

# Lumiere: A Space-Time Diffusion Model for Video Generation

Authors:

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann,  
Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur,  
Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein,  
Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, Inbar  
Mosseri

SIGGRAPH 2024, 259 citations

Presented by:

Anantapadmanaabha Prasannakumar

# Outline

- Motivation
- Method
- Evaluations
- Applications
- Societal Impact
- Limitations
- Conclusion

# Motivation

- ▶ Restricted capability of existing models
  - ▶ Sensitive to error
  - ▶ Suffers from memory and computing constraints
  - ▶ Obtaining large-scale data is cumbersome
  - ▶ Training large-scale T2V is challenging

# Motivation

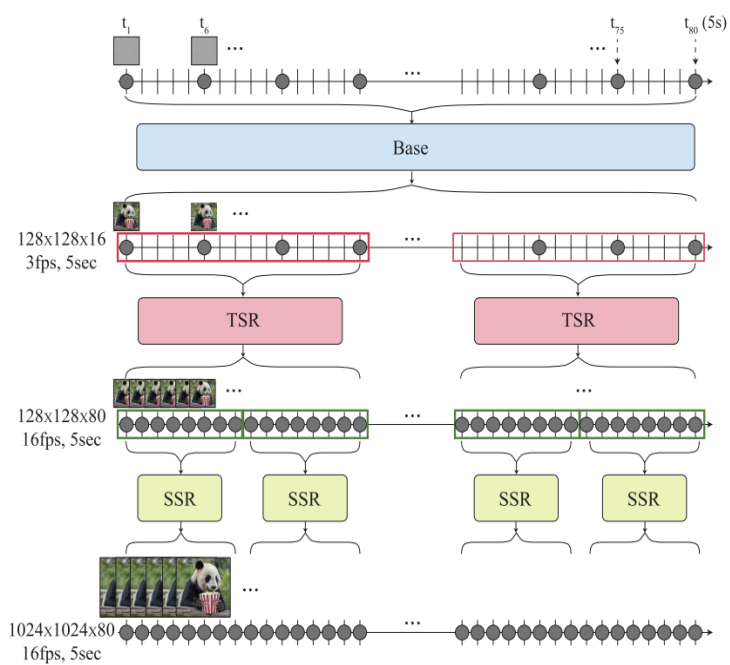
- ▶ Employing temporal cascades is hindersome
  - ▶ Generates aggressively sub-sampled set of keyframes
  - ▶ TSR modules are constrained to fixed, small temporal context
  - ▶ Cascaded training suffers from domain gap

# Method - Lumiere

- ▶ Utilizes Diffusion Probabilistic Models
  - ▶ Through denoising steps, trained to approximate data distribution
  - ▶ Starting from noise, a clean sample is drawn from the targeted distribution
- ▶ Incorporates additional guiding signals

# Method - Pipeline

(a) Common Approach with TSR model(s)

