Microsoft

CVPR
JUNE 17-21, 2024
SEATTLE, WA

Video and 3D Generation

Kevin Lin

6/17/2024

Evolution of Video Diffusion Model

Google

VDM 2022

Meta AI

Emu Video 2023

OpenAI

Sora 2024

Timeline

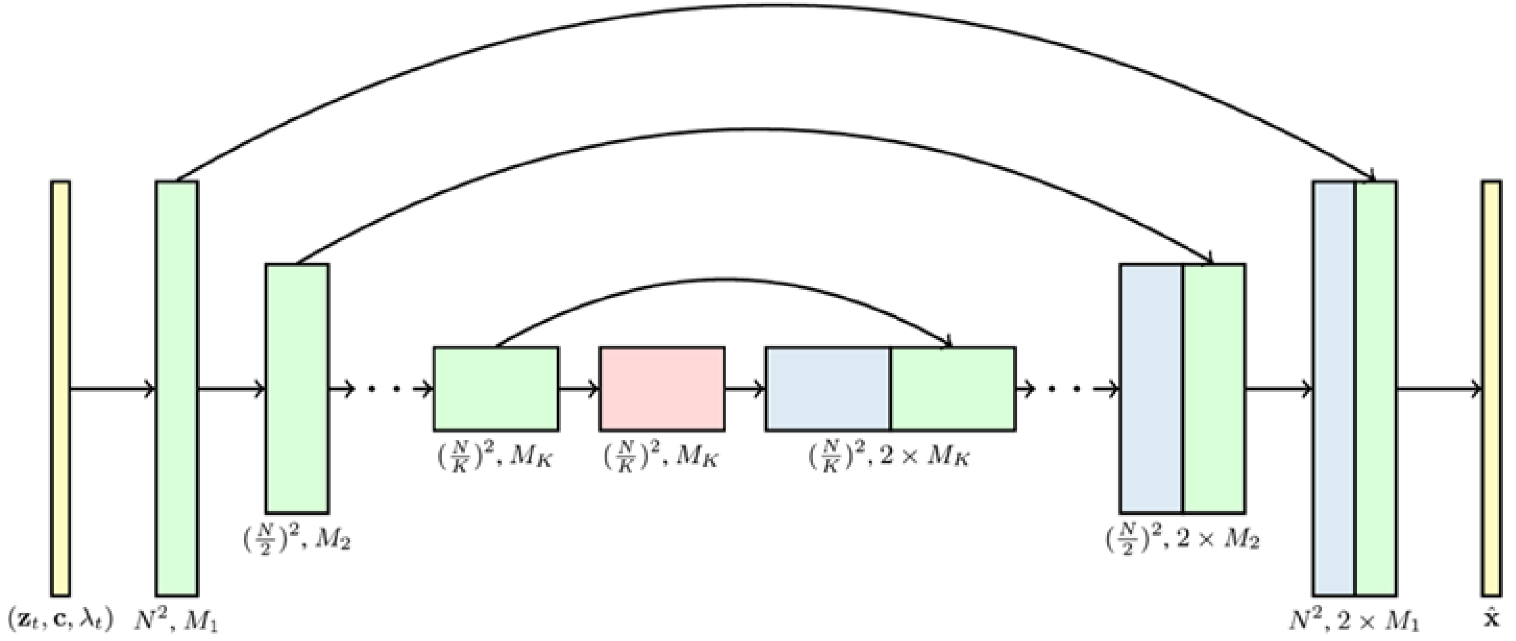
- Video-Gen Foundation Model
- Faster training
- Faster inference
- Diverse content creation

Creating Video from Text

Prompt: A flock of paper airplanes flutters through a dense jungle, weaving around trees as if they were migrating birds.

Video Diffusion Models: *Pioneer Work*

- 3D UNet factorized over space and time
  - 2D conv is inflated to 3D
- Insert temporal attention layer that attends across the temporal dimension

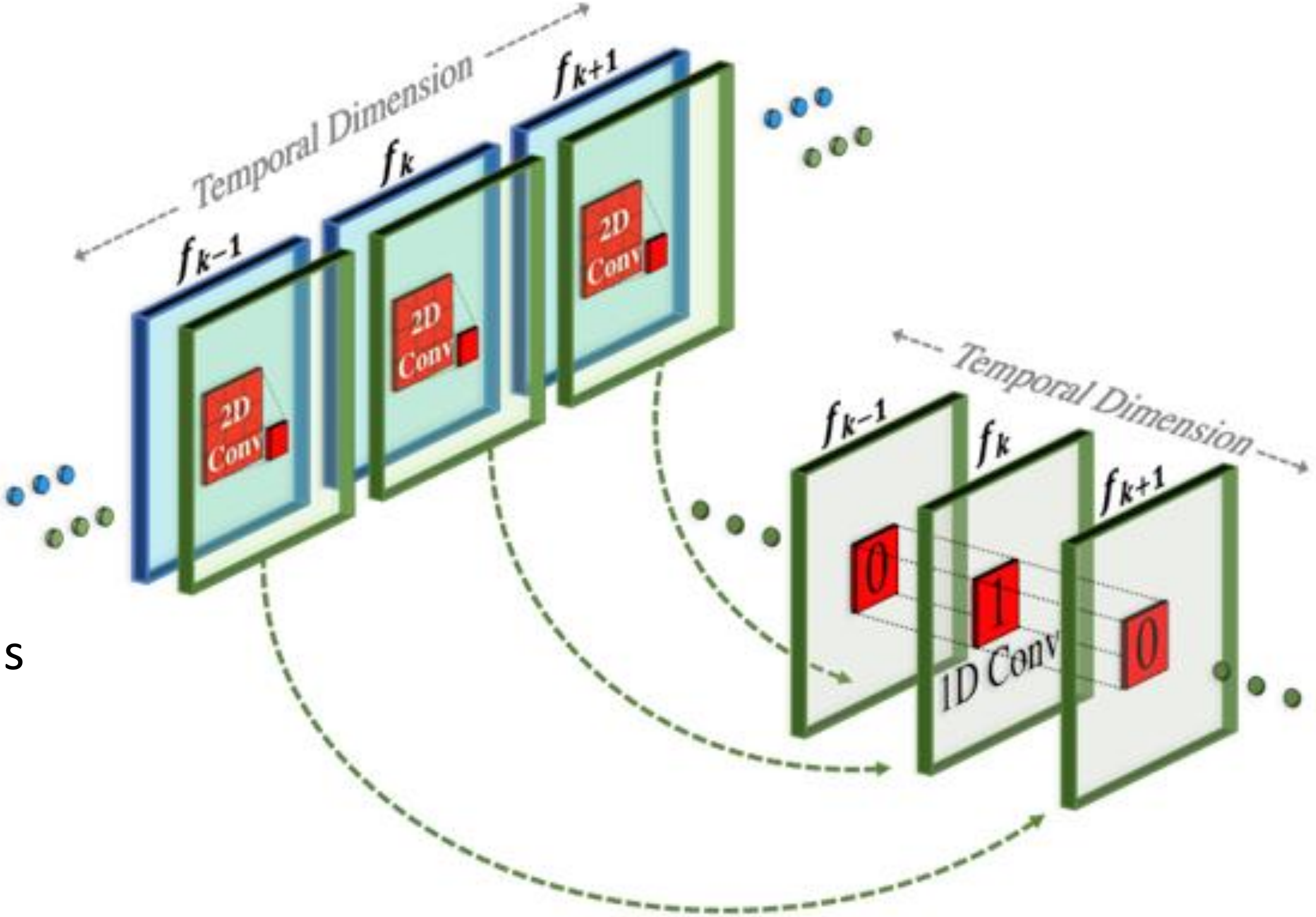


Video Diffusion Models (Ho et al. NeurIPS'22)

