Lumiere: A Space-Time Diffusion Model for Video Generation

Omer Bar-Tal* 1 2 Hila Chefer* 1 3 Omer Tov* 1 Charles Herrmann †1 Roni Paiss †1 Shiran Zada †1
Ariel Ephrat †1 Junhwa Hur †1 Guanghui Liu 1 Amit Raj 1 Yuanzhen Li 1 Michael Rubinstein 1
Tomer Michaeli 1 4 Oliver Wang 1 Deqing Sun 1 Tali Dekel 1 2 Inbar Mosseri †1

Google Research

날짜 : 2024-11-26

발표자 : 홍진욱

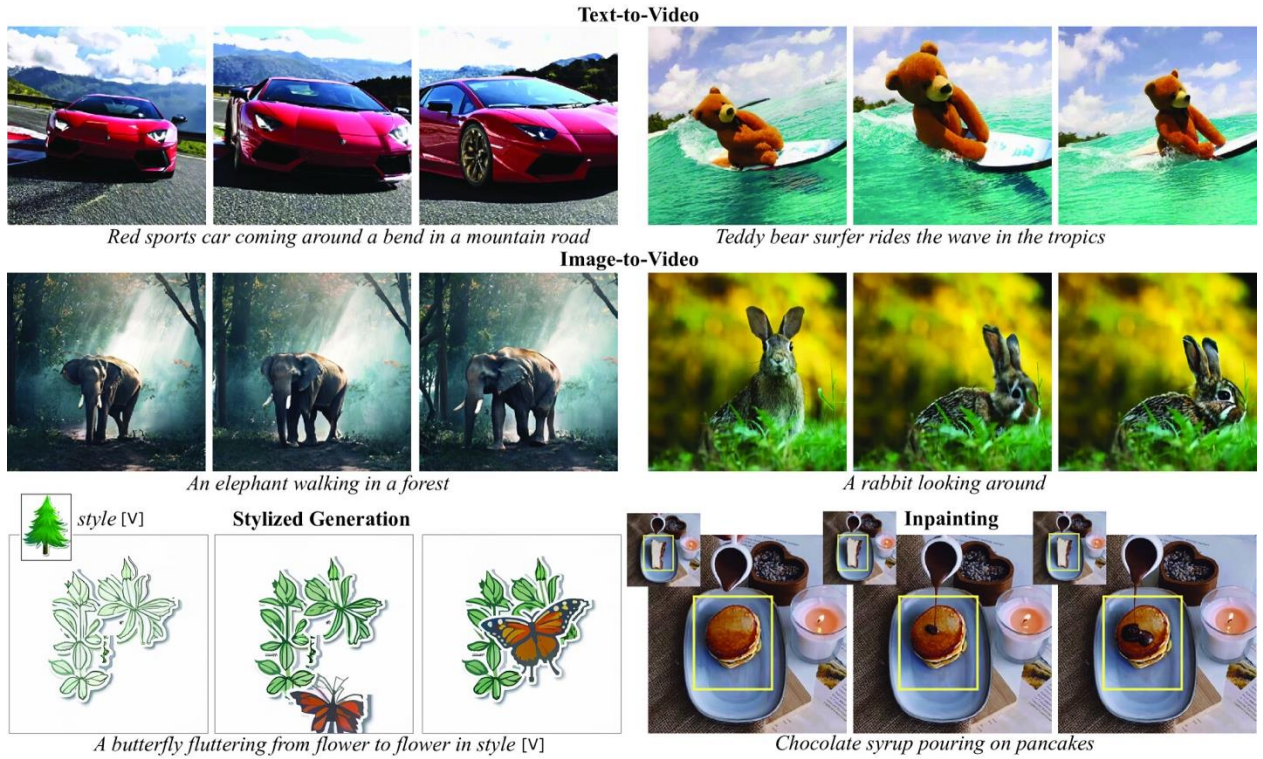
9th-together-RL

Contents

Introduction

Background

- While deep generative models have shown astonishing results in the visual domain.
- Recent progress achieves promising results of video generation such as text-to-video, image-to-video, and video-to-video.



Introduction

Summary

- Lumiere successfully extend the capability of a pretrained text-to-image model (Imagen) into video generation for various applications.
- Different from common approaches, which adopt a cascade of temporal super-resolution, Lumiere can generate full frames at once (80 frames, 16 fps, 5s).
- Space-Time UNet (STUNet) + Temporal super-resolution w/ multi-diffusion
- Competitive results on benchmark (UCF101) and user-study.

