Tutorial: Video Diffusion Models

Mike Shou

Asst Prof, National U. of Singapore

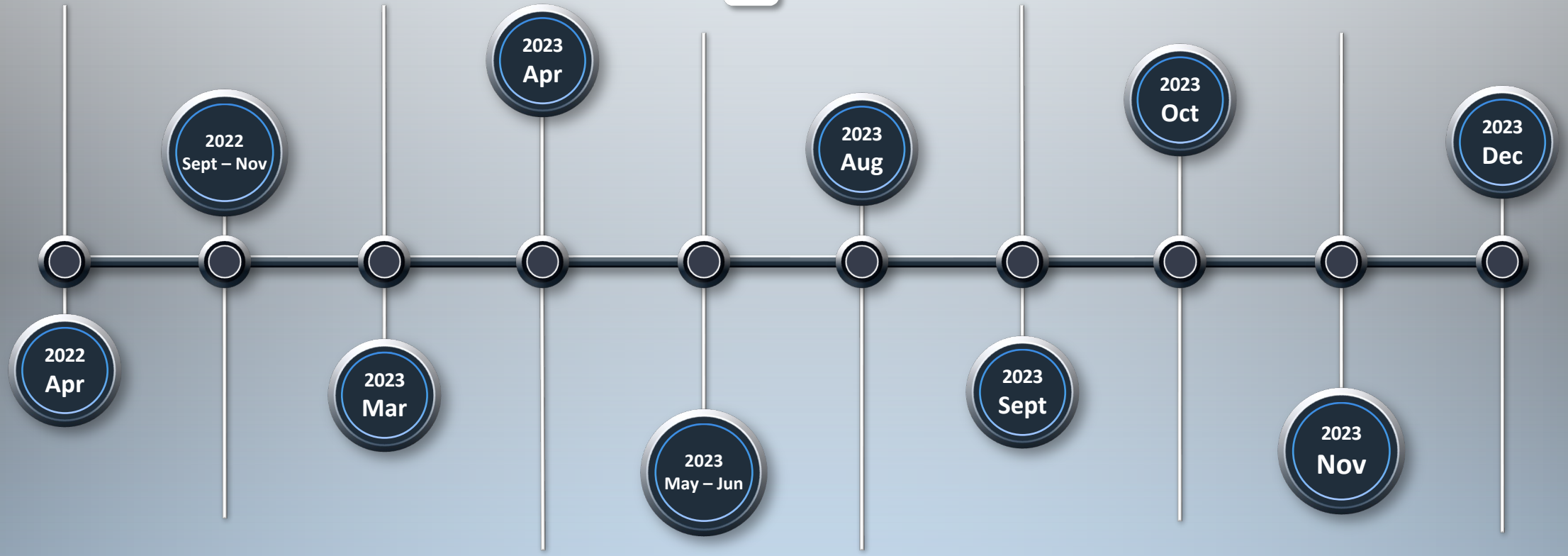
Joint work with Pei Yang & Jay Wu

NUS National University of Singapore

Show

Slides: https://sites.google.com/view/showlab/tutorial

Video Foundation Model



VDM Google

Gen-1

VideoFusion

ModelScope Alibaba

VideoFactory

PYoCo NVIDIA

Gen-2

Show-1

NUS National University of Singapore

Show

LaVie

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Hotshot-XL

Stable Video Diffusion stability.ai

Make-A-Video Meta

ImagenVideo Google

MagicVideo ByteDance

Video LDM NVIDIA

I2VGen-XL Alibaba

Emu Video Meta

VideoCrafter1 Tencent

Gentron Meta

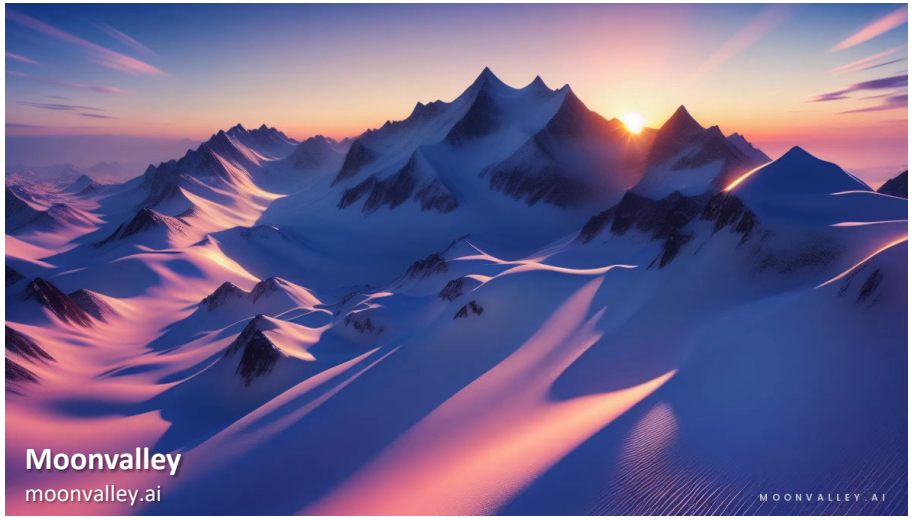
W.A.L.T. Stanford

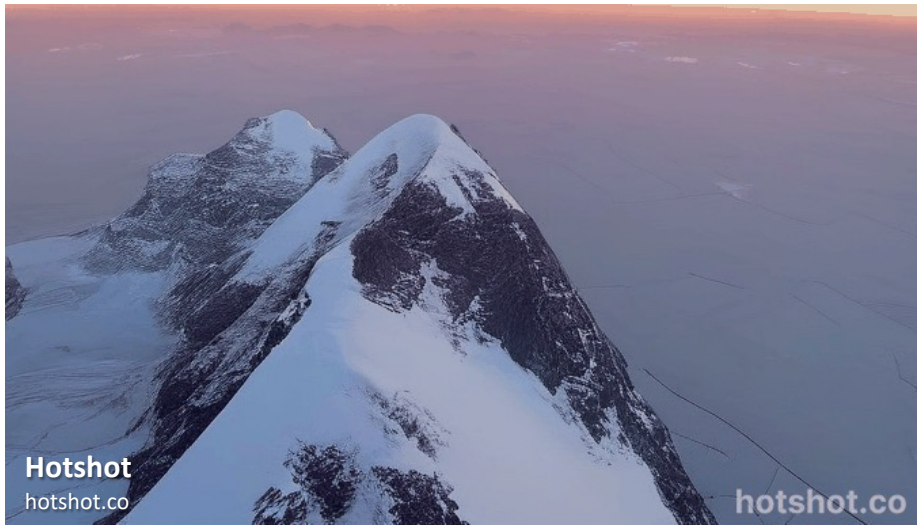
= Open source works

Video Generation/Editing Products

Input text: an aerial footage of snow mountains, sunset

Video Generation/Editing Products



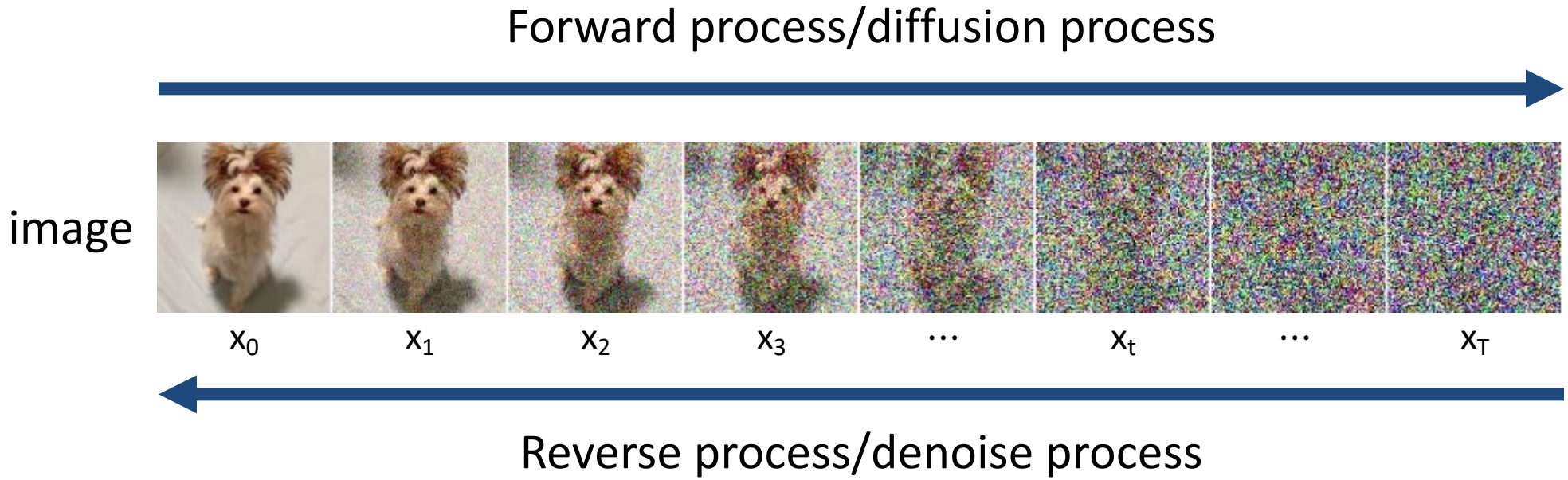


Outline of “Tutorial: Video Diffusion Models”

1. Fundamentals of Diffusion Models
2. **Video Generation**
3. **Video Editing**
4. Summary

1 Fundamentals of Diffusion Models

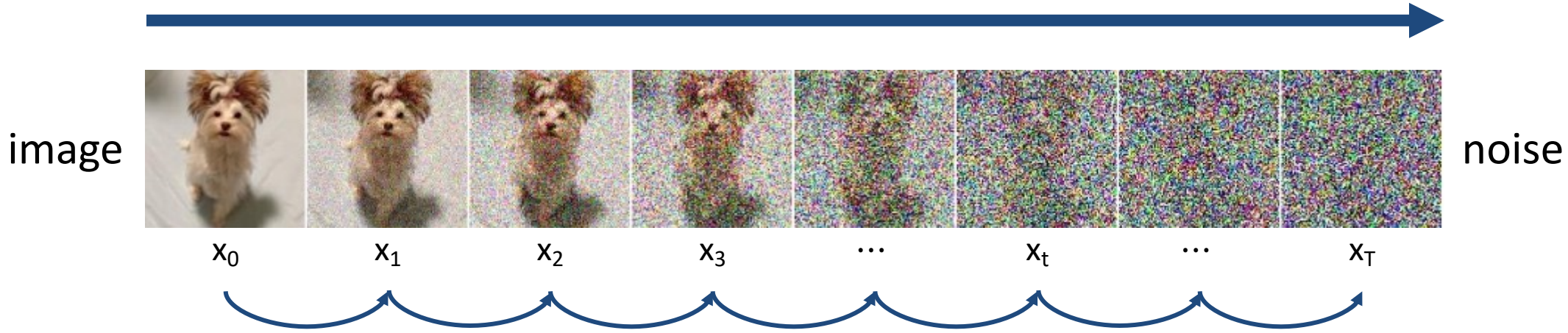
DDPM (Denoising Diffusion Probabilistic Models)



Ho et al., “Denoising Diffusion Probabilistic Models,” NeurIPS 2020.
Sohl-Dickstein et al., “Deep Unsupervised Learning using Nonequilibrium Thermodynamics,” ICML 2015.
Song et al., “Score-Based Generative Modeling through Stochastic Differential Equations,” ICLR 2021.
Vahdat et al., “Denoising Diffusion Models: A Generative Learning Big Bang,” CVPR 2023 Tutorial.

DDPM (Denoising Diffusion Probabilistic Models)

Forward process/diffusion process: add noise



$$\Pr(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}\left(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right)$$

β_t is the variance (strength) of noise

Ho et al., "Denoising Diffusion Probabilistic Models," NeurIPS 2020.
Sohl-Dickstein et al., "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," ICML 2015.
Song et al., "Score-Based Generative Modeling through Stochastic Differential Equations," ICLR 2021.
Vahdat et al., "Denoising Diffusion Models: A Generative Learning Big Bang," CVPR 2023 Tutorial.

DDPM (Denoising Diffusion Probabilistic Models)

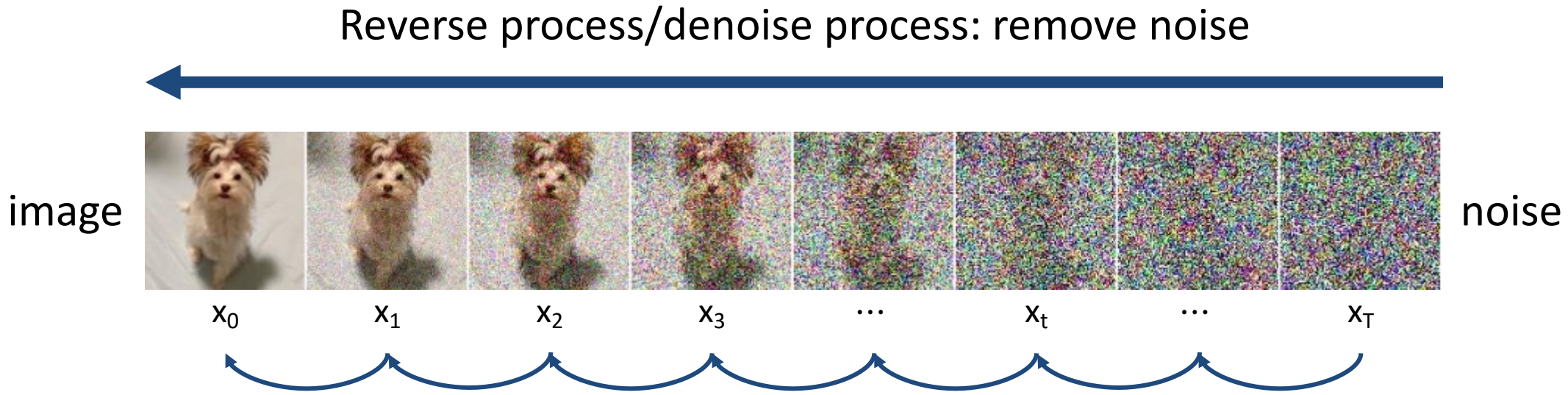
Forward process/diffusion process: add noise

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \qquad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$$

Ho et al., “Denoising Diffusion Probabilistic Models,” NeurIPS 2020.
Sohl-Dickstein et al., “Deep Unsupervised Learning using Nonequilibrium Thermodynamics,” ICML 2015.
Song et al., “Score-Based Generative Modeling through Stochastic Differential Equations,” ICLR 2021.
Vahdat et al., “Denoising Diffusion Models: A Generative Learning Big Bang,” CVPR 2023 Tutorial.

DDPM (Denoising Diffusion Probabilistic Models)

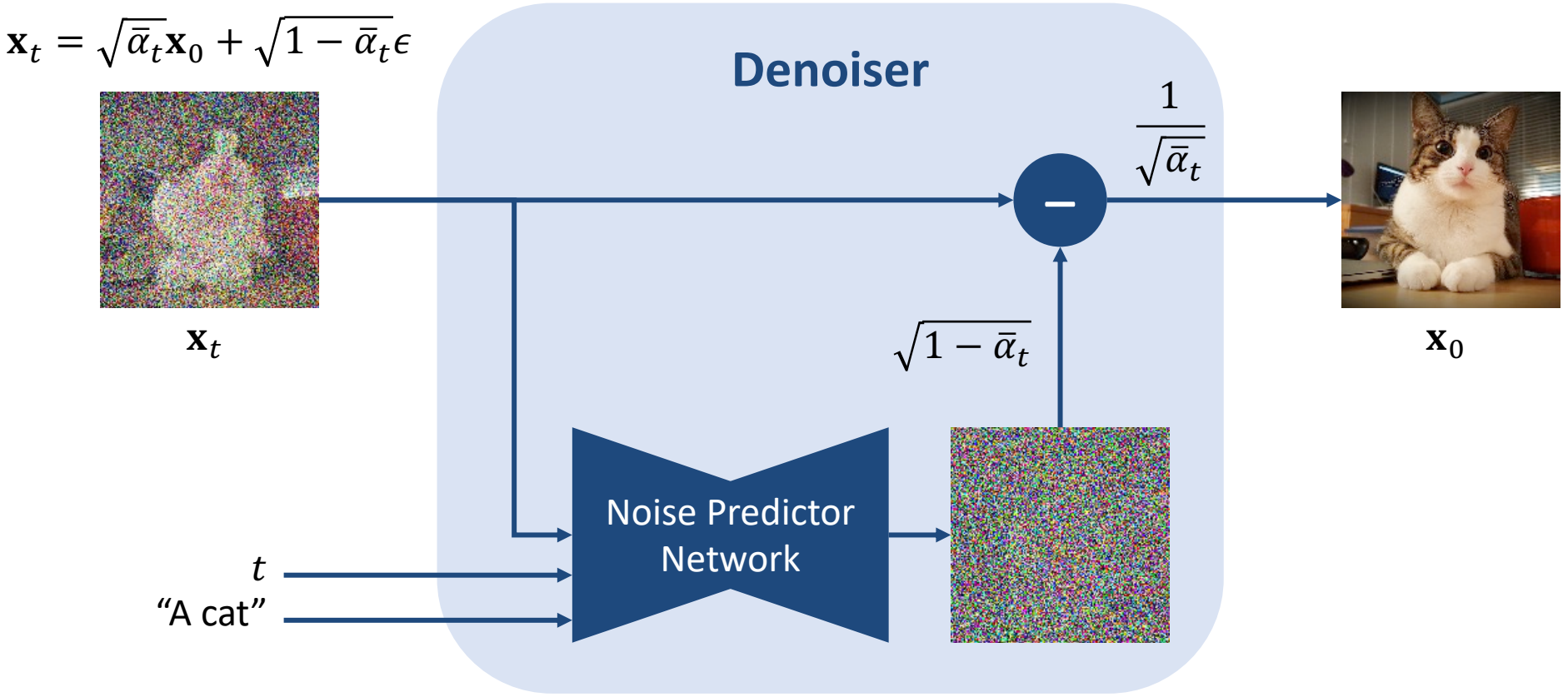


Train a neural network model to remove noise gradually

Ho et al., “Denoising Diffusion Probabilistic Models,” NeurIPS 2020.
Sohl-Dickstein et al., “Deep Unsupervised Learning using Nonequilibrium Thermodynamics,” ICML 2015.
Song et al., “Score-Based Generative Modeling through Stochastic Differential Equations,” ICLR 2021.
Vahdat et al., “Denoising Diffusion Models: A Generative Learning Big Bang,” CVPR 2023 Tutorial.

