NATIONAL INSTITUTE OF TECHNOLOGY, AP

Consistenti2v: Elevating Image-to-Video Generation

EC449 Major Project Work A – I Mid Review

PROJECT GUIDE : **Dr. ARUN KUMAR A**

TEAM MEMBERS	ROLL NUMBER
Mamilla Vishnu Teja	621208
Kuruva Pushparaj	621174
Vabbina Deveswar Naidu	621262

Literature Review

Title	Methodology	Remarks
[1]ImaGINator: Conditional Spatio-Temporal GAN for Video Generation	• **Spatio-Temporal GAN**: Merges spatial image features with temporal dynamics for realistic video generation. • **Two-Stream Network**: Processes spatial and temporal information separately to ensure frame coherence. • **Attention Mechanisms**: Enhances video quality by dynamically focusing on relevant features while preserving motion continuity.	(+) The use of conditional architectures and attention mechanisms significantly enhances the realism and coherence of generated videos, making them more suitable for practical applications (-) Maintaining **long-term consistency** in video generation (where objects and scenes must persist coherently over many frames) can be challenging for GAN-based models, which are often better at handling short video clips than long sequences.
[2]Spatiotemporal Consistency Enhancement for Video Representation Learning	• **Self-Supervised Learning**: Enhances video representation for spatiotemporal consistency. • **Contrastive Learning**: Maximizes similarity between views of the same video to learn invariant features. • **Temporal Transformations**: Employs temporal augmentations to extract robust features that are invariant to motion, improving consistency in recognizing motion patterns.	(+) The integration of self-supervised learning and novel conditioning mechanisms represents a significant step forward in the field, providing new insights into video representation and generation (-) Since the model is self-supervised, it might not learn certain task-specific features that would be learned through supervised methods
[3]Faster Image2Video Generation: Impact of CLIP Image Embedding	• **CLIP Embedding**:Uses CLIP (Contrastive Language-Image Pretraining) to extract rich, semantically meaningful features from the input image, providing a strong foundation for video generation. • **Computational Efficiency**:Reducing computations by removing TCA, and replacing SCA by linear layer , improving speed of video generation.	(+) CLIP embeddings capture rich visual and semantic features from images, contributing to the generation of videos with improved aesthetic appeal. (-) While CLIP embeddings enhance the visual quality of individual frames, they may not significantly improve temporal consistency across frames.

Motivation to work on this project

- Growing demand for **advanced video generation techniques** to produce high-quality, coherent videos from static images.
- **Current models** often lack visual consistency and smooth motion, limiting their effectiveness in practical use.
- Our model aims to address these challenges by improving motion fluidity and consistency in video generation.
- The project will contribute to the **advancement of video generation technology**, providing more effective tools for **creative professionals**.

Problem Statement

•**Visual Consistency Issues**: Current image-to-video (I2V) generation methods struggle to maintain consistent subjects, backgrounds, and styles across video frames, leading to visual inconsistencies and degraded video quality.

•**Abrupt Motions**: Generated videos often suffer from abrupt motions and disjointed transitions, reducing the realism and smoothness of motion.

•**Spatiotemporal Challenges**: Traditional models lack the ability to effectively capture both spatial and temporal relationships in video sequences.

•**Need for Improved Framework**: A more robust model is required to generate high-quality, coherent video sequences from an initial image, maintaining visual consistency and storytelling fluidity.