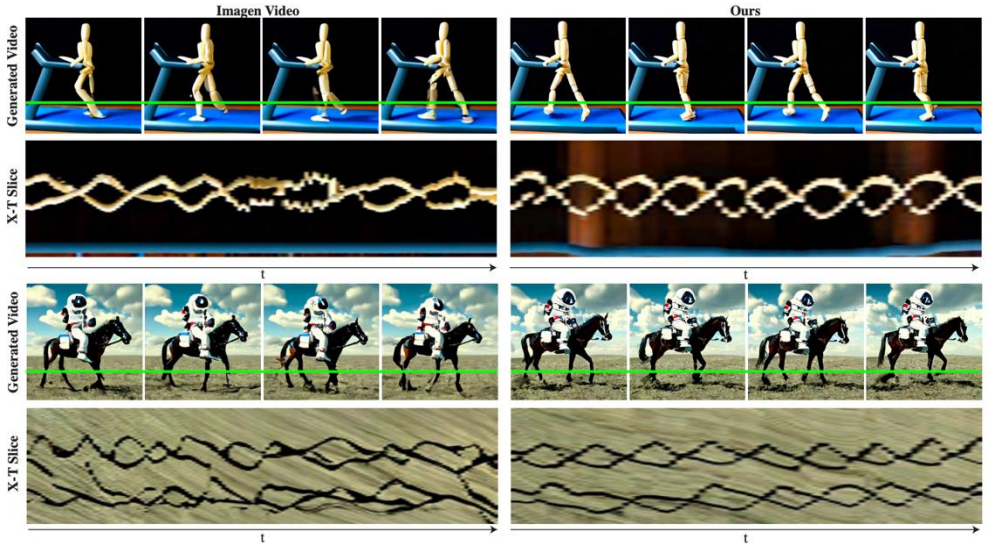
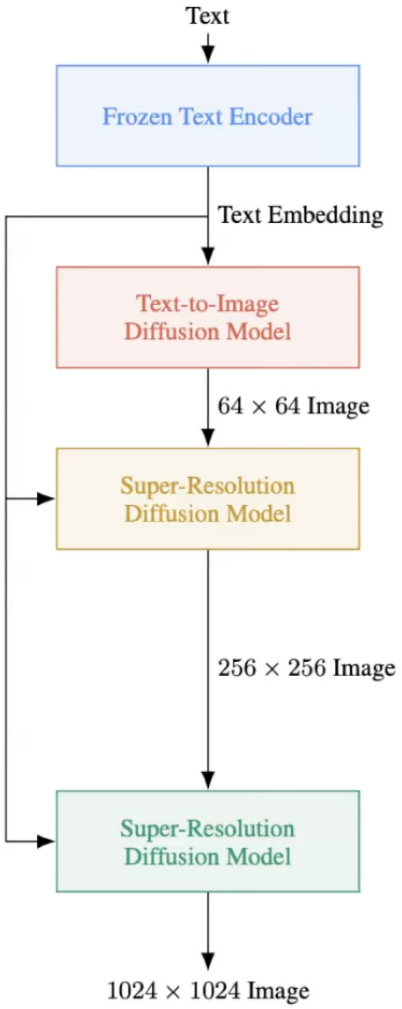


Imagen (NeurIPS '22)



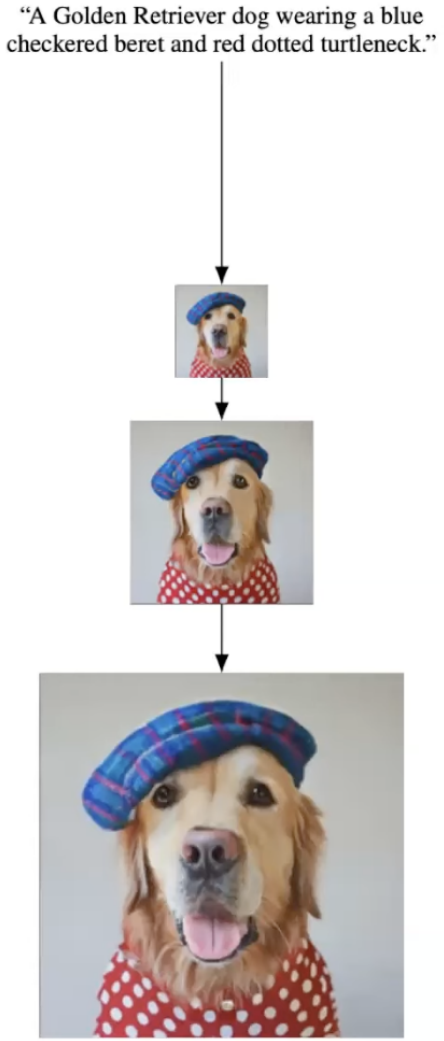
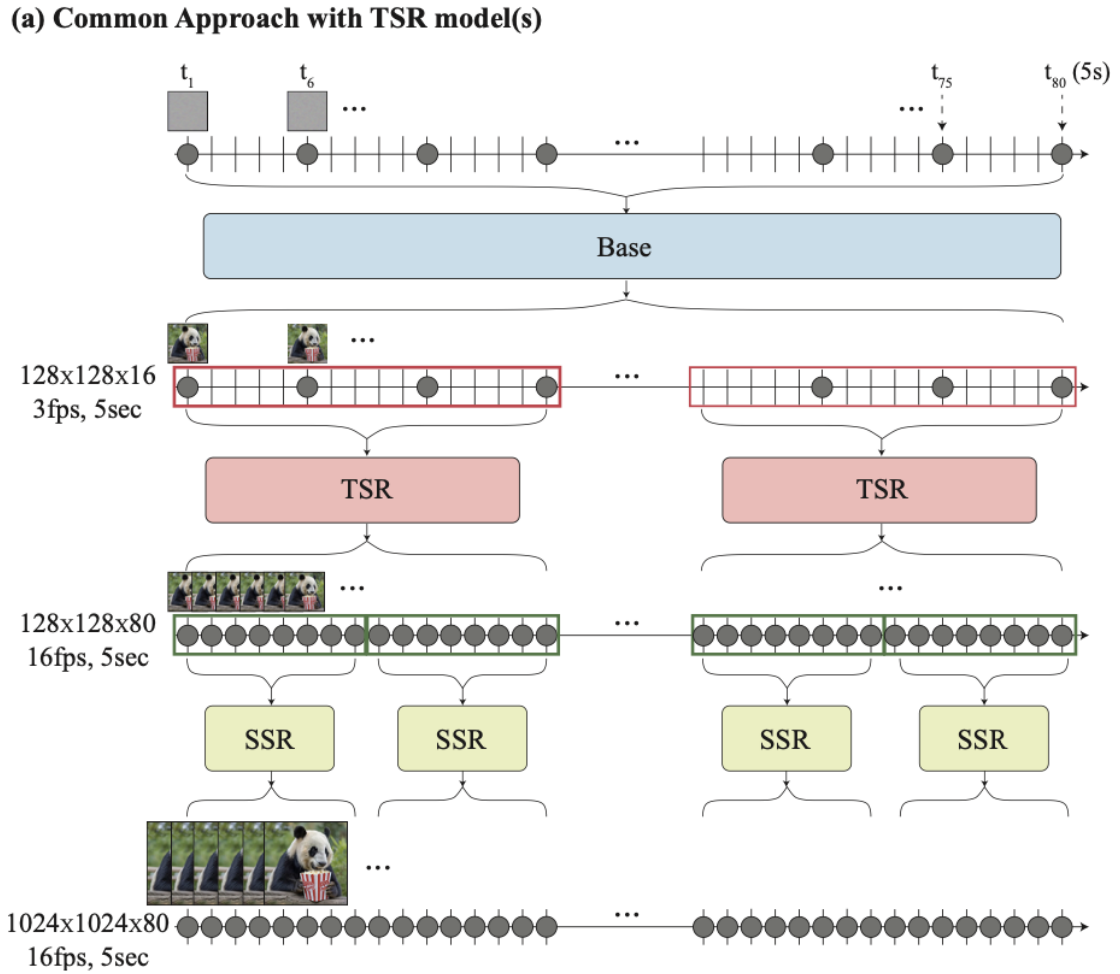