Make-A-Video

Spatial Convolution

Initialized from
pre-trained T2I models



Temporal Convolution

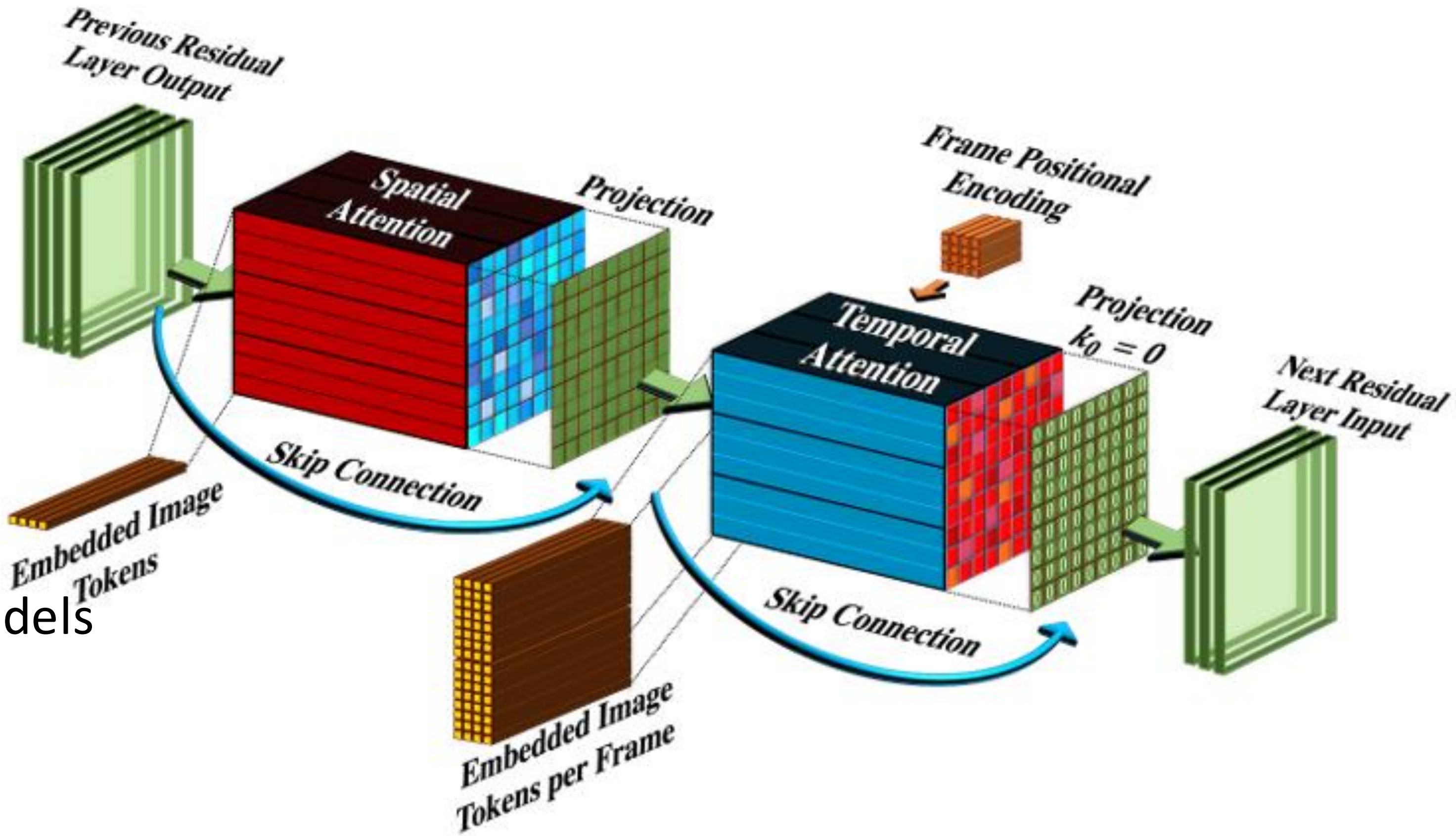
Initialized with
identity function

Make-A-Video (Singer et al. ‘22)

Make-A-Video

Spatial Attention

Initialized from pre-trained T2I models



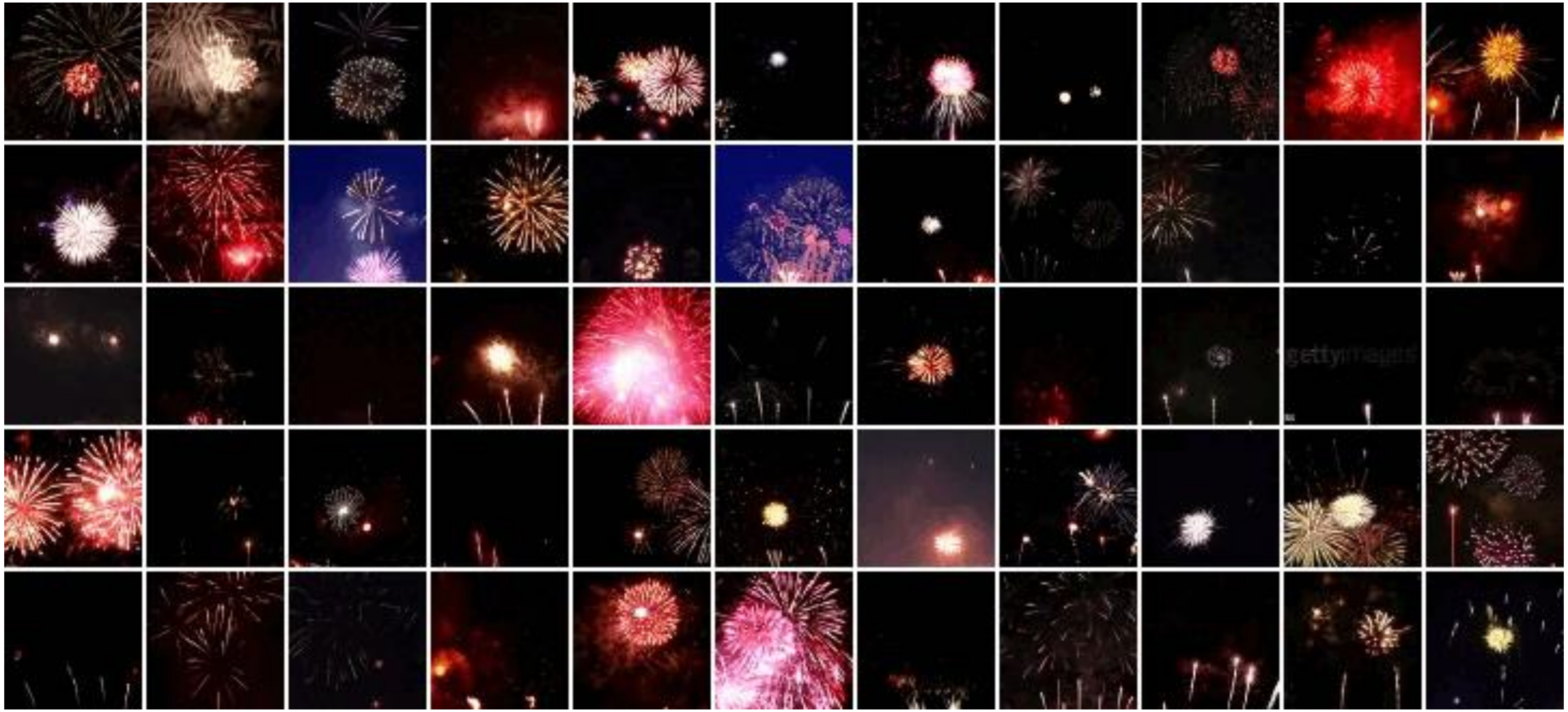
Temporal Attention

Initialized with zero projection (resulting in identity function)

Make-A-Video (Singer et al. ‘22)

Preliminary Results

Prompt: Firework



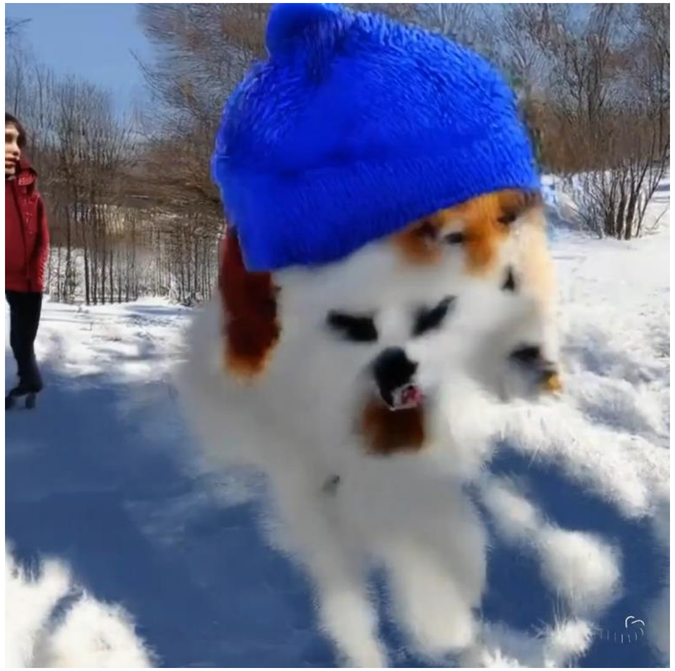
Video Diffusion Models (Ho et al. ‘22)

Curated Training Data Improves Performance

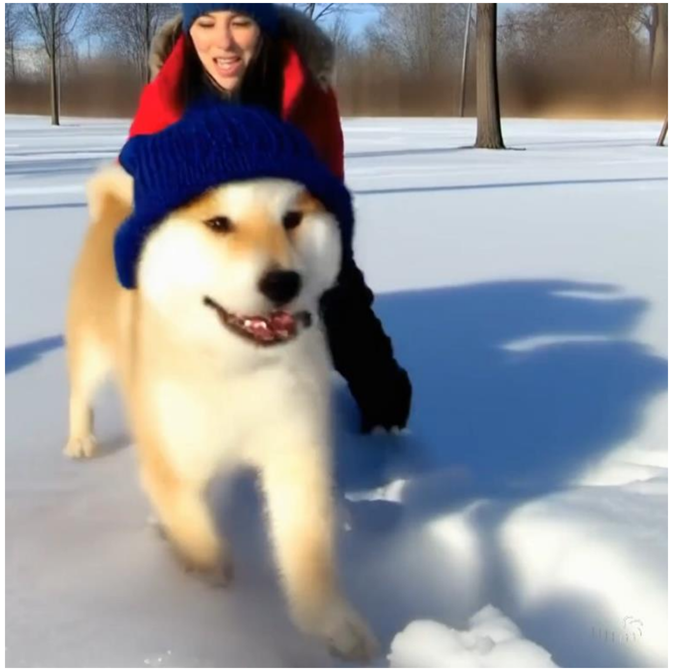
- Scaling training data from 10M to 577M video clips

Stable Video Diffusion (Blattmann et al. ‘23)