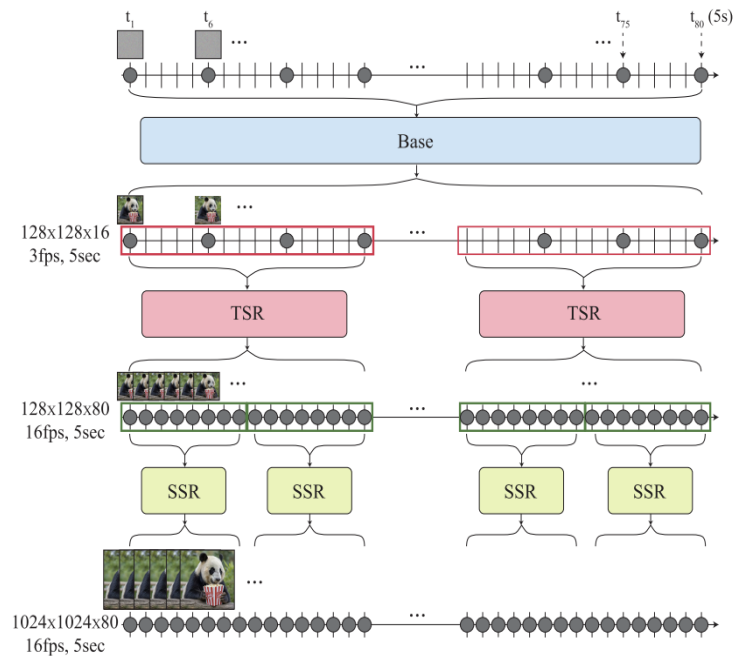


## Common T2V Framework:

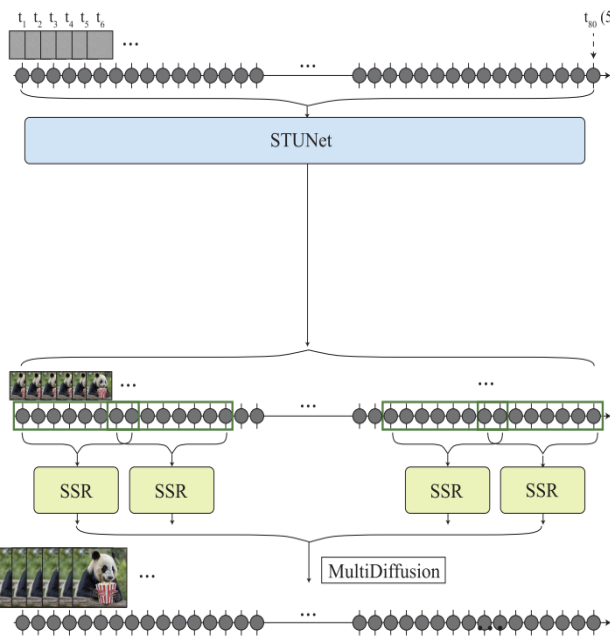
- Base Model
- Temporal Super-resolution Model (TSR)
- Spatial Super-resolution Model (SSR)

# Method - Pipeline

(a) Common Approach with TSR model(s)



(b) Our Approach

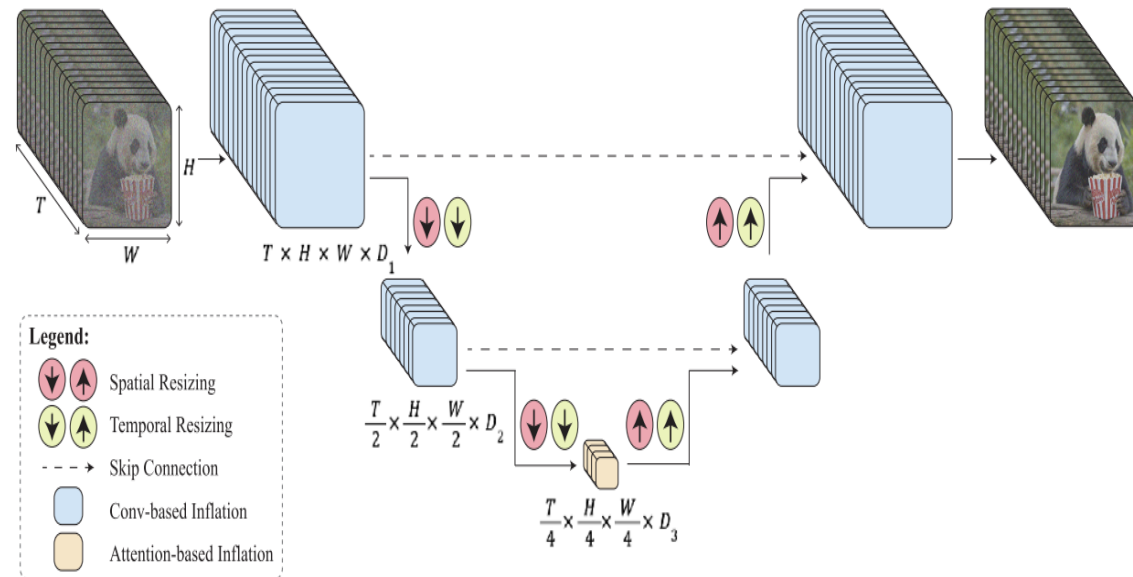


Lumiere Framework:

- Base Model
- Spatial Super-resolution Model (SSR)
- Multidiffusion

# Method - STUnet

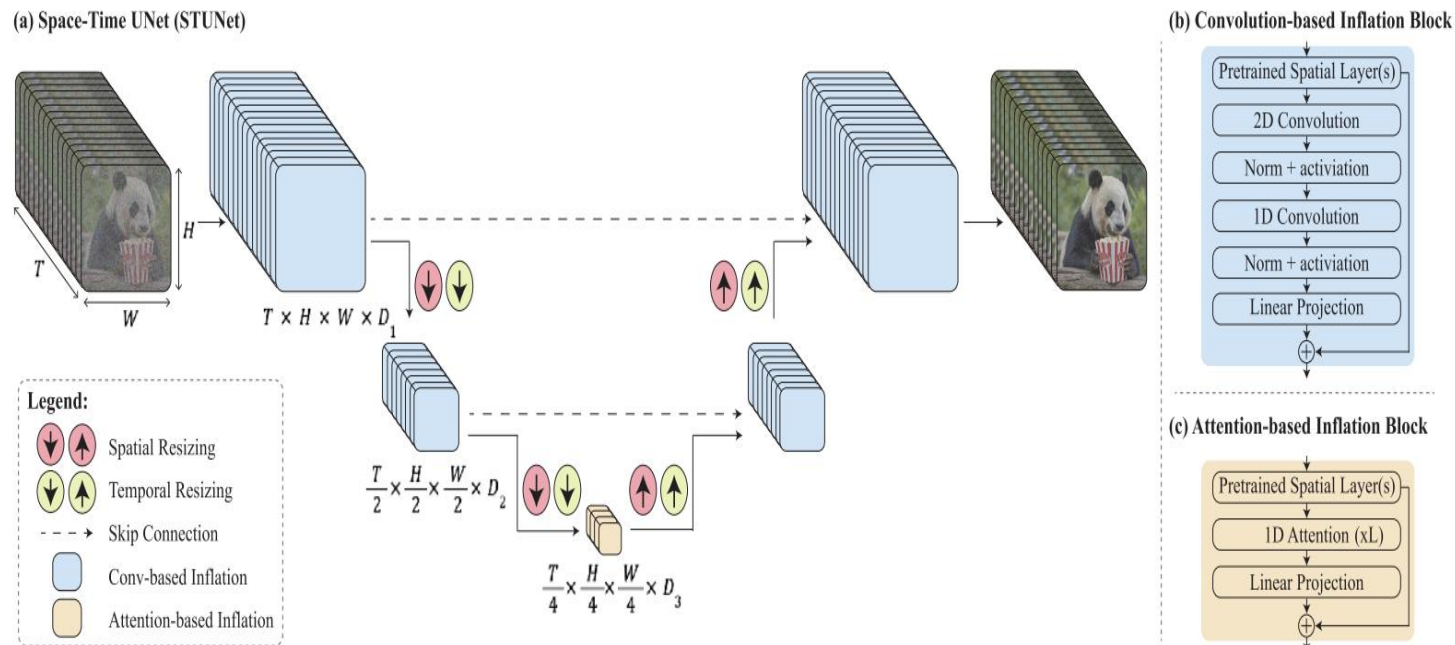
(a) Space-Time UNet (STUnet)