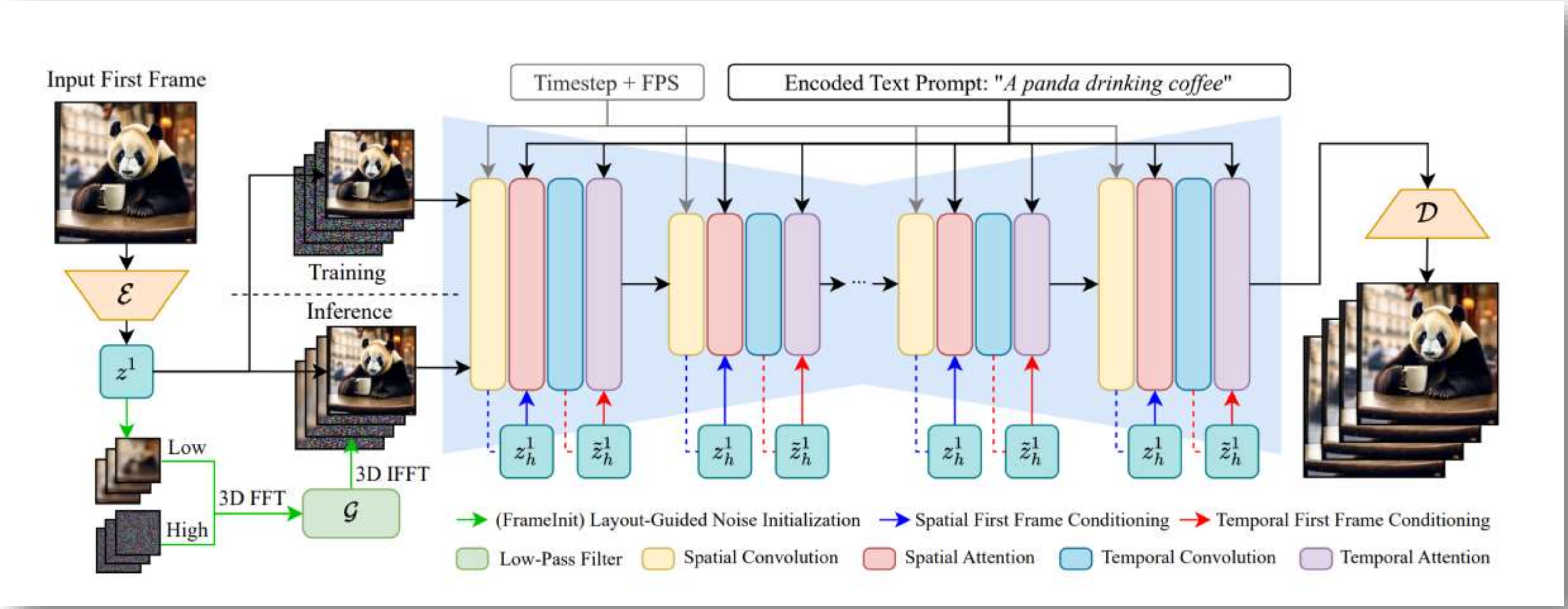
Objectives

1. Enhance Visual Consistency: Develop a new model to maintain visual coherence throughout the video generation process, ensuring the integrity of subjects, backgrounds, and styles.

2. Implement Spatiotemporal Attention Mechanisms: Utilize advanced spatiotemporal attention layers to ensure smooth transitions and spatial coherence across generated video frames.

3. Optimize Noise Initialization: Introduce the FrameInit strategy to leverage low-frequency components from the initial frame, stabilizing video generation and improving layout consistency.

Block Diagram



Methodology

Base Architecture (Text-to-Image U-Net):

The video generation model builds on **latent diffusion models (LDMs)** for text-to-image generation, specifically referring to the work of **Rombach et al., 2022**.

- **U-Net** is a common architecture for image generation tasks. It consists of:
  - **Downsampling blocks**:
  - **Upsampling blocks**
  - **Skip connections**:

Since video generation requires temporal consistency (across frames), the architecture is modified to handle not just **spatial dimensions** (height and width) but also the **temporal dimension** (time across frames).

Standard Temporal Self-Attention:

The intermediate hidden state of the video is represented by $\bar{z} \in \mathbb{R}^{(H \times W) \times N \times C}$, where:

- $H \times W$ are the spatial dimensions (height and width).
- N is the number of frames.
- C is the number of channels (features).

RoPE (Rotary Positional Embedding):

RoPE (Rotary Position Embedding), introduced by **Su et al., 2024**, is used to inject positional information to temporal layers.This helps to understand where a frame is located in the time sequence.

Limitation of standard temporal attention mechanism and how its addressed

•Traditional temporal attention mechanisms in video models track individual pixels over time but fail to consider nearby areas, risking loss of tracking moving objects.
•**Proposed Solution**: The window-based temporal feature conditioning approach incorporates features from a K×K window arouneach spatial position in the first frame into the query, key, and value matrices, enabling the model to better track object movement by considering neighboring pixels.

Spatial Self-Attention:

- **Self-Attention Layers**: The U-Net architecture contains **spatial self-attention layers**, which calculate attention across different spatial positions within each frame independently. This allows the model to focus on important features (like a panda's face or cup in the video) when generating a new frame.

- **Cross-Frame Attention**: In the process, the model also employs **cross-frame attention mechanisms**, which help maintain consistency by comparing features across multiple frames.

Fine-Grained Spatial Feature Conditioning:

The process described above provides **fine-grained conditioning** of future frames based on the first frame. This means that during the generation of subsequent frames, the model has access to all the detailed spatial information from the first frame.

Frequency Decomposition in Video Generation

Videos can be decomposed into different **frequency bands**:

- The **high-frequency component** contains the fine details and captures fast-moving objects.
- The **low-frequency component** represents the **coarse layout**, slow-moving parts, and general structure of the video.