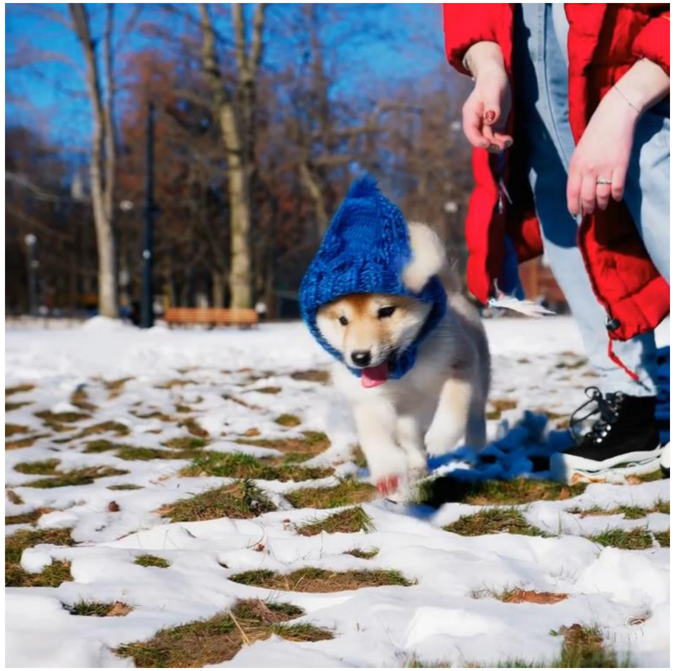
Quality Improves as Training Compute Increases



Base compute



4x compute



32x compute

Sora: Video generation models as world simulators (OpenAI ‘24)

- Video-Gen Foundation Model
- Faster training
- Faster inference
- Diverse content creation

Training a Video Diffusion Model is Expensive!

SOTA Video Diffusion Model

(Case Study: Stable Video Diffusion)

😭💸

577M Video Clips

1521M Parameters

Dedicated GPU clusters

Slow and complex training recipe

Faster Training

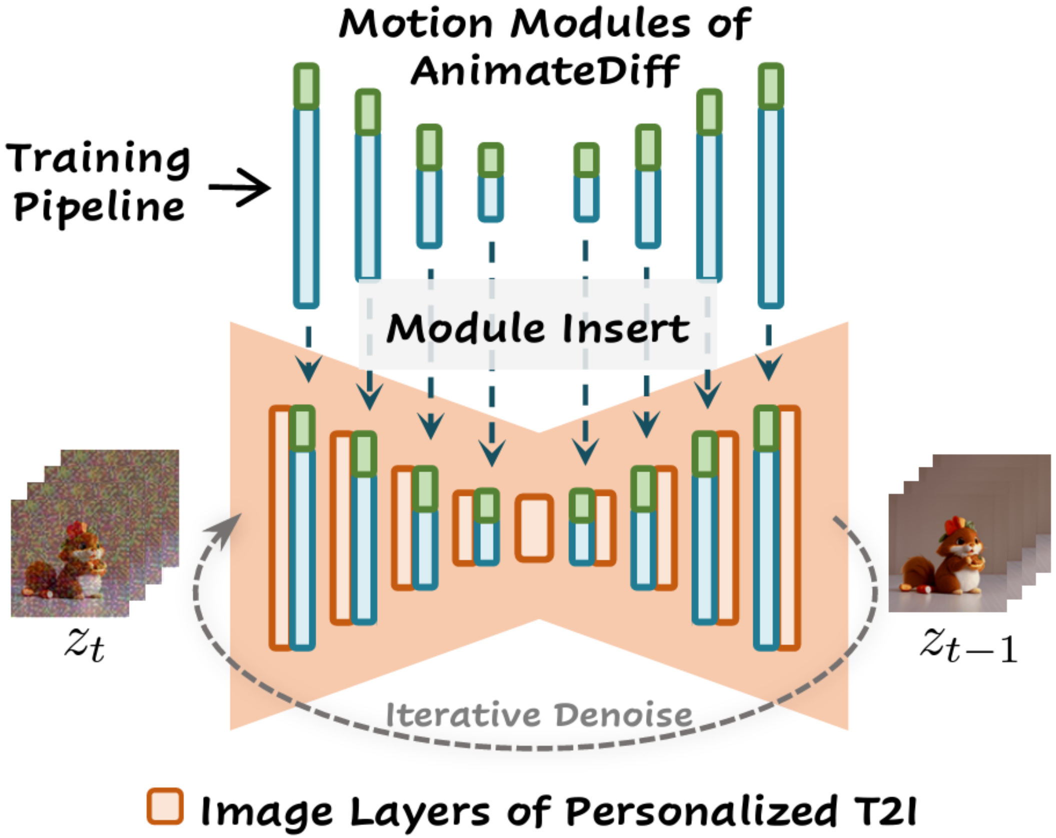
20~50 Video Clips

+30M Parameters

~2 hours training

😁

Faster Training with Motion Modules


Motion Modules of AnimateDiff
Training Pipeline
Module Insert
z_t
z_{t-1}
Iterative Denoise
🟧 Image Layers of Personalized T2I

20~50 Video Clips

+30M Parameters

~2 hours training

AnimateDiff (Guo et al. ICLR‘24)

We have faster training now. How about inference?

SOTA Video Diffusion Model

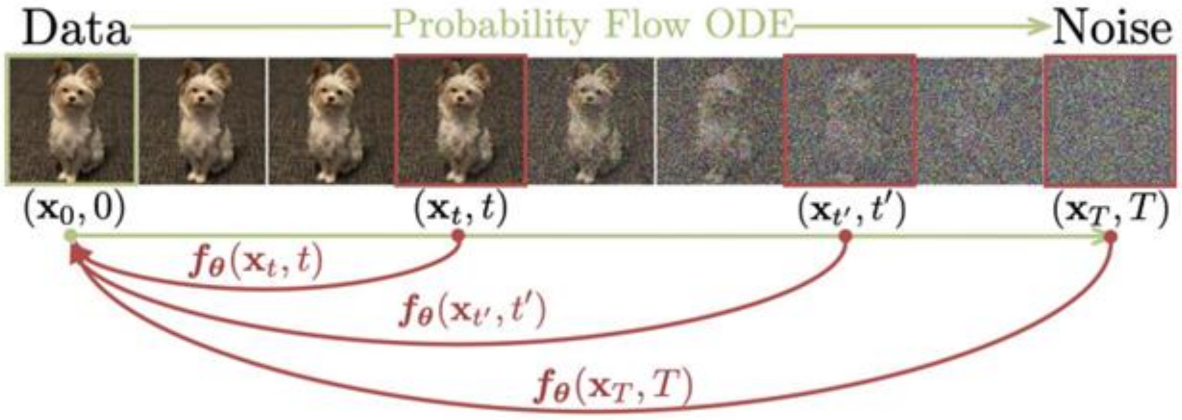
250 Steps

Faster Sampling

4 Steps

Challenges in Video Diffusion Distillation

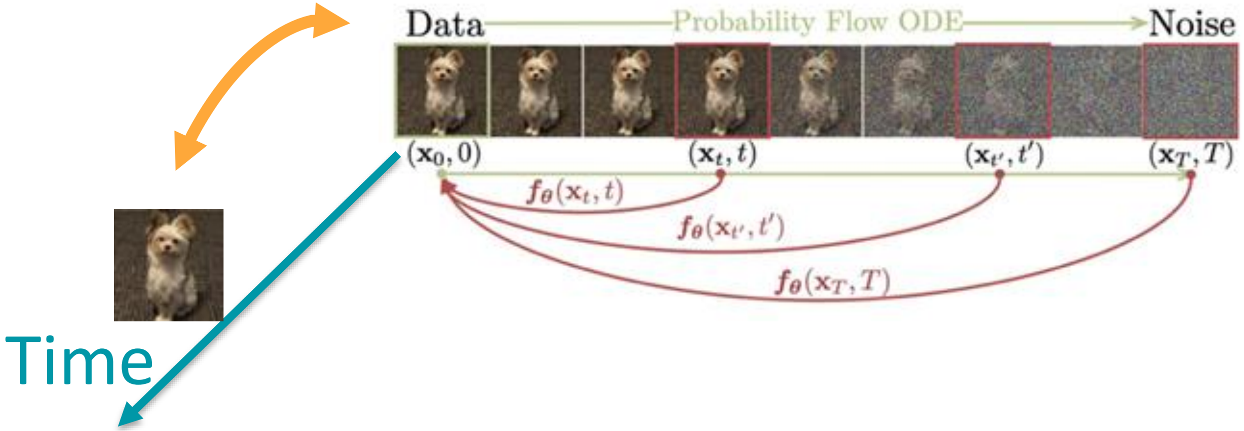
Image Diffusion Distillation


Data
Probability Flow ODE
Noise
$(\mathbf{x}_0, 0)$
$(\mathbf{x}_t, t)$
$(\mathbf{x}_{t'}, t')$
$(\mathbf{x}_T, T)$
$\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$
$\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}_{t'}, t')$
$\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}_T, T)$

Consistency Model (Song et al. ‘23)

Video Diffusion Distillation

Appearance Consistency


Data
Probability Flow ODE
Noise
$(\mathbf{x}_0, 0)$
$(\mathbf{x}_t, t)$
$(\mathbf{x}_{t'}, t')$
$(\mathbf{x}_T, T)$
$\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$
$\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}_{t'}, t')$
$\boldsymbol{f}_{\boldsymbol{\theta}}(\mathbf{x}_T, T)$

