Winter Internship Project
Report

Advances in Video Generation Models: Comparative analysis and application in 3D/4D reconstruction

Submitted by

Anmol Agarwal
Bachelors of Science in Engineering Sciences
Indian Institute of Science, Education and Research, Bhopal

Under the guidance of

Dr. Venkatesh Babu
PhD, Indian Institute of Science, Bangalore

Computational and Data Science Department
Indian Institute of Science
Bangalore

Winter Internship 2024

Abstract

In this internship report, I present a comparative analysis of current text-to-video (T2V) and image-to-video (I2V) generation models, focusing on CogVideoX-5B, Mochi-1, and HunyuanVideo. Through an extensive literature review, I studied recent innovations like CausVid's approach to real-time video generation and DimensionX's novel ST-Director for 3D/4D scene generation from single images.

My evaluation of these models across various conditions reveals that HunyuanVideo achieves superior photorealistic generation and physics understanding, even under FP8 quantization, while CogVideoX-5B performs well in simpler scenes despite its smaller size.

Through exploring different techniques, I propose future work for combining ReconX's point cloud conditioning from Dust3r with CausVid's real-time generation for fast, autoregressive 3D-consistent I2V generation, while exploring depth map integration for better 4D reconstruction.

Contents

1 Objective 1

1.1 Objective 1

2 Introduction 2

2.1 Background for Latent Diffusion models 2

2.2 Transformer based backbone 2

2.3 Emergence of T2V models 2

3 Work Done 3

3.1 Qualitative Analysis and comparison 3

3.1.1 Test Conditions 3

3.1.2 Comparison Study 4

3.1.3 Image-to-Video testing 5

3.1.4 Visual Comparison of Generated Videos 5

3.2 Detailed Study of Model Architectures 6

4 Literature Survey 7

4.1 Papers Studied 7

4.1.1 CausVid - From Slow Bidirectional to Fast Causal Video Generators [7] 7

4.1.2 DimensionX: High-Fidelity 3D and 4D Scene Generation from a Single Image [6] 8

5 Conclusion 9

5.1 T2V Comparison Results 9

5.2 Future Work 9

Acknowledgements 10

References 11

Chapter 1

Objective

1.1 Objective

- Conduct an exhaustive literature review on existing video diffusion models to understand their evolution and current state.
- Explore public text-to-video (T2V) and image-to- models for temporally and spatially consistent generation
- Evaluate model performance and architectural innovations
- Analyze model strengths, limitations, and potential improvements
- Find potential future work ideas for application in 3D tasks inlcuding 3D consistent generation

Chapter 2

Introduction

2.1 Background for Latent Diffusion models

Latent Diffusion models [1] are generative models used primarily for image generation and other computer vision tasks. Diffusion-based neural networks are trained through deep learning to progressively “diffuse” samples with random noise, then reverse that diffusion process to generate high-quality images.

2.2 Transformer based backbone

Following the paper ”Scalable Diffusion Models with Transformers” [2], which introduced DiT as a backbone for noise prediction replacing the classic U-Net approach, results showed that DiT performance scales effectively with compute, enabling more consistent image generation.

2.3 Emergence of T2V models

The introduction of papers like ”Latte: Latent Diffusion Transformer for Video Generation” [3] and ”Sora: Video Generation Models as World Simulators” marked early explorations of Diffusion Transformers (DiT) in video generation. The recent emergence of models like Sora has demonstrated the true potential of transformer architectures in this domain. Operating at unprecedented scale with extensive training data, these models exhibit remarkable capabilities in simulating physical world dynamics, suggesting that transformer-based architectures can develop strong world priors for coherent video generation. -

Chapter 3

Work Done

3.1 Qualitative Analysis and comparison

3.1.1 Test Conditions

Model evaluation was performed under various test conditions to assess generation capabilities.

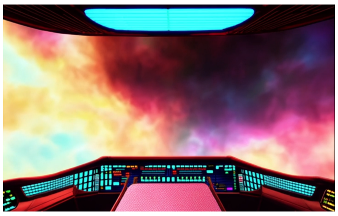
- Static scenes: Landscapes, still-life compositions
- Dynamic scenes: Moving objects, natural phenomena
- Subject-action pairs: Specific actions performed by subjects
- Human subjects: People in various poses and activities
- Physics understanding: Object Interactions and Natural Motion

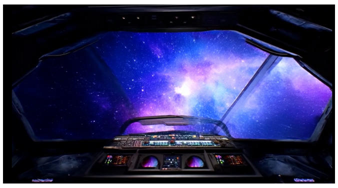
Following this systematic testing approach, a detailed analysis was performed on:

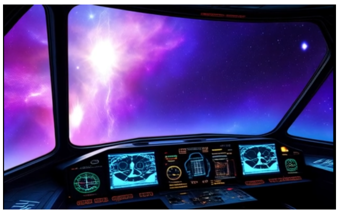
- Comparison of three open-source text-to-video generation models:
  - CogVideoX-5B (fp16) [4]
  - Mochi-1 (fp8)
  - HunyuanVideo (fp8) [5]
- Run on a RTX 3090 with 24GB VRAM

3.1.2 Comparison Study

- **Static Scene Generation**

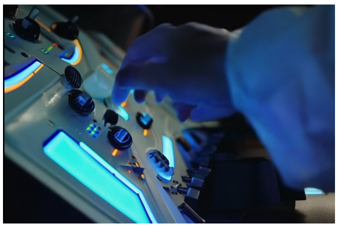

CogVideoX-5B

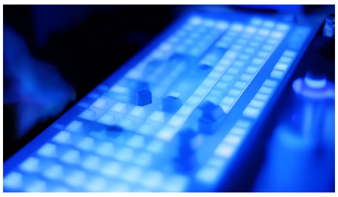

Mochi 1-preview

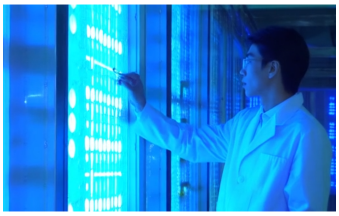

HunyuanVideo

Prompt: A spaceship cockpit overlooking a vibrant nebula through a massive wind

- **Subject → Action**

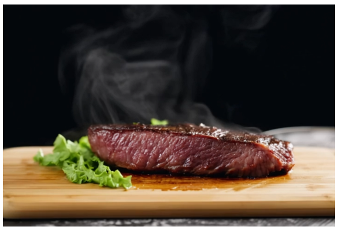

CogVideoX-5B

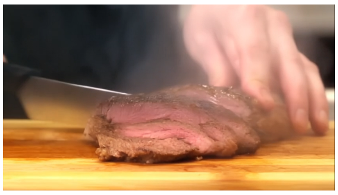

Mochi 1-preview

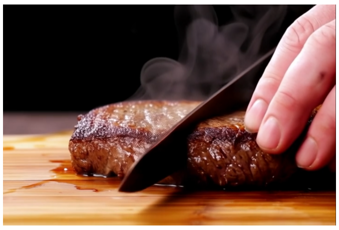

HunyuanVideo

Prompt: A scientist adjusts the dials on a glowing control panel inside a high-tech lab.

- **Complex Scene - Physics Understanding**


CogVideoX-5B

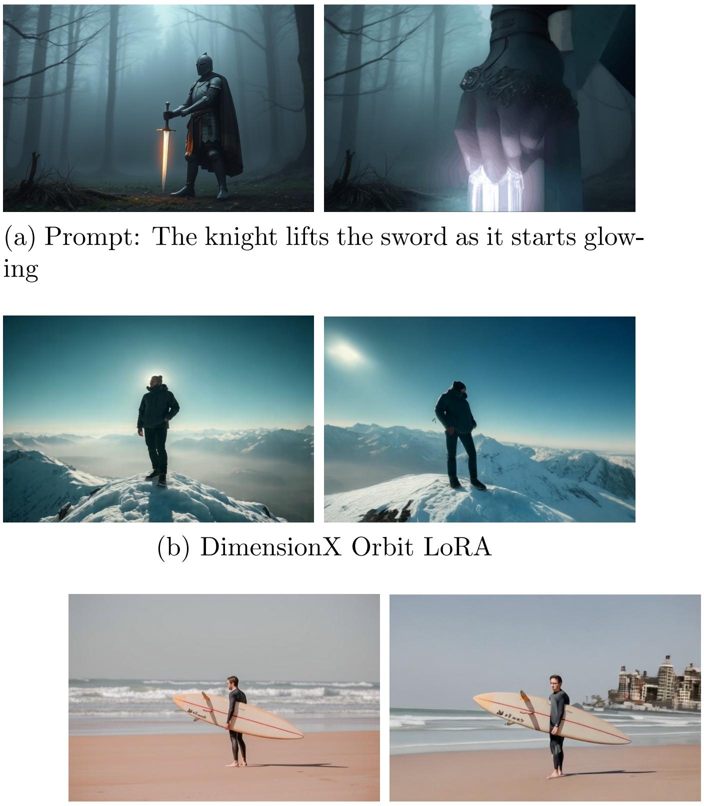

Mochi 1-preview


HunyuanVideo

Prompt: A pair of hands skillfully slicing a perfectly cooked steak on a wooden cutting board, faint steam rising from it.