Imagen Video for Text-to-Video Generation (ArXiv '22)

- Imagen video generates the contents of videos from a user-provided text prompt, sequentially increasing the spatial and temporal resolutions.
- Here, spatial and temporal super-resolutions are separately processed.

Common Approaches of T2V vs. Lumiere



First, generate key frames at a low resolution and a low frame ratio (fps)

Divide the total duration into some chunks, and conduct temporal super-resolution (TSR).

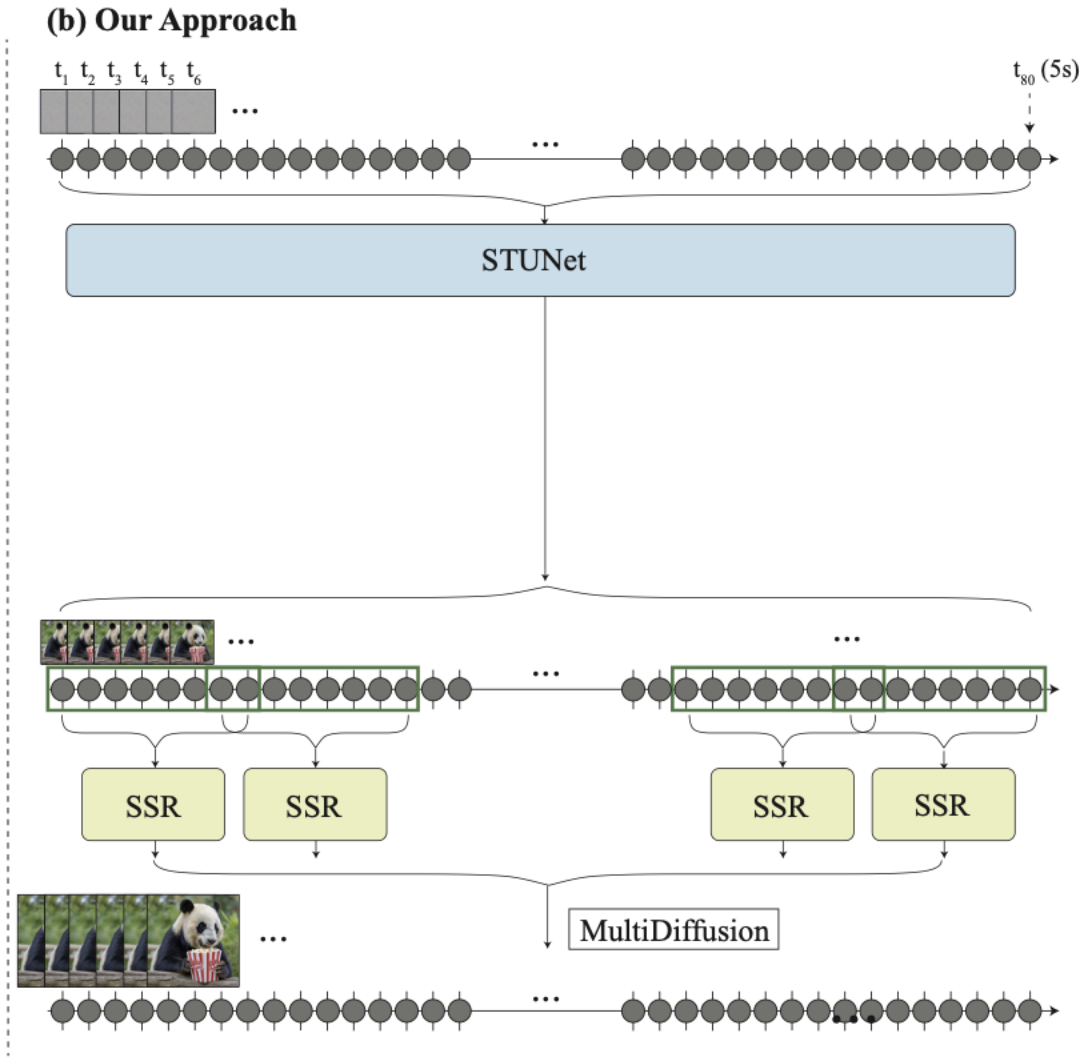
- (-) base model: aliasing on fast motions
- (-) TSR model: inconsistent motions, covariate shifts

For each chunk at the high frame ratio, conduct spatial super-resolutions.