Time

Motion Consistency

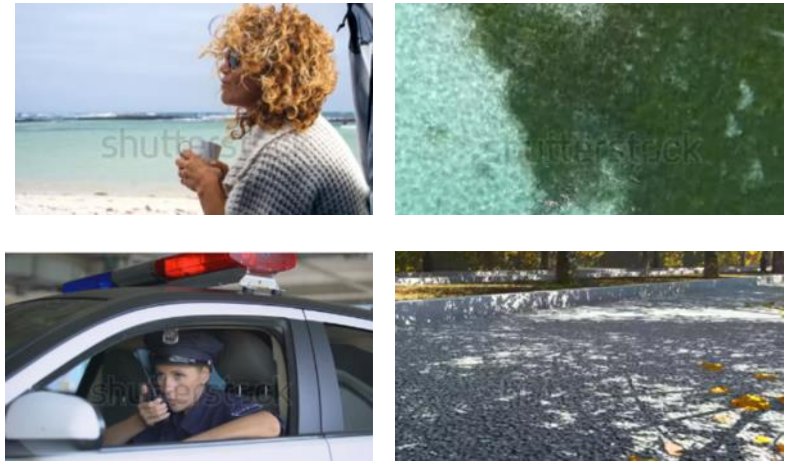
Motion Consistency Model (Zhai et al. ‘24)

Challenges in Video Diffusion Distillation

High Quality Image Data

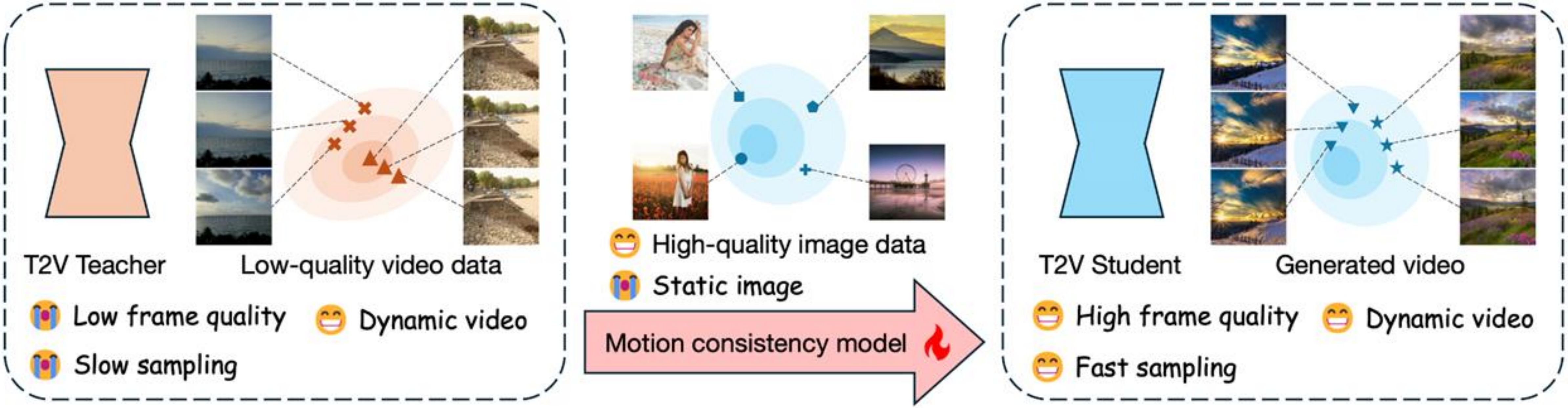
LAION-Aesthetics

Low Quality Video Data



WebVid-10M

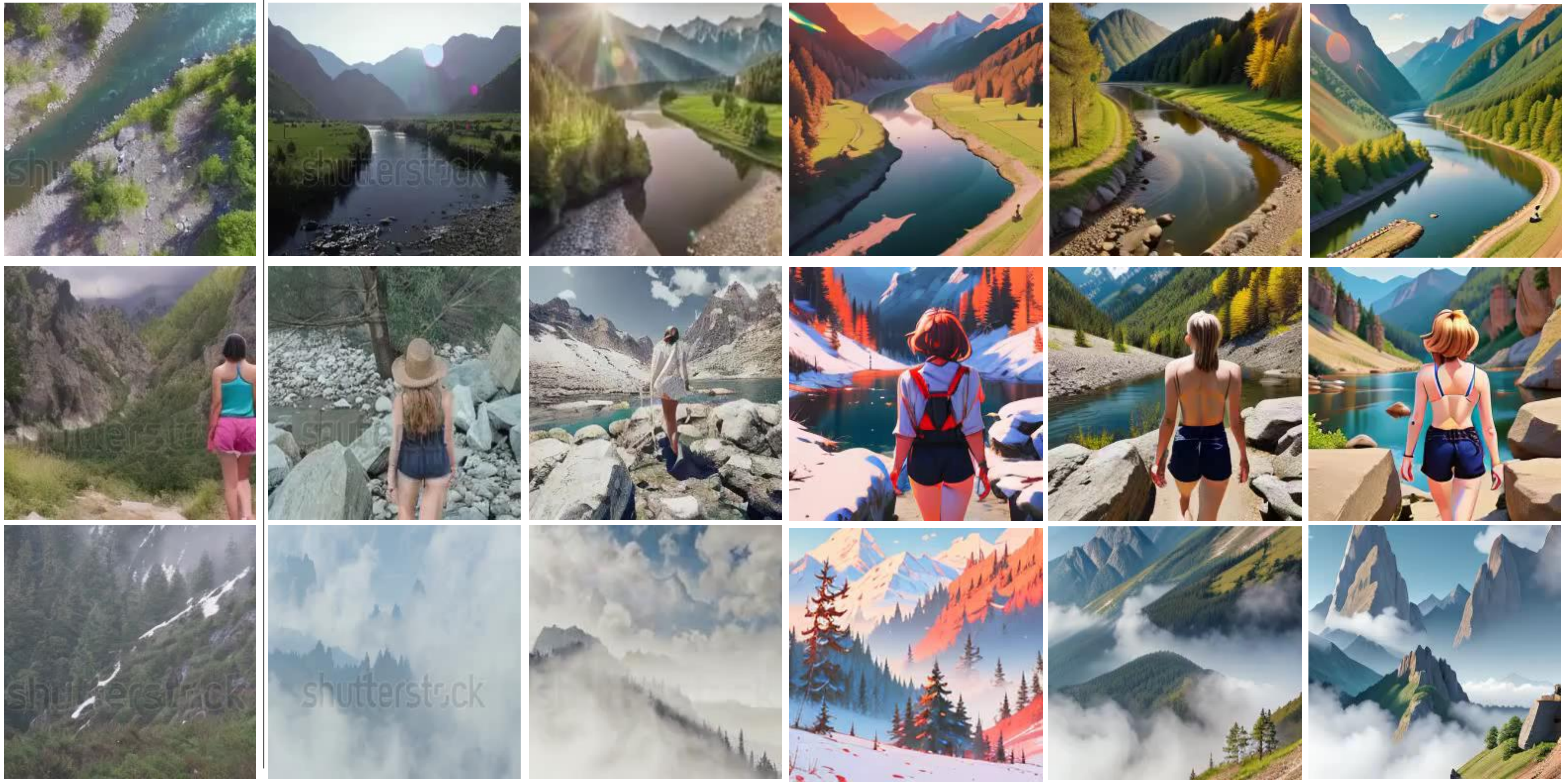
Motion Consistency Model



Our motion consistency model not only distill the motion prior from the teacher to accelerate sampling, but also can benefit from an additional high-quality image dataset to improve the frame quality of generated videos.

Motion Consistency Model (Zhai et al. ‘24)

Motion Consistency Model



Teacher
50 steps

MCM + WebVid
4 steps

MCM + LAION-aes
4 steps

MCM + Anime
4 steps

MCM + Realistic
4 steps

MCM + 3D Cartoon
4 steps

Aerial uhd 4k view. mid-air flight over fresh and clean mountain river at sunny summer morning. Green trees and sun rays on horizon. Direct on sun.

Back of woman in shorts going near pure creek in beautiful mountains.

Misty mountain landscape

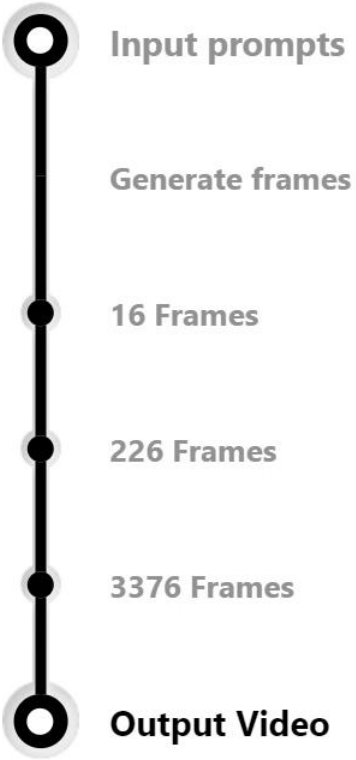
Motion Consistency Model (Zhai et al. ‘24)

- ✓ Video-Gen Foundation Model
- ✓ Faster training
- ✓ Faster inference
- ✓ Diverse content creation:

  Variable durations Controllability Consistency

LONG VIDEO

Given the prompts of a script, NUWA-XL can generate an extremely long video that conforms to it in a "coarse-to-fine" process.



