Lumiere: A Space-Time Diffusion Model for Video Generation

Authors: Bar-Tal et al.

SIGGRAPH 2024, 259 citations

PRESENTED BY:
ANANTAPADMANAABHA PRASANNAKUMAR

Outline

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

Motivation

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

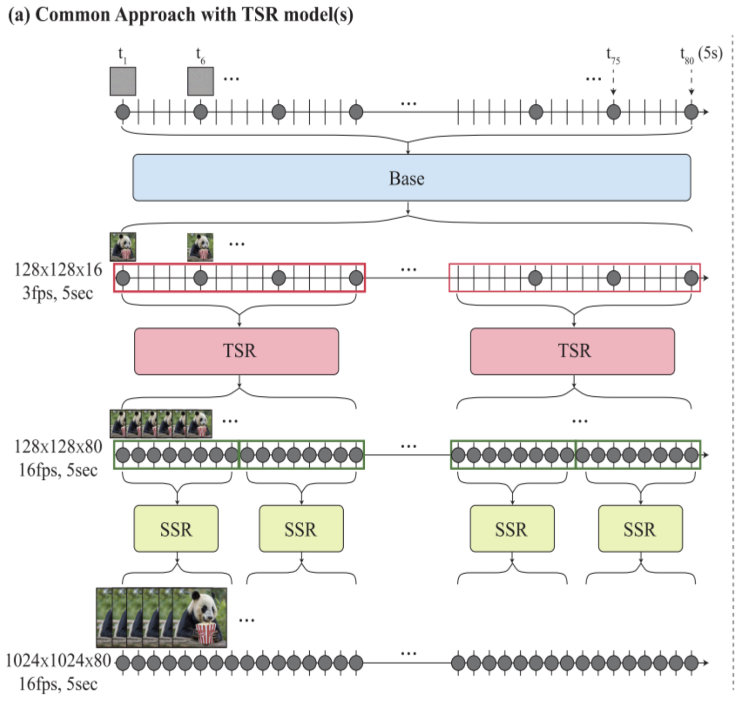
Motivation

- 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, the model obtains target distribution
- Incorporates additional guiding signals

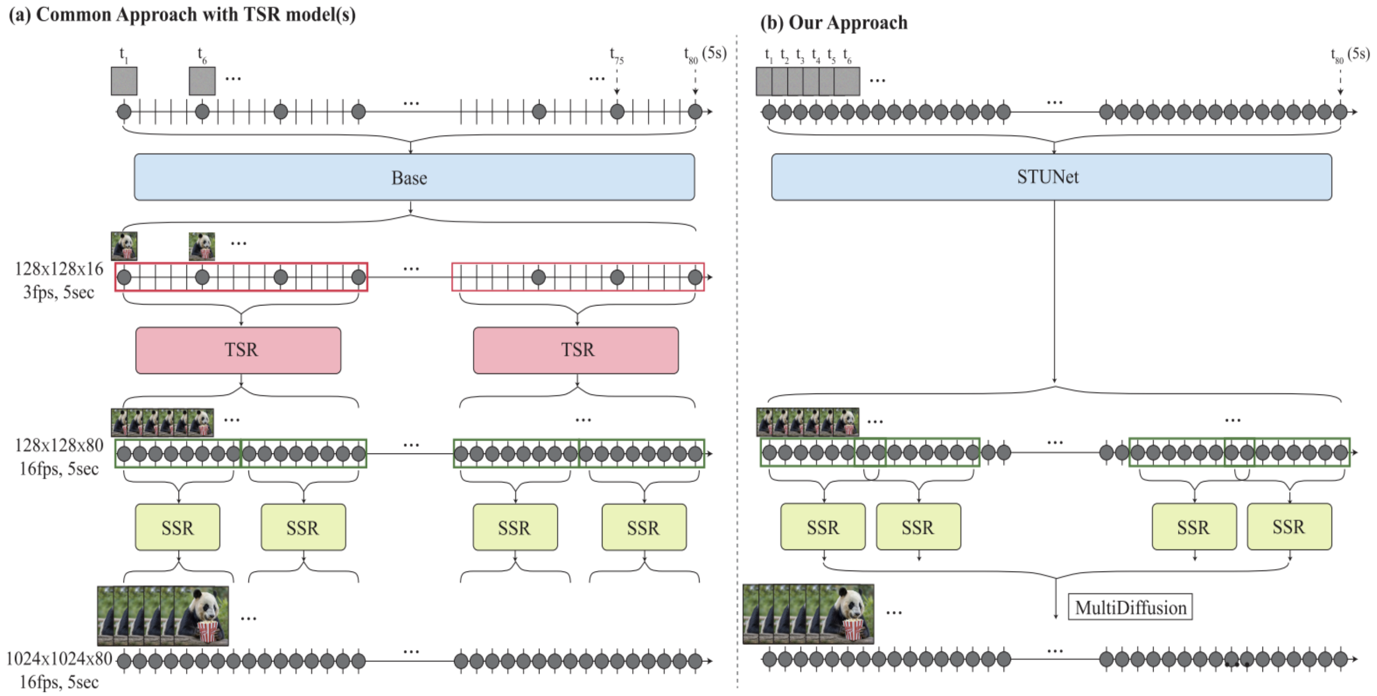
Method - Pipeline



Common Framework:

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

Method - Pipeline

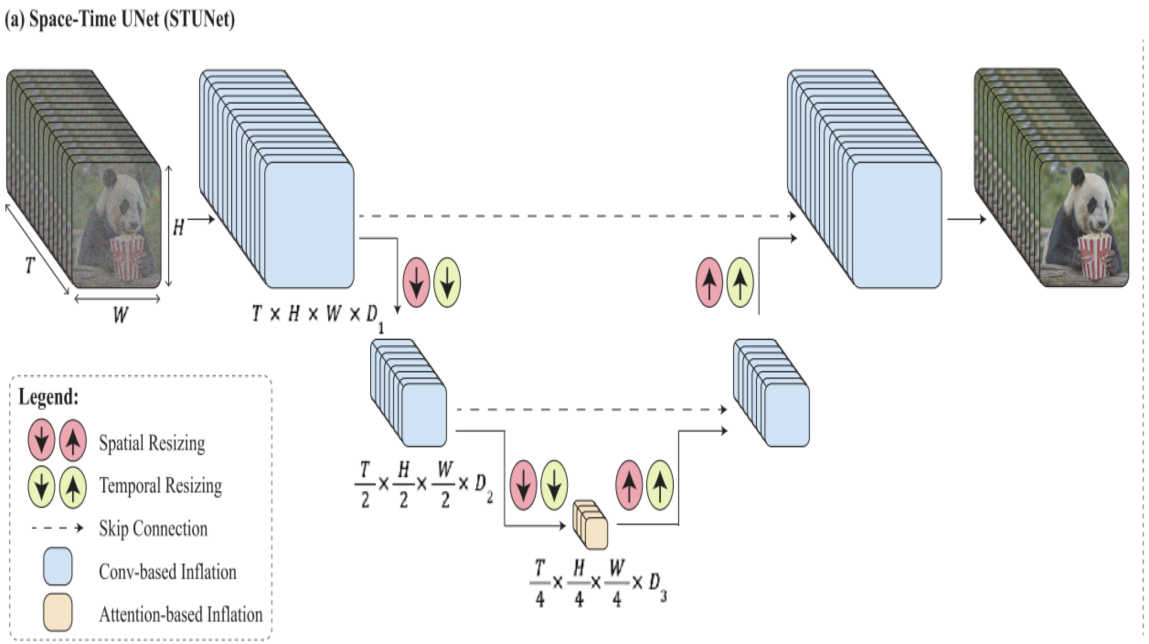


Lumiere Framework:

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

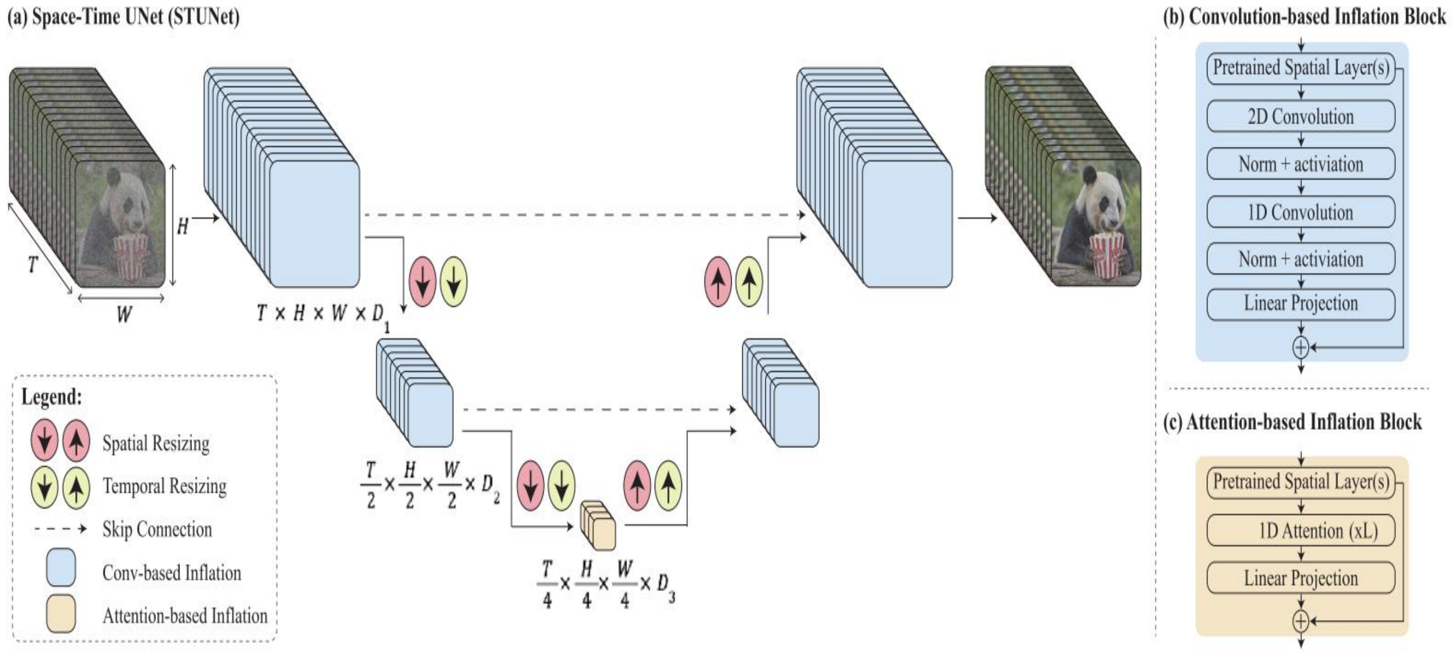
Lumiere: A Space-Time Diffusion Model for Video Generation, SIGGRAPH 2024, https://arxiv.org/pdf/2401.12945