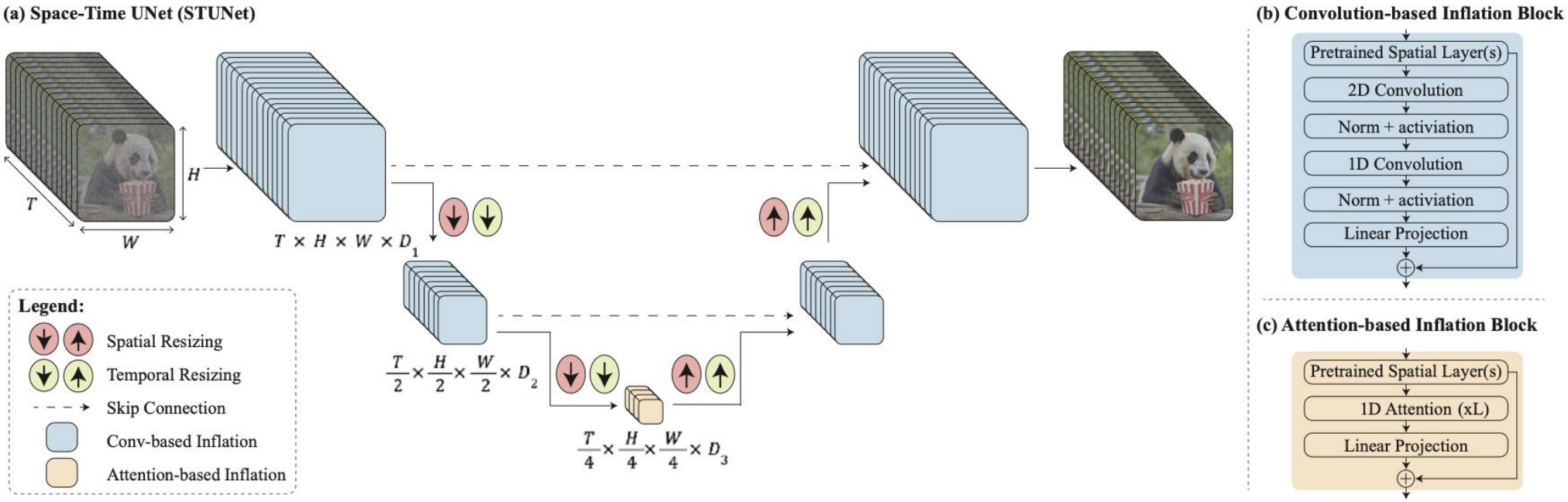
Common Approaches of T2V vs. Lumiere

Generate whole frames at a low spatial resolution (128x128, 80 frames, 16 fps, 5 seconds)

Conduct spatial super-resolutions for overlapped video clips
using the technique of multi-diffusions.

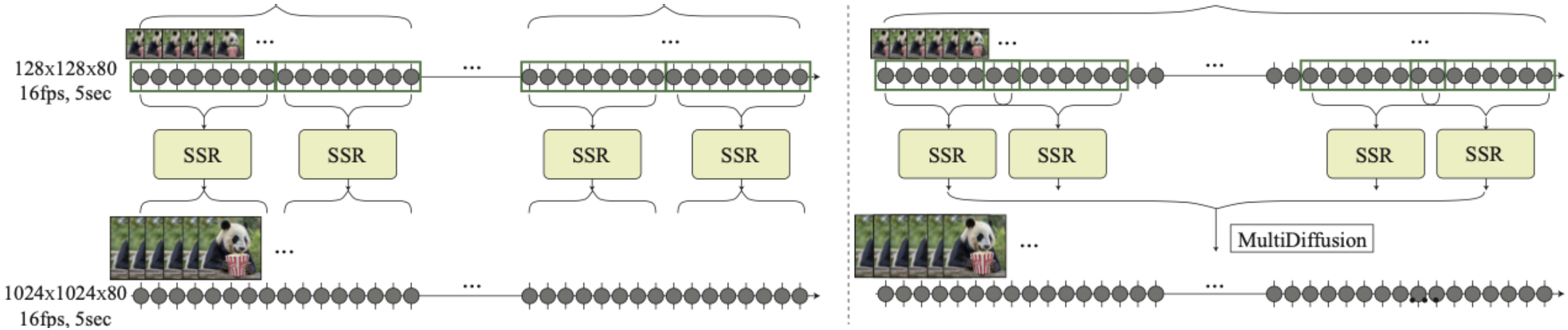


Space-Time UNet (STUNet)