Input prompts

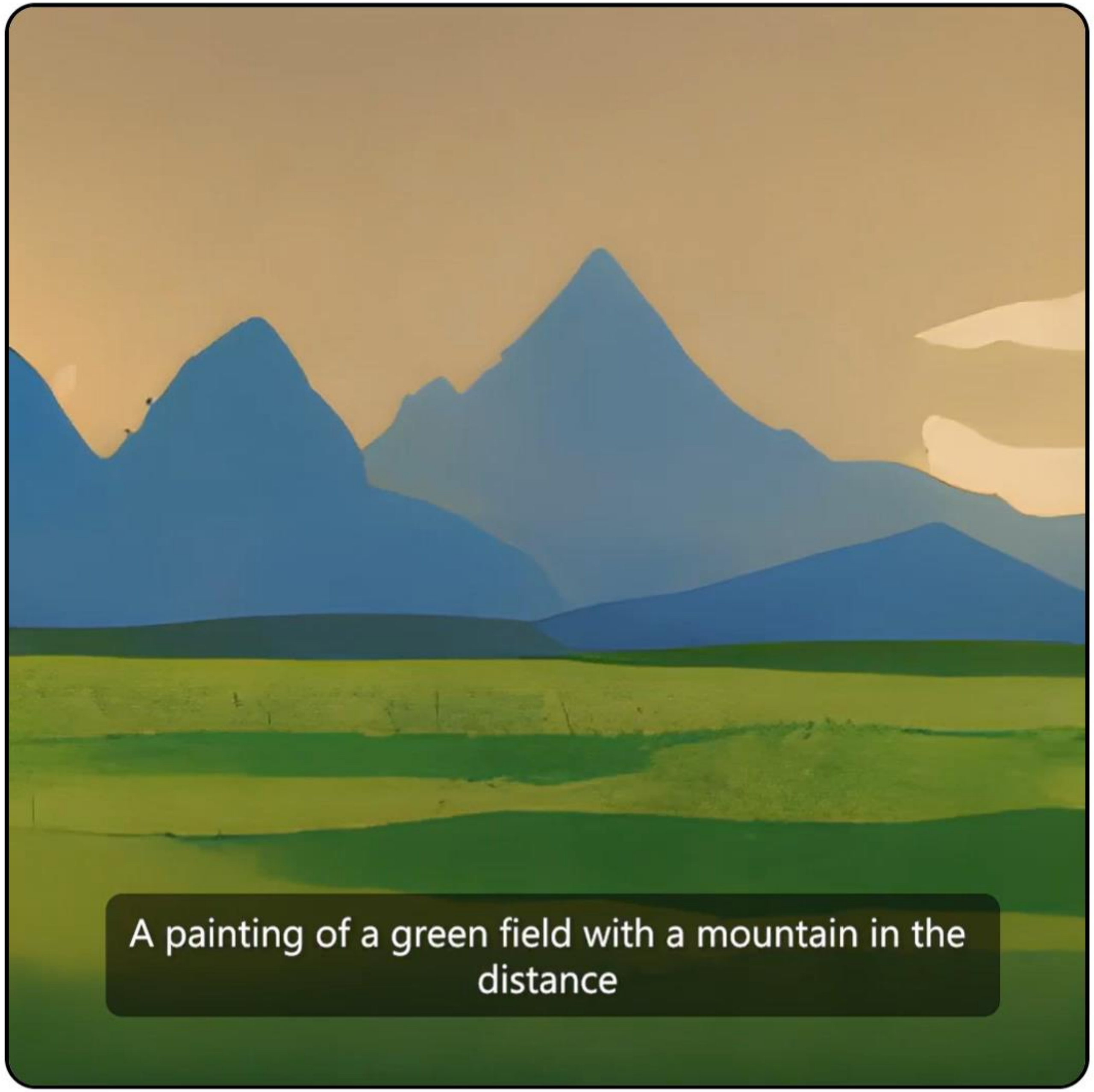
Generate frames

16 Frames

226 Frames

3376 Frames

Output Video



A painting of a green field with a mountain in the distance

Diffusion over diffusion

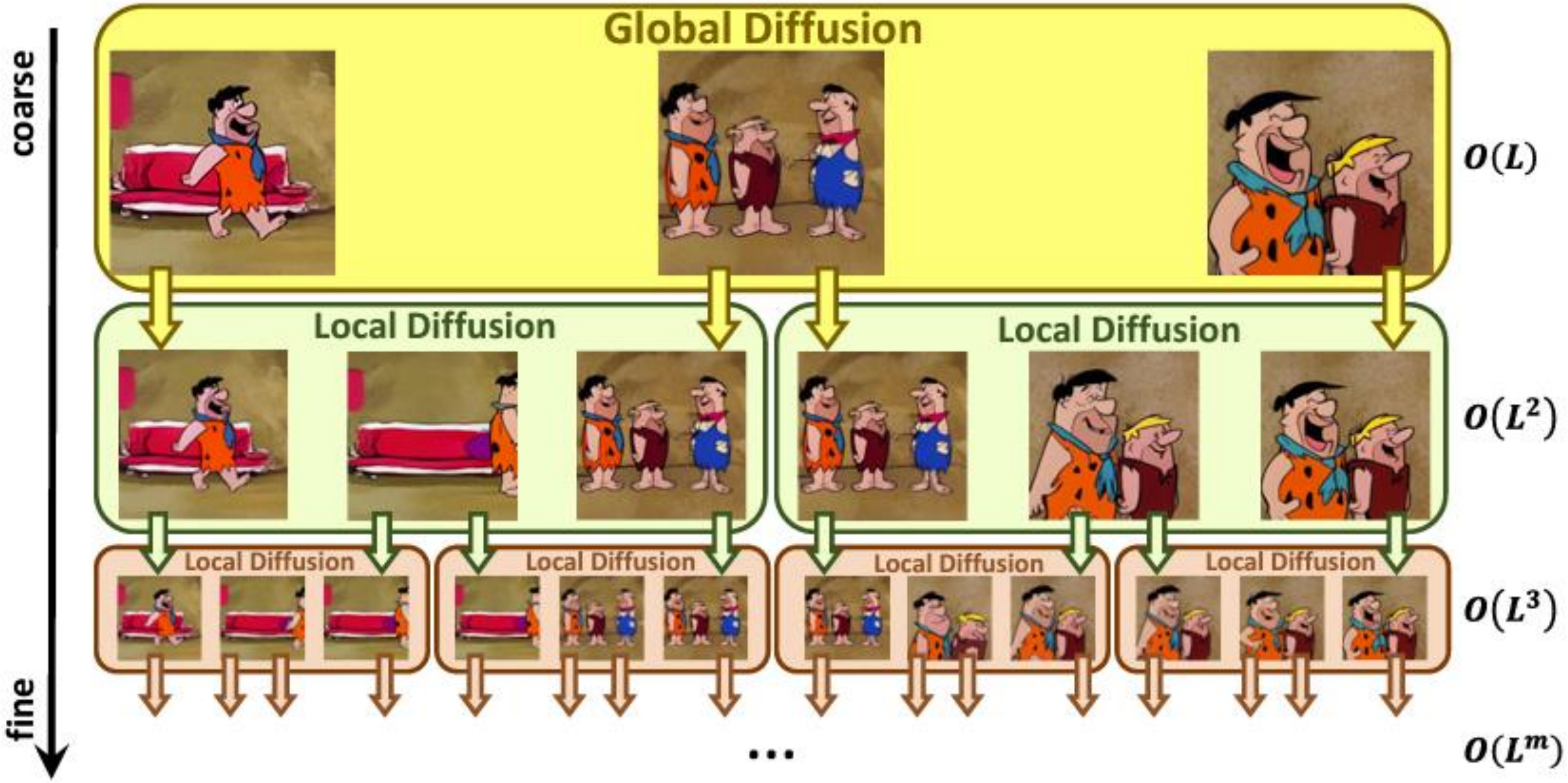
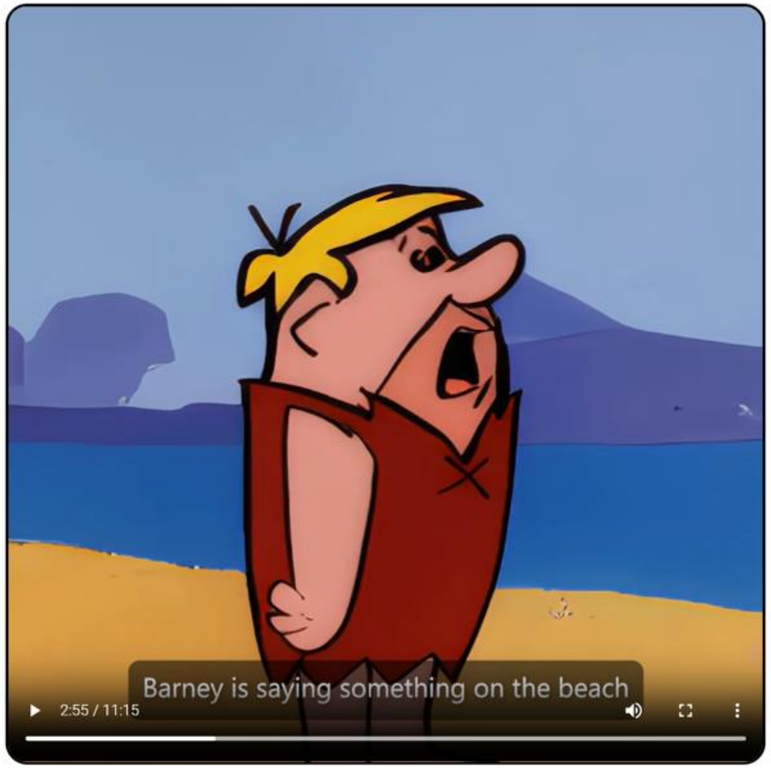


Figure 1: Overview of NUWA-XL for extremely long video generation in a "coarse-to-fine" process. A global diffusion model first generates L keyframes which form a "coarse" storyline of the video, a series of local diffusion models are then applied to the adjacent frames, treated as the first and the last frames, to iteratively complete the middle frames resulting $O(L^m)$ "fine" frames in total.

