NATIONAL INSTITUTE OF TECHNOLOGY, AP

Consistenti2v: Elevating Image-to-Video Generation

EC449 Major Project Work A – I End 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.

Literature Review (2)

Title	Methodology	Issues solved by our model
[4] Emu-Video (Girdhar et al., 2023): Latent features concatenation for I2V conditioning.	• Both focus on first-frame conditioning mechanisms to guide video generation. • Emu-Video uses simple latent feature concatenation for conditioning, which is extended and improved in ConsistentI2V with cross-frame attention for better spatial and temporal consistency.	**Weak Fine-Grained Control**: ConsistentI2V introduces spatiotemporal attention mechanisms for fine-grained first-frame conditioning. **Jittery Motion**: Temporal layers in ConsistentI2V use local windows of first-frame features to improve motion coherence.
[5] Dynamicrafter (Xing et al., 2023): Cross-attention layers for improved consistency.	• Both methods incorporate cross-attention mechanisms to address consistency issues in I2V generation. • Dynamicrafter emphasizes smoother frame transitions, a challenge directly targeted by ConsistentI2V.	**Training Complexity:** ConsistentI2V's FrameInit reduces the need for complex temporal conditioning designs. **Resource Intensive**: Modular design optimizes efficiency, reducing the computational burden seen in Dynamicrafter.
[6] Moonshot (Zhang et al., 2024): Similar I2V enhancement techniques.	• Both use advanced conditioning mechanisms and focus on noise initialization for temporal stability. • ConsistentI2V builds on similar ideas with its FrameInit strategy to further stabilize training and inference.	**Complex Implementation**: ConsistentI2V introduces simpler, more modular methods to achieve temporal smoothness and spatial alignment. **Inference Speed**: FrameInit ensures efficient inference by leveraging low-frequency components, reducing computational demand.

Problem Statement

•**Inconsistent Visual and Motion Quality**: Existing Image-to-Video (I2V) generation methods struggle to maintain the integrity of subjects, backgrounds, and styles, leading to flickering and abrupt motion transitions that compromise the video narrative.

•**Limitations of Current Conditioning Techniques:** Current approaches to incorporating first-frame conditioning often fail to preserve local details and spatial-temporal coherence, resulting in appearance and motion inconsistencies in generated videos.