Employs traditional U-Net Model

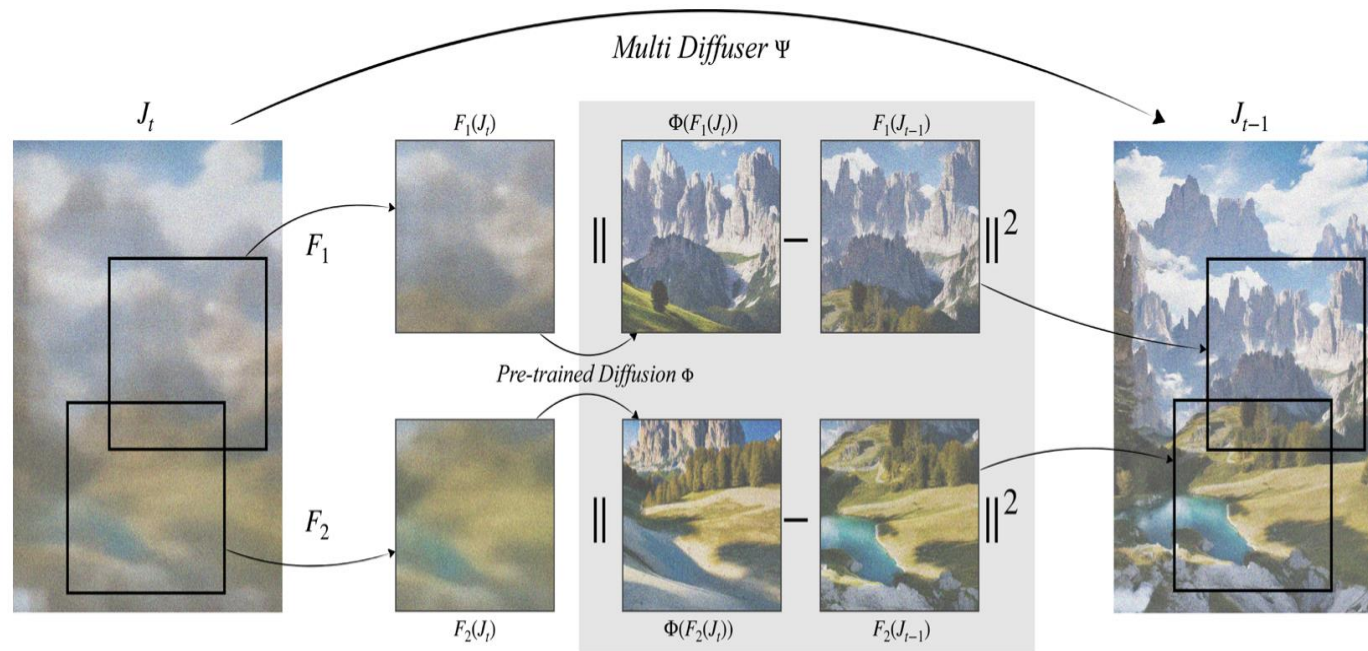


# Method - STUnet



- Interleave temporal blocks to T2I Architecture
- Trains only new parameters
- Performs identity Initialization
- Low computational Overhead

# Method - Multidiffusion



## Multidiffusion Framework:

- New generation process from a pre-trained model
- Fuses inconsistent directions into one global denoising step

# SSR with Multidiffusion

- ▶ An inflated SSR network can only operate on short videos
- ▶ Employ multidiffusion for smooth temporal transition
- ▶ Multidiffusion prevents temporal artifacts
  - ▶ Resolved by linearly combining video segments