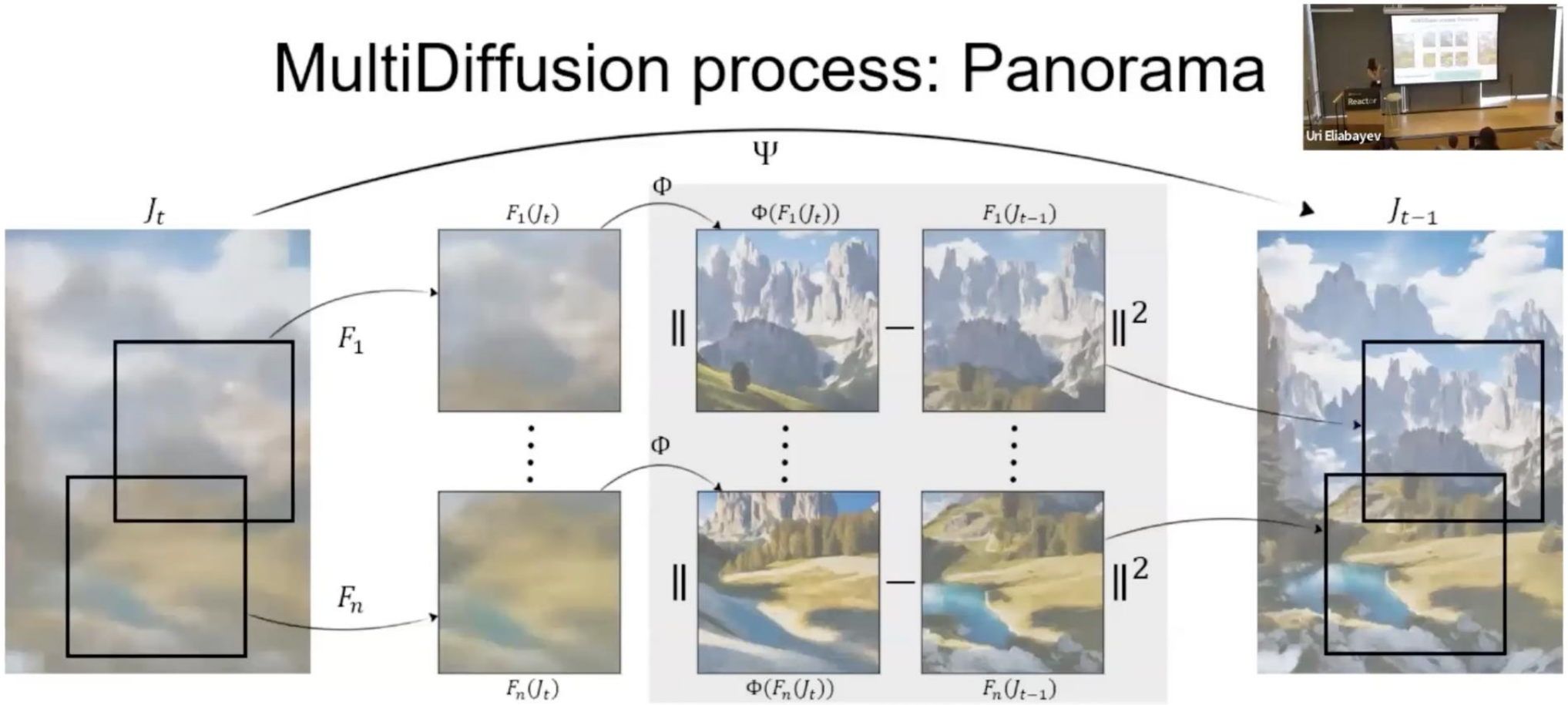
- Temporal layers are added into the base model of Imagen that freeze its weights.
- For the computational efficiency, temporal attentions are used only for coarsest level.

Common Approaches of T2V vs. Lumiere



- Due to memory constraints, the inflated SSR network can operate only on short segments of the video.
- To avoid temporal boundary artifacts, Lumiere employs Multidiffusion along the temporal axis for smooth transitions.

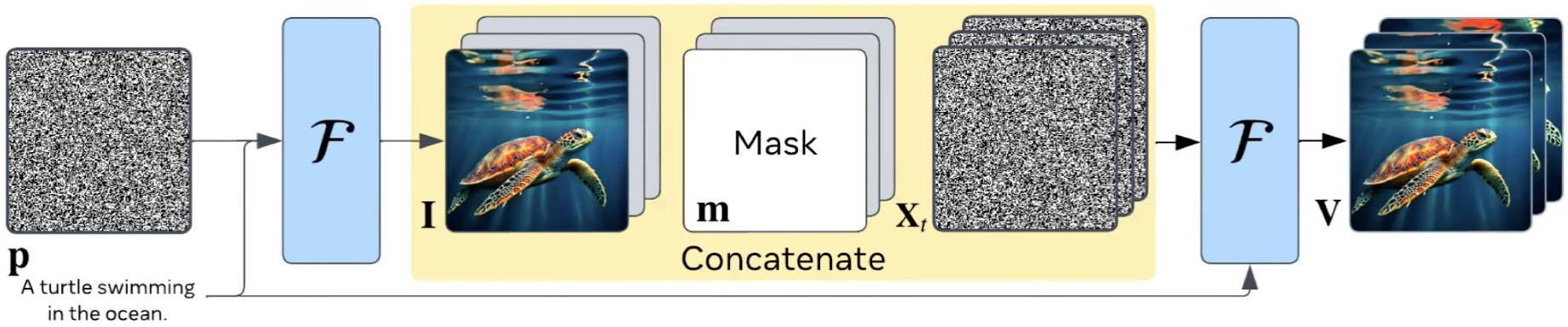
Space-Time UNet (STUNet)