DDPM (Denoising Diffusion Probabilistic Models)

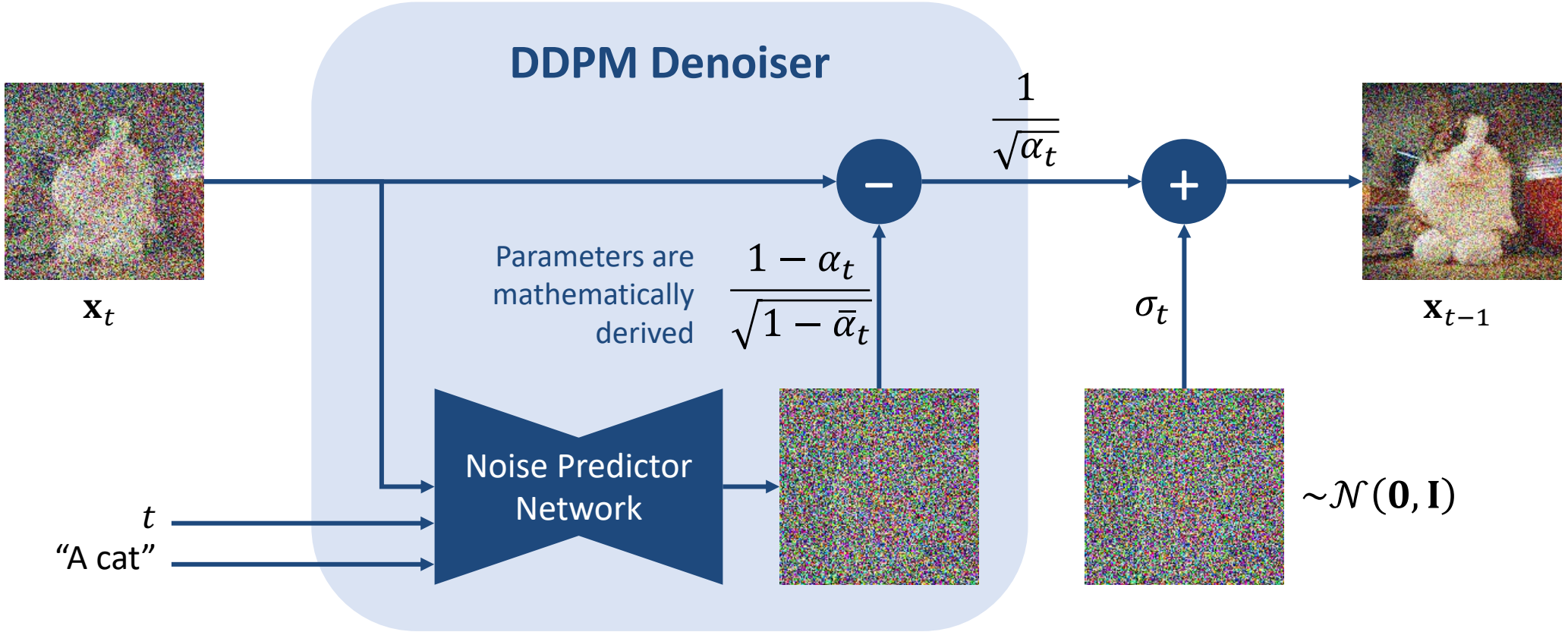
Training objective: one-step predict the noise w.r.t. the original image



Ho et al., "Denoising Diffusion Probabilistic Models," NeurIPS 2020.
Sohl-Dickstein et al., "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," ICML 2015.
Song et al., "Score-Based Generative Modeling through Stochastic Differential Equations," ICLR 2021.
Vahdat et al., "Denoising Diffusion Models: A Generative Learning Big Bang," CVPR 2023 Tutorial.

DDPM (Denoising Diffusion Probabilistic Models)

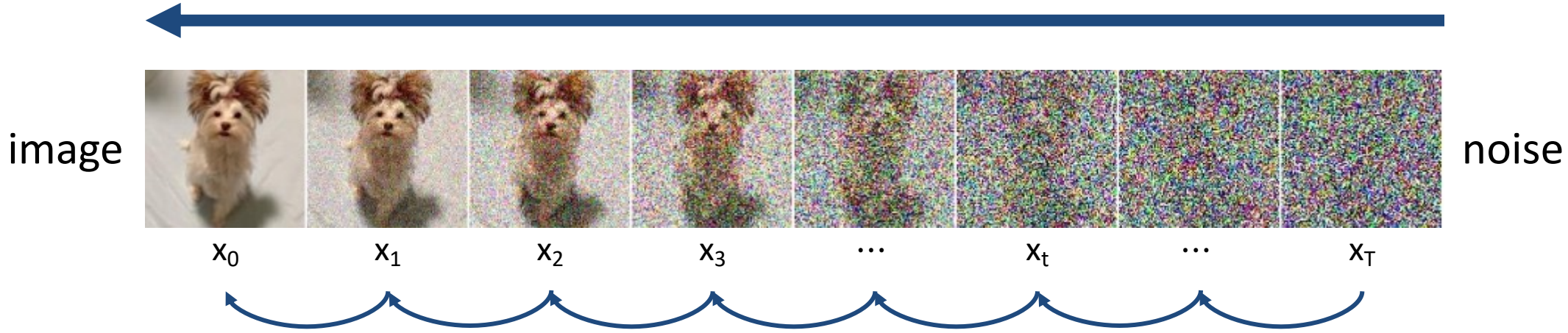
During generation: denoise step-by-step, in each step, add noise after noise removal



Ho et al., "Denoising Diffusion Probabilistic Models," NeurIPS 2020.
Sohl-Dickstein et al., "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," ICML 2015.
Song et al., "Score-Based Generative Modeling through Stochastic Differential Equations," ICLR 2021.
Vahdat et al., "Denoising Diffusion Models: A Generative Learning Big Bang," CVPR 2023 Tutorial.

DDPM (Denoising Diffusion Probabilistic Models)

Reverse process/denoise process: remove noise



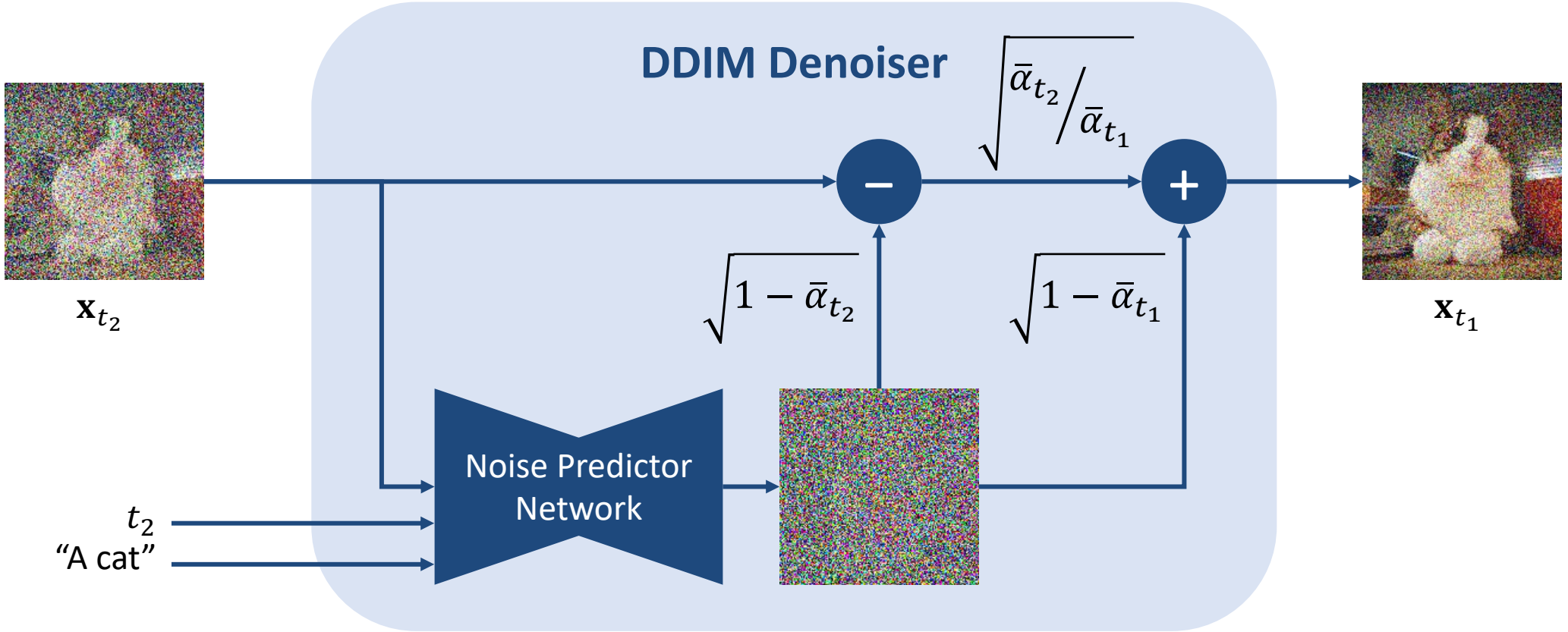
Train a neural network model to remove noise gradually

Speed up? Skip steps during denoising!

Ho et al., “Denoising Diffusion Probabilistic Models,” NeurIPS 2020.
Sohl-Dickstein et al., “Deep Unsupervised Learning using Nonequilibrium Thermodynamics,” ICML 2015.
Song et al., “Score-Based Generative Modeling through Stochastic Differential Equations,” ICLR 2021.
Vahdat et al., “Denoising Diffusion Models: A Generative Learning Big Bang,” CVPR 2023 Tutorial.

DDIM (Denoising Diffusion Implicit Models)

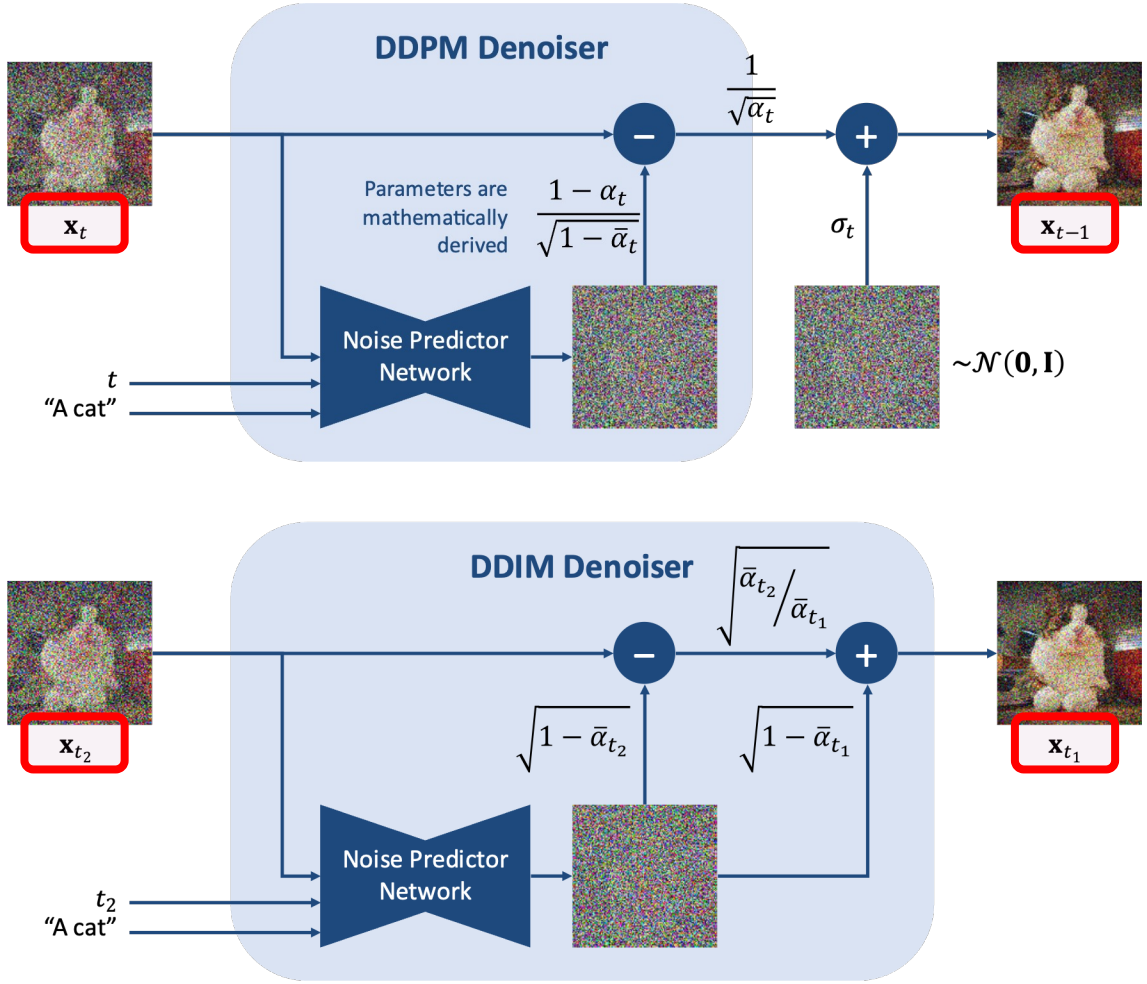
During generation: can skip steps from t_2 directly to t_1



Song et al., "Score-Based Generative Modeling through Stochastic Differential Equations," ICLR 2021.
Song et all, "Denoising Diffusion Implicit Models," ICLR 2021.

Denoising Diffusion Models

DDPM vs DDIM



DDPM cannot skip timesteps

A few hundreds steps to generate an image

DDIM can skip timesteps

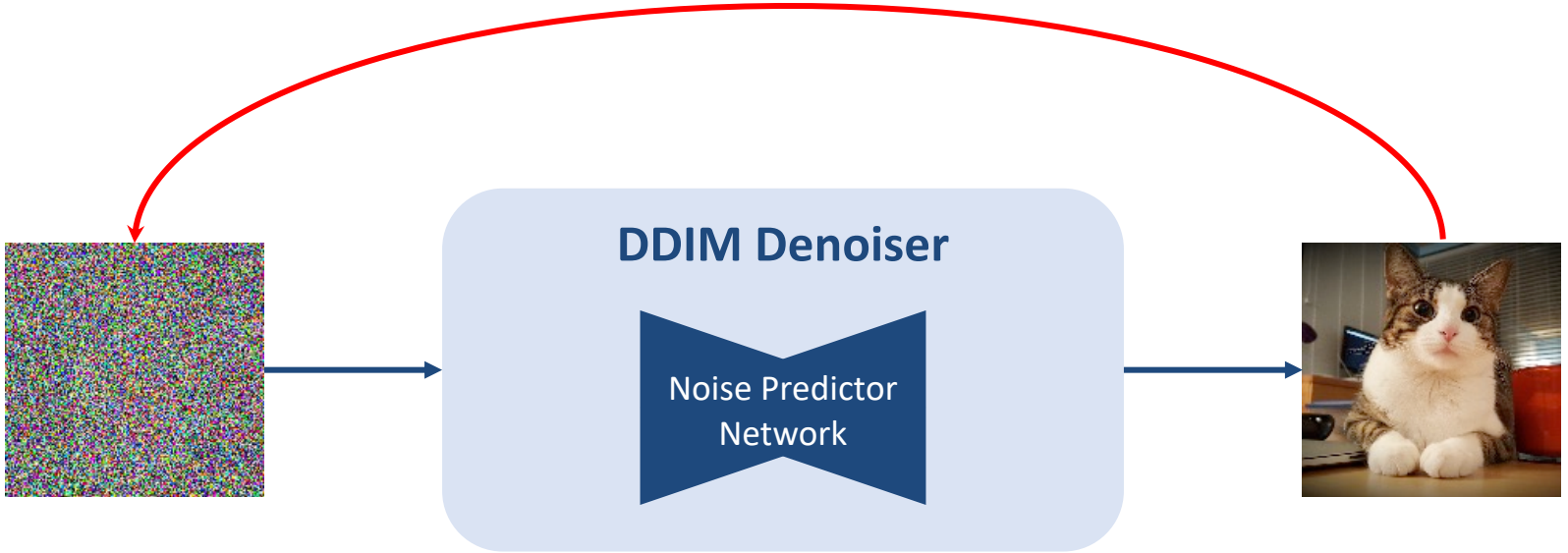
Say 50 steps to generate an image

Song et al., "Score-Based Generative Modeling through Stochastic Differential Equations," ICLR 2021.
Song et all, "Denoising Diffusion Implicit Models," ICLR 2021.

DDIM Inversion

The task of Inversion

Given an image & trained denoiser, find out which initial noise can be denoised into this image?

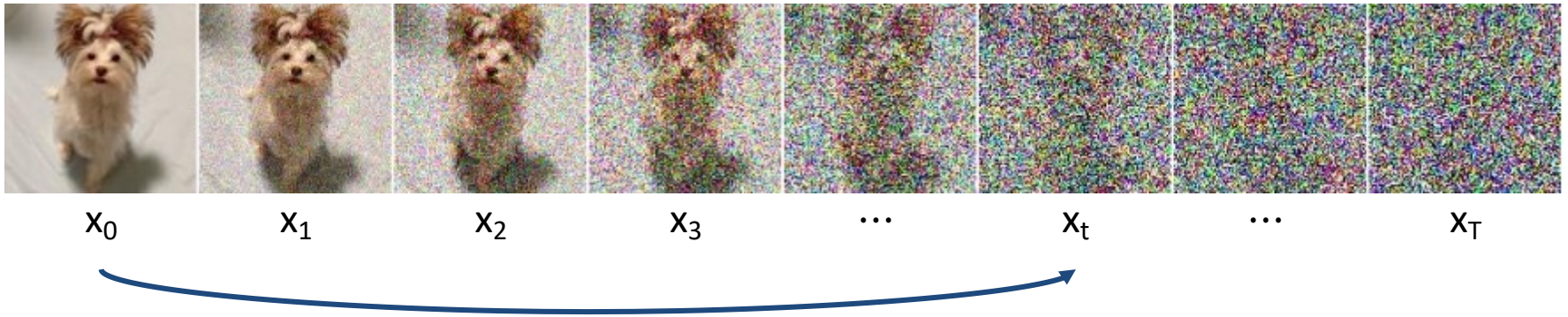


Song et al., “Denoising Diffusion Implicit Models,” ICLR 2021.
Su et al., “Dual Diffusion Implicit Bridges for Image-to-Image Translation,” ICLR 2023.
Mokadi et al., “Null-text Inversion for Editing Real Images using Guided Diffusion Models,” CVPR 2023.

DDIM Inversion

Based on the assumption that the ODE process can be reversed in the limit of small steps

Forward Diffusion Process: Add $\mathcal{N}(0, \mathbf{I})$ Noise



DDIM Inversion Process: Add Noise inverted by the trained DDIM denoiser

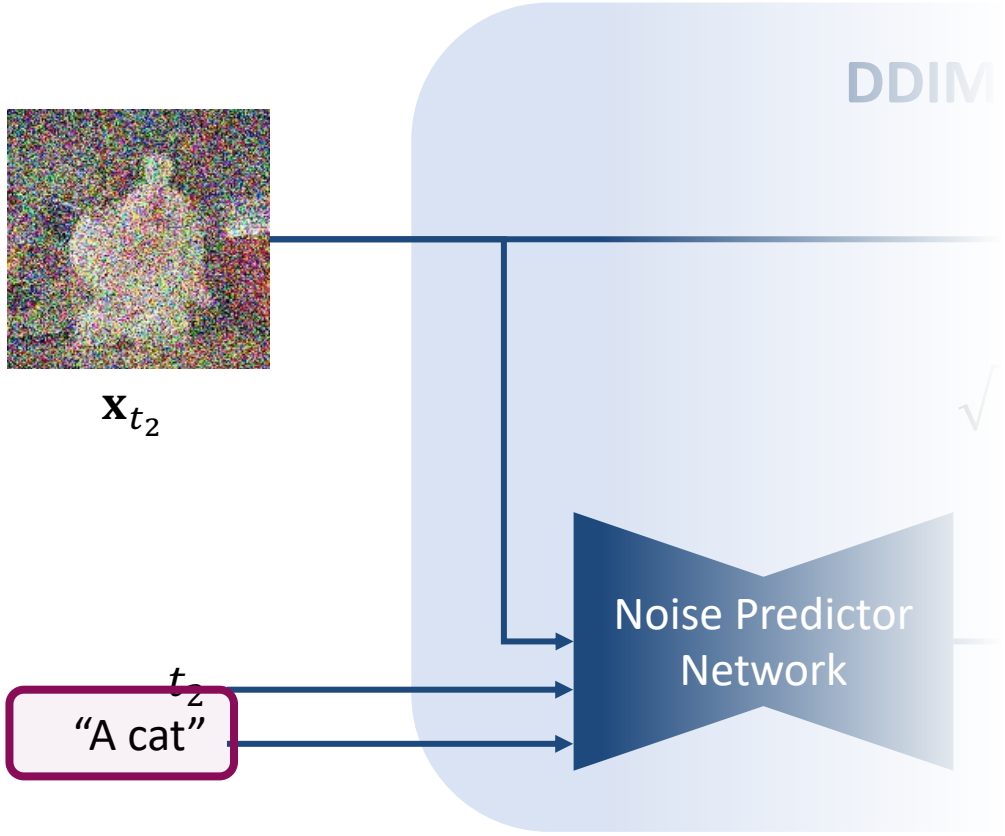
Song et al., "Denoising Diffusion Implicit Models," ICLR 2021.
Su et al., "Dual Diffusion Implicit Bridges for Image-to-Image Translation," ICLR 2023.
Mokadi et al., "Null-text Inversion for Editing Real Images using Guided Diffusion Models," CVPR 2023.