NUWA-XL (Yu et al., ACL'23)

Simple Cartoon Videos

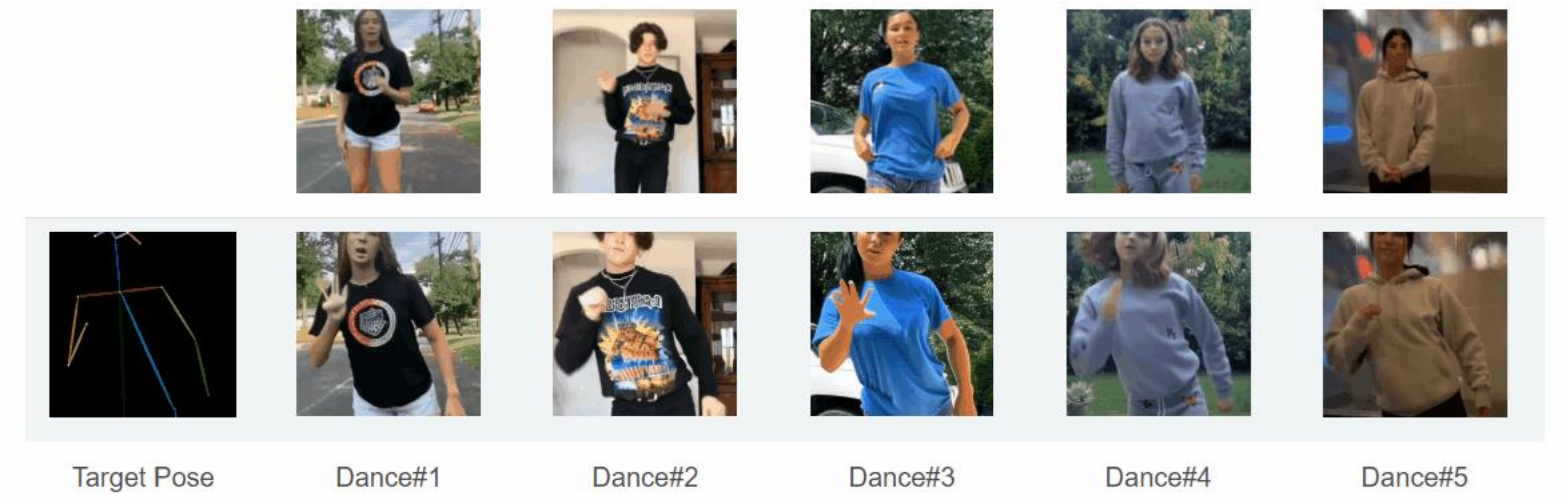
Rich and Diverse Contents?



Barney is saying something on the beach

Human Dance Generation

- Different subject – Same pose

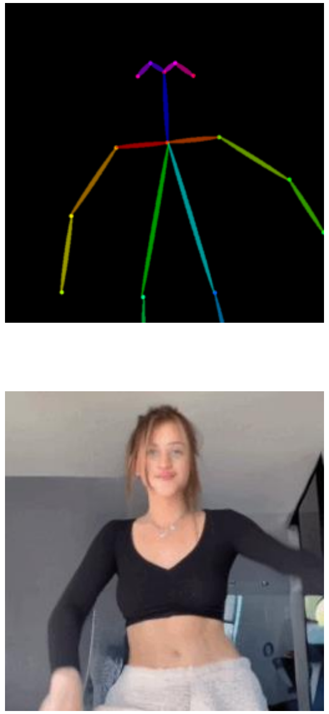


DisCo (Wang et al. CVPR‘24)

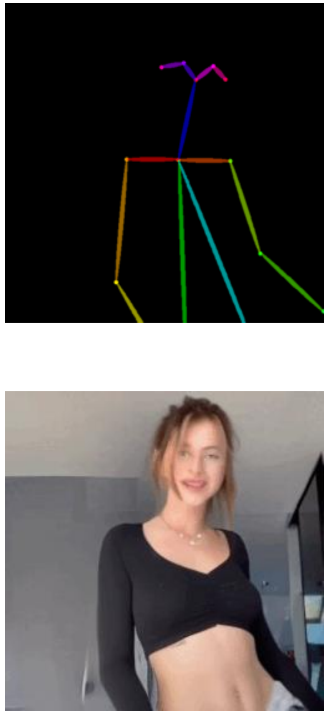
Human Dance Generation

- Same subject – Different pose

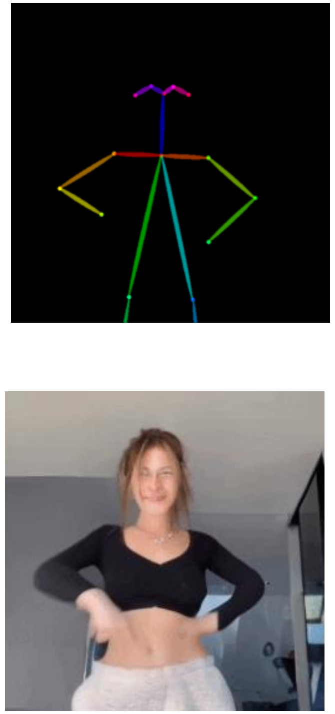
Reference Image



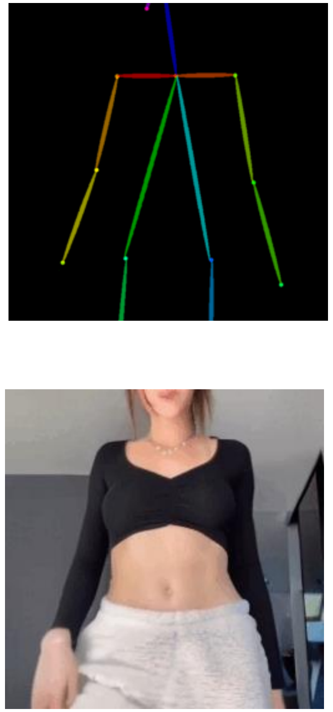
Dance #1



Dance #2



Dance #3

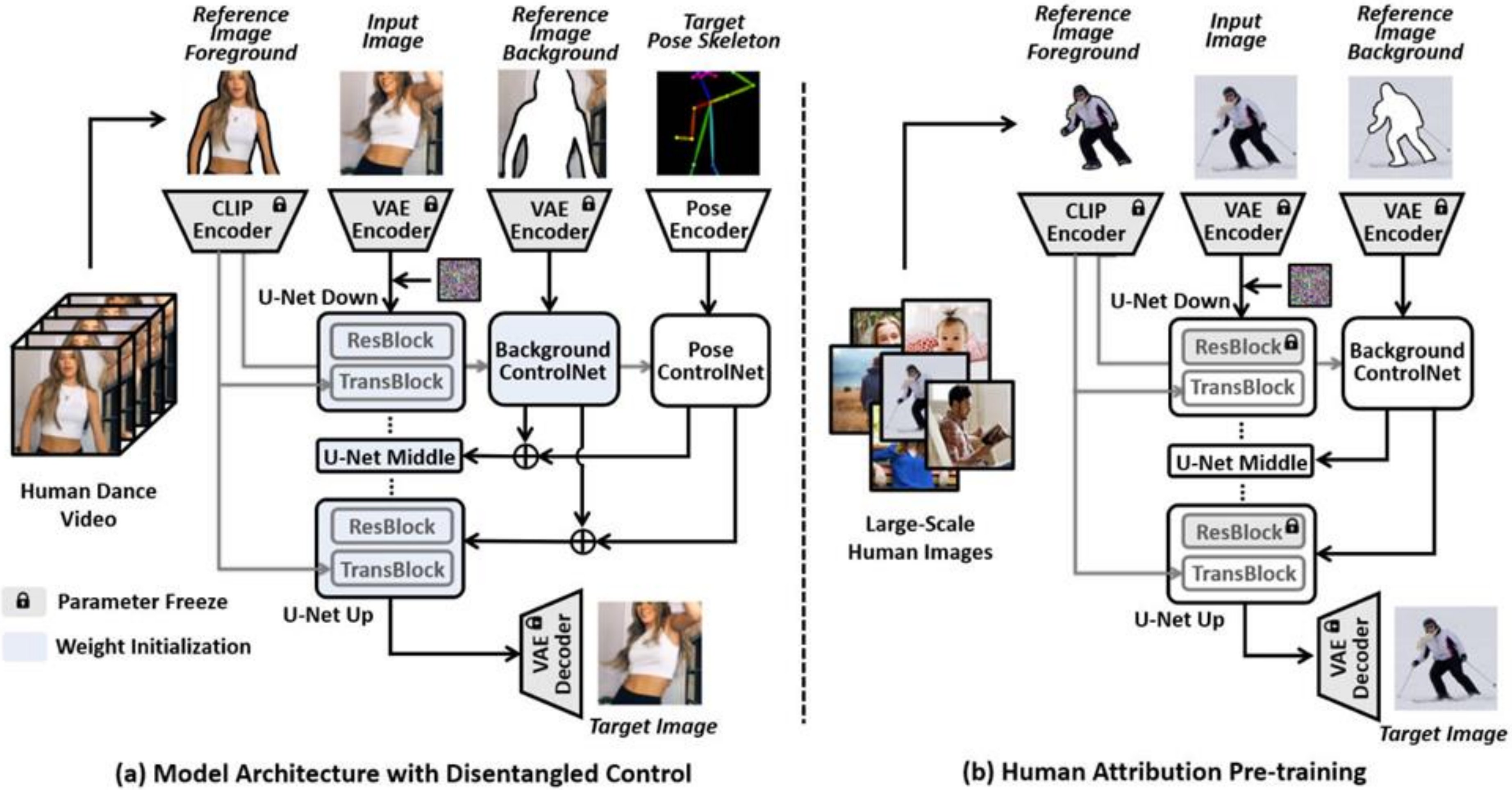


Dance #4

DisCo (Wang et al. CVPR‘24)