Method - STUnet



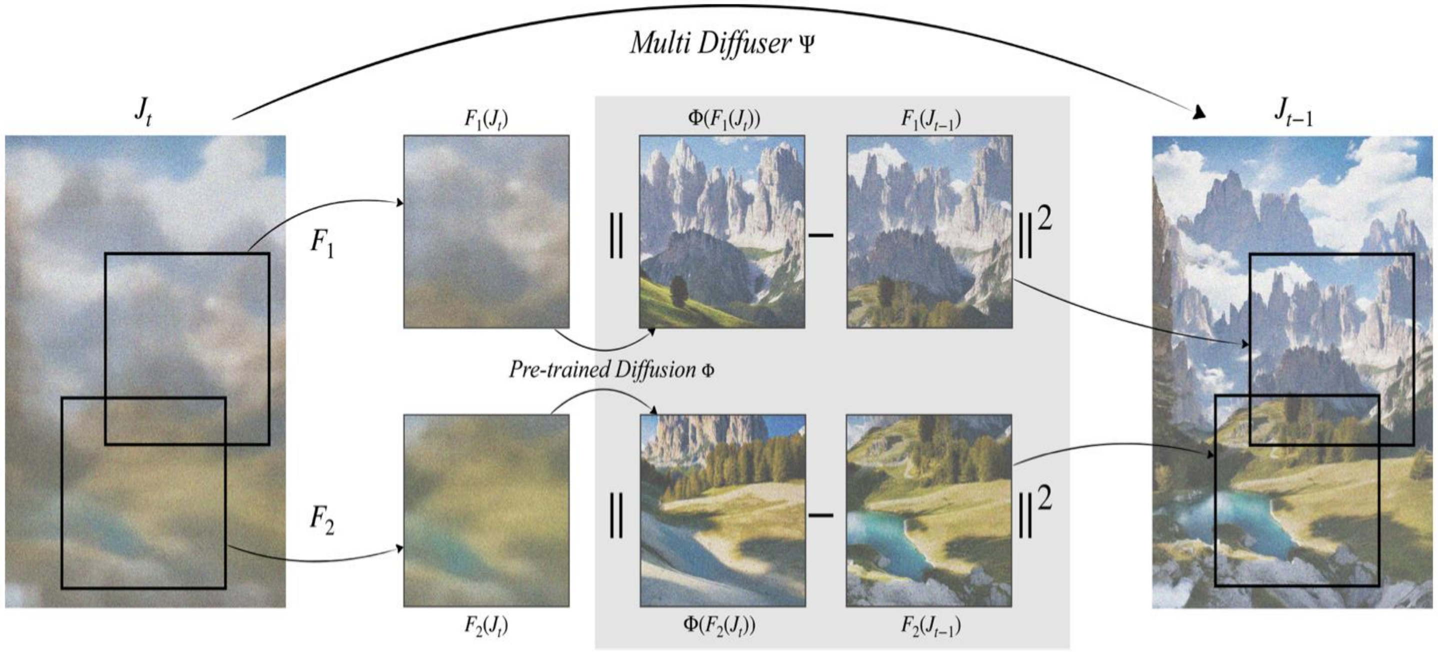
Employs traditional U-Net Model

Method - STUnet



- Convolution-based Inflation block
- Attention-based Inflation block

Method - Multidiffusion



Multidiffusion Framework:

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

MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation, ICML 2023, https://arxiv.org/pdf/2302.08113

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

Mathematically:

- At each generation step:
  - split noisy input video $J \in \mathbb{R}^{H \times W \times T \times 3}$ into $i \dots N$ 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 prediction during denoising step:

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

Evaluation Setup

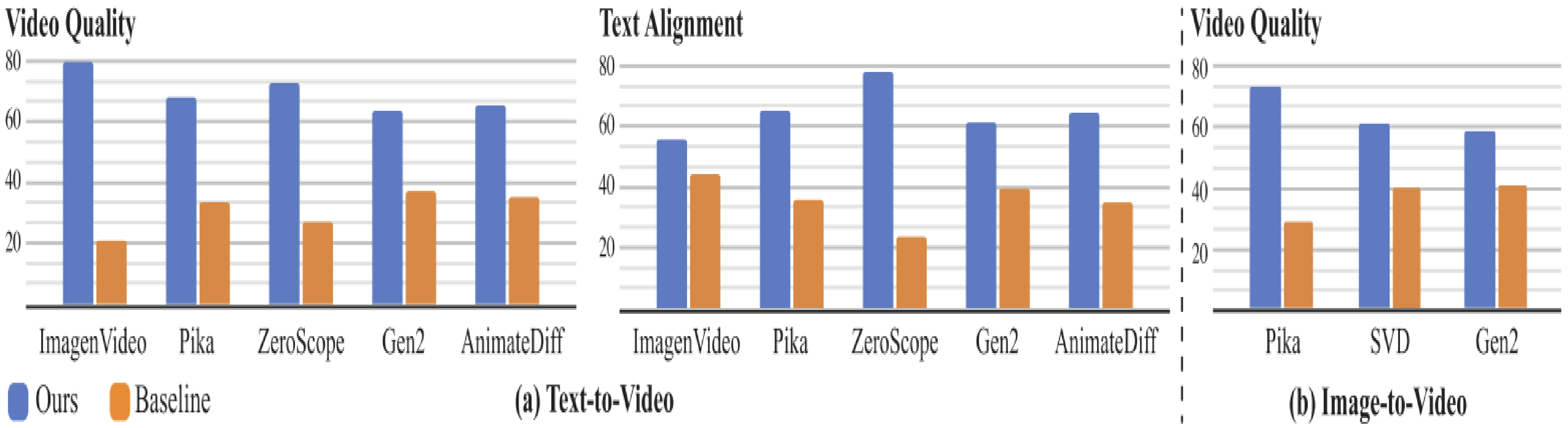
- Train T2V model on 30M videos with text prompts
  - Videos are 80 frames long at 16 fps
  - 109 text prompts
  - Base model dimension: 128 x 128 frames
  - SSR dimension: 1024 x 1024 frames

Quantitative Evaluation

Zero-shot evaluation on UCF-101

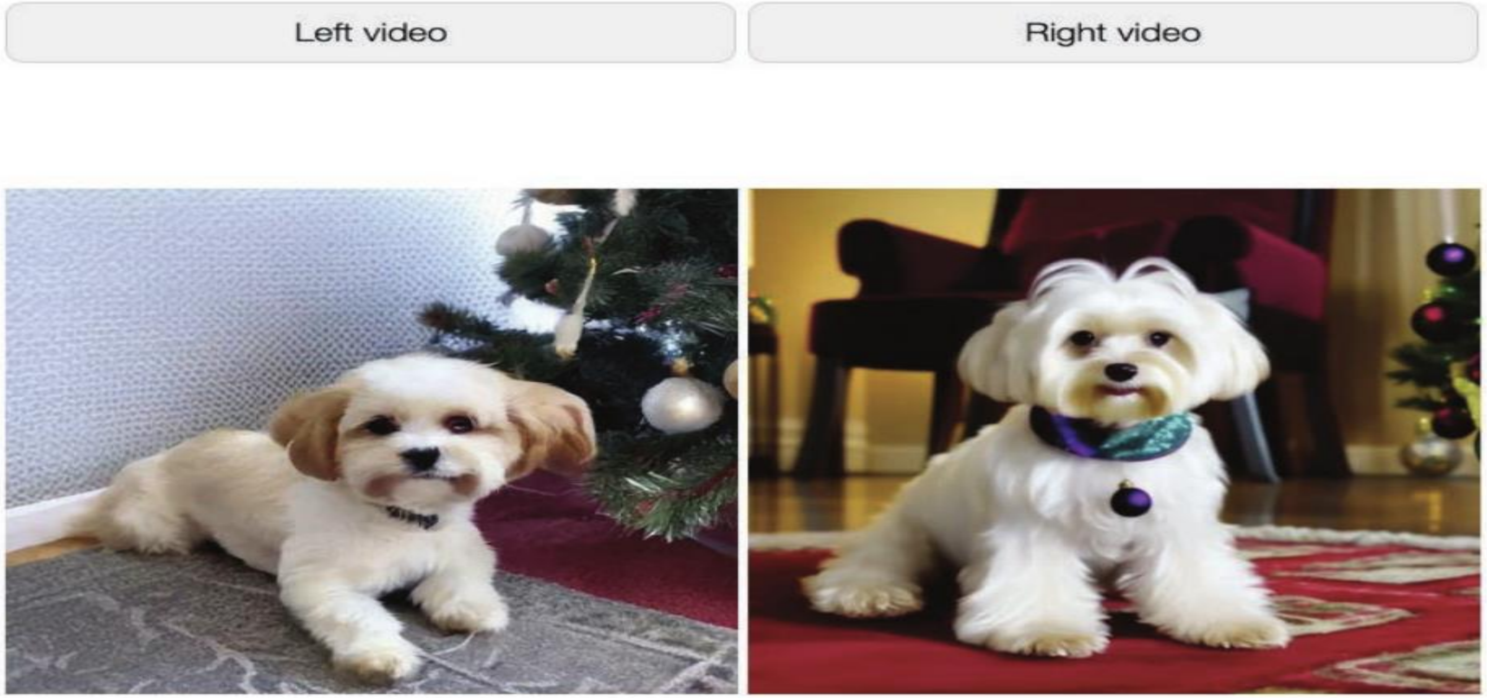
Method	FVD ↓	IS ↑
MagicVideo (Zhou et al., 2022)	655.00	-
Emu Video (Girdhar et al., 2023)	606.20	42.70
Video LDM (Blattmann et al., 2023b)	550.61	33.45
Show-1 (Zhang et al., 2023a)	394.46	35.42
Make-A-Video (Singer et al., 2022)	367.23	33.00
PYoCo (Ge et al., 2023)	355.19	47.76
SVD (Blattmann et al.,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