Wanted to learn more?

- CVPR Tutorial (English): https://www.youtube.com/watch?v=cS6JQpEY9cs
- Lil's blog: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/
- Hung-yi Lee (Chinese):
  - https://www.youtube.com/watch?v=azBugJzmz-o
  - https://www.youtube.com/watch?v=ifCDXFdeaaM
- Checkout codes -- Always associate theory and implementation!

Conditional Generation

Explicit conditions



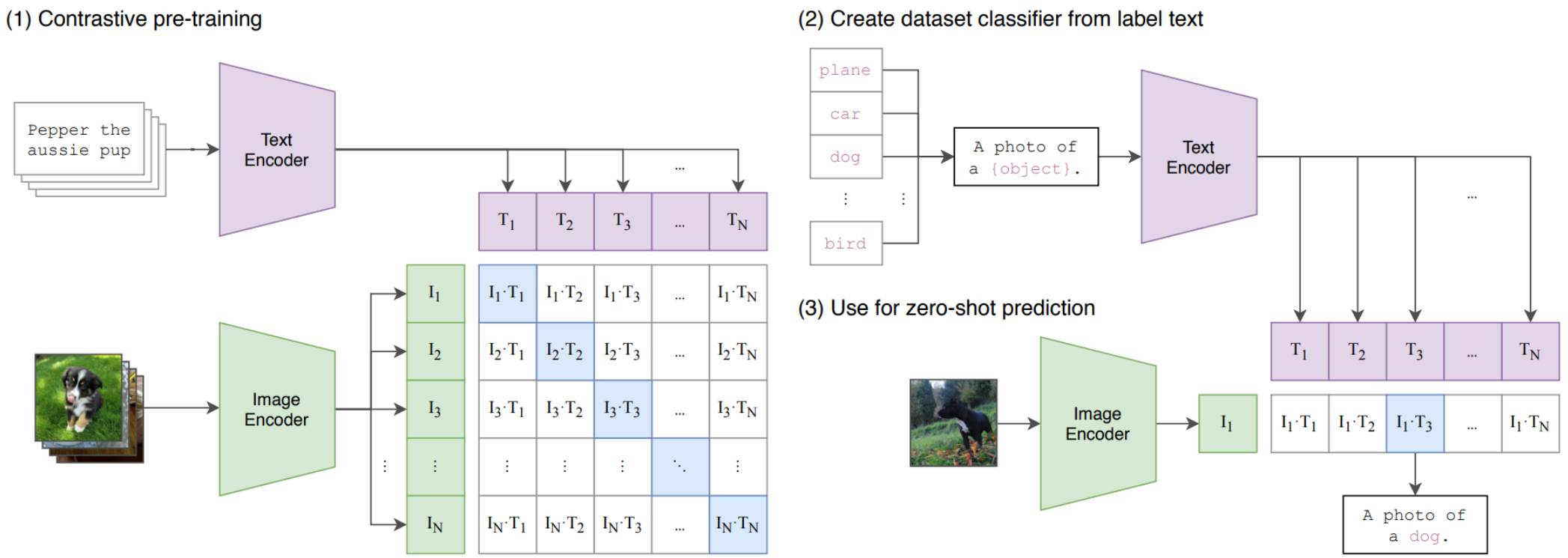
explicit conditions

Dhariwal et al., "Diffusion Models Beat GANs on Image Synthesis," NeurIPS 2021.
Ho et al., "Classifier-Free Diffusion Guidance," arXiv 2022.

CLIP

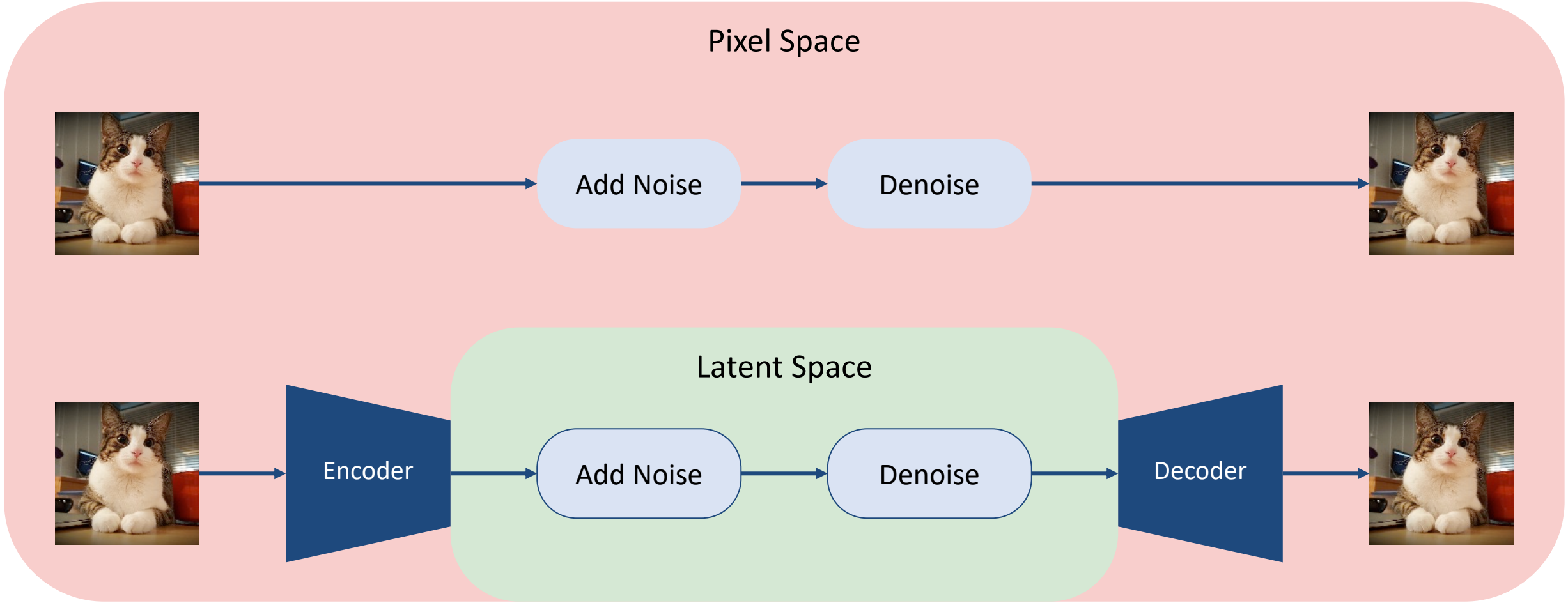
Encoders bridge vision and language

- CLIP text-/image-embeddings are commonly used in diffusion models for conditional generation



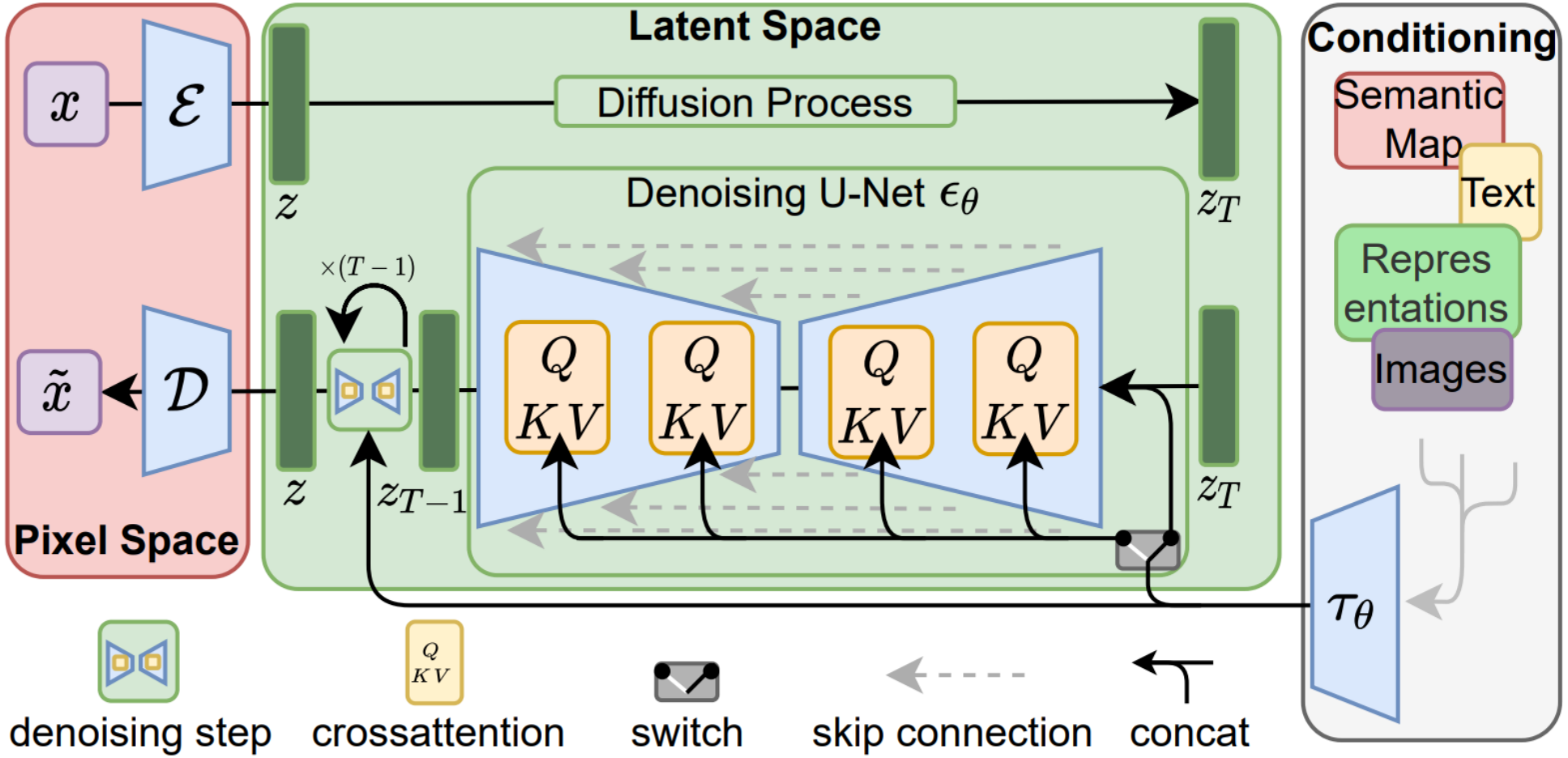
Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML 2021.

Latent Diffusion



Stable Diffusion

Conditional/unconditional image generation



Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models," CVPR 2022.

Stable Diffusion

Conditional/unconditional image generation

child's crayon drawing of minions

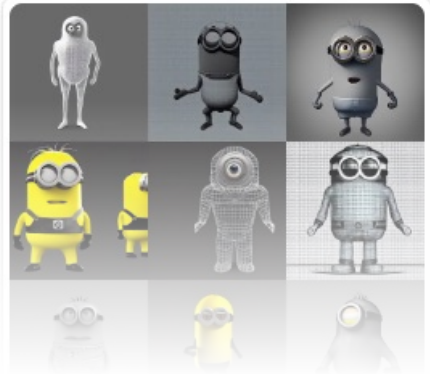
Minion

Totoro from my Neighbour

Biblically accurate Minion

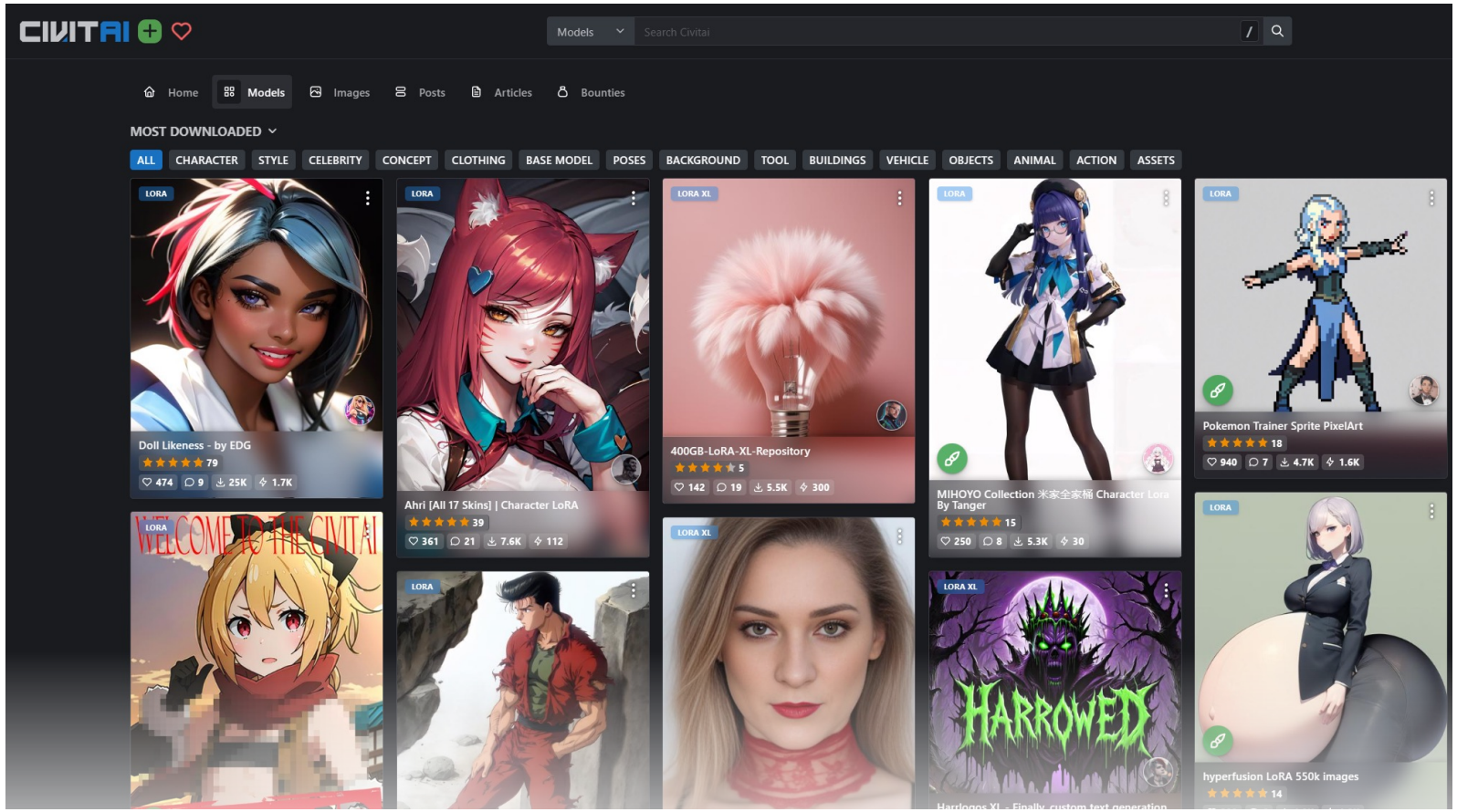
vintage photograph of minions operating a tank during ww2

Historical footage of a riot caused by Minions, Nuremberg 1930s, grainy, detailed



LoRA: Low-Rank Adaptation

Few-shot finetuning of large models for personalized generation



A Parameter-Efficient Fine-Tuning Method

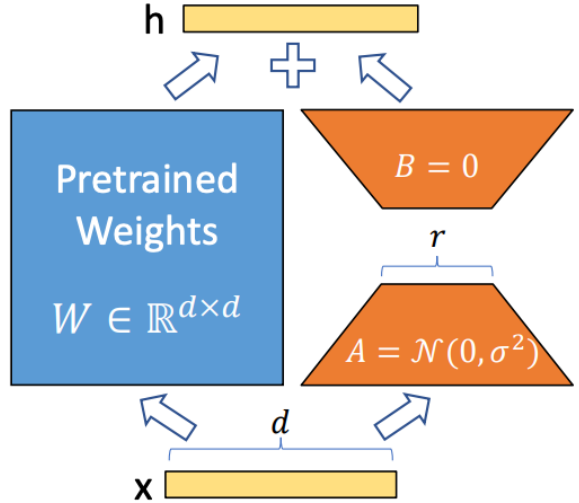
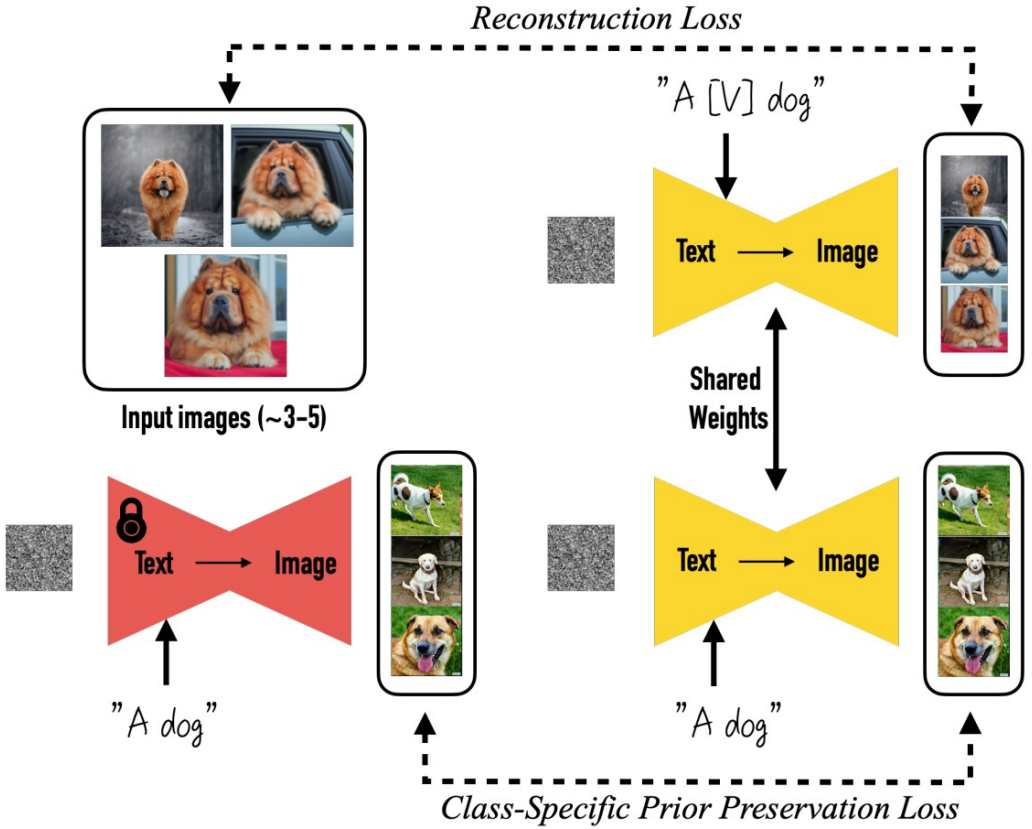


Figure 1: Our reparametrization. We only train A and B.

Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv 2021.
Gu et al., "Mix-of-Show: Decentralized Low-Rank Adaptation for Multi-Concept Customization of Diffusion Models," arXiv 2023.

DreamBooth

Few-shot finetuning of large models for generating personalized concepts



Ruiz et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," CVPR 2023.

ControlNet

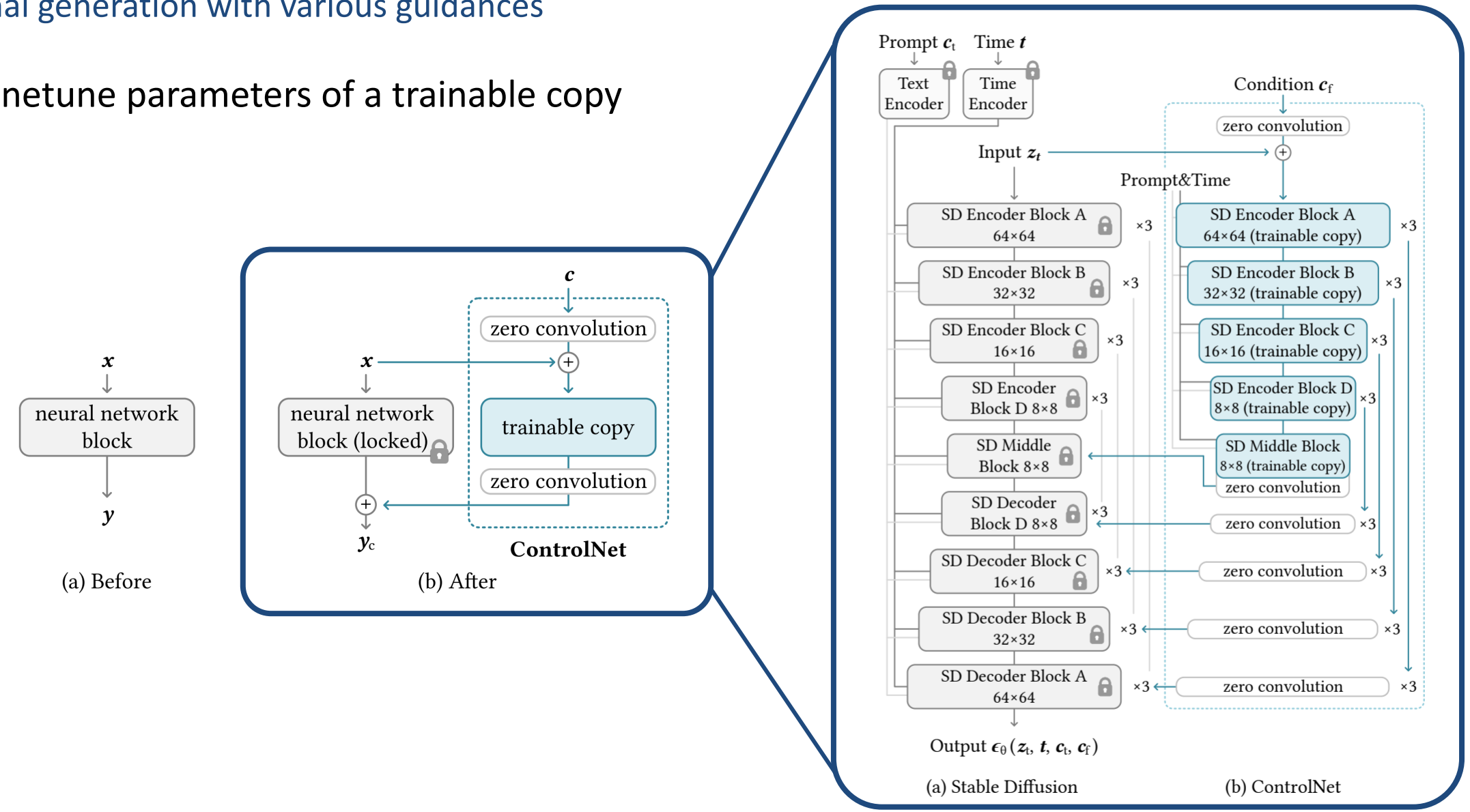
Conditional generation with various guidances

Zhang et al., "Adding Conditional Control to Text-to-Image Diffusion Models," ICCV 2023.

ControlNet

Conditional generation with various guidances

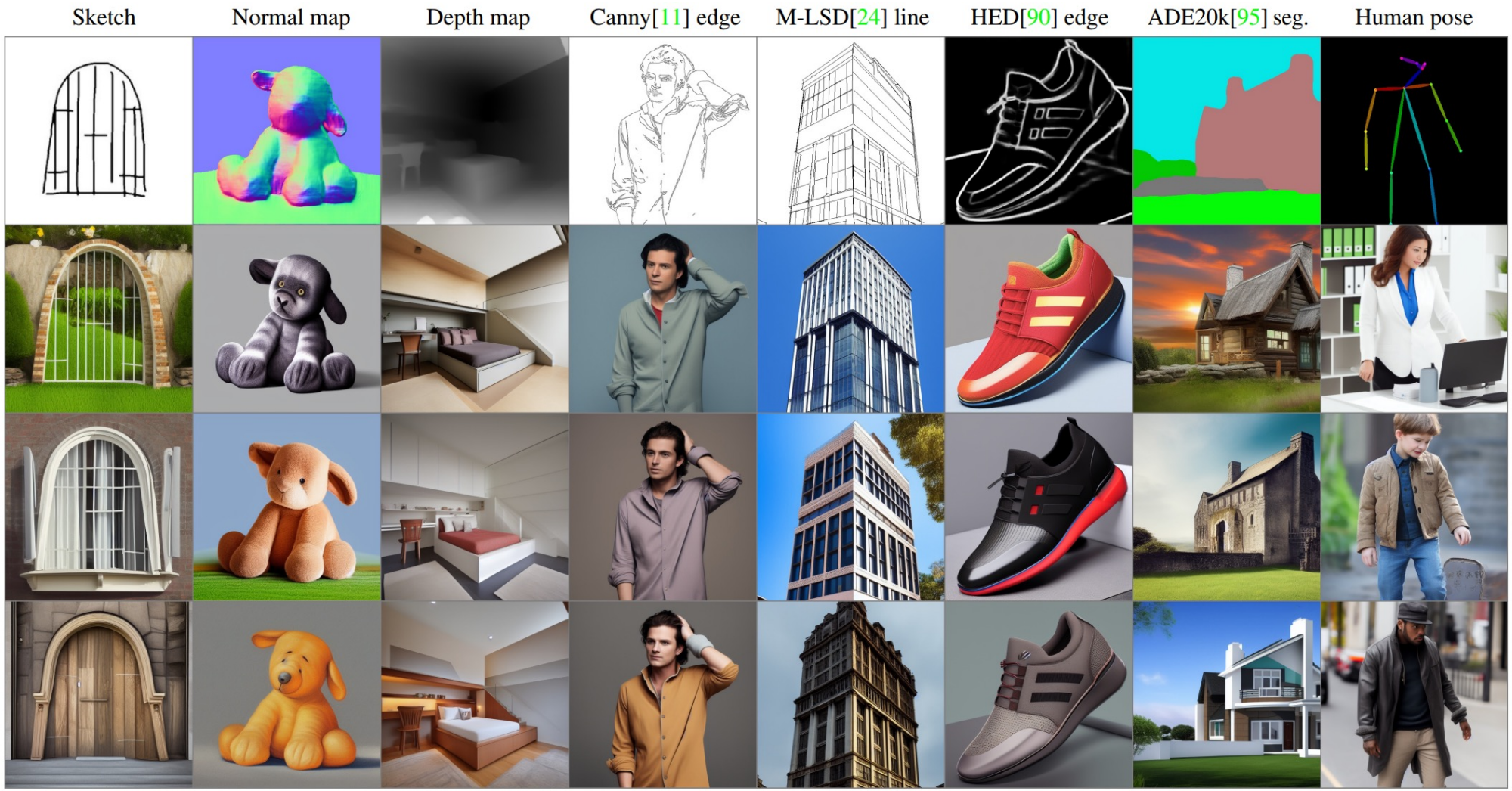
- Finetune parameters of a trainable copy



Zhang et al., "Adding Conditional Control to Text-to-Image Diffusion Models," ICCV 2023.

ControlNet

Conditional generation with various guidances



Zhang et al., "Adding Conditional Control to Text-to-Image Diffusion Models," ICCV 2023.

2 Video Generation

Video Foundation Model

VDM Google

Gen-1

VideoFusion

ModelScope Alibaba

VideoFactory

PYoCo NVIDIA

Gen-2

Show-1

NUS National University of Singapore

Show

Hotshot-XL

LaVie

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Stable Video Diffusion stability.ai

Make-A-Video Meta

ImagenVideo Google

MagicVideo ByteDance

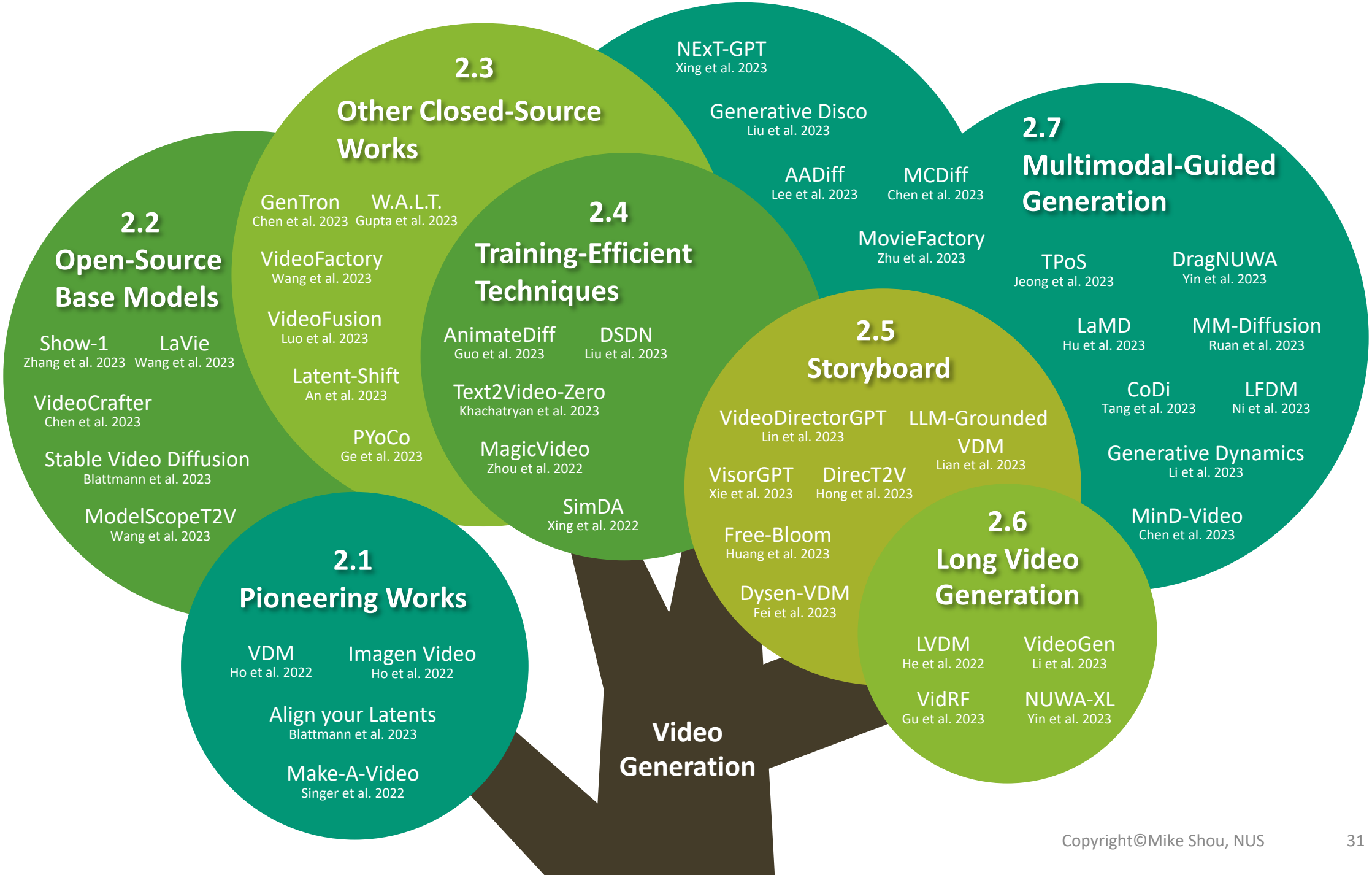
Video LDM NVIDIA

I2VGen-XL Alibaba

Emu Video Meta

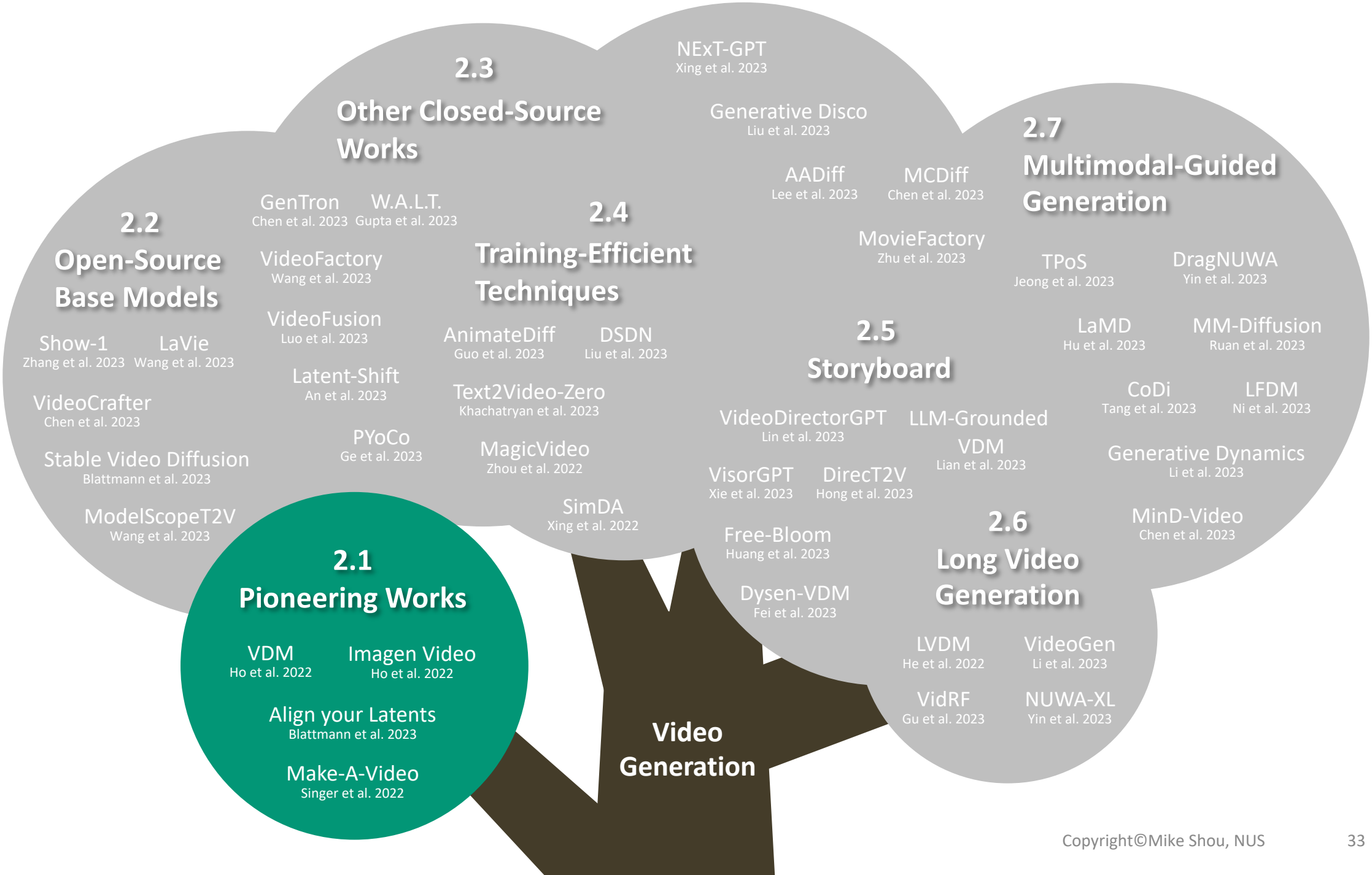
VideoCrafter1 Tencent

= Open source works



2 Video Generation

2.1 Pioneering/early works



Problem Definition

Text-Guided Image Generation

Text prompt → image

"Toad practicing karate."

Text-Guided Video Generation

Text prompt → video

"Toad practicing karate."

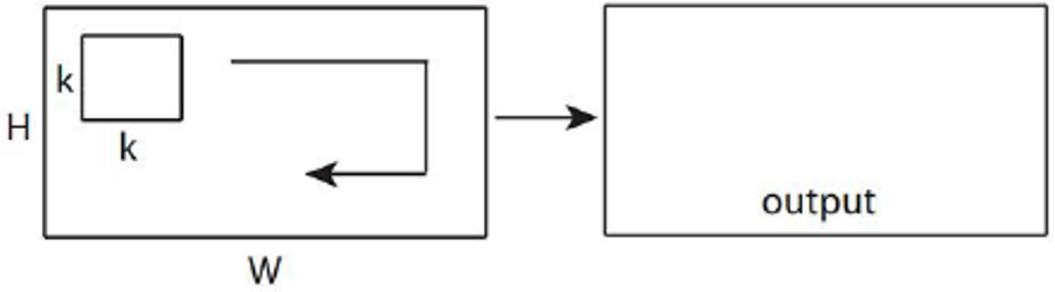
Video from Zhang et al., "Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation," arXiv 2023.