3.1.3 Image-to-Video testing

3.1.4 Visual Comparison of Generated Videos



(a) Prompt: The knight lifts the sword as it starts glowing

(b) DimensionX Orbit LoRA

(c) DimensionX Orbit LoRA

Figure 3.1: Visual examples of generated videos showing the starting frame and a sample from the generated frames.

I explored image-to-video (I2V) generation using these two approaches.

- **CogVideoX Native I2V**: Since CogVideoX has native support for image-to-video generation, I performed initial tests by generating videos from various single image inputs. These tests provided a baseline understanding of CogVideoX's capability in the I2V task.

- **DimensionX LoRA I2V**: Additionally, I experimented with the DimensionX LoRA for single-image-to-4D video generation. The LoRA enabled the creation of rotating camera views from static images. However, I observed that this process is prone to certain limitations:

- **Deformations:** Deformations were commonly encountered, particularly in the generated motion patterns of objects within the video.
- **Instability:** The generated videos exhibited noticeable instability, often appearing shaky and containing visual artifacts.

Despite the instabilities and deformations observed, these experiments demonstrated the potential of image-to-video generation using a combination of the base model and external LoRA approaches.

3.2 Detailed Study of Model Architectures

Architectural Details

- Side-by-side comparison of model attributes (table or list):

Attribute	**CogVideoX-5B**	**Mochi-1**	**HunyuanVideo**
Core Architecture	Expert transformer with 3D attention	AssymDiT with multi-modal Self Attention	Dual-to-Single-stream hybrid model design
Params Size	5B	13B	10B
VAE compression	$8p \times 8p \times 4f \times 16channels$	$8p \times 8p \times 6f \times 12channels$	$8p \times 8p \times 4f \times 16channels$
VRAM usage (without optimisations)	24GB	60GB	45GB - 60GB

Trade-Offs and Conclusion:

- Having a bigger model makes fine-tuning on a single machien and inference on consumer hardware very difficult. The inference time also increases (around 300 seconds for 60 frames) which isn't suitable.
- Small models struggle with temporal consistency and object permanence, but running big models like hunyuan and mochi quantized is also unstable.

Chapter 4

Literature Survey

4.1 Papers Studied

4.1.1 CausVid - From Slow Bidirectional to Fast Causal Video Generators [7]

- Problem: High-quality video generators are slow and not interactive, taking minutes to create short videos.
- Why Existing Solutions Fail: Speed-up attempts often sacrifice quality or limit video length to 2-5 seconds.
- CausVid Solution:
  - Faster: Processes videos in small ”chunks,” one after another.
  - ”Teacher-Student”: A slow, high-quality generator teaches a faster one.
  - Causal Training: The faster generator learns to only use past video information.
  - Speed Memory: Uses a ”memory” (KV-caching) for faster generation.
- Results:
  - Real-time Speed: Generates videos at 9 FPS on one GPU.
  - Longer Videos: Creates videos longer than training data.
  - Versatile: Handles image-to-video and live editing.
  - Training: Took 2 days on 64 GPUs.
  - Minor Issues: Some visual glitches and smoothness limitations.

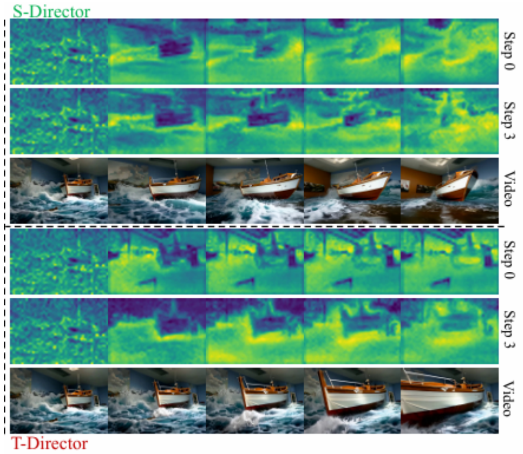


Figure 4.1: Observation - early denoising steps capture spatial information!

- Spatial director - moving camera/still scene
- Temporal director - still camera/moving objects
- No expensive fine tuning needed! - Combine different LoRAs

4.1.2 DimensionX: High-Fidelity 3D and 4D Scene Generation from a Single Image [6]

- **Problem:** Generating consistent, high-fidelity 3D and 4D scenes from a single image using controllable video diffusion is challenging due to inconsistencies, limited control, and a lack of suitable datasets.

