SOVEREIGN: For video diffusion transformers fine-tuned on high-resolution images, how does CAT's token allocation strateg

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

We present a practical pipeline for fine-tuning open-source video diffusion transformers to synthesize cinematic scenes for television and film production from small datasets. The proposed two-stage process decouples visual style learning from motion generation. In the first stage, Low-Rank Adaptation (LoRA) modules are integrated into the cross-attention layers of the Wan2.1 I2V-14B model to adapt its visual representations using a compact dataset of short clips from Ay Yapim's historical television film El Turco. This enables efficient domain transfer within hours on a single GPU. In the sec

1 Introduction

Analysis of: Fine-Tuning Open Video Generators for Cinematic Scene Synthesis: A Small-Data Pipeline with LoRA and Wan2.1 I2V. Research goal: For video diffusion transformers fine-tuned on high-resolution images, how does CAT's token allocation strategy affect end-to-end inference time and output quality compared to uniform tokenization under different spatial-temporal redundancy levels, measured by throughput and LPIPS on UHD benchmarks?.

2 Methodology

Multi-query arXiv search (4 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

11 papers retrieved. 16 claims extracted, 3 verified. Tribunal: 6.1/10 → 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
Diffusion transformers have evolved into powerful spatio-temporal generators capable of producing coherent multi-second	×	0.06
Open-source efforts such as VideoCrafter, ModelScope, and Wan2.x have narrowed the gap with commercial systems like Runw	×	0.05
Cinematic generation remains mostly inaccessible to small studios or independent creators.	×	0.05
The complete training and inference pipeline is released to support reproducibility and adaptation across cinematic doma	✓	0.27
Quantitative and qualitative evaluations using FVD, CLIP-SIM, and LPIPS metrics demonstrate measurable improvements in c	✓	0.32
Lightweight parallelization and sequence partitioning strategies are applied to accelerate inference without quality deg	✓	0.19
The model uses LoRA rank 8 / α 16 configuration.	×	0.07
Learning rate is set at 3×10.	×	0.03
The optimizer used is AdamW ($\beta 1 = 0.9$, $\beta 2 = 0.999$, wd=0.01).	×	0.01
Batch size is configured as 1 video × grad-acc 4 = 2 effective.	×	0.03
Precision is set to bf16.	×	0.02
Activation checkpointing is enabled.	×	0.02
Framework used is PyTorch + DeepSpeed (FSDP).	×	0.02
Cosine schedule with 5% warm-up is used.	×	0.02
Early stopping applied at LPIPS plateau.	×	0.02
Inference time on a single A100-80GB is 187 seconds with 1.0× speedup.	×	0.03

References

- http://arxiv.org/abs/2510.27364v1
- http://arxiv.org/abs/2601.06142v1
- http://arxiv.org/abs/2412.16117v1