# SSR with Multidiffusion

- ▶ At each generation step:
  - ▶ split noisy input video  $J \in \mathbb{R}^{H \times W \times T \times 3}$  into  $1 \dots N$  overlapping segments
  - ▶ Where  $J_i \in \mathbb{R}^{H \times W \times T' \times 3}$  is the  $i^{th}$  segment
  - ▶ Temporal duration:  $T' < T$
- ▶ To reconcile per-segment SSR predictions:

$$\arg \min_{J'} \sum_{i=1}^n \|J' - \Phi(J_i)\|^2.$$

# Evaluation Setup

- ▶ Train T2V model on 30M videos with text prompts
  - ▶ Videos are 80 frames long at 16 fps
  - ▶ 109 text prompts
  - ▶ Base model dimension: 128 x 128 frames
  - ▶ SSR dimension: 1024 x 1024 frames

# Zero-shot on UCF-101

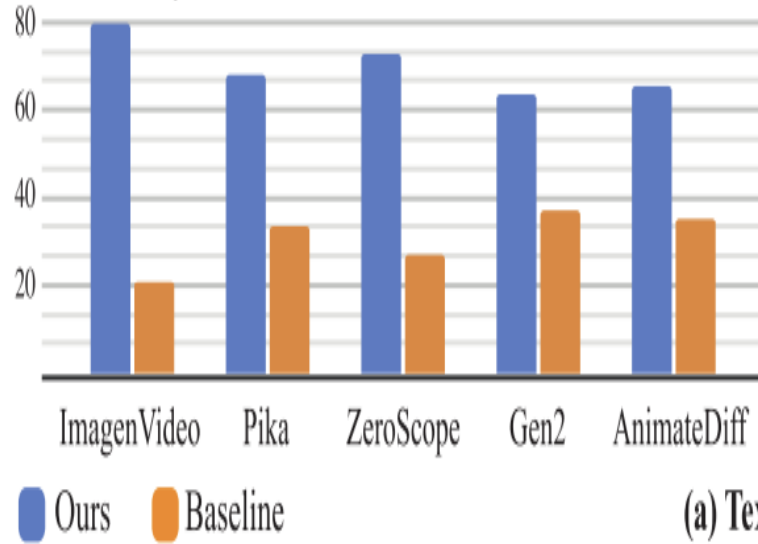
Method	FVD ↓	IS ↑
MagicVideo (Zhou et al., 2022)	655.00	-
Emu Video (Girdhar et al., 2023)	606.20	42.70
Video LDM (Blattmann et al., 2023b)	550.61	33.45
Show-1 (Zhang et al., 2023a)	394.46	35.42
Make-A-Video (Singer et al., 2022)	367.23	33.00
PYoCo (Ge et al., 2023)	355.19	47.76
SVD (Blattmann et al., 2023a)	242.02	-
<b>Lumiere (Ours)</b>	<b>332.49</b>	<b>37.54</b>

# User Study

- ▶ Two-alternative Forced Choice protocol Adopted
  - ▶ Randomly ordered pairs of videos are provided
  - ▶ 400 user judgments obtained
  - ▶ 109 prompts were utilized
  - ▶ Fixed random seed
  - ▶ Spatial and Temporal alignment