Problem Definition

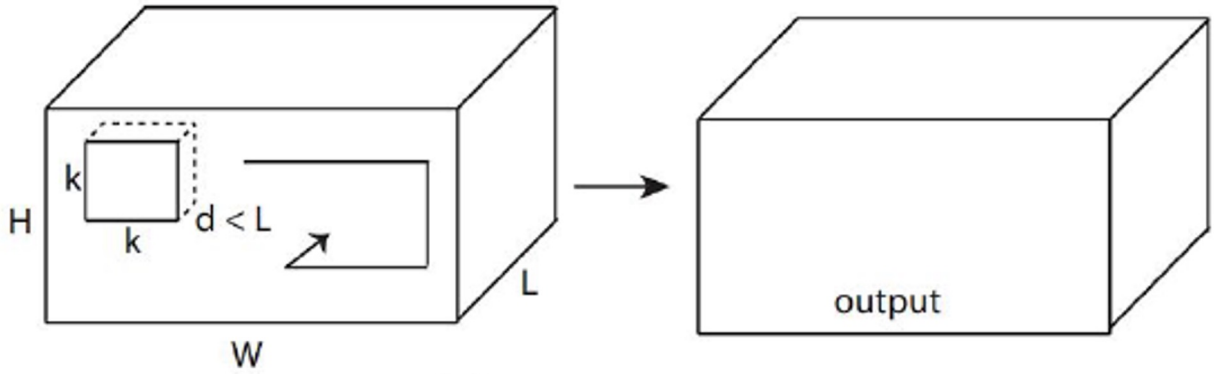
2D output → 3D output

Video Diffusion Models

Recap 3D Conv



(a) 2D convolution

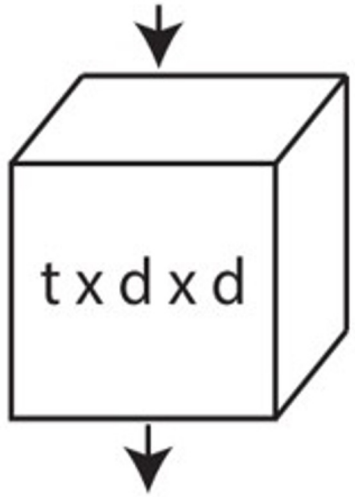


(c) 3D convolution

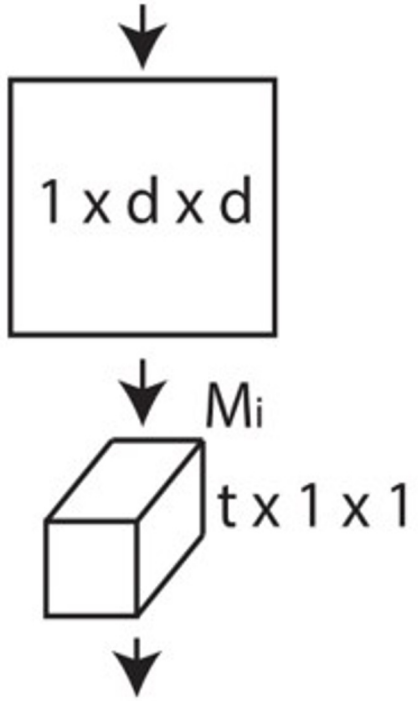
Du et al., "Learning Spatiotemporal Features with 3D Convolutional Networks," ICCV 2015.

Video Diffusion Models

Recap (2+1)D Conv



3D Conv



(2+1)D Conv

Du et al., "A Closer Look at Spatiotemporal Convolutions for Action Recognition," CVPR 2018.

Video Diffusion Models

Early work on video generation

16-frame "firework" videos

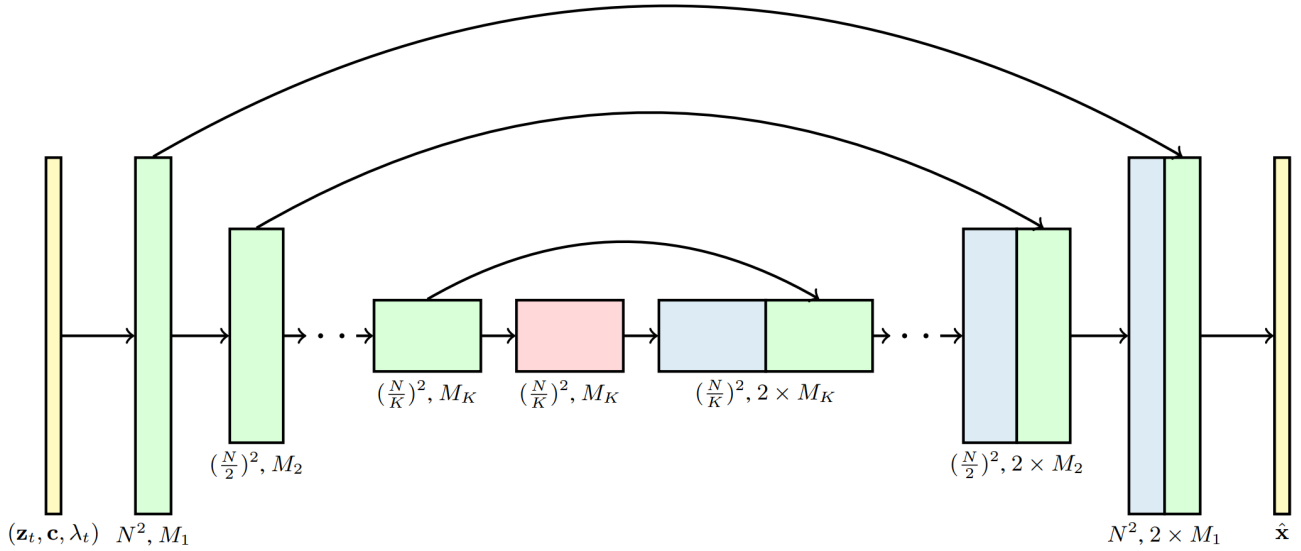
These fireworks does not exist in the real world!

Ho et al., "Video Diffusion Models," NeurIPS 2022.

Video Diffusion Models

Early work on video generation

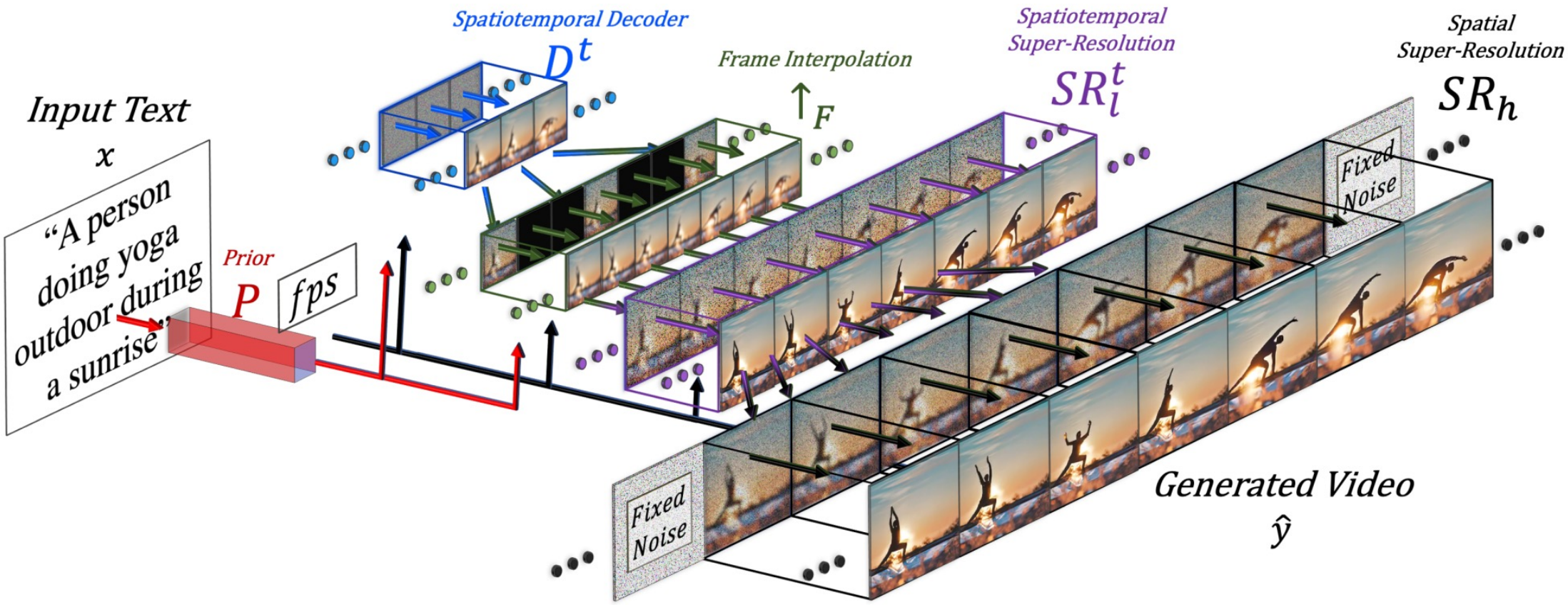
- 3D U-Net factorized over space and time
- Image 2D conv inflated as → space-only 3D conv, i.e., 2 in (2+1)D Conv
  - Kernel size: (3 × 3) → (1 × 3 × 3)
  - Feature vectors: (height × weight × channel) → (frame × height × width × channel)
- Spatial attention: remain the same
- Insert temporal attention layer: attend across the temporal dimension (spatial axes as batch)



Ho et al., "Video Diffusion Models," NeurIPS 2022.

Make-A-Video

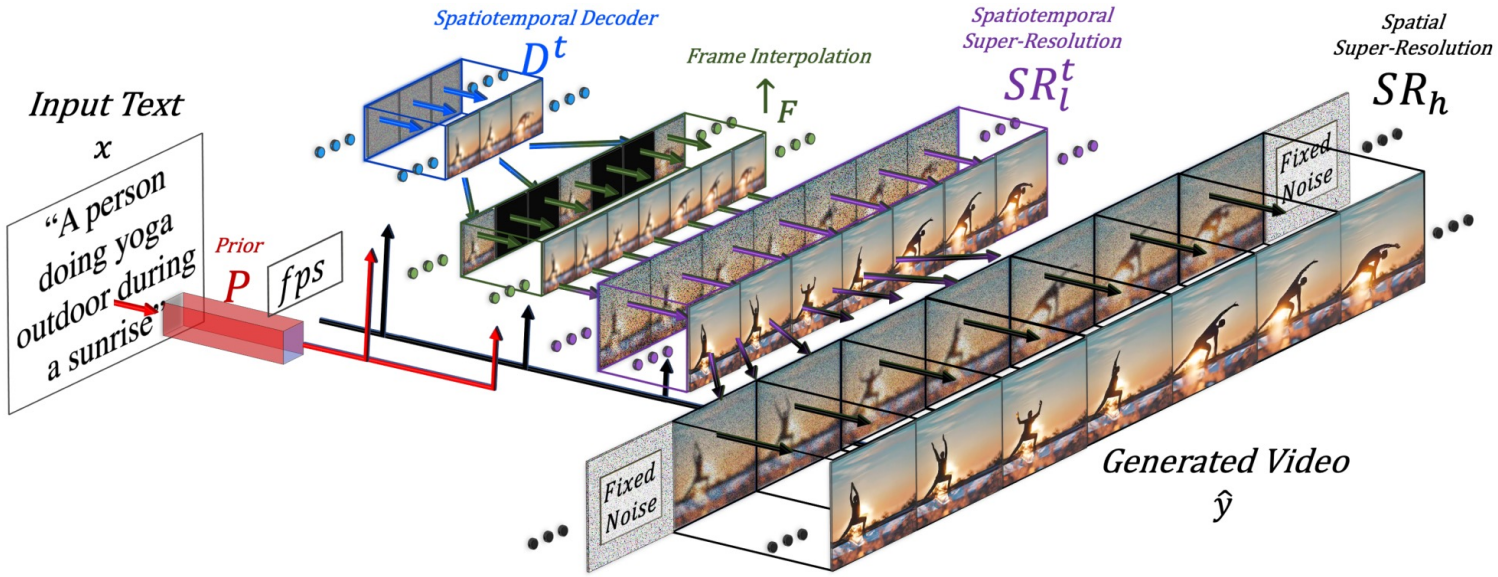
Cascaded generation

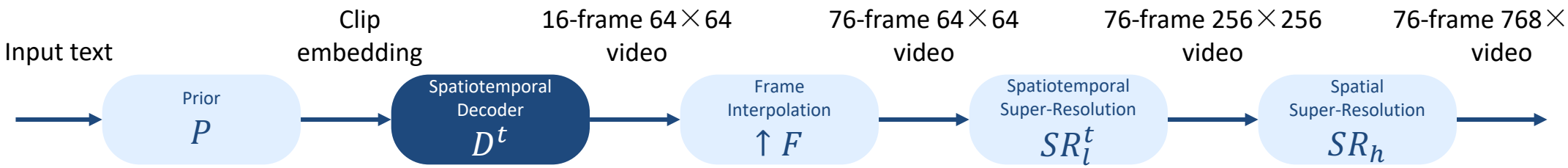


Singer et al., "Make-A-Video: Text-to-Video Generation without Text-Video Data," arXiv 2022.

Make-A-Video

Cascaded generation

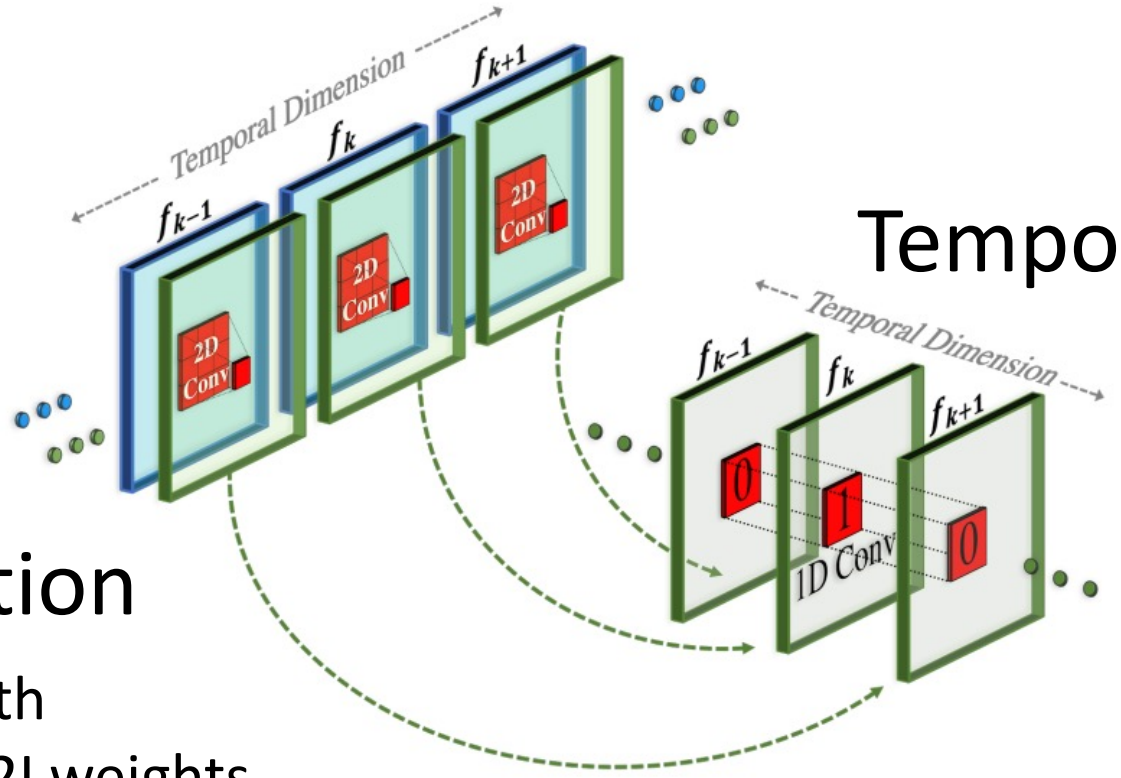




Make-A-Video

Cascaded generation

Architecture and Initialization Scheme of Pseudo-3D Convolution Layers



Spatial Convolution

Initialized with pretrained T2I weights

Temporal Convolution

Initialized with

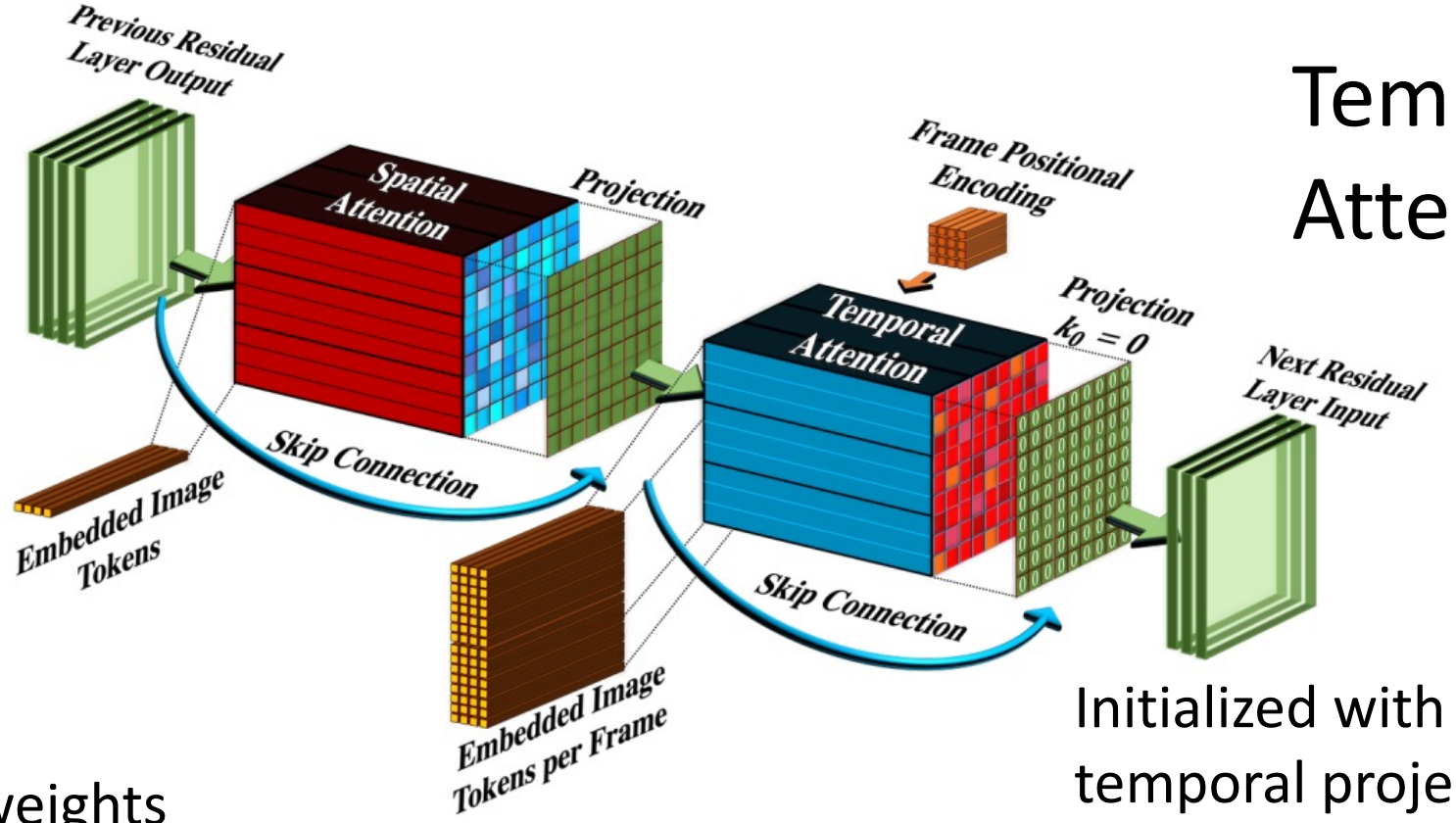
0 × {previous frame}
+ 1 × {current frame}
+ 0 × {next frame}

Singer et al., "Make-A-Video: Text-to-Video Generation without Text-Video Data," arXiv 2022.

