202503.1432.v1

4. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **2020**, *33*, 6840–6851.
5. Nichol, A.Q.; Dhariwal, P. Improved denoising diffusion probabilistic models. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 8162–8171.
6. Karras, T.; Aittala, M.; Aila, T.; Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems* **2022**, *35*, 26565–26577.
7. Gonzalez, R.C. *Digital image processing*; Pearson education india, 2009.
8. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **2004**, *13*, 600–612.
9. Netflix. VMAF: Video Multi-Method Assessment Fusion. https://github.com/Netflix/vmaf, 2016. Accessed: 2023-10-01.