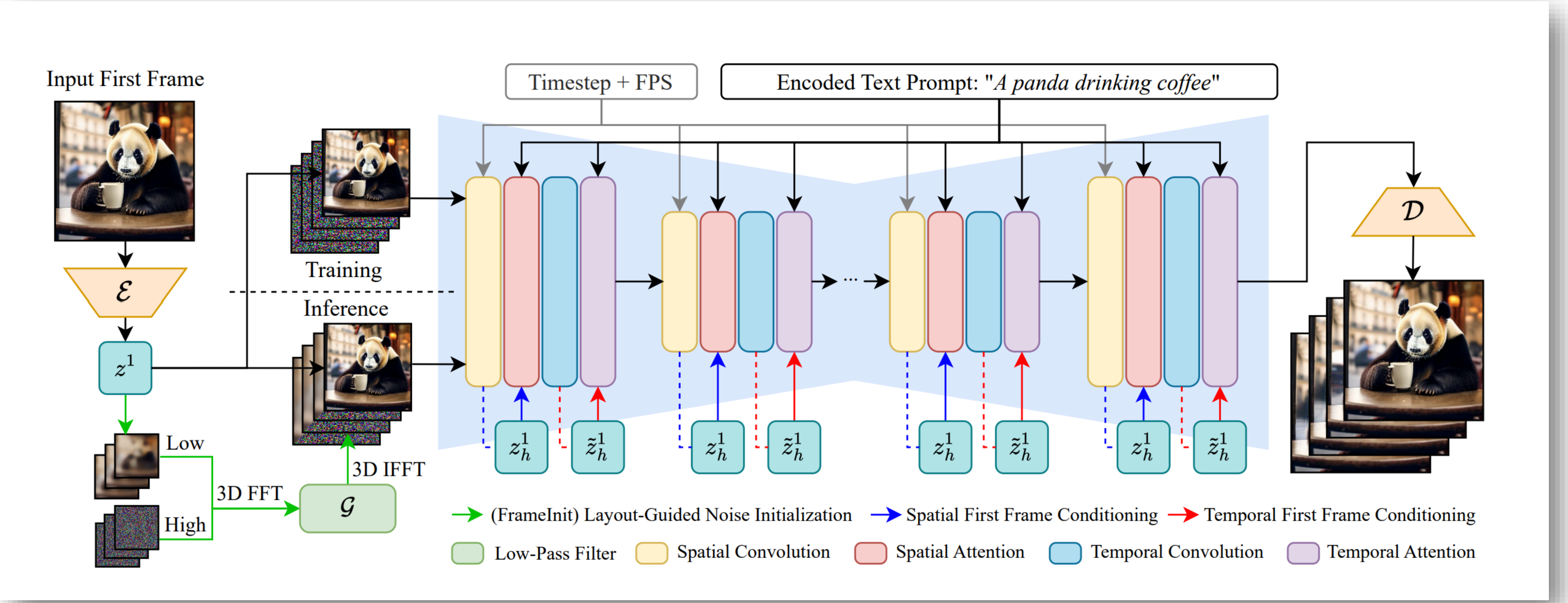
Objectives

1. **Enhancing Visual Consistency:** Develop a diffusion-based approach using spatiotemporal attention and low-frequency noise initialization to maintain subject integrity, background stability, and motion consistency