Ψ as close as possible to Φ

$$\Psi(J_t) = \text{argmin}_J \sum_{i=1}^{n} ||F_i(J) - \Phi(F_i(J_t)||^2$$

Conditional Video Generation

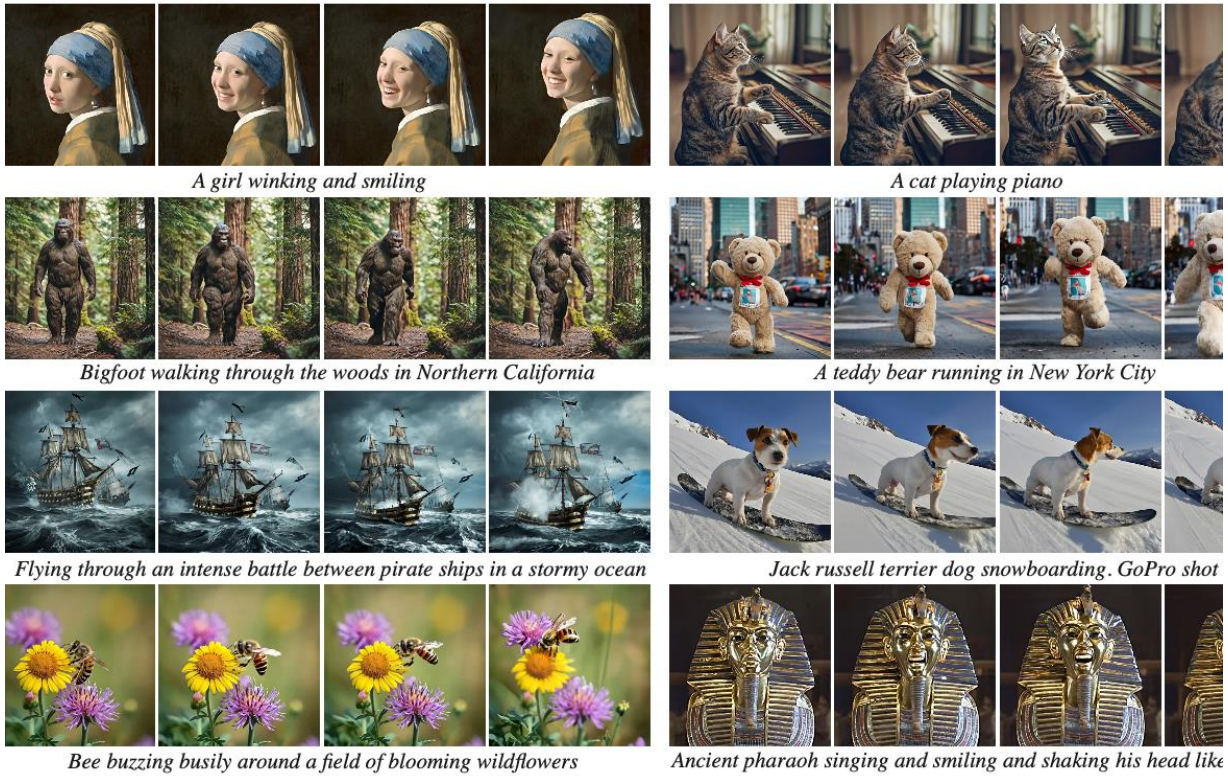


- For conditional video generation, an input condition and a binary mask map are concatenated into the noised input of STUNet.
- Various applications can be implemented with different making patterns : Image-to-Video, Inpainting, Cinemagraphs