Frequency Decomposition

$$F_{low}(z_T) = FFT_3D(z_T) * G(D)$$

$$F_{high}(z_T) = FFT_3D(z_T) * (1 - G(D))$$

Here, G(D0) is a **Gaussian low-pass filter** that isolates the low-frequency part of the signal, while (1−G(D0)) isolates the high-frequency part.

Combining Frequencies: The low-frequency information is combined with the high-frequency noise:

$$\epsilon' = IFFT_3D(F_{low}(z_\tau) + F_{high}(\epsilon))$$

IFFT 3 D is the inverse FFT that transforms the combined frequency components back into the spatial domain.

INPUT DATASET : *WebVid-10M Dataset* Overview ;

It comprises 10 million diverse low-resolution videos, each paired with descriptive text and featuring a fixed-position watermark. It serves as a key resource for training models in video synthesis, captioning, and video-to-text learning.

EVALUATION PARAMETERS

- **Fréchet Video Distance (FVD)**
- **CLIP Similarity (CLIPSIM)**
- **Temporal Flickering**
- **Dynamic Degree**
- **Background Consistency , Subject Consistency**

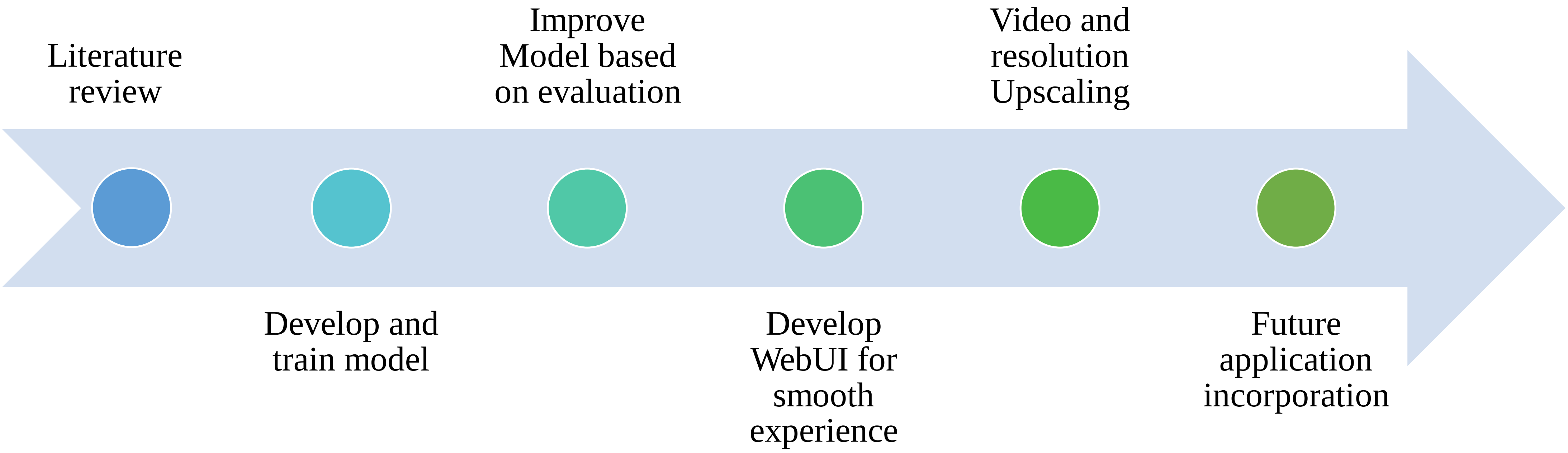
RESULTS OBTAINED SO FAR

Input Frame

Generated video by our skeleton model

Resolution : 256 * 256
Time taken to generate: 15 minutes
System config used GPU : RTX 4060 (max wattage 50W)

ROADMAP


Literature review
Develop and train model
Improve Model based on evaluation
Develop WebUI for smooth experience
Video and resolution Upscaling
Future application incorporation

References

- **[1]Fox et al., 2021**: *Towards Realistic Video Generation with GANs*
- **[2]Brooks et al., 2022**: *Generative Video with Autoregressive Transformers*
- **[3]Tian et al., 2021**: *Towards Real-Time Video Generation*
- **[4]Wang et al., 2020**: *ImaGINator: Conditional Spatio-Temporal GAN for Video Generation*
- **[5]Bi, S., Hu, Z., Zhao, M. *et al*.** *Spatiotemporal consistency enhancement self-supervised representation learning for action recognition*
- **[6]Taghipour, A., Ghahremani, M., Bennamoun, M., Rekavandi, A. M., Li, Z., Laga, H., & Boussaid, F.** (2024). *Faster Image2Video Generation: A Closer Look at CLIP Image Embedding's Impact on Spatio-Temporal Cross-Attentions.*

THANK YOU

Thank You