Make-A-Video

Cascaded generation

Architecture and Initialization Scheme of Attention Layers



Temporal Attention

Spatial Attention

Initialized with pretrained T2I weights

Initialized with zero temporal projection (identity attn blocks)

Singer et al., "Make-A-Video: Text-to-Video Generation without Text-Video Data," arXiv 2022.

Make-A-Video

Cascaded generation

Training

- 4 main networks (decoder + interpolation + 2 super-res)
  - First trained on images alone
  - Insert and finetune temporal layers on videos
- Train on WebVid-10M and 10M subset from HD-VILA-100M

Singer et al., “Make-A-Video: Text-to-Video Generation without Text-Video Data,” arXiv 2022.

The WebVid-10M Dataset

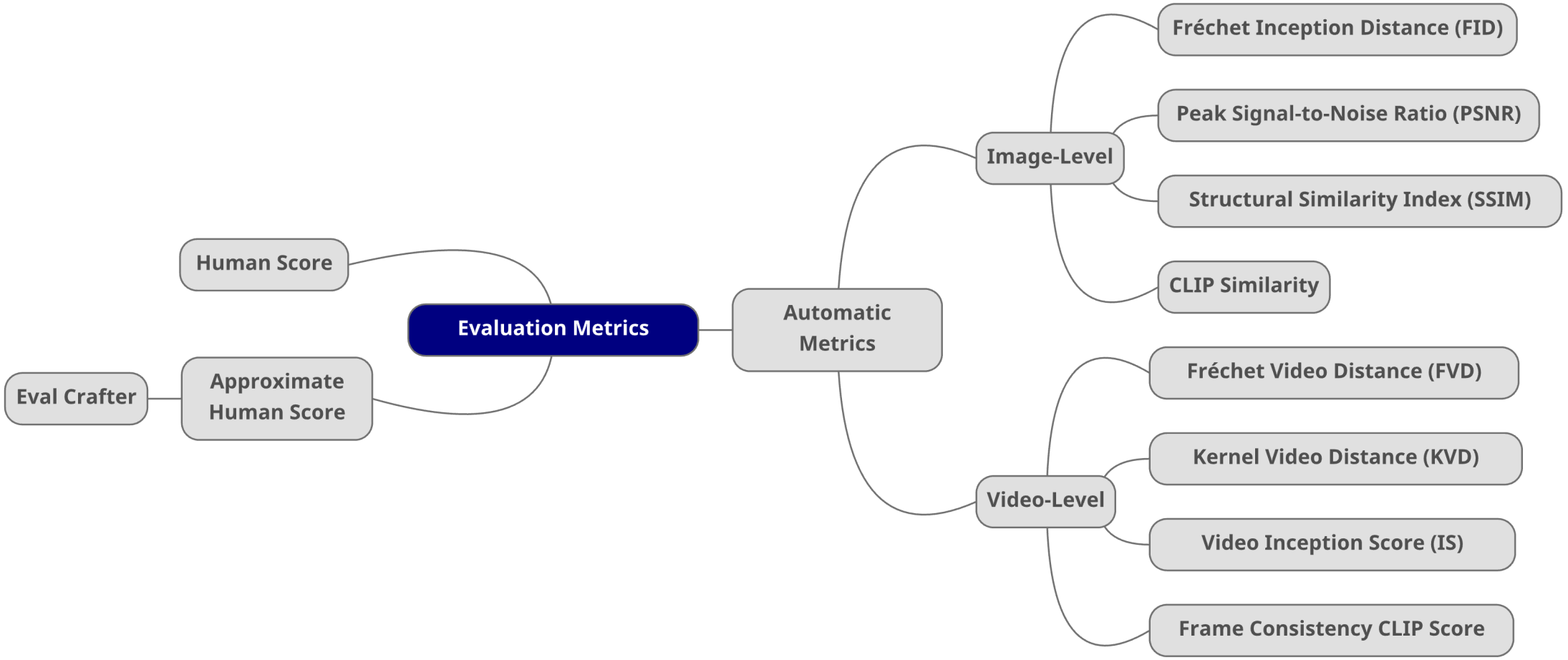
Video-caption pairs

Lonely beautiful woman sitting on the tent looking outside. wind on the hair and camping on the beach near the colors of water and shore. freedom and alternative tiny house for traveler lady drinking.

Female cop talking on walkietalkie, responding emergency call, crime prevention

Billiards, concentrated young woman playing in club

Evaluation Metrics



Evaluation Metrics

Quantitative evaluations

Image-level Evaluation Metrics

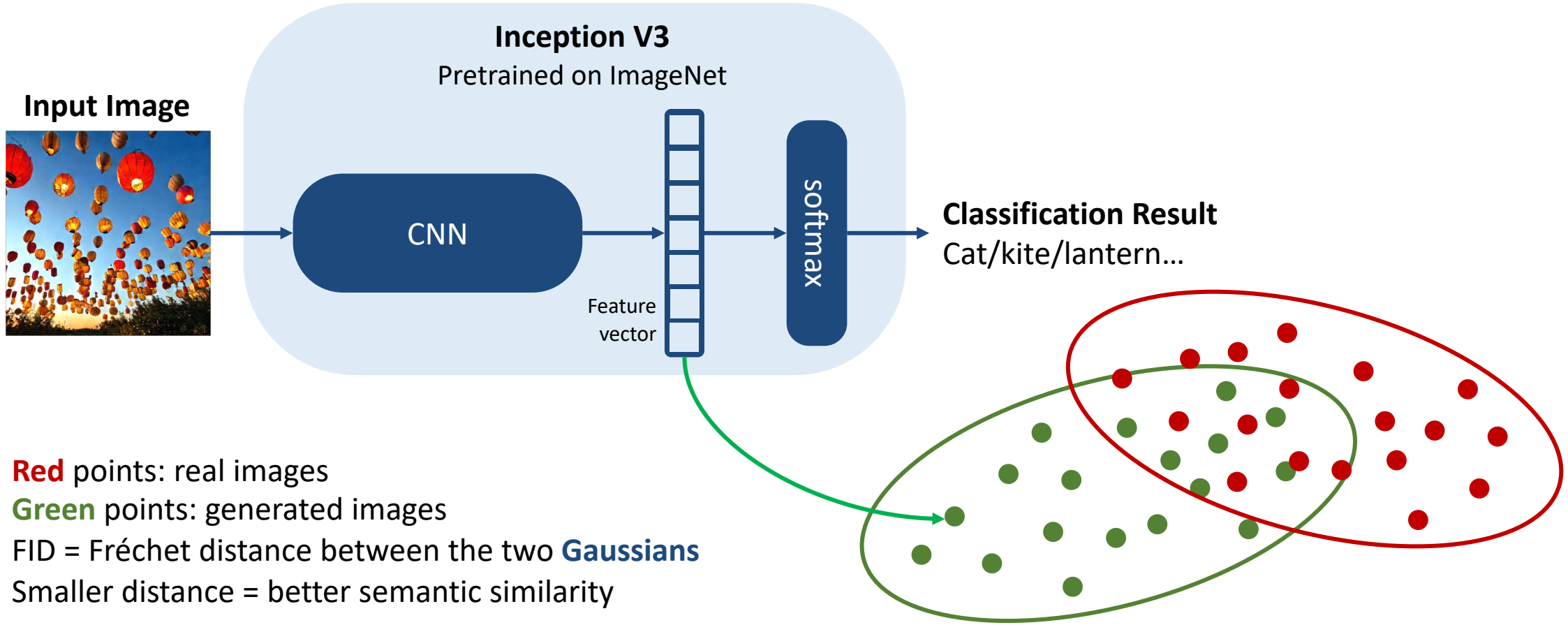
- Fréchet Inception Distance (FID, ↓): semantic similarity between images
- Peak Signal-to-Noise Ratio (PSNR, ↑): pixel-level similarity between images
- Structural Similarity Index (SSIM, ↓): pixel-level similarity between images
- CLIPSIM (↑): image-text relevance

Video-level Evaluation Metrics

- Fréchet Video Distance (FVD, ↓): semantic similarity & temporal coherence
- Kernel Video Distance (KVD, ↓): video quality (via semantic features and MMD)
- Video Inception Score (IS, ↑): video quality and diversity
- Frame Consistency CLIP Score (↑): frame temporal semantic consistency

Fréchet Inception Distance (FID)

Semantic similarity between images



Red points: real images
Green points: generated images
FID = Fréchet distance between the two **Gaussians**
Smaller distance = better semantic similarity

Lantern image generated with Stable Diffusion 2.1.

Heusel et al., "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," NeurIPS 2017.
Hung-Yi Lee, "Machine Learning 2023 Spring," National Taiwan University.

Peak Signal-to-Noise Ratio (PSNR)

Pixel-level similarity between images

- For two images x, y of shape $M \times N$:

$$\text{PSNR}(x, y) = 10 \log_{10} \frac{255^2}{\text{MSE}(x, y)}$$

where

$$\text{MSE}(x, y) = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(x_{ij} - y_{ij}\right)^2$$

Horé et al., "Image Quality Metrics: PSNR vs. SSIM," ICPR 2010.

Structural Similarity Index Measure (SSIM)

Pixel-level similarity between images

- Model any image distortion as a combination of:
  (1) loss of correlation, (2) luminance distortion, (3) contrast distortion

- For two images x, y of shape $M \times N$:

$$\mathrm{SSIM}(x, y) = l(x, y) \cdot c(x, y) \cdot s(x, y)$$

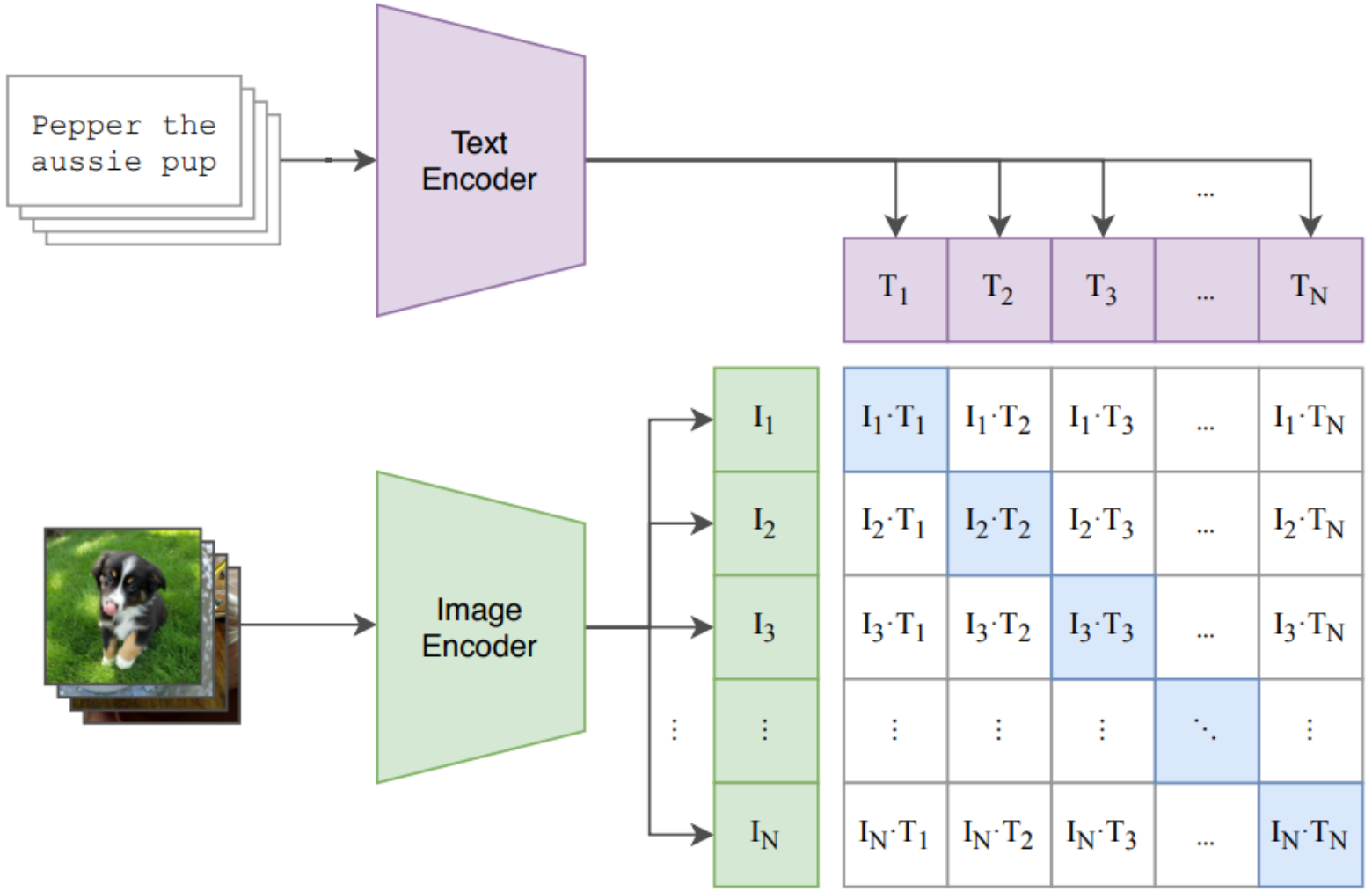
where

$$\begin{cases} \text{Lumiannce Comparison Function: } l(x, y) = \dfrac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \\ \text{Contrast Comparison Function: } c(x, y) = \dfrac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \\ \text{Structure Comparison Function: } s(x, y) = \dfrac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \end{cases}$$

Wang et al., “Image Quality Assessment: from Error Visibility to Structural Similarity,” IEEE Transactions on Image Processing, April 2004.
Horé et al., “Image Quality Metrics: PSNR vs. SSIM,” ICPR 2010.

CLIP Similarity

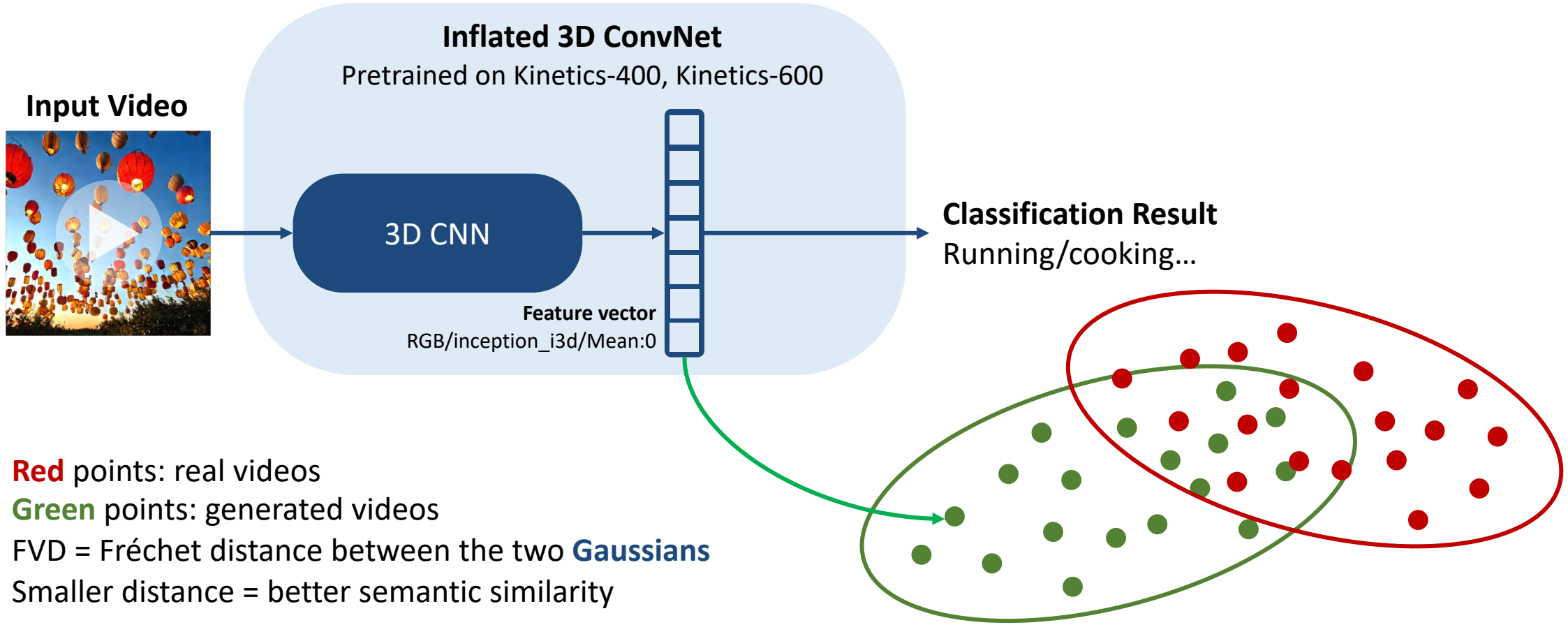
Image-caption similarity



Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML 2021.

Fréchet Video Distance (FVD)

Semantic similarity and temporal coherence between two videos

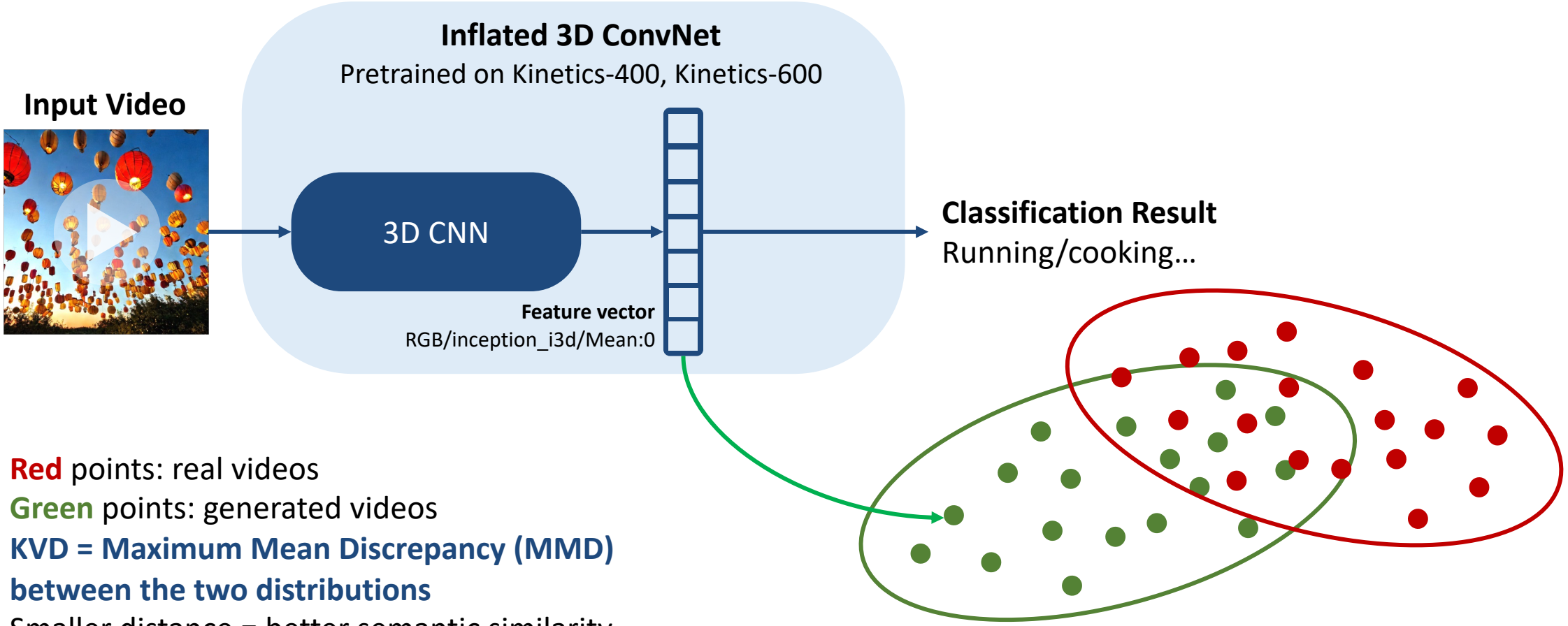


Red points: real videos
Green points: generated videos
FVD = Fréchet distance between the two **Gaussians**
Smaller distance = better semantic similarity

Unterthiner et al., "FVD: A new Metric for Video Generation," ICLR 2019.
Unterthiner et al., "Towards Accurate Generative Models of Video: A New Metric & Challenges," arXiv 2018.