in Image-to-Video (I2V) generation.

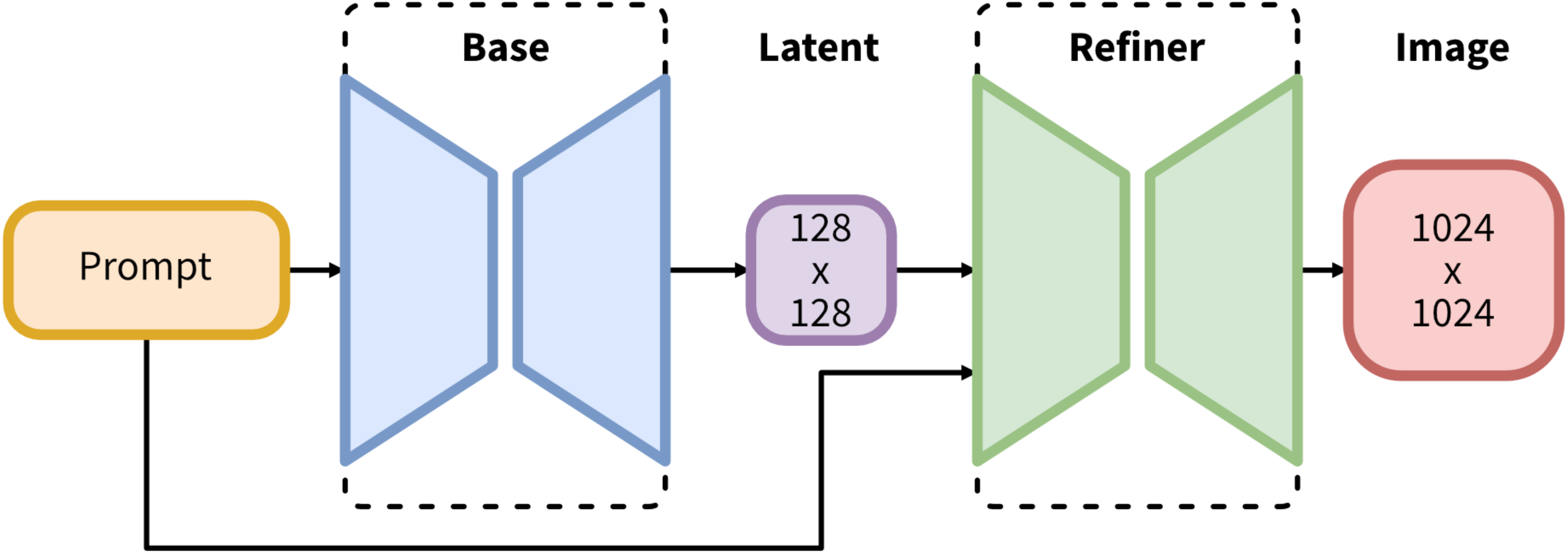
2. Implement Spatiotemporal Attention Mechanisms: Utilize advanced spatiotemporal attention layers to ensure smooth transitions and spatial coherence across generated video frames.