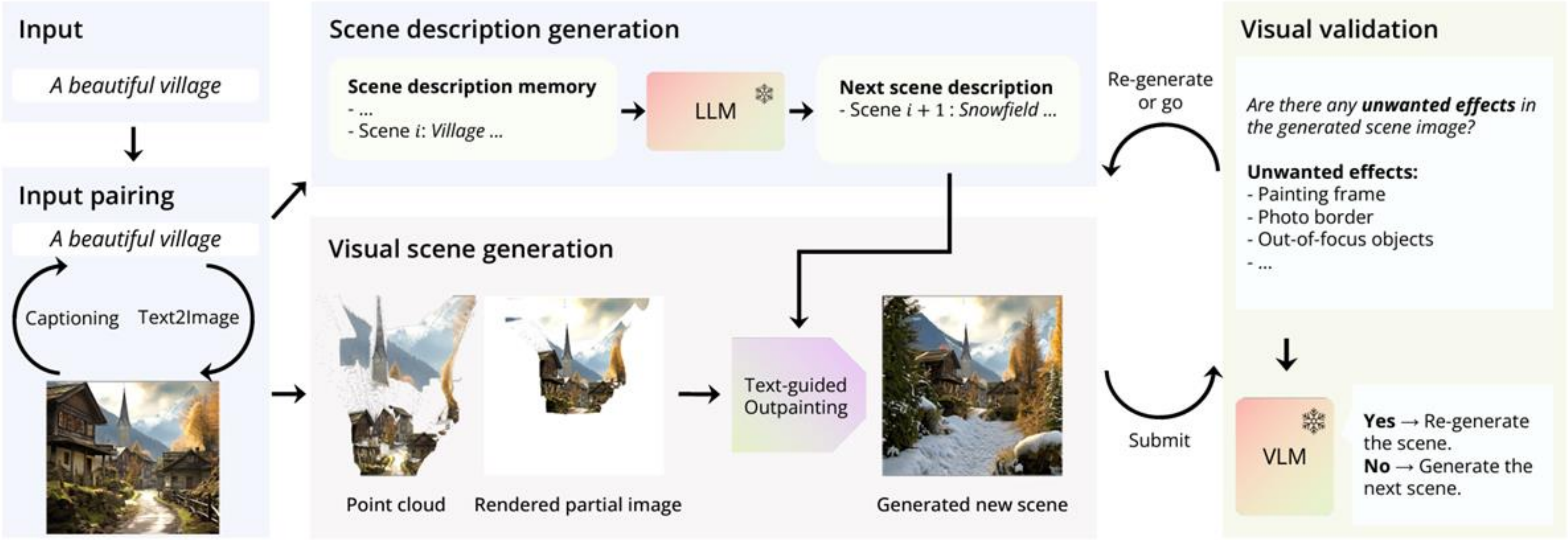
DisCo for Human Dance Generation

- By disentangling the control from all three conditions, DisCo enable arbitrary compositionality of human subjects, backgrounds, and dance-moves.



DisCo (Wang et al. CVPR‘24)

Wonderjourney



Input

A beautiful village

Input pairing

A beautiful village

Captioning

Text2Image

Scene description generation

Scene description memory
- ...
- Scene i: *Village* ...

LLM

Next scene description
- Scene $i+1$: *Snowfield* ...

Visual scene generation

Point cloud

Rendered partial image

Text-guided Outpainting

Generated new scene

Re-generate or go

Submit

Visual validation

*Are there any **unwanted effects** in the generated scene image?*

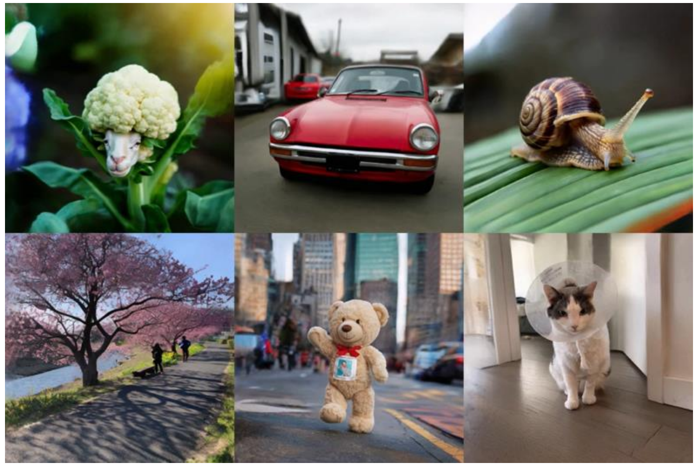
Unwanted effects:
- Painting frame
- Photo border
- Out-of-focus objects
- ...

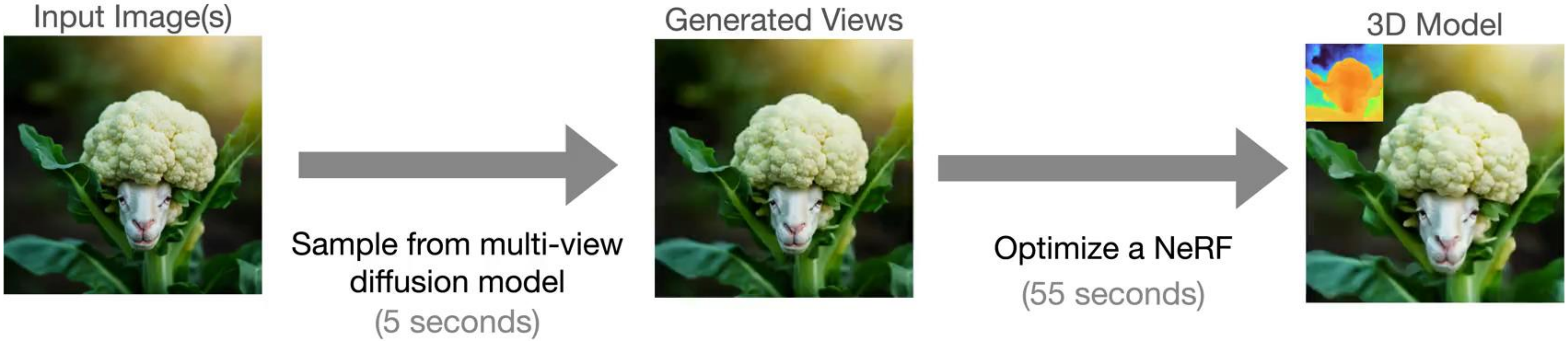
VLM

Yes → Re-generate the scene.
No → Generate the next scene.

Wonderjourney: Going from Anywhere to Everywhere (Yu et al. CVPR‘24)

CAT3D: Multi-View Latent Diffusion Model





NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. (Mildenhall et al., ECCV '22)
CAT3D: Create Anything in 3D with Multi-View Diffusion Models (Gao et al. '24)

✓ Video-Gen Foundation Model

✓ Faster training

✓ Faster inference

✓ Coherent video + Realistic contents + 3D consistency

→ World Model!

Video Generation Models as World Simulators

- ✓ 3D Consistency
- ✓ Coherence

Sora: Video generation models as world simulators (OpenAI ‘24)