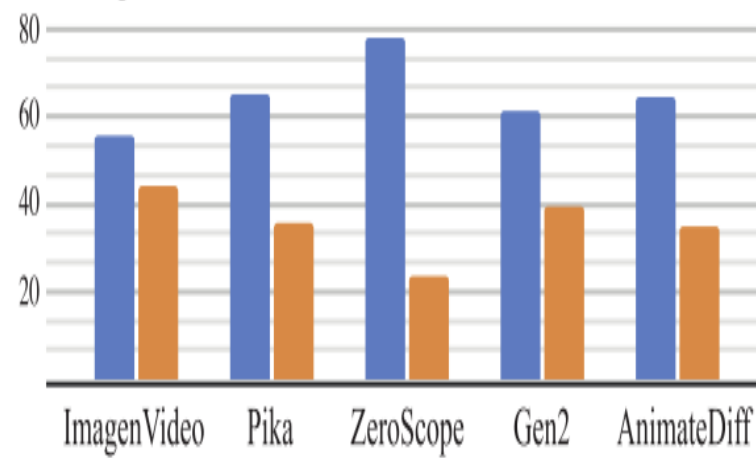
# User Study

Video Quality

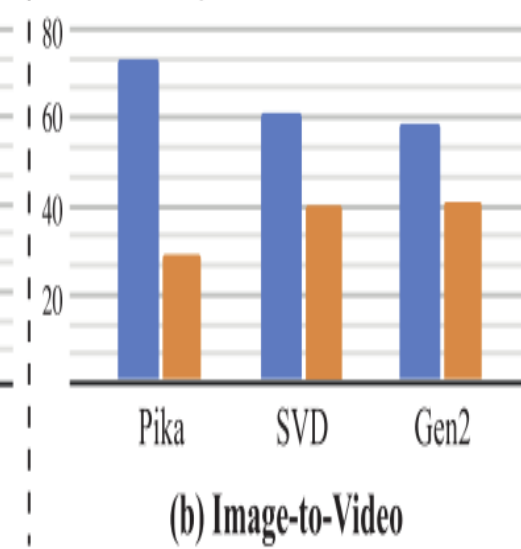


(a) Text-to-Video

Text Alignment



Video Quality



(b) Image-to-Video