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 of ConsistentI2V



Block Diagram of Diffusion model for image gen



Methodology

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

Latent Diffusion models and Unet

Standard Temporal Self-Attention:

RoPE (Rotary Positional Embedding):

Limitation of standard temporal attention mechanism and how its addressed

Spatial Self-Attention:

Self attention layers and Cross attention layers

Fine-Grained Spatial Feature Conditioning

Guided Noise Initialization

Frequency Decomposition in Video Generation

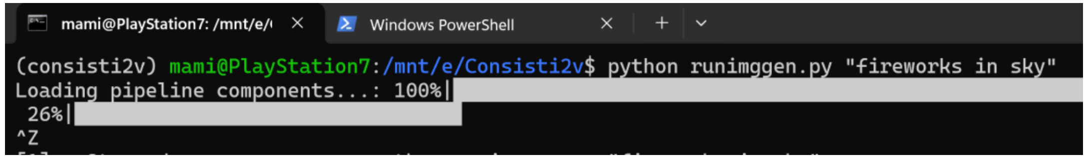
High Frequency and low frequency

Frequency Decomposition and Combining Frequencies

IMAGE GENERATED USING MODEL

"Fireworks in sky"

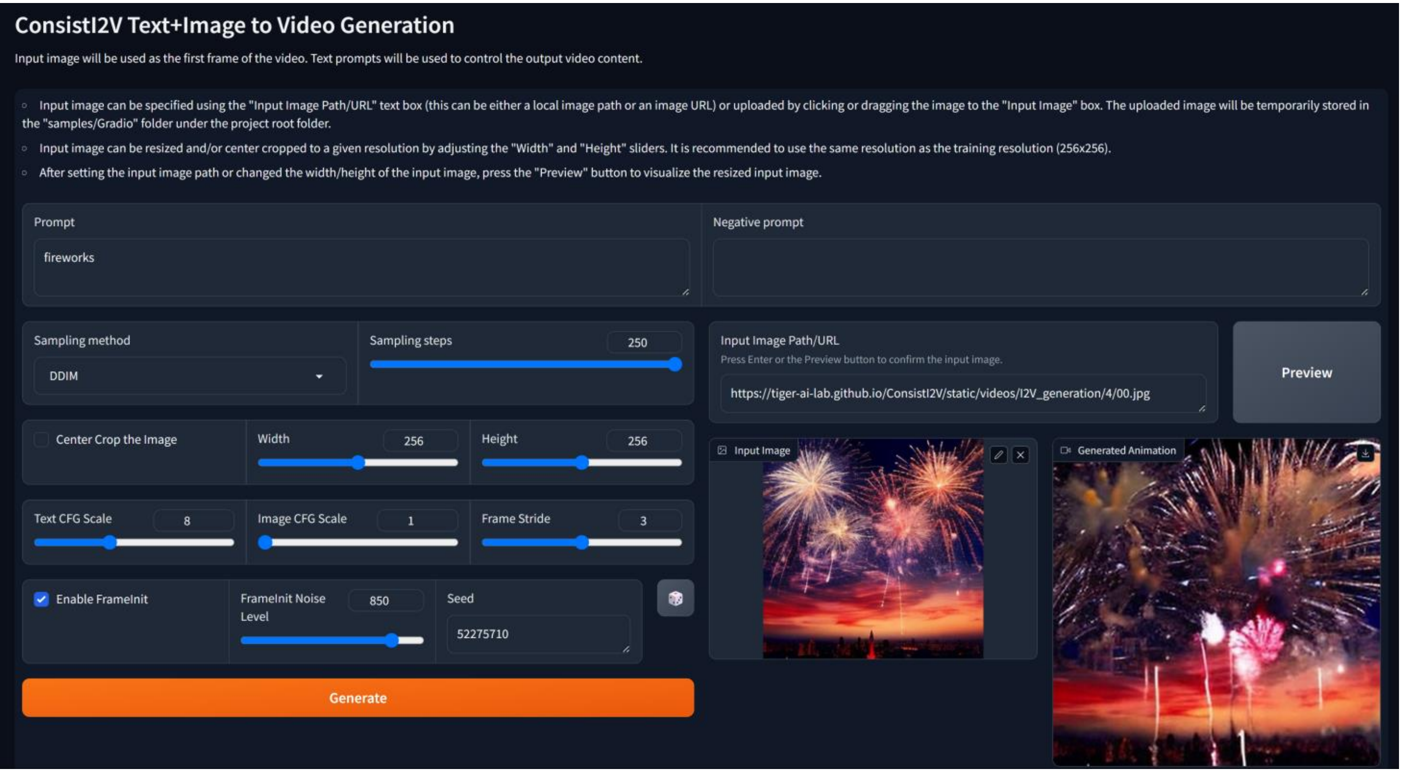
Given Prompt



Generated image by diffusion model

Resolution : 1024x1024 Time taken to generate: 32 minutes
System config used CPU : Intel Ultra 9 , RAM: 32 Gb 1024x1024