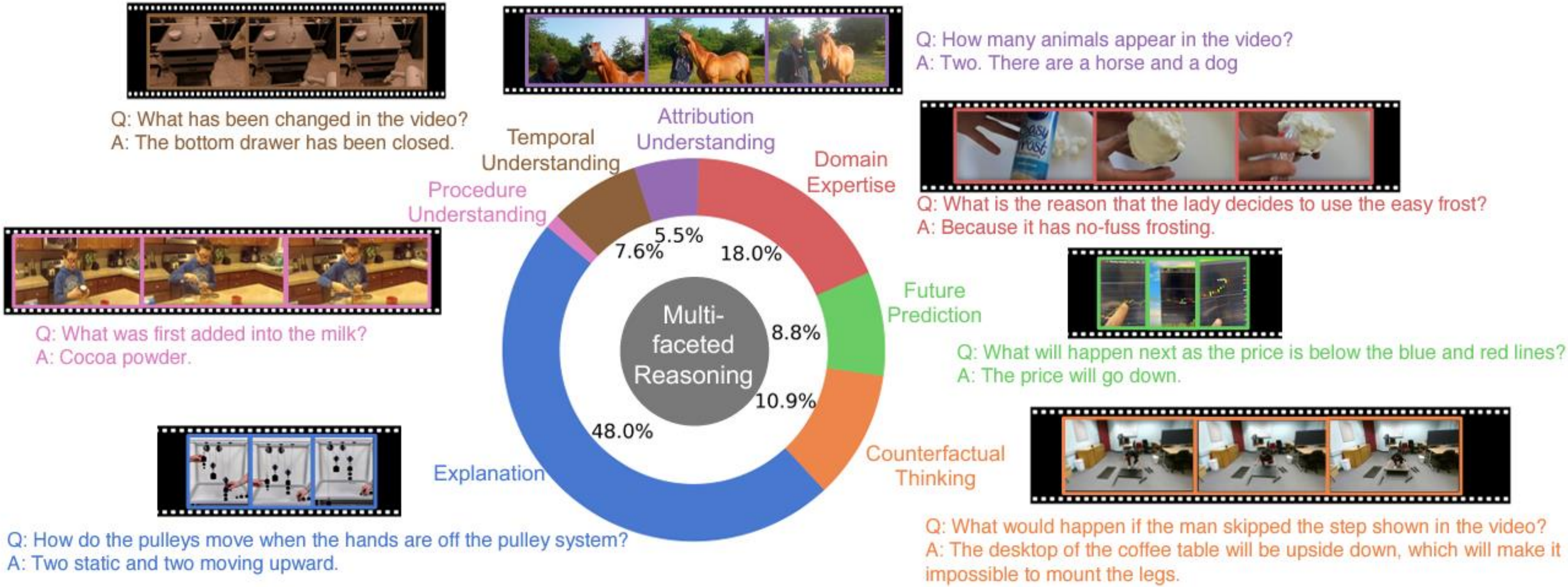
Discussion

- How to accurately model the physical and digital world?
  Physics, object states, and things beyond languages

- How to effectively evaluate the emerging capabilities?
  Needs of exploration and new benchmark

- Safety
  Learning from real-world use is a critical component

MMWorld: World Model Evaluation in Videos



Q: What has been changed in the video?
A: The bottom drawer has been closed.

Q: How many animals appear in the video?
A: Two. There are a horse and a dog

Q: What is the reason that the lady decides to use the easy frost?
A: Because it has no-fuss frosting.

Q: What was first added into the milk?
A: Cocoa powder.

Q: What will happen next as the price is below the blue and red lines?
A: The price will go down.

Q: How do the pulleys move when the hands are off the pulley system?
A: Two static and two moving upward.

Q: What would happen if the man skipped the step shown in the video?
A: The desktop of the coffee table will be upside down, which will make it impossible to mount the legs.

MMWorld: Towards Multi-discipline Multi-faceted World Model Evaluation in Videos (He et al. '24)

Discussion: World Model

- Build internal representations of the 3D world
- Predict and simulate future events within the internal representation
- Reasoning and planning: governed by our brain’s prediction of the future based on our internal world model

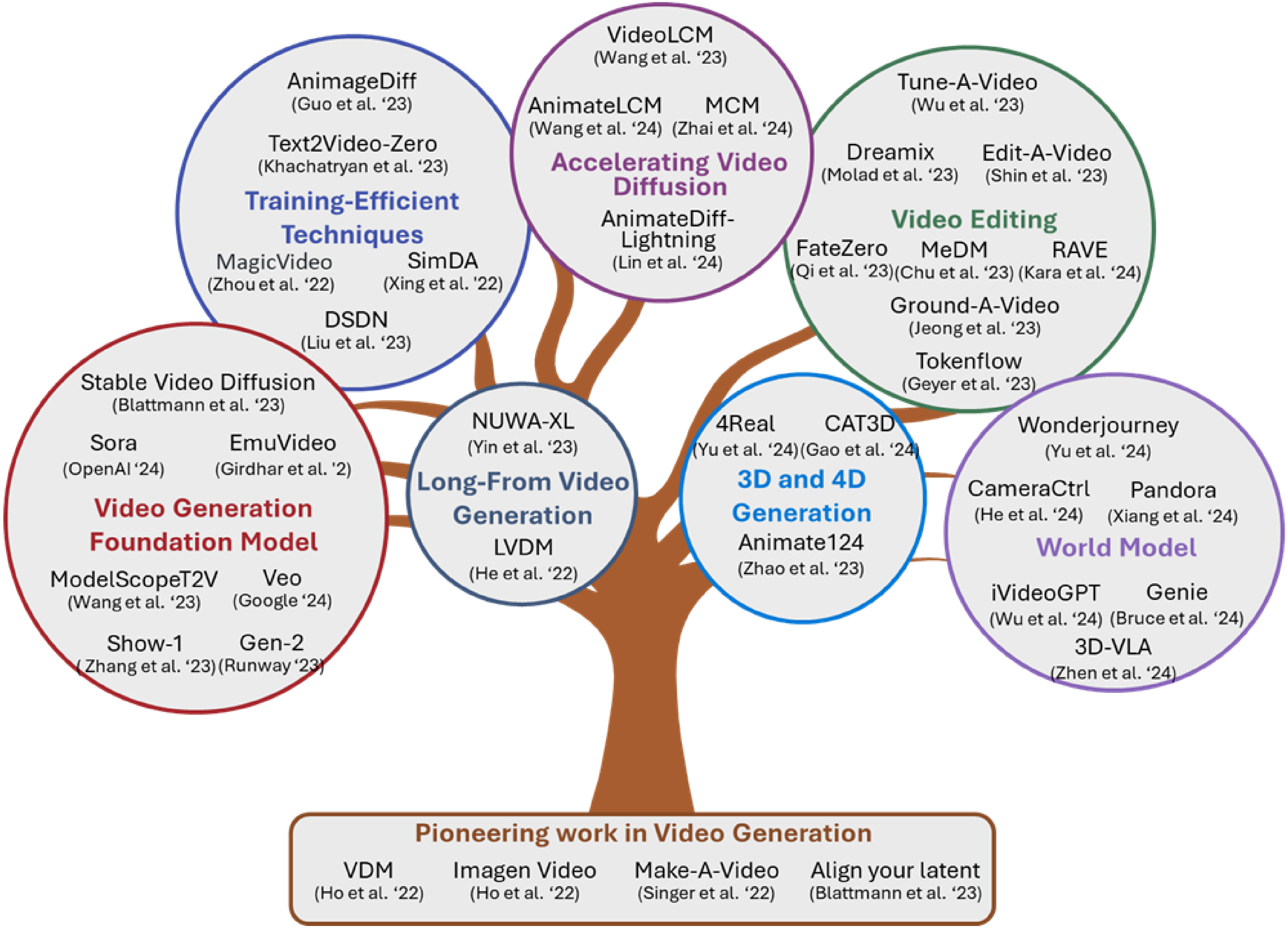
[1] Primary Visual Cortex Represents the Difference Between Past and Present. N. Nortmann et al. 2015
[2] Counterintuitive behavior of social systems. J.W. Forrester. 1971.
[3] Motion-Dependent Representation of Space in Area MT+. M. Gerrit et al. 2013

Discussion: World Model

- Build internal representations of the 3D world
- Predict and simulate future events within the internal representation
- Reasoning and planning: governed by our brain’s prediction of the future based on our internal world model

The image of the world around us, which we carry in our head, is just a model. He selects concepts and relationships, and uses those to represent and simulate the real system.

- Jay Wright Forrester, Father of System Dynamics talks about mental world models



VideoLCM
(Wang et al. '23)
AnimateLCM
(Wang et al. '24)
MCM
(Zhai et al. '24)
Accelerating Video Diffusion
AnimateDiff-Lightning
(Lin et al. '24)
AnimageDiff
(Guo et al. '23)
Text2Video-Zero
(Khachatryan et al. '23)
Training-Efficient Techniques
MagicVideo
(Zhou et al. '22)
SimDA
(Xing et al. '22)
DSDN
(Liu et al. '23)
Tune-A-Video
(Wu et al. '23)
Dreamix
(Molad et al. '23)
Edit-A-Video
(Shin et al. '23)
Video Editing
FateZero
(Qi et al. '23)
MeDM
(Chu et al. '23)
RAVE
(Kara et al. '24)
Ground-A-Video
(Jeong et al. '23)
Tokenflow
(Geyer et al. '23)
Stable Video Diffusion
(Blattmann et al. '23)
Sora
(OpenAI '24)
EmuVideo
(Girdhar et al. '2)
Video Generation Foundation Model
ModelScopeT2V
(Wang et al. '23)
Veo
(Google '24)
Show-1
(Zhang et al. '23)
Gen-2
(Runway '23)
NUWA-XL
(Yin et al. '23)
Long-From Video Generation
LVDM
(He et al. '22)
4Real
(Yu et al. '24)
CAT3D
(Gao et al. '24)
3D and 4D Generation
Animate124
(Zhao et al. '23)
Wonderjourney
(Yu et al. '24)
CameraCtrl
(He et al. '24)
Pandora
(Xiang et al. '24)
World Model
iVideoGPT
(Wu et al. '24)
Genie
(Bruce et al. '24)
3D-VLA
(Zhen et al. '24)
Pioneering work in Video Generation
VDM
(Ho et al. '22)
Imagen Video
(Ho et al. '22)
Make-A-Video
(Singer et al. '22)
Align your latent
(Blattmann et al. '23)

Acknowledgment

Mike Shou's Youtube video

- Tutorial: Video Diffusion Models

Lilian Weng’s blog

- What are Diffusion Models?

Hung-Yi Lee's Youtube video

- Introduction to Diffusion Models

Yuyang Zhao

Yining Hong

Yuanhao Zhai

Thank you!