# User Study

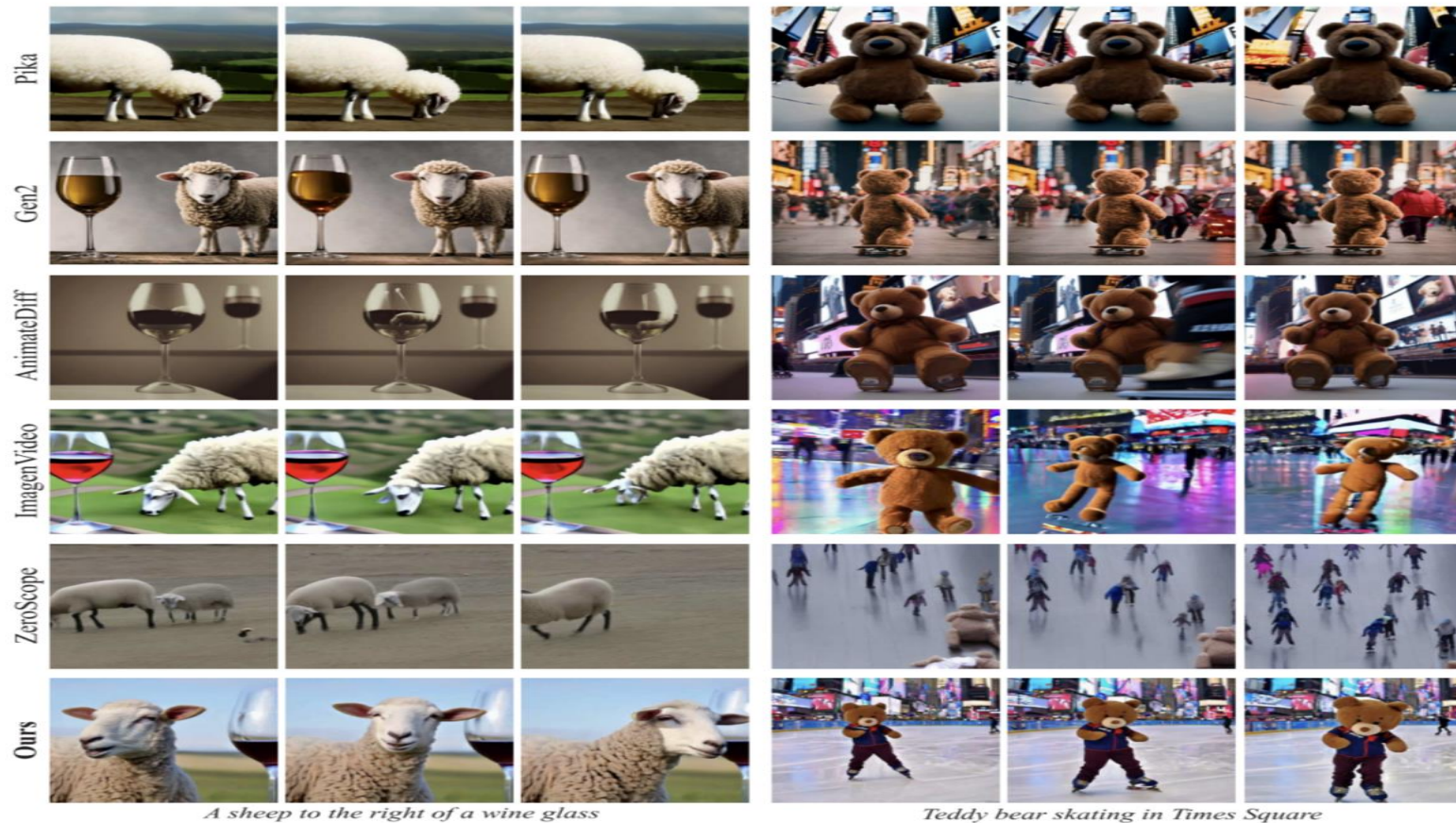
Left video



Right video

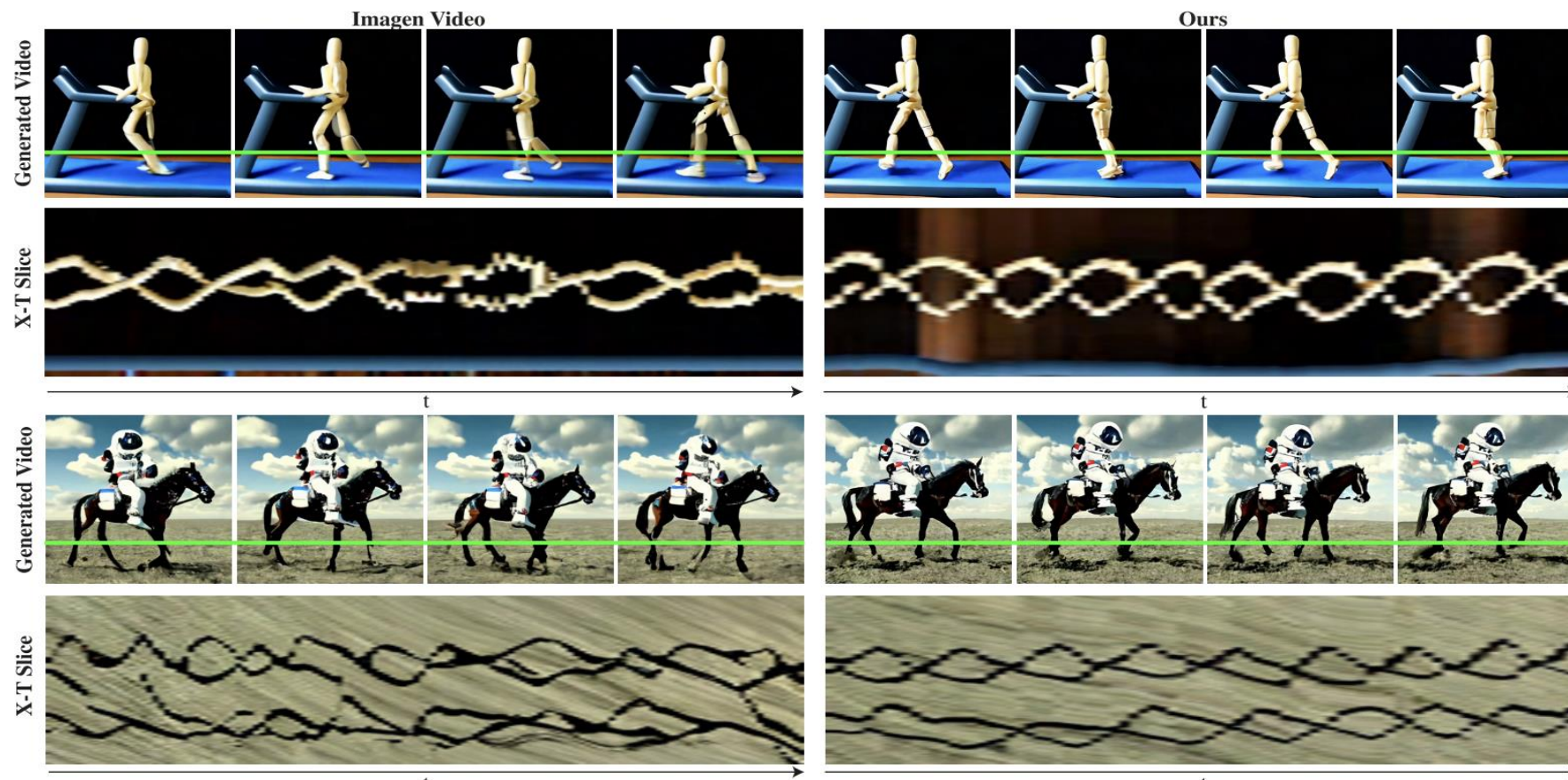


# Qualitative Evaluation

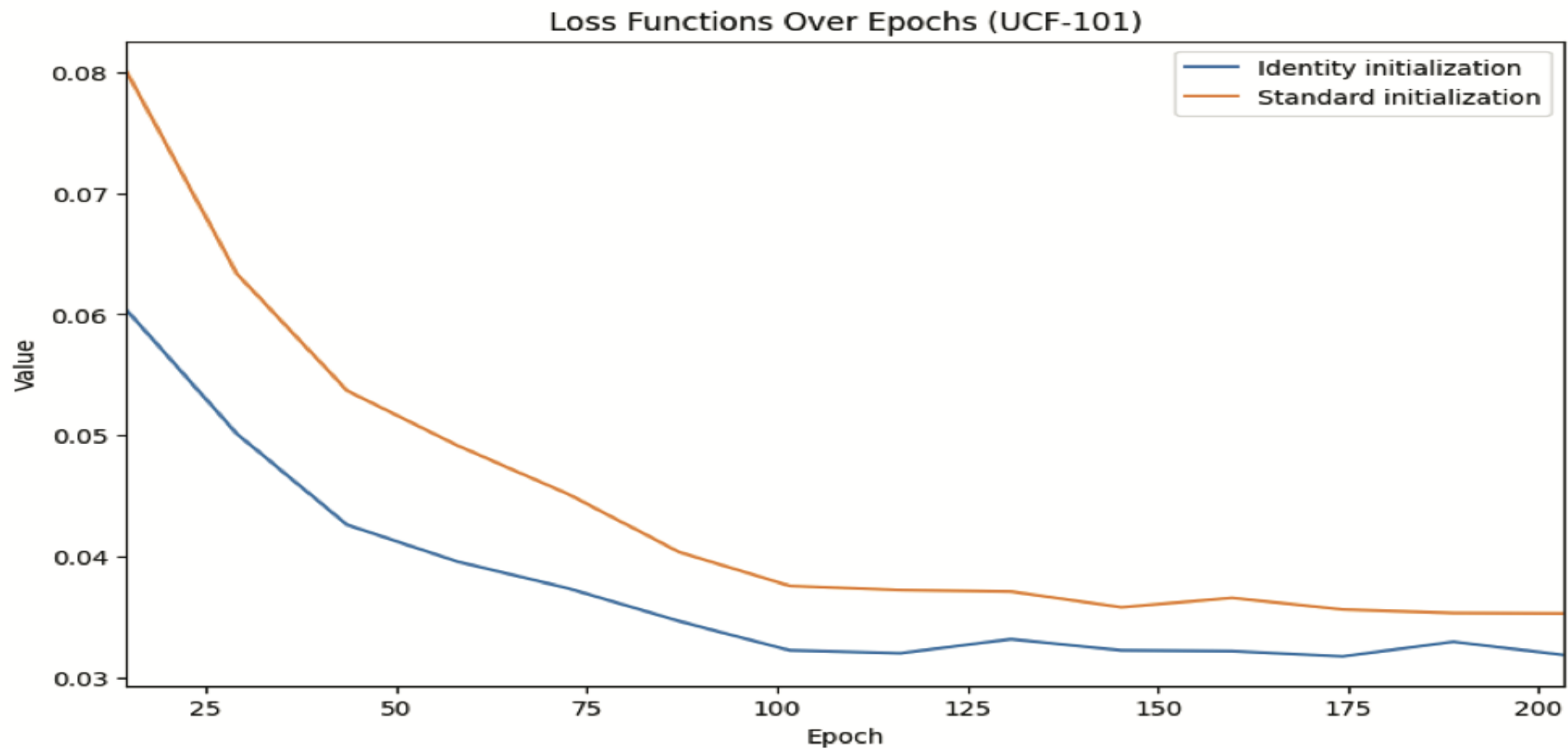