IMAGE IS FED TO CONSISTENTI2V MODEL



Used Gradio interface of the model

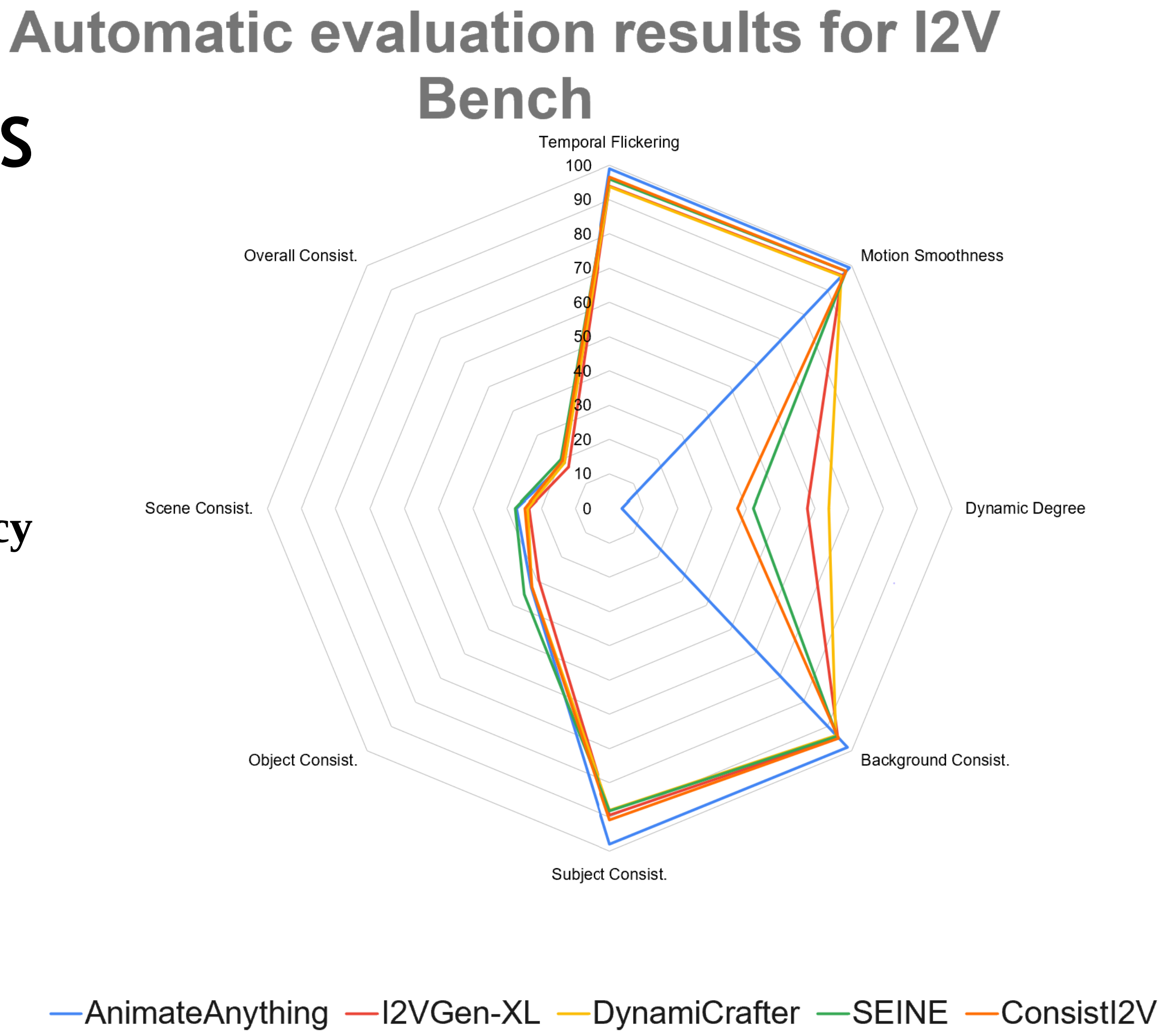
Generated video by ConsistentI2V

Resolution : 256x256 Time taken to generate: 32 minutes
System config used CPU : Intel Ultra 9 , RAM: 32 Gb
Running at 8 frames/sec

```
(consisti2v) mami@PlayStation7:/mnt/e/ConsistI2v$ ./run1.sh "fishes are swimming in sea"
Loading pipeline components...: 100%|                                        | 7/7 [00:00<00:00,  7
100%|                                                                  | 50/50 [27:11<00:00, 32
/mnt/e/ConsistI2v/consisti2v/models/rotary_embedding.py:35: FutureWarning: `torch.cuda.amp.autocast(args...)` is
ted. Please use `torch.amp.autocast('cuda', args...)` instead.
  @autocast(enabled = False)
/mnt/e/ConsistI2v/consisti2v/models/rotary_embedding.py:252: FutureWarning: `torch.cuda.amp.autocast(args...)` i
ated. Please use `torch.amp.autocast('cuda', args...)` instead.
  @autocast(enabled = False)
Loading pipeline components...: 100%|                                        | 5/5 [00:14<00:00,  2
100%|                                                                | 100/100 [01:21<00:00,  1
100%|                                                                  | 16/16 [00:00<00:00, 28
Video saved at: /mnt/e/ConsistI2v/generated_videos/2024-12-01T16-56-24/generated_video_16-58-04.mp4
Configuration saved at: /mnt/e/ConsistI2v/generated_videos/2024-12-01T16-56-24/config.json
(consisti2v) mami@PlayStation7:/mnt/e/ConsistI2v$
```