User Study

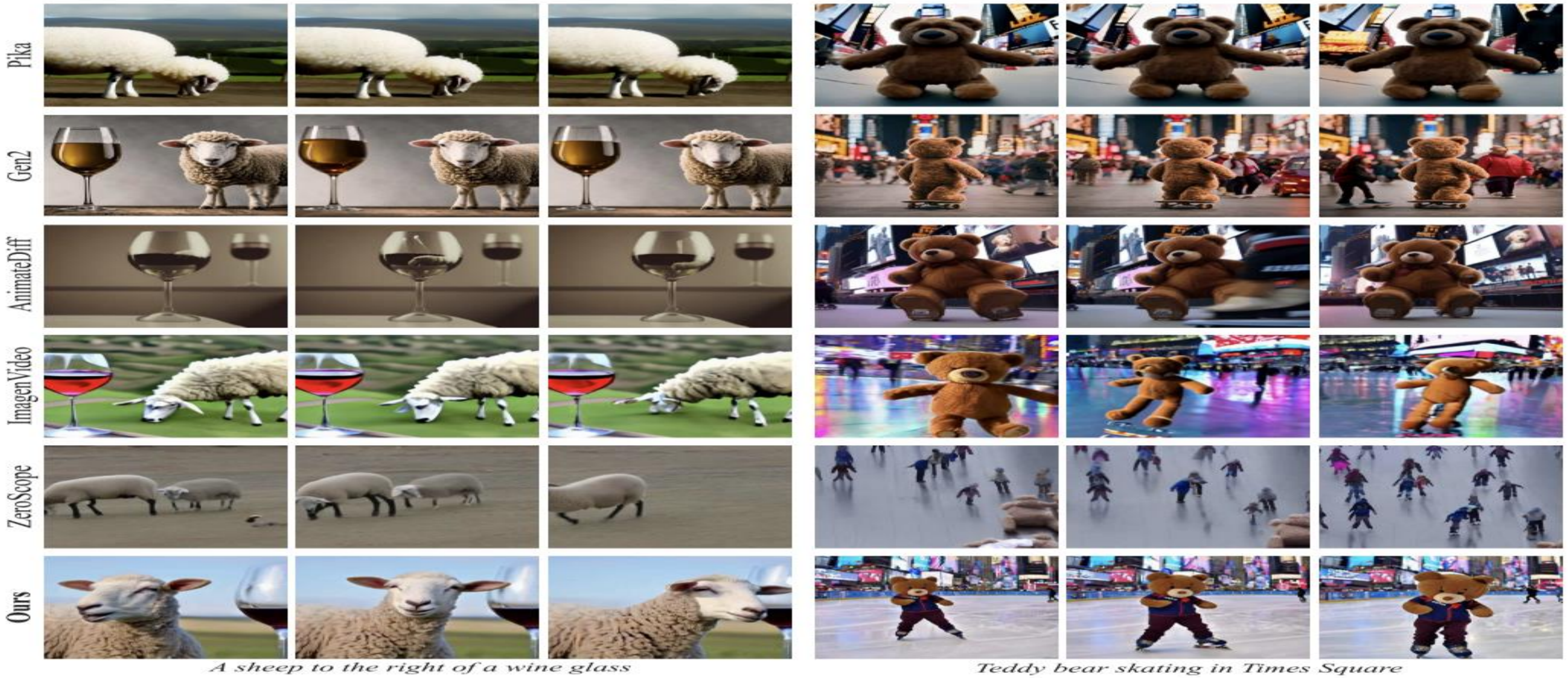


Video Quality
80
60
40
20
ImagenVideo
Pika
ZeroScope
Gen2
AnimateDiff
Text Alignment
80
60
40
20
ImagenVideo
Pika
ZeroScope
Gen2
AnimateDiff
Video Quality
80
60
40
20
Pika
SVD
Gen2
Ours
Baseline
(a) Text-to-Video
(b) Image-to-Video