# Temporal Consistency



# Ablation - Initialization



# Visualize Initialization Schemes

Standard

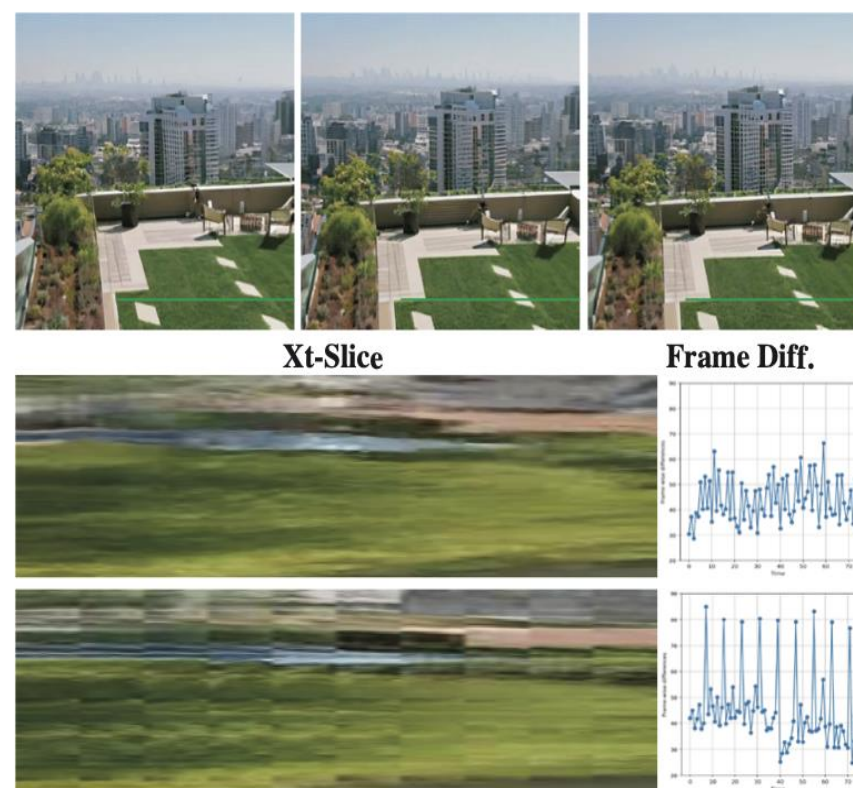
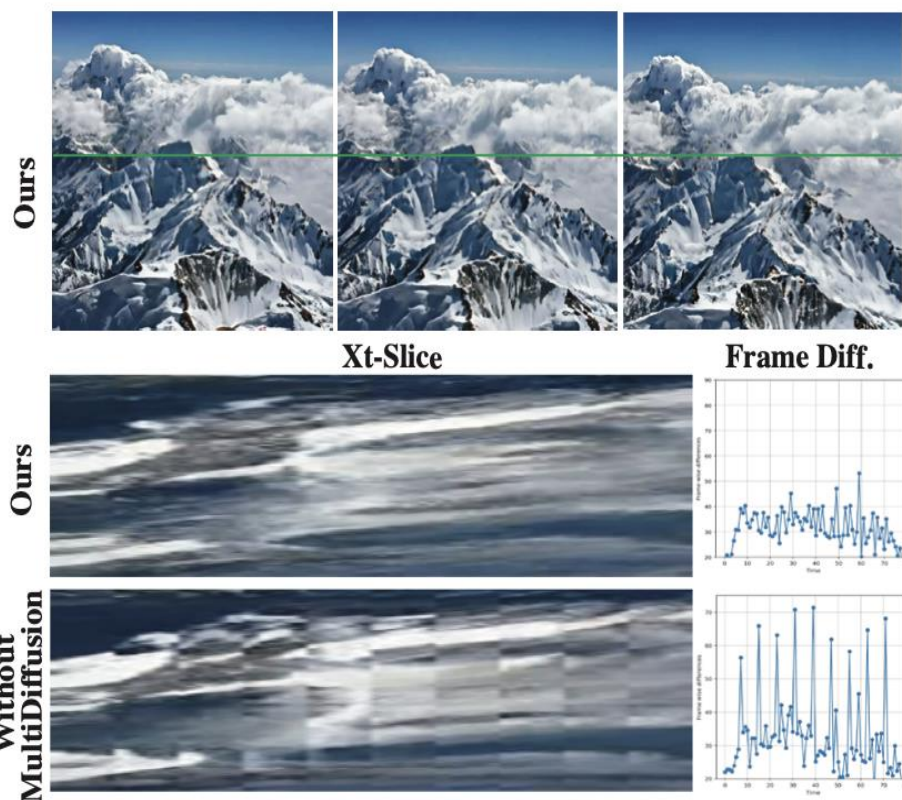


