Parameter-Efficient Fine-Tuning and Temporal Stability in Text-to-Video Generation Models

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Do parameter-efficient fine-tuning methods like LoRA in text-to-video generation models achieve comparable temporal stability metrics (e.g., FVD-128, FID-128) to full fine-tuning when evaluated on. 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. 14 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Fine-Tuning Open Video Generators for Cinematic Scene Synthesis: A Small-Data Pipeline with LoRA and Wan2.1 I2V. Research question: Do parameter-efficient fine-tuning methods like LoRA in text-to-video generation models achieve comparable temporal stability metrics (e.g., FVD-128, FID-128) to full fine-tuning when evaluated on the VidComp benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.8/10.

3 Results

12 papers retrieved. 14 claims extracted; 2 independently verified. Quality review score: 4.8/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The fine-tuning pipeline uses a LoRA rank of 8 and an alpha value of 16.	×	0.08
The learning rate used for training is 3 $\times$ 10-5 with a cosine schedule and 5% warm-up.	×	0.02
The optimizer used is AdamW with beta1=0.9, beta2=0.999, and weight decay=0.01.	×	0.00
The effective batch size is 2, calculated as 1 video multiplied by a gradient accumulation of 4.	×	0.02
The model was trained for 4000 steps using bf16 precision.	×	0.04
Activation checkpointing is enabled to reduce the VRAM footprint.	×	0.02
The framework used is PyTorch combined with DeepSpeed (FSDP).	×	0.01
Training employs early stopping based on the LPIPS plateau.	×	0.02
Configuration time on a single A100-80GB GPU is 187 seconds.	×	0.02
The method generates coherent 720p video sequences.	×	0.08
Evaluations were conducted using FVD, CLIP-SIM, and LPIPS metrics.	✓	0.16
The approach demonstrates measurable improvements in cinematic fidelity and temporal stability over the base model.	✓	0.20
Diffusion transformers have evolved to produce coherent multi-second videos from textual descriptions.	×	0.05
Open-source efforts such as VideoCrafter, ModelScope, and Wan2.x have narrowed the performance gap with commercial syste	×	0.05

References

- http://arxiv.org/abs/2510.27364v1
- http://arxiv.org/abs/2411.14961v3
- http://arxiv.org/abs/2602.05988v1