Conditional Video Generation

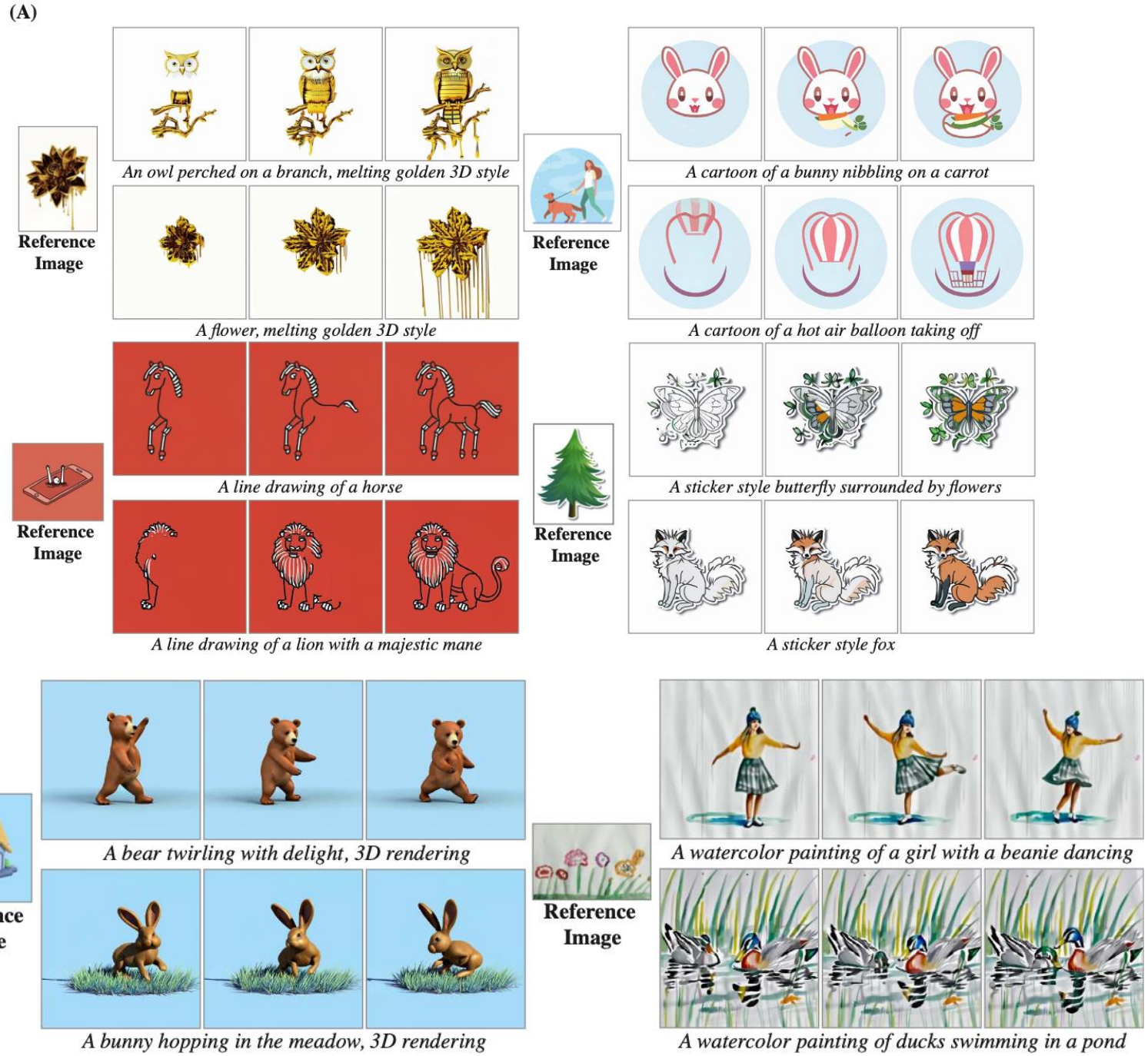
Text-to-Video

Image-to-Video

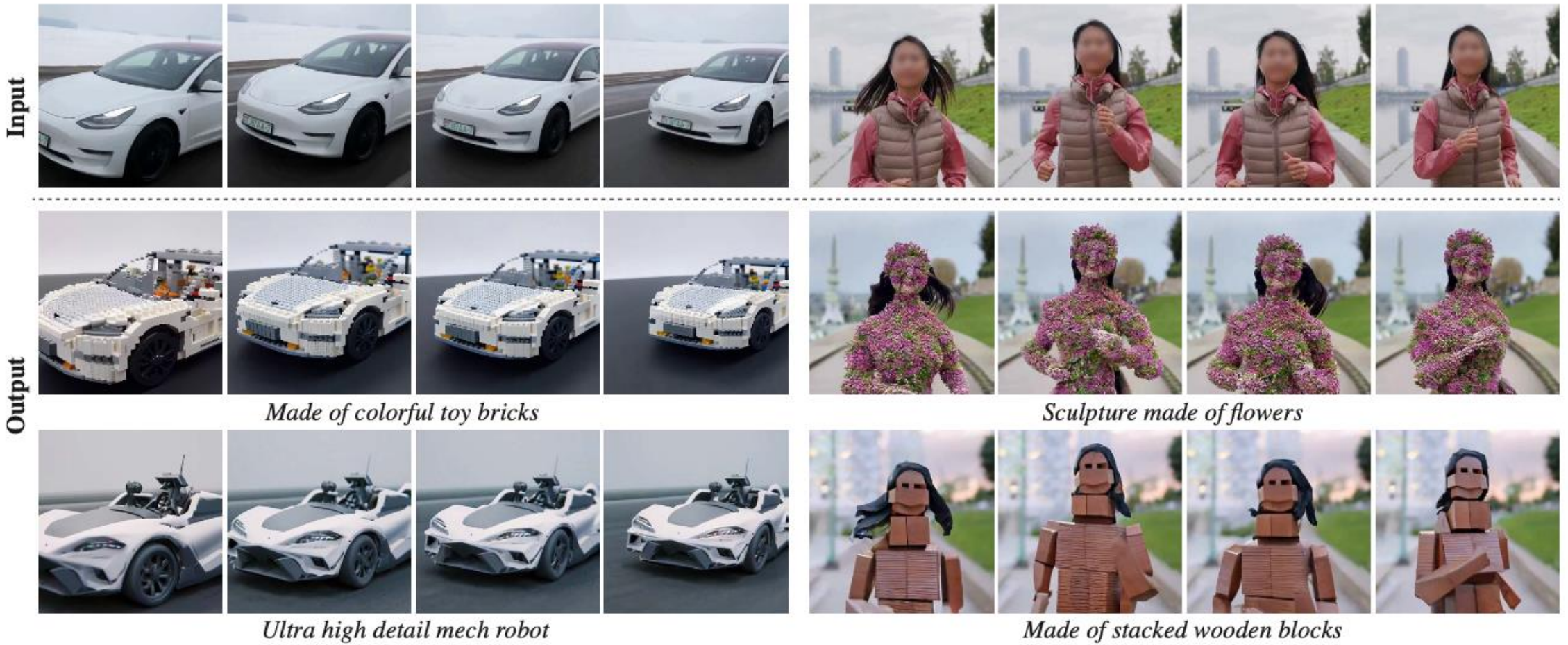


Stylized Generation

$$W_{\text{interpolate}} = \alpha \cdot W_{\text{style}} + (1-\alpha) \cdot W_{\text{orig}}.$$