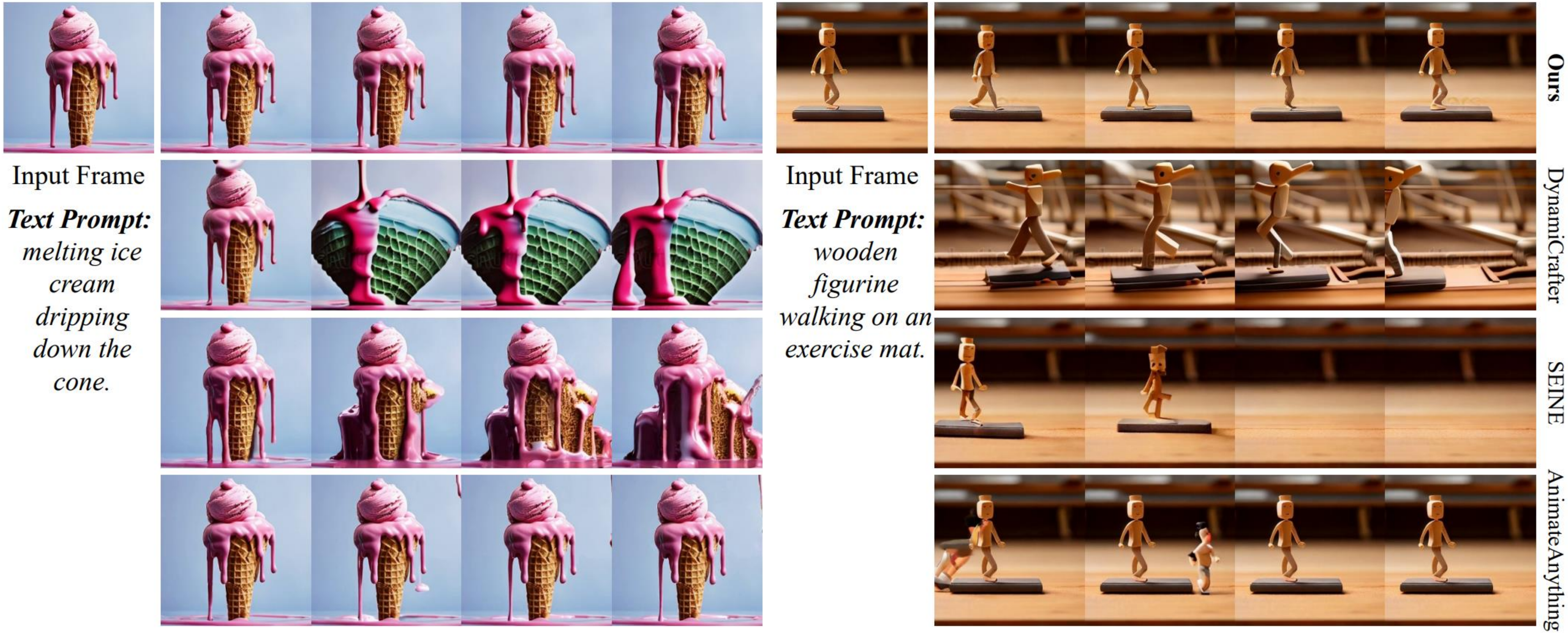
EVALUATION PARAMETERS

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



Method	#Data	UCF-101			MSR-VTT		Human Eval: Consistency	
		FVD ↓	IS ↑	FID ↓	FVD ↓	CLIPSIM ↑	Appearance ↑	Motion ↑
AnimateAnything	10M+20K†	642.64	**63.87**	**10.00**	218.10	0.2661	43.07%	20.26%
I2VGen-XL	35M	597.42	18.20	42.39	270.78	0.2541	1.79%	9.43%
DynamiCrafter	10M+10M†	404.50	41.97	32.35	219.31	0.2659	44.49%	31.10%
SEINE	25M+10M†	<u>306.49</u>	54.02	26.00	<u>152.63</u>	**0.2774**	<u>48.16%</u>	<u>36.76%</u>
ConsistI2V	10M	**177.66**	<u>56.22</u>	<u>15.74</u>	**104.58**	<u>0.2674</u>	**53.62%**	**37.04%**

Comparision of models



Limitations

- Dataset and Resolution: The WebVid-10M dataset contains low-resolution videos with fixed watermarks, leading to similar artifacts and low-resolution outputs in generated videos.

- Limited Motion: FrameInit enhances stability but often restricts subject movement, limiting motion magnitude.

- Training Constraints: Spatial conditioning requires U-Net layer tuning, increasing training costs and limiting adaptability to personalized T2I models.-

- Base Model Flaws: Inherits limitations from Stable Diffusion, such as inaccuracies in rendering faces and text.

FUTURE APPLICATIONS

- **Content Creation and Marketing**

  Ad Campaigns: Generate short, engaging video clips based on brand-related keywords and images for digital marketing.

  Social Media: Real-time generation of videos for trending topics or events to boost engagement.

- **Entertainment and Media**

  Creative Storytelling: Generate video snippets to match a story or script, allowing authors to visualize scenes dynamically.

- **Cyclone Prediction**

Consisitient2V is designed to model temporal dynamics in videos, but it can also be applied to spatially varying data like weather patterns. By incorporating spatial features and relationships into the model, you could potentially use Consistient2V to predict cyclone movement or intensity based on the current state of atmospheric conditions.

This approach leverages the spatiotemporal dynamics inherent in weather patterns, allowing the model to capture complex interactions between different variables.

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