User Study

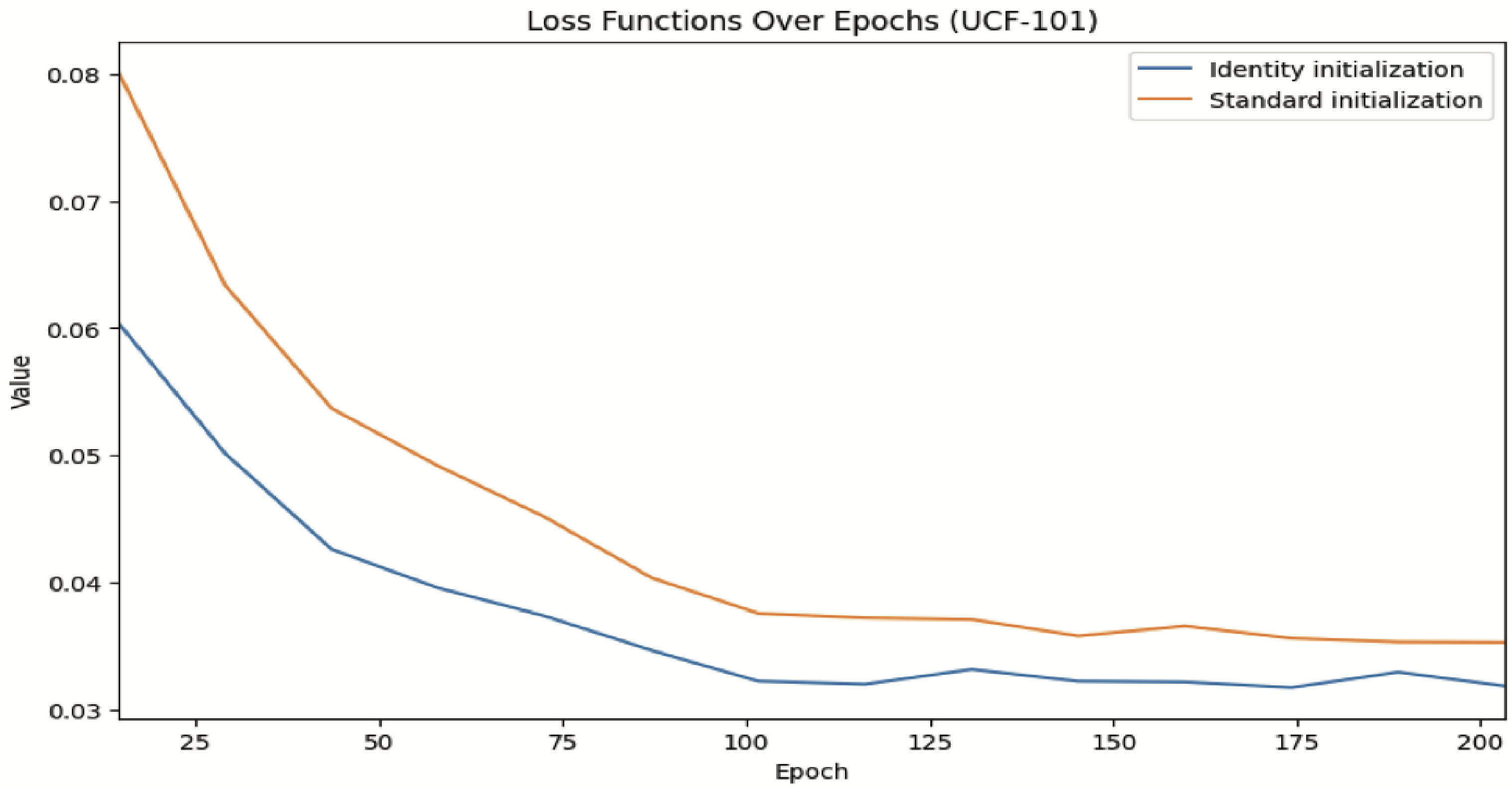


Qualitative Evaluation



Ablation Studies

Initialization Ablation

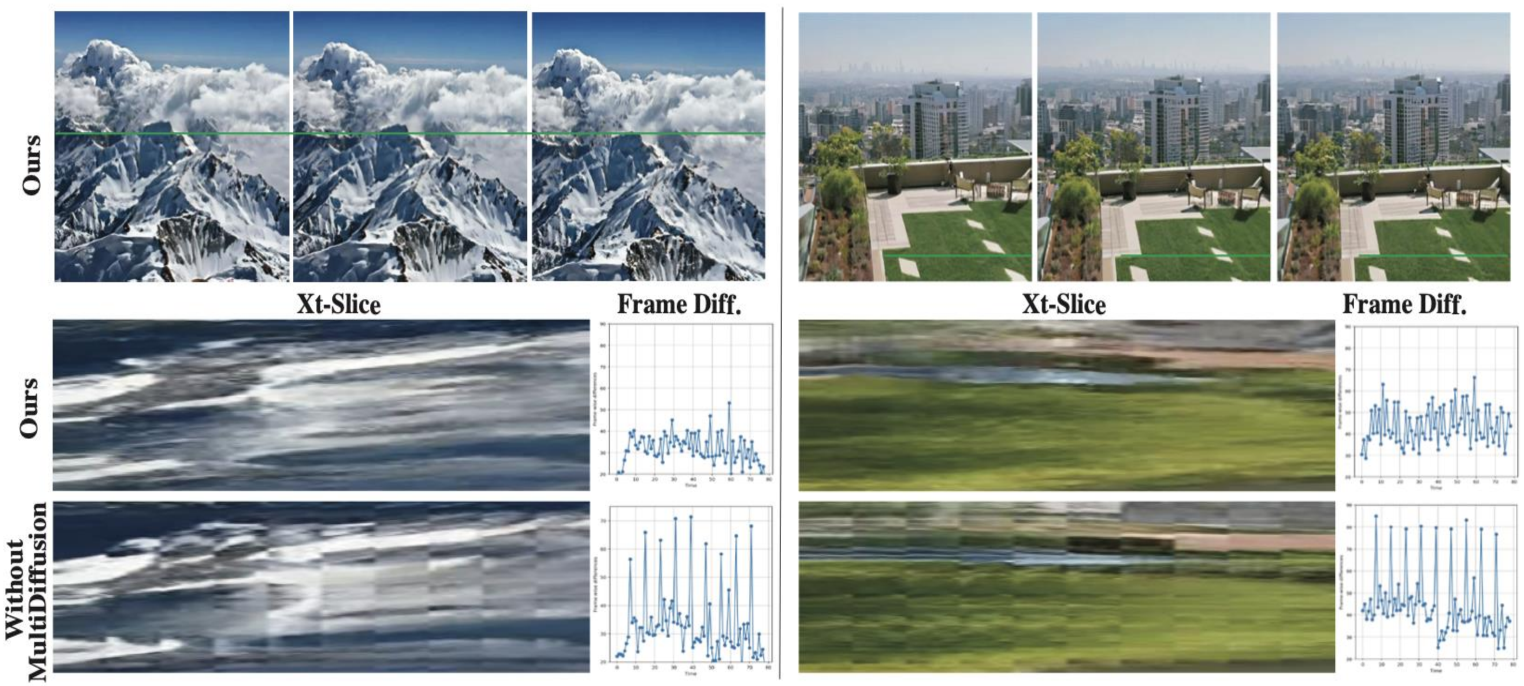


Ablation Studies

Different Initialization Schemes

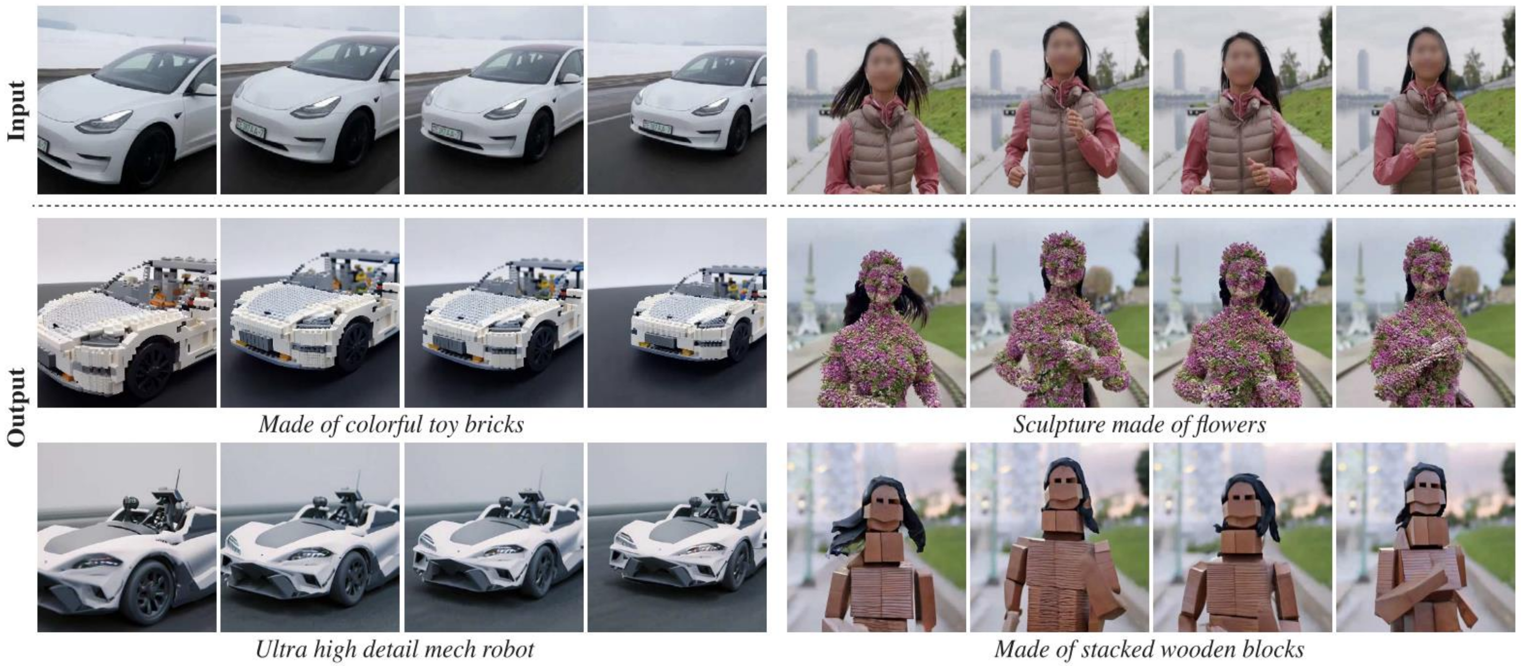
Ablation Studies

Multidiffusion Ablation



Applications

Video-to-video Editing



- Lack of TSR ideal for V2V editing
- Employs SDEdit

SDEDIT: GUIDED IMAGE SYNTHESIS AND EDITING WITH STOCHASTIC DIFFERENTIAL EQUATIONS, Jan 2022, https://arxiv.org/pdf/2108.01073