Kernel Video Distance

Video quality assessment via semantic features and MMD



Red points: real videos
Green points: generated videos
KVD = Maximum Mean Discrepancy (MMD) between the two distributions
Smaller distance = better semantic similarity

Unterthiner et al., "FVD: A new Metric for Video Generation," ICLR 2019.
Unterthiner et al., "Towards Accurate Generative Models of Video: A New Metric & Challenges," arXiv 2018.

Video Inception Score (IS)

Video quality and diversity

KL divergence

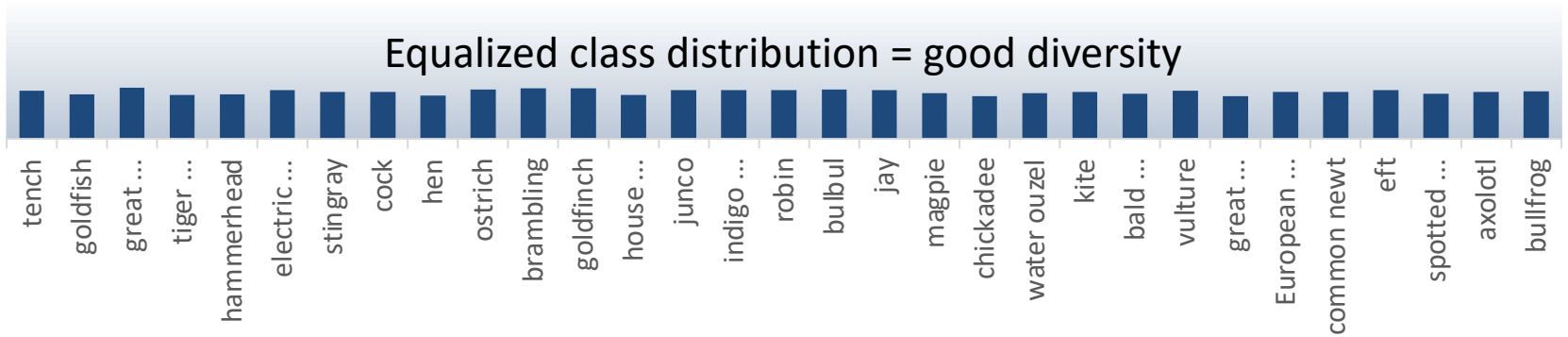
$$D_{KL}\left(\Pr(y|x) \,\|\, \Pr(y)\right)$$

$$\mathrm{IS}(G) \approx \exp\left(\frac{1}{N}\sum_{i=1}^{N} D_{KL}\right)$$

Larger IS = generated distribution is a "sharper and more distinct" collection of images

$\Pr(y|x)$: **quality** of containing clear objects; 3D-Convnets softmax logits

$\Pr(y) \approx \frac{1}{N}\sum_{i=1}^{N} \Pr(y|x_i)$: class **diversity** of generated images

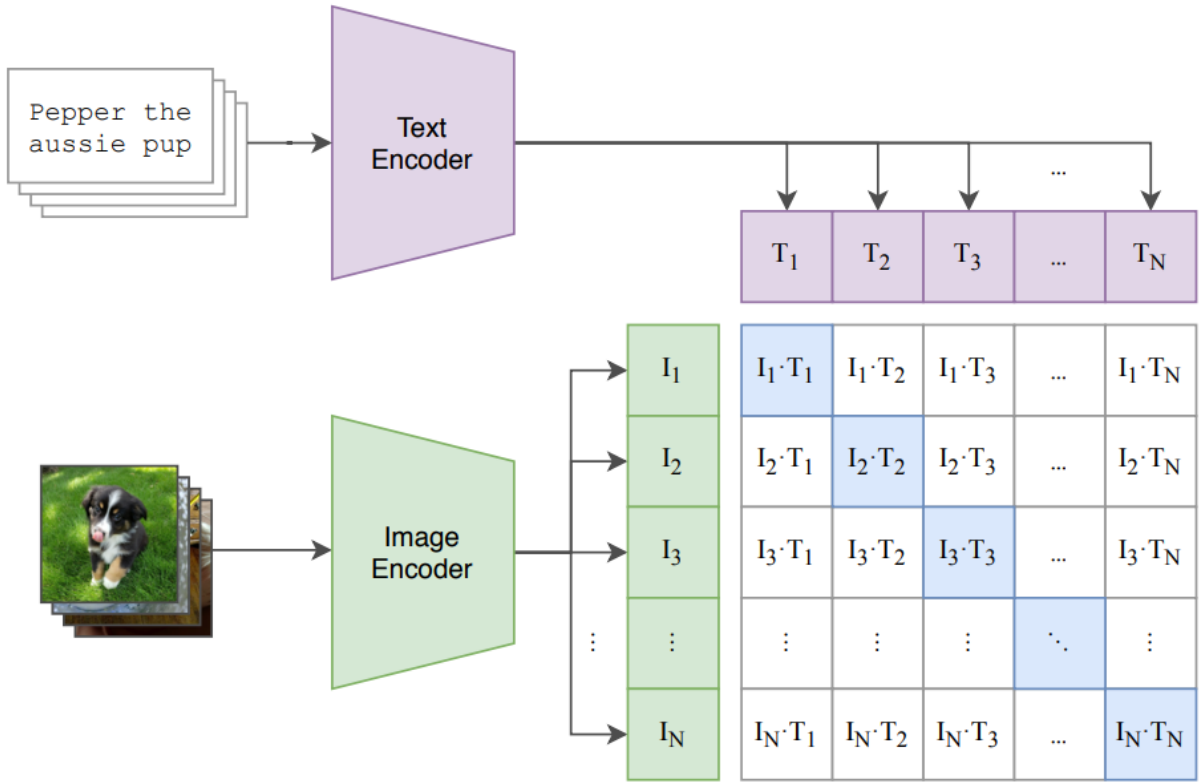


Salimans et al., "Improved Techniques for Training GANs," NeurIPS 2016.
Barratt et al., "A Note on the Inception Score," ICML 2018.
Saito et al., "Train Sparsely, Generated Densely: Memory-Efficient Unsupervised Training of High-Resolution Temporal GAN," IJCV 2020.

Frame Consistence CLIP scores

Frame temporal semantic consistency

- Compute CLIP image embeddings for all frames
- Report average cosine similarity between all pairs of frames



Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML 2021.

Evaluation Metrics

Qualitative evaluations

And…

Human ratings

Which image has better quality? Faithfulness?

Comparison	Benchmark	Quality	Faithfulness
Make-A-Video (ours) vs. VDM	VDM prompts (28)	84.38	78.13
Make-A-Video (ours) vs. CogVideo (Chinese)	DrawBench (200)	76.88	73.37
Make-A-Video (ours) vs. CogVideo (English)	DrawBench (200)	74.48	68.75
Make-A-Video (ours) vs. CogVideo (Chinese)	Our Eval. Set (300)	73.44	75.74
Make-A-Video (ours) vs. CogVideo (English)	Our Eval. Set (300)	77.15	71.19

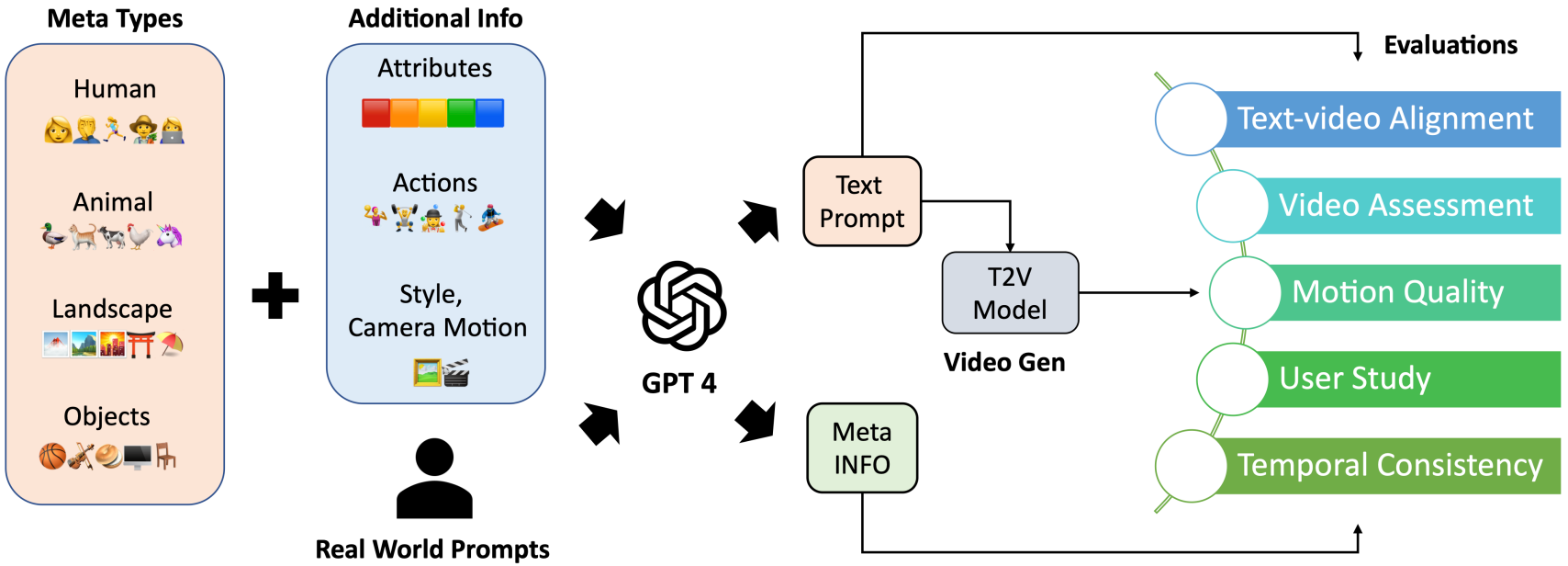
Example results from Singer et al., "Make-A-Video: Text-to-Video Generation without Text-Video Data," arXiv 2023.

Evaluation Metrics

Hybrid evaluation

EvalCrafter

- Creates a balanced prompt list for evaluation
- **Multi-criteria decision analysis** on 18 metrics: visual quality, content quality…
- Regress the coefficients of all metrics to generate an overall score aligned with user opinions



Liu et al., "EvalCrafter: Benchmarking and Evaluating Large Video Generation Models," arXiv 2023.

Make-A-Video

Cascaded generation

Table 1: T2V generation evaluation on MSR-VTT. Zero-Shot means no training is conducted on MSR-VTT. Samples/Input means how many samples are generated (and then ranked) for each input.

Method	Zero-Shot	Samples/Input	FID (↓)	CLIPSIM (↑)
GODIVA (Wu et al., 2021a)	No	30	–	0.2402
NÜWA (Wu et al., 2021b)	No	–	47.68	0.2439
CogVideo (Hong et al., 2022) (Chinese)	Yes	1	24.78	0.2614
CogVideo (Hong et al., 2022) (English)	Yes	1	23.59	0.2631
Make-A-Video (ours)	Yes	1	**13.17**	**0.3049**

Singer et al., "Make-A-Video: Text-to-Video Generation without Text-Video Data," arXiv 2022.

Make-A-Video

Cascaded generation

Table 2: Video generation evaluation on UCF-101 for both zero-shot and fine-tuning settings.

Method	Pretrain	Class	Resolution	IS (↑)	FVD (↓)
Zero-Shot Setting					
CogVideo (Chinese)	No	Yes	480 × 480	23.55	751.34
CogVideo (English)	No	Yes	480 × 480	25.27	701.59
Make-A-Video (ours)	No	Yes	256 × 256	**33.00**	**367.23**
Finetuning Setting					
TGANv2(Saito et al., 2020)	No	No	128 × 128	26.60 ± 0.47	-
DIGAN(Yu et al., 2022b)	No	No		32.70 ± 0.35	577 ± 22
MoCoGAN-HD(Tian et al., 2021)	No	No	256 × 256	33.95 ± 0.25	700 ± 24
CogVideo (Hong et al., 2022)	Yes	Yes	160 × 160	50.46	626
VDM (Ho et al., 2022)	No	No	64 × 64	57.80 ± 1.3	-
TATS-base(Ge et al., 2022)	No	Yes	128 × 128	79.28 ± 0.38	278 ± 11
Make-A-Video (ours)	Yes	Yes	256 × 256	**82.55**	**81.25**

Singer et al., "Make-A-Video: Text-to-Video Generation without Text-Video Data," arXiv 2022.

Make-A-Video

Cascaded generation

Table 3: Human evaluation results compared to CogVideo (Hong et al., 2022) on DrawBench and our test set, and to VDM (Ho et al., 2022) on the 28 examples from their website. The numbers show the percentage of raters that prefer the results of our Make-A-Video model.

Comparison	Benchmark	Quality	Faithfulness
Make-A-Video (ours) vs. VDM	VDM prompts (28)	84.38	78.13
Make-A-Video (ours) vs. CogVideo (Chinese)	DrawBench (200)	76.88	73.37
Make-A-Video (ours) vs. CogVideo (English)	DrawBench (200)	74.48	68.75
Make-A-Video (ours) vs. CogVideo (Chinese)	Our Eval. Set (300)	73.44	75.74
Make-A-Video (ours) vs. CogVideo (English)	Our Eval. Set (300)	77.15	71.19

Singer et al., “Make-A-Video: Text-to-Video Generation without Text-Video Data,” arXiv 2022.

Make-A-Video

Cascaded generation

" A dog wearing a Superhero outfit with red cape flying through the sky"

" A teddy bear painting a portrait"

Singer et al., "Make-A-Video: Text-to-Video Generation without Text-Video Data," arXiv 2022.

Make-A-Video

Cascaded generation

From static to magic

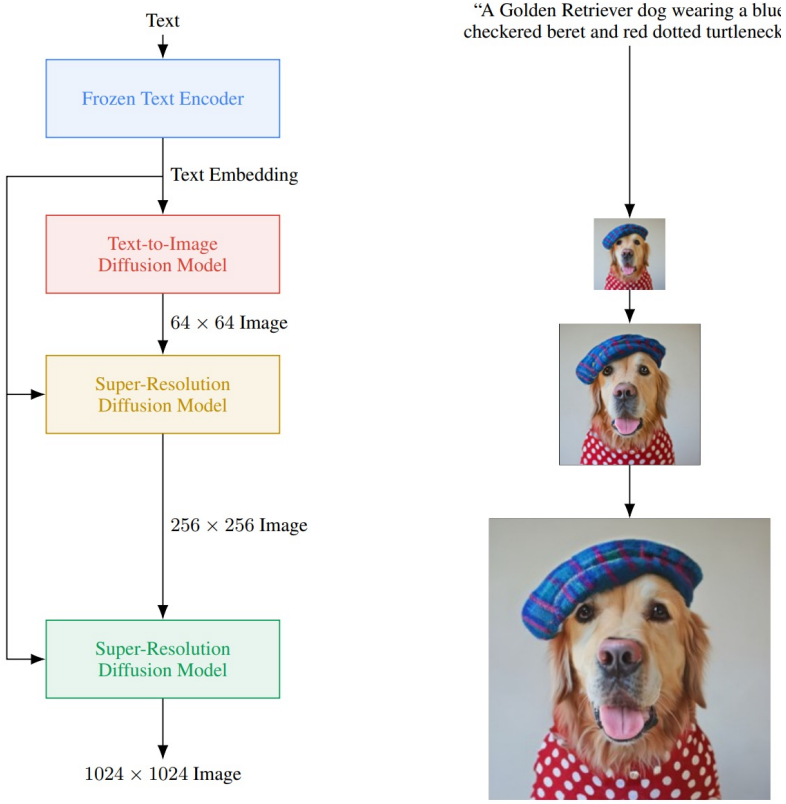
Add motion to a single image or fill-in the in-between motion to two images

Singer et al., “Make-A-Video: Text-to-Video Generation without Text-Video Data,” arXiv 2022.

Imagen & Imagen Video

Leverage pretrained T2I models for video generation; Cascaded generation

Imagen (for images)



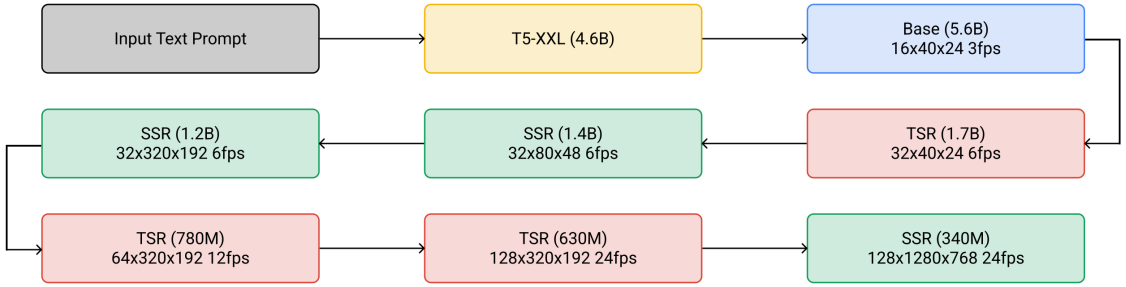
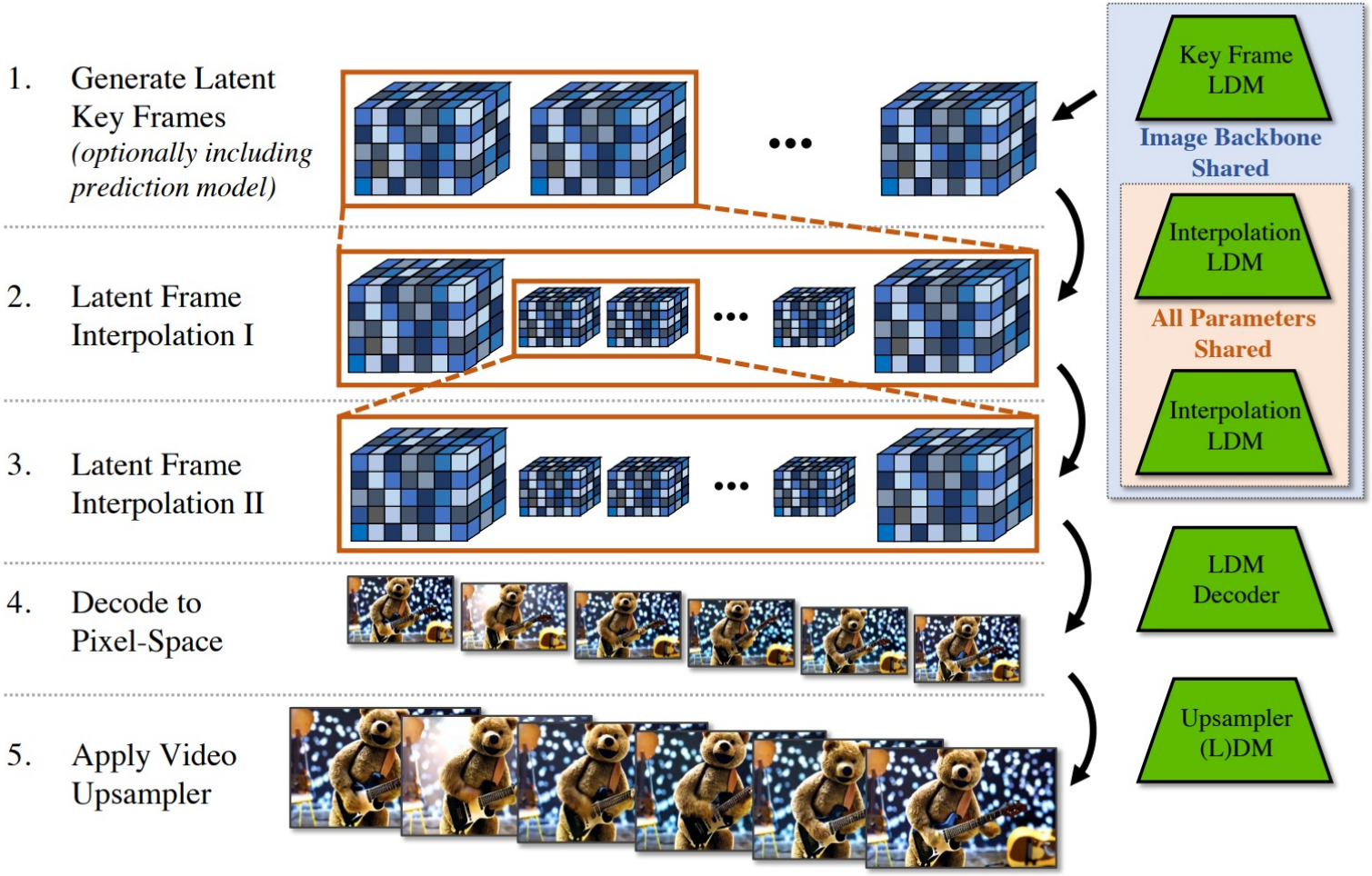


Imagen: Saharia et al., "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," arXiv 2022.
Imagen Video: Ho et al., "Imagen Video: High Definition Video Generation with Diffusion Models," arXiv 2022.