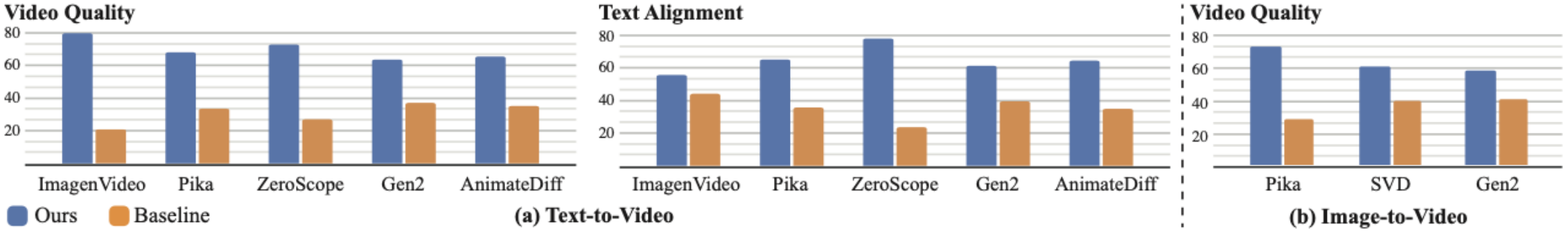
Video-to-Video via SDEdit



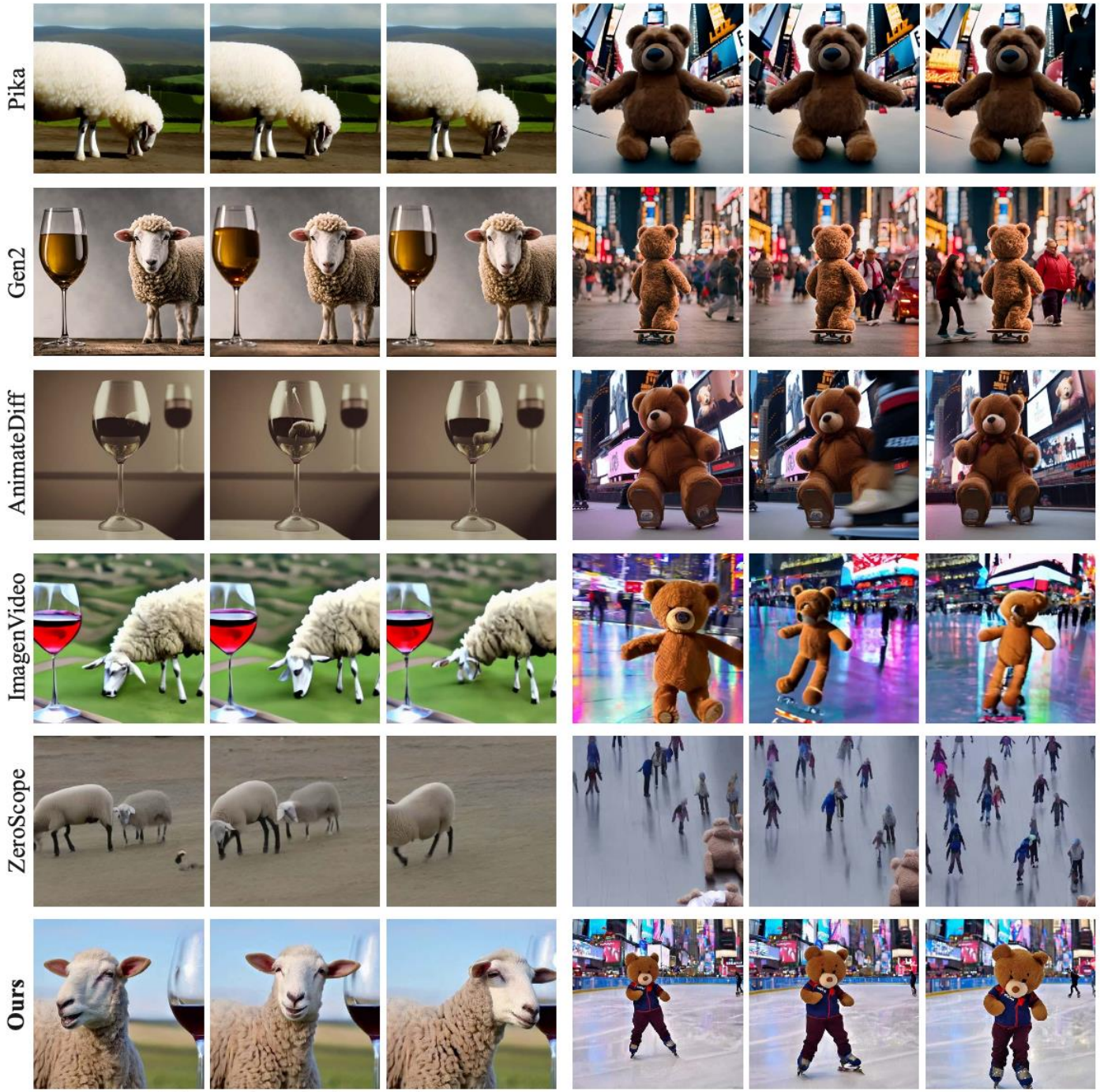
Results

- Zero-shot text-to-video on UCF101.
- User study on 113 text prompts (~ 400 user judgements).

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



Results


Pika
Gen2
AnimateDiff
ImagenVideo
ZeroScope
Ours
A sheep to the right of a wine glass
Teddy bear skating in Times Square

Conclusion

- A new text-to-video generation framework is presented using a pretrained T2I model.
- To learn globally-coherent motion, a space-time U-Net architecture design enables to generate the whole frames at once with a low resolution.
- Lumiere shows State-of-the-art results and is applicable for various applications.