LoRA Rank Effects on Temporal Consistency and Perceptual Quality in Wan2.1 I2V-14B Video Synthesis

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the choice of LoRA rank in Wan2.1 I2V-14B influence its ability to preserve temporal consistency (measured via FVD) versus perceptual quality (measured via R-LPIPS) in long-form human video. 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. 17 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.3/10. 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: How does the choice of LoRA rank in Wan2.1 I2V-14B influence its ability to preserve temporal consistency (measured via FVD) versus perceptual quality (measured via R-LPIPS) in long-form human video synthesis tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.3/10.

3 Results

14 papers retrieved. 17 claims extracted; 2 independently verified. Quality review score: 5.3/10.

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 study utilizes a LoRA rank of 8 and an alpha value of 16 for fine-tuning.	×	0.09
The learning rate used in the experiments is 3 $\times$ 10-5 with a cosine schedule and 5% warm-up.	×	0.06
The optimizer configured is AdamW with beta1=0.9, beta2=0.999, and weight decay=0.01.	×	0.01
The effective batch size is 2, calculated as 1 video multiplied by a gradient accumulation of 4.	×	0.03
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, utilizing Fully Sharded Data Parallelism (FSDP).	×	0.03
Training employs early stopping based on the LPIPS plateau.	×	0.02
On a single A100-80GB GPU, the configuration time is 187 seconds.	×	0.03
The single A100-80GB setup serves as the 1.0$\times$ speedup baseline.	×	0.01
The pipeline expands inputs into coherent 720p video sequences.	×	0.09
Evaluations were conducted using FVD, CLIP-SIM, and LPIPS metrics.	✓	0.16
A small expert user study was conducted to support the quantitative evaluations.	×	0.14
The results demonstrate measurable improvements in cinematic fidelity and temporal stability over the base model.	✓	0.23
Diffusion transformers have evolved to produce coherent multi-second videos from textual descriptions.	×	0.05
Open-source efforts including VideoCrafter, ModelScope, and Wan2.x have narrowed the performance gap with commercial sys	×	0.04
Cinematic generation capabilities, such as film-like motion and controlled lighting, remain mostly inaccessible to small	×	0.08

References

- http://arxiv.org/abs/2410.05203v2
- http://arxiv.org/abs/2510.27364v1
- http://arxiv.org/abs/2405.03150v2