SOVEREIGN: How does the choice of LoRA rank in cross-attention layers influence the trade-off between FVD and LPIPS score

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

We present W.A.L.T, a transformer-based approach for photorealistic video generation via diffusion modeling. Our approach has two key design decisions. First, we use a causal encoder to jointly compress images and videos within a unified latent space, enabling training and generation across modalities. Second, for memory and training efficiency, we use a window attention architecture tailored for joint spatial and spatiotemporal generative modeling. Taken together these design decisions enable us to achieve state-of-the-art performance on established video (UCF-101 and Kinetics-600) and image

1 Introduction

Analysis of: Photorealistic Video Generation with Diffusion Models. Research goal: How does the choice of LoRA rank in cross-attention layers influence the trade-off between FVD and LPIPS scores across different temporal consistency metrics in Wan2.1 I2V-14B?.

2 Methodology

Multi-query arXiv search (1 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

6 papers retrieved. 12 claims extracted, 8 verified. Tribunal: 7.2/10 $\rightarrow$ REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
W.A.L.T is a transformer-based approach for photorealistic video generation via diffusion modeling.	✓	0.34
W.A.L.T uses a causal encoder to jointly compress images and videos within a unified latent space.	✓	0.27
The unified latent space in W.A.L.T enables training and generation across modalities.	✓	0.19
W.A.L.T uses a window attention architecture tailored for joint spatial and spatiotemporal generative modeling.	✓	0.27
W.A.L.T achieves state-of-the-art performance on the UCF-101 video generation benchmark.	✓	0.16
W.A.L.T achieves state-of-the-art performance on the Kinetics-600 video generation benchmark.	✓	0.16
W.A.L.T achieves state-of-the-art performance on the ImageNet image generation benchmark.	×	0.10
W.A.L.T achieves state-of-the-art performance without using classifier-free guidance.	✓	0.19
A cascade of three models was trained for text-to-video generation using W.A.L.T.	×	0.15
The text-to-video cascade consists of a base latent video diffusion model and two video super-resolution diffusion model	✓	0.31
The text-to-video cascade generates videos at a resolution of 512x896.	×	0.09
The text-to-video cascade generates videos at 8 frames per second.	×	0.15

References

- https://doi.org/10.48550/arxiv.2312.06662
- https://doi.org/10.48550/arxiv.2406.07686

- https://openalex.org/W7160969225