- **DimensionX Approach:** DimensionX introduces a novel ST-Director to decouple spatial and temporal control within a video diffusion model, using learned LoRAs.
  - It generates multi-view videos for complete scene reconstruction, employing latent sharing and appearance refinement to ensure 4D consistency.
  - Efficient 3D/4D scene construction is achieved through Gaussian splatting.
  - A "Switch-Once" strategy efficiently combines spatial and temporal control.

- **Results:** Produces photorealistic videos and 3D/4D scenes with precise control over spatial and temporal dimensions, outperforming prior work on various datasets.

- **Limitations:** Inference speed and fine-grained camera control require further improvement.

Chapter 5

Conclusion

5.1 T2V Comparison Results

- No single model consistently excels in photorealism and 3D-coherence.
- **Hunyuan:** Consistently produces photorealistic results, even at FP8 quantization. Shows strong physics understanding (see steak-cutting example – *Section 3.1.2*).
- **Mochi:** Highly inconsistent performance; sometimes excellent, sometimes worse than CogVideoX, possibly due to quantization artifacts.
- **CogVideoX:** Good T2V generations relative to its smaller size; good prompt following in simpler scenes. Struggles with complex scenes where Hunyuan excels.
  I2V generations are unstable, with poor prompt following, frequent distortions, and struggles with object permanence.
  However, using DimensionX LoRA improves 3D scene understanding, shadows, and physics (e.g., coffee latte art pouring).

5.2 Future Work

- Address DimensionX's struggles with 4D consistency in multi-view videos, especially at edges. Explore using depth maps as input or conditioning to improve 4D reconstruction.
- Investigate combining ReconX (using Dust3r point clouds) with CausVid for real-time, autoregressive, 3D-consistent I2V generation.

Acknowledgment

I would like to express my sincere gratitude to **Dr. Venkatesh Babu** for providing me with the opportunity to undertake this internship at the **Video Analytics Lab, IISc Bangalore**. His support enabled me to explore advancements in video diffusion models and attend my first computer vision conference, **IVGIP 2024**, which was an inspiring experience.

I extend my heartfelt thanks to my mentor, **Mr. Ankit Dhiman**, Ph.D. scholar at VAL Lab, for his invaluable guidance, mentorship, and support throughout this internship. His insights and constructive feedback were instrumental in shaping my understanding of the research process.

This internship has been a transformative experience, offering me invaluable learning opportunities such as reading research papers critically, understanding how research is conducted in a serious laboratory environment, formulating meaningful problem statements, and sharing ideas with accomplished researchers. The exposure to diverse perspectives and interactions with experts has greatly enriched my academic journey.

I would also like to express my gratitude to the **Indian Institute of Science (IISc)** for fostering an environment that promotes cutting-edge research and learning. Finally, I am thankful to all the colleagues and researchers I met during this journey for their insightful discussions and encouragement.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, et al., ”High-Resolution Image Synthesis with Latent Diffusion Models,” *arXiv preprint arXiv:2112.10752*, 2021. Available: `https://arxiv.org/abs/2112.10752`

[2] William Peebles and Saining Xie, ”Scalable Diffusion Models with Transformers,” *arXiv preprint arXiv:2212.09748*, 2023. Available: `https://arxiv.org/abs/2212.09748`

[3] Xin Ma, Yaohui Wang, Gengyun Jia, et al., ”Latte: Latent Diffusion Transformer for Video Generation,” *arXiv preprint arXiv:2401.03048*, 2024. Available: `https://arxiv.org/abs/2401.03048`

[4] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, et al., ”CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer,” *arXiv preprint arXiv:2408.06072*, 2024. Available: `https://arxiv.org/abs/2408.06072`

[5] Weijie Kong, Qi Tian, Zijian Zhang, et al., ”HunyuanVideo: A Systematic Framework For Large Video Generative Models,” *arXiv preprint arXiv:2412.03603*, 2025. Available: `https://arxiv.org/abs/2412.03603`

[6] Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, Yikai Wang, ”DimensionX: Create Any 3D and 4D Scenes from a Single Image with Controllable Video Diffusion” *arXiv preprint arXiv:2411.04928*, 2024. Available: `https://arxiv.org/abs/2411.04928`

[7] Tianwei Yin, Qiang Zhang, Richard Zhang, et al., ”From Slow Bidirectional to Fast Autoregressive Video Diffusion Models,” *arXiv preprint arXiv:2412.07772*, 2025. Available: `https://arxiv.org/abs/2412.07772`