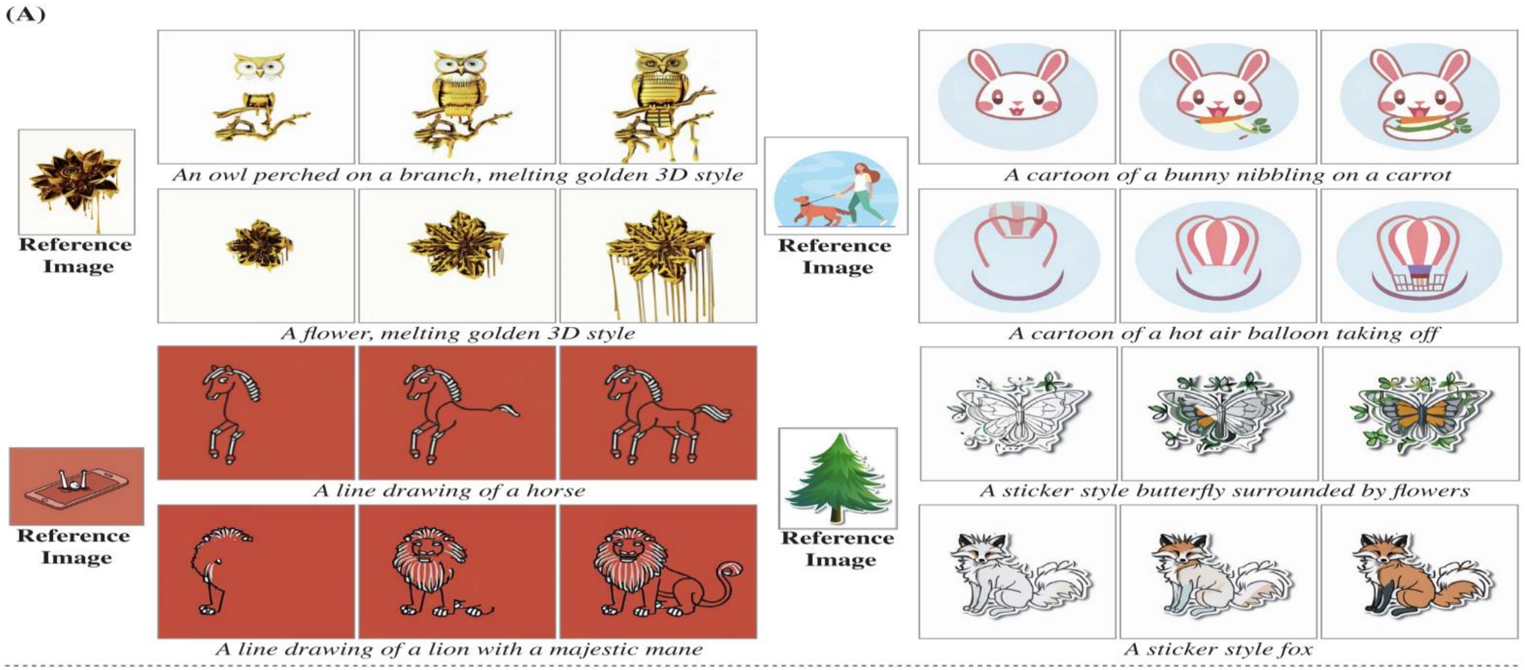
Application

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

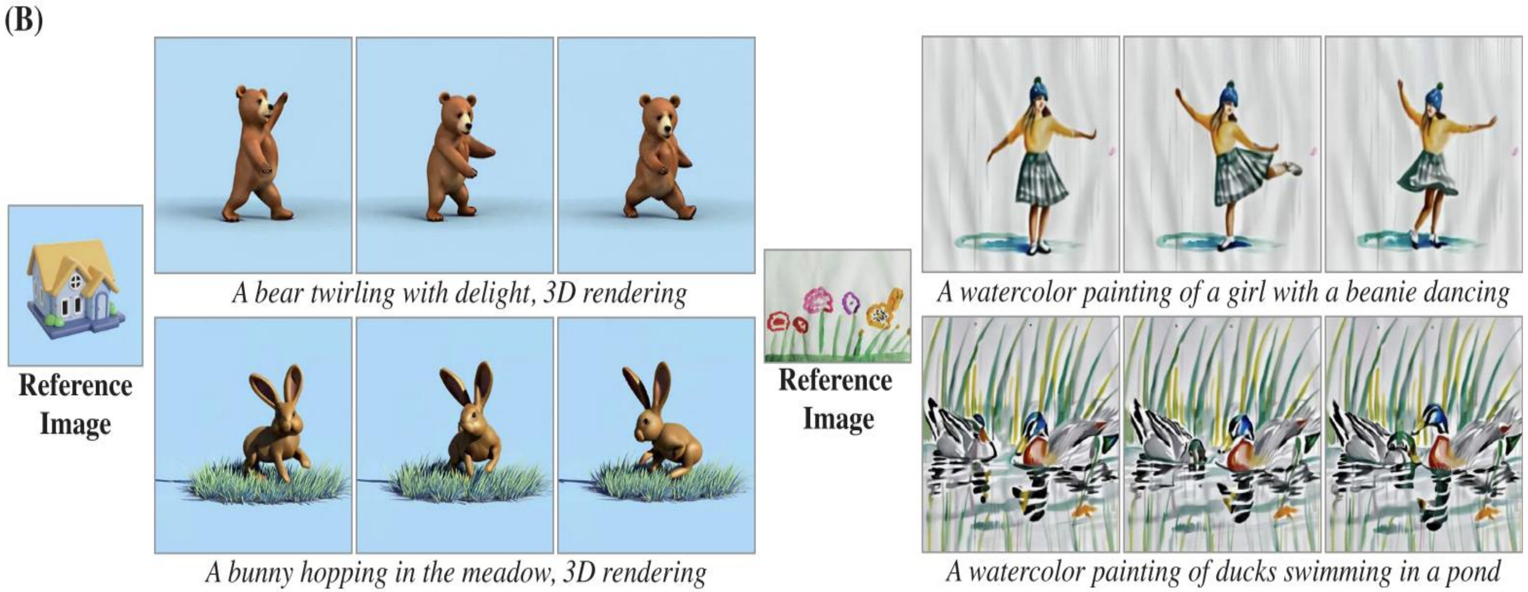
Stylized Generation



Vector art styles

Application

Stylized Generation



Realistic styles

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 $C \in \mathbb{R}^{H \times W \times T \times 3}$
  - Masked video $M \in \mathbb{R}^{H \times W \times T \times 1}$

- Concatenated Tensor $\langle J, C, M \rangle = \mathbb{R}^{T \times H \times W \times 7}$