Identity





# Ablation - Multidiffusion



# Applications – Video Editing



Original Video



Generated Video

# Application – Stylized Generation

- ▶ Pre-trained T2I weights remain fixed
- ▶ Newly added temporal layers are trained
- ▶ Linear interpolation between fixed and fine-tuned T2I weights
  - ▶  $W_{interpolate} = \alpha \cdot W_{style} + (1 - \alpha) \cdot W_{orig}$
  - ▶ Where  $\alpha \in [0.5, 1]$



# Application – Stylized Generation

## Vector art styles



Reference Image



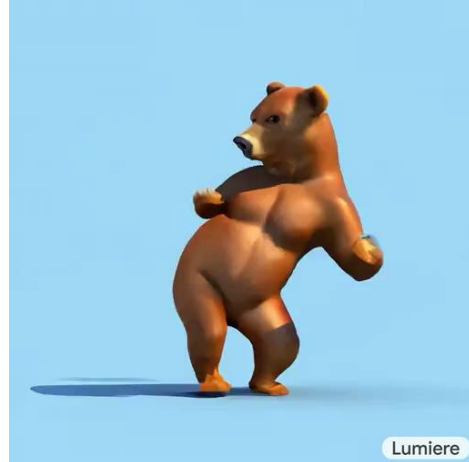
Output

# Application – Stylized Generation

Realistic styles



Reference Image



Output

# Application – Conditional Generation

- ▶ Model conditioned on additional input signals
  - ▶ Noisy video  $J \in \mathbb{R}^{H \times W \times T \times 3}$
  - ▶ Text prompt
  - ▶ Masked conditioning video  $C \in \mathbb{R}^{H \times W \times T \times 3}$
  - ▶ Binary Mask  $M \in \mathbb{R}^{H \times W \times T \times 1}$
- ▶ Concatenated Tensor  $\langle J, C, M \rangle = \mathbb{R}^{T \times H \times W \times 7}$

# Application – Image to Video



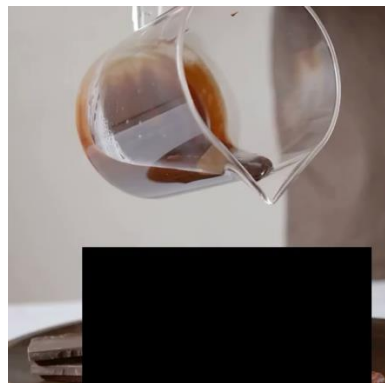
# Application - Inpainting



Video + Mask



Output





# Application - Cinemagraphs



Source Image + Mask



Output



# Societal Impact

- ▶ Risk of misuse
  - ▶ Tools for detecting biases and malicious use cases
  - ▶ To ensure safe and fair use

# Limitations

- ▶ The model cannot generate videos
  - ▶ Multiple shots
  - ▶ Transition between scenes
- ▶ The model operates in pixel space



# Conclusion

- ▶ Presents a novel T2V framework
  - ▶ Built on a pre-trained T2I model
  - ▶ Introduces space-time U-Net Architecture
  - ▶ Utilizes Multidiffusion framework
- ▶ Demonstrates state-of-the-art generation results
- ▶ Showcases applicability to various downstream tasks



Thank you