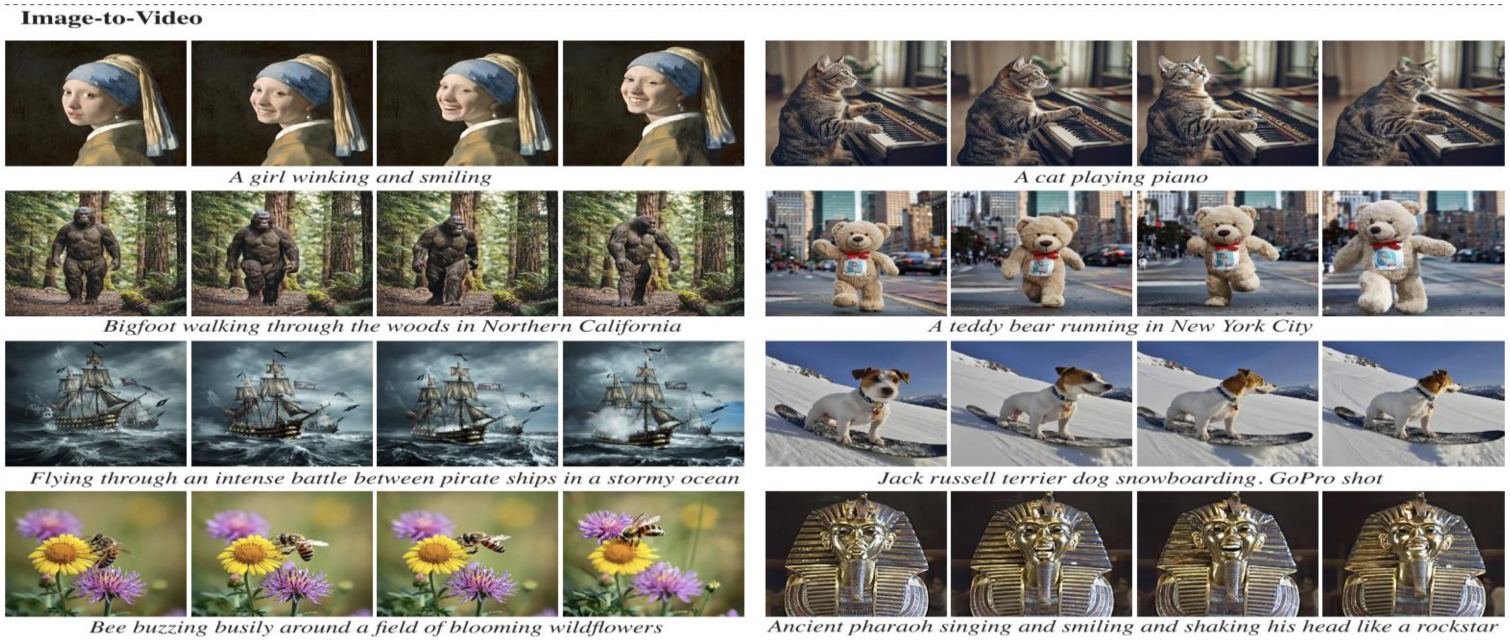
Application

Image-to-Video Generation

Sample I2V results

Application

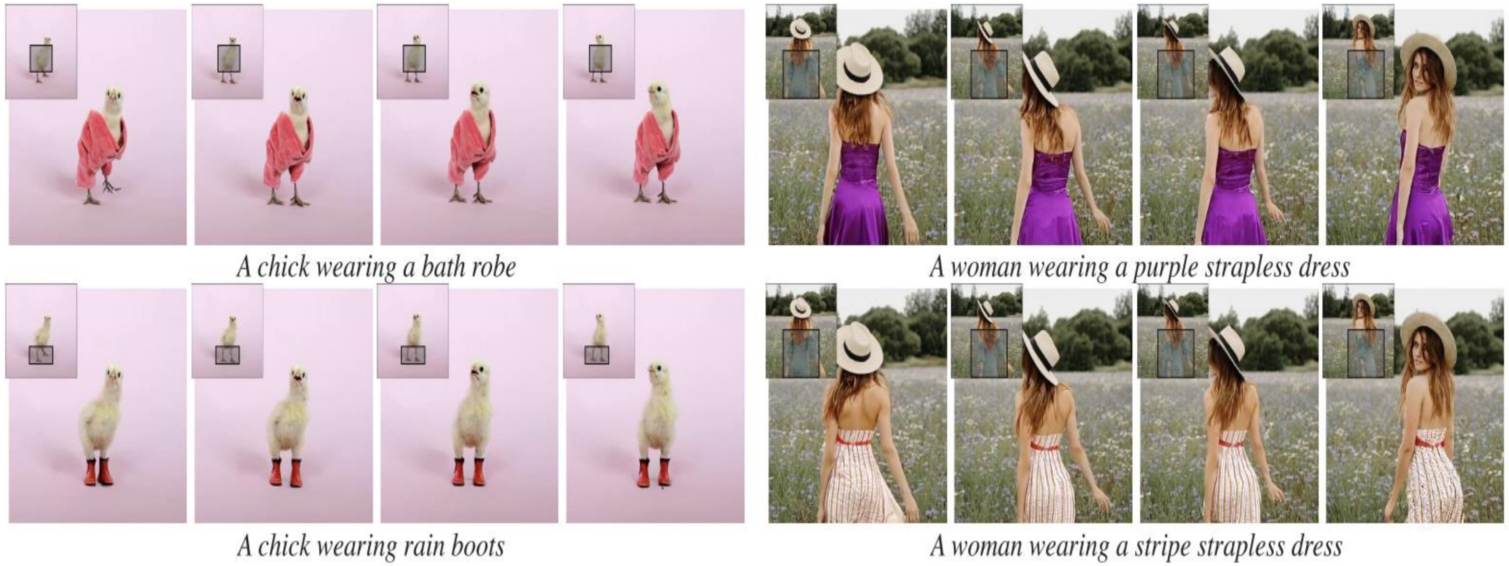
Image-to-Video Generation



Sample I2V results

Application

Inpainting

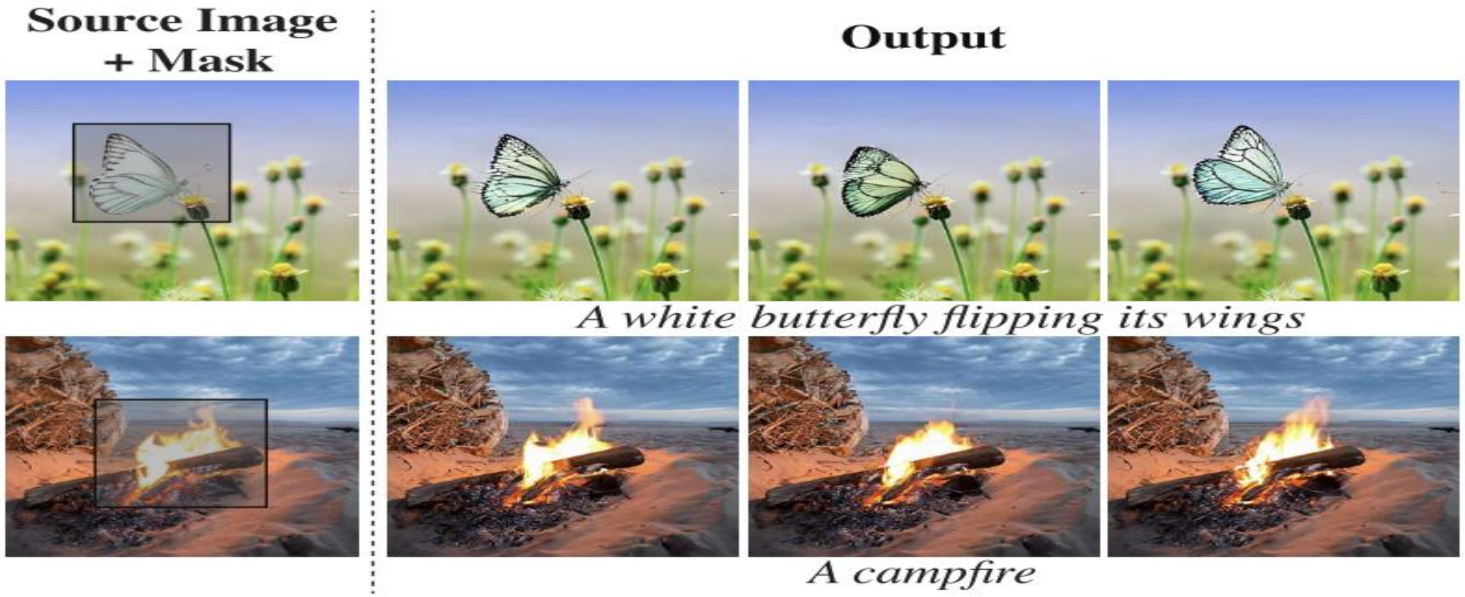


For a given video C:

- Animate the masked region

Application

Cinemagraphs



For a given input image C and mask M

- Generate a masked video

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