Align your Latents

Leverage pretrained T2I models for video generation; Cascaded generation



Blattmann et al., "Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models," CVPR 2023.

Align your Latents

Leverage pretrained T2I models for video generation

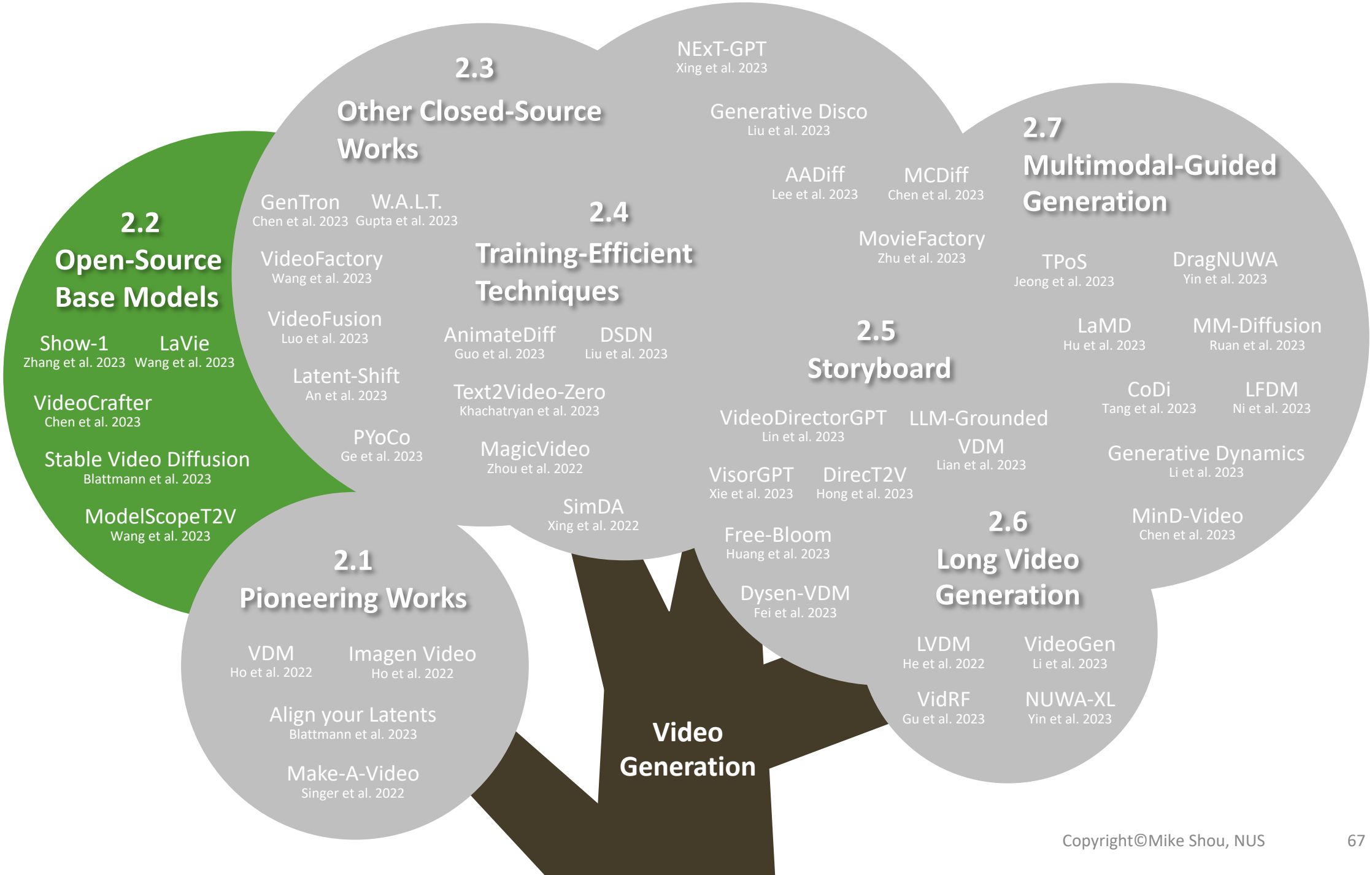
Inserting Temporal Layers

- Latent space diffusion model: insert temporal convolutional & 3D attention layers
- Decoder: add 3D convolutional layers
- Upsampler diffusion model: add 3D convolution layers

Blattmann et al., “Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models,” CVPR 2023.

2 Video Generation

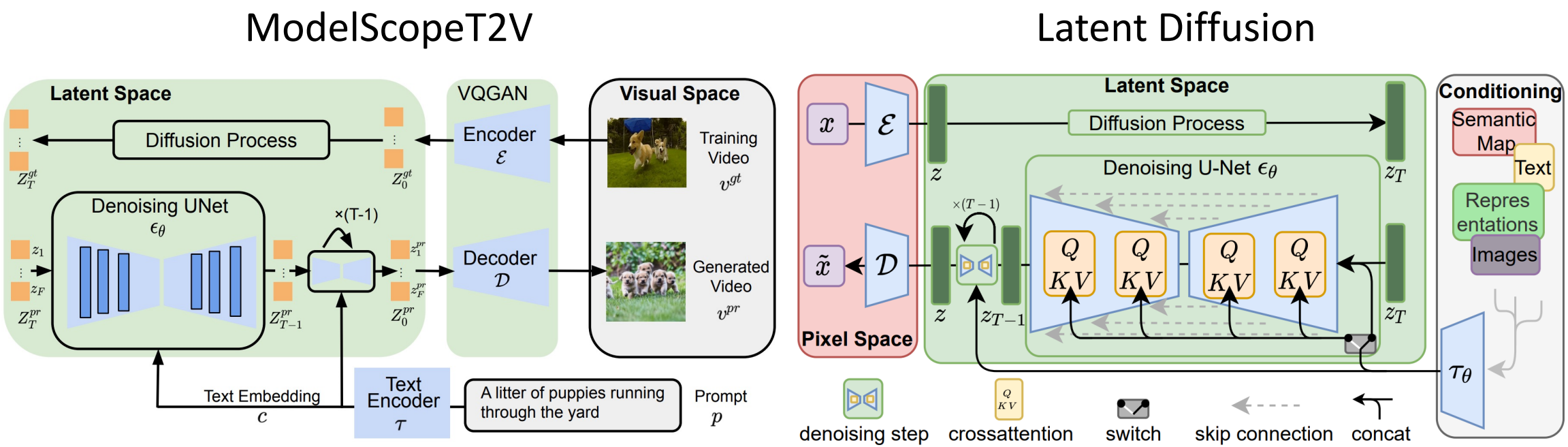
2.2 Open-source base models



ModelScopeT2V

Leverage pretrained T2I models for video generation

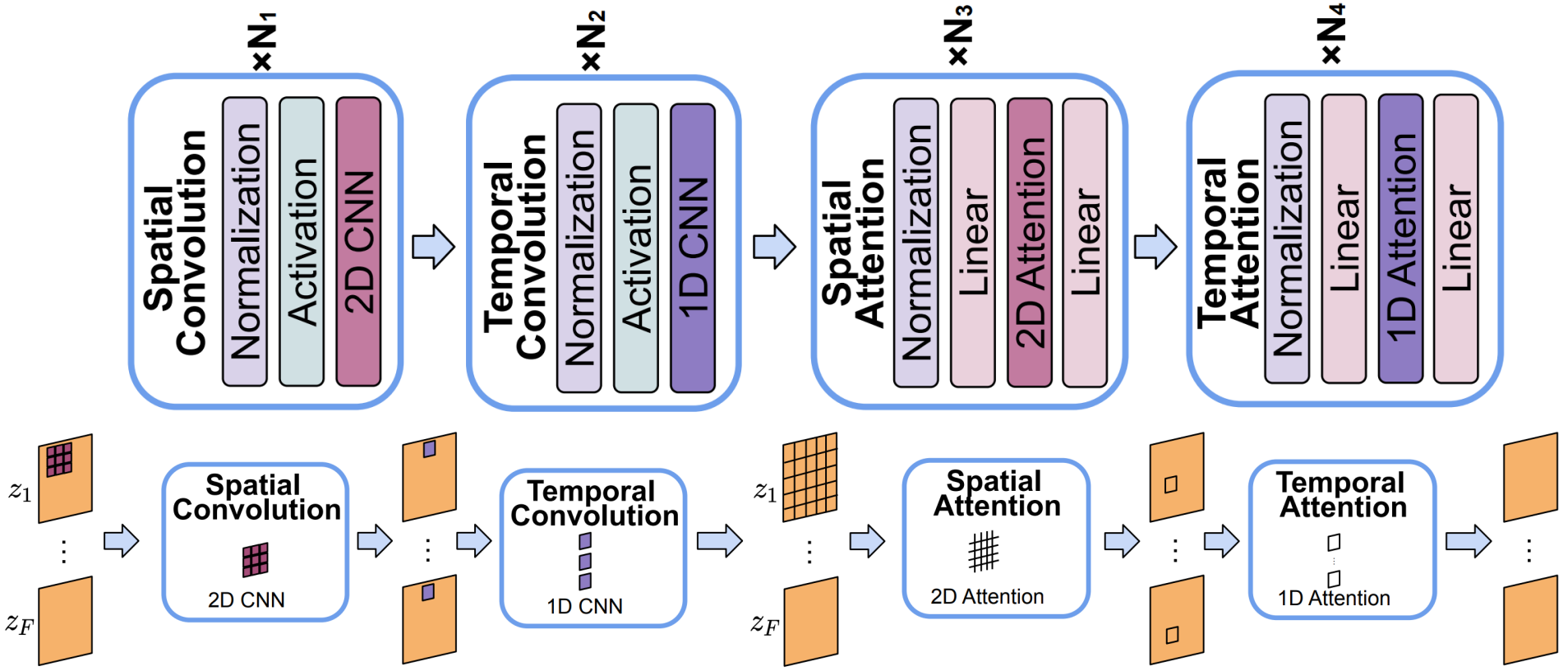
- Inflate Stable Diffusion to a 3D model, preserving pretrained weights
- Insert spatio-temporal blocks, can handle varying number of frames



ModelScopeT2V

Leverage pretrained T2I models for video generation

- Inflate Stable Diffusion to a 3D model, preserving pretrained weights
- Insert spatio-temporal blocks

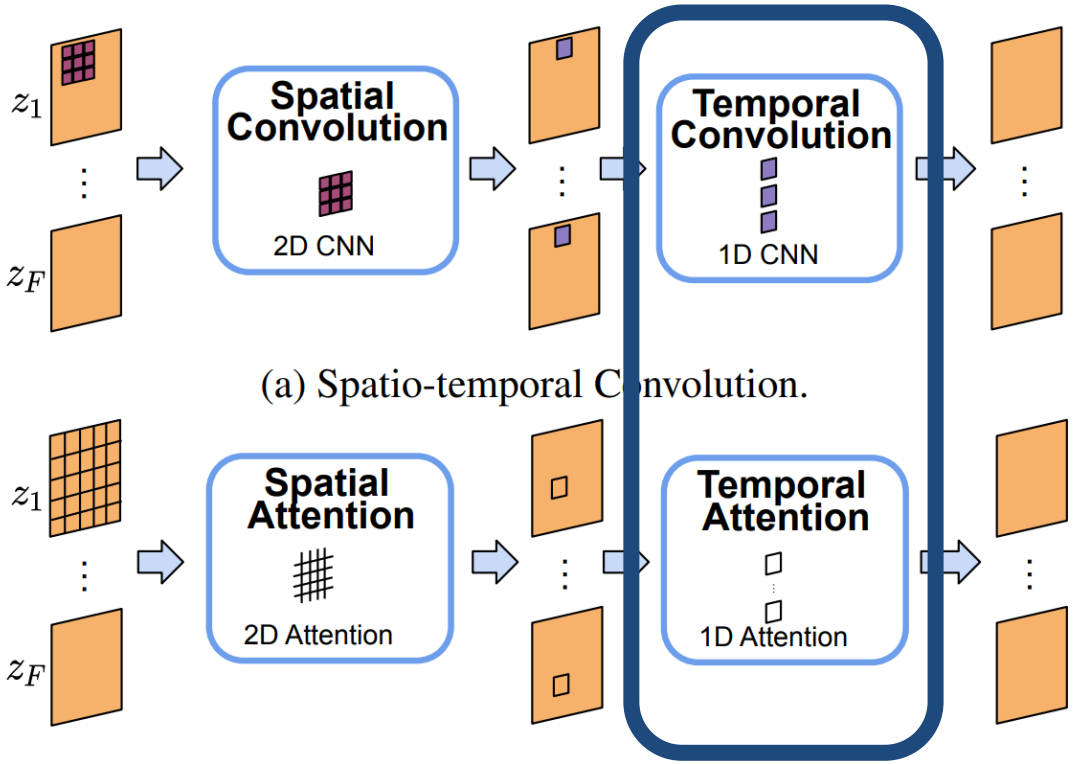


Wang et al., "ModelScope Text-to-Video Technical Report," arXiv 2023.

ModelScopeT2V

Leverage pretrained T2I models for video generation

- Inflate Stable Diffusion to a 3D model, preserving pretrained weights
- Insert spatio-temporal blocks, **can handle varying number of frames**



Temporal convolution block handles

$1 \times \{\text{varying length}\}$

Sequences

Temporal attention block handles

$1 \times \{\text{varying length}\}$

Sequences

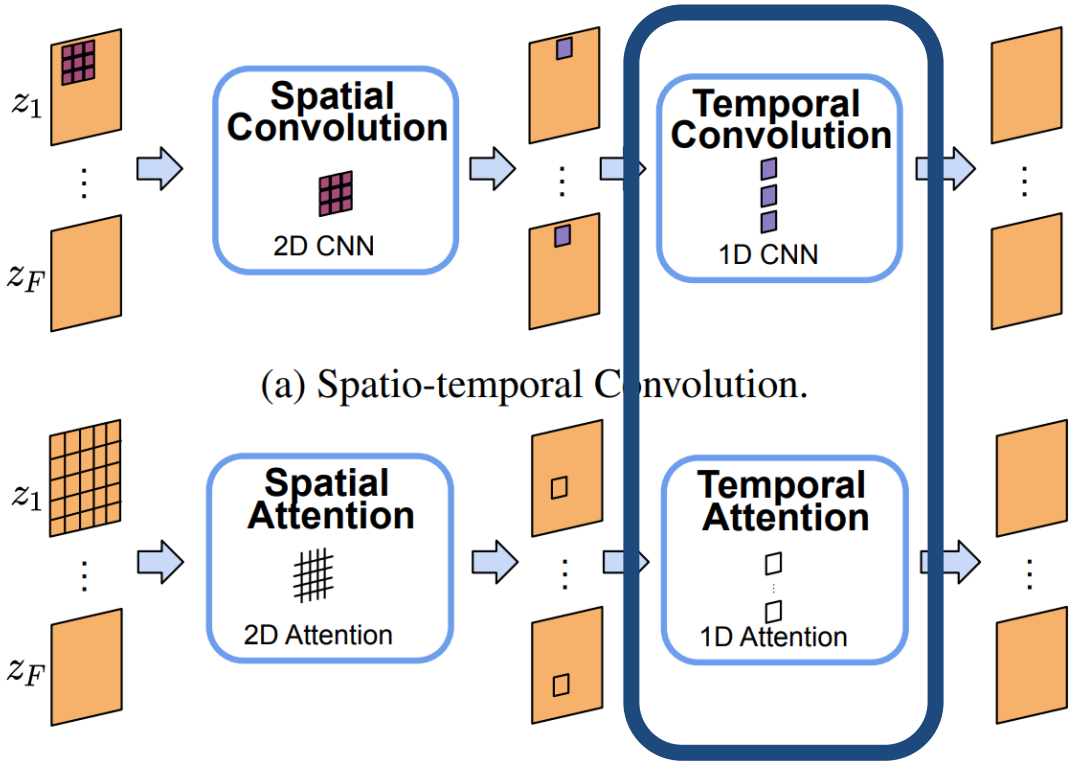
(Size of attn map changes)

Wang et al., "ModelScope Text-to-Video Technical Report," arXiv 2023.

ModelScopeT2V

Leverage pretrained T2I models for video generation

- Inflate Stable Diffusion to a 3D model, preserving pretrained weights
- Insert spatio-temporal blocks, **can handle varying number of frames**



Length = 1

Model generate images

Wang et al., "ModelScope Text-to-Video Technical Report," arXiv 2023.

ModelScopeT2V

Leverage pretrained T2I models for video generation

Models	FID-vid (↓)	FVD (↓)	CLIPSIM (↑)
NÜWA [62]	47.68	-	0.2439
CogVideo (Chinese) [20]	24.78	-	0.2614
CogVideo (English) [20]	23.59	1294	0.2631
MagicVideo [71]	-	1290	-
Video LDM [3]	-	-	0.2929
Make-A-Video [51]	13.17	-	**0.3049**
ModelScopeT2V (ours)	**11.09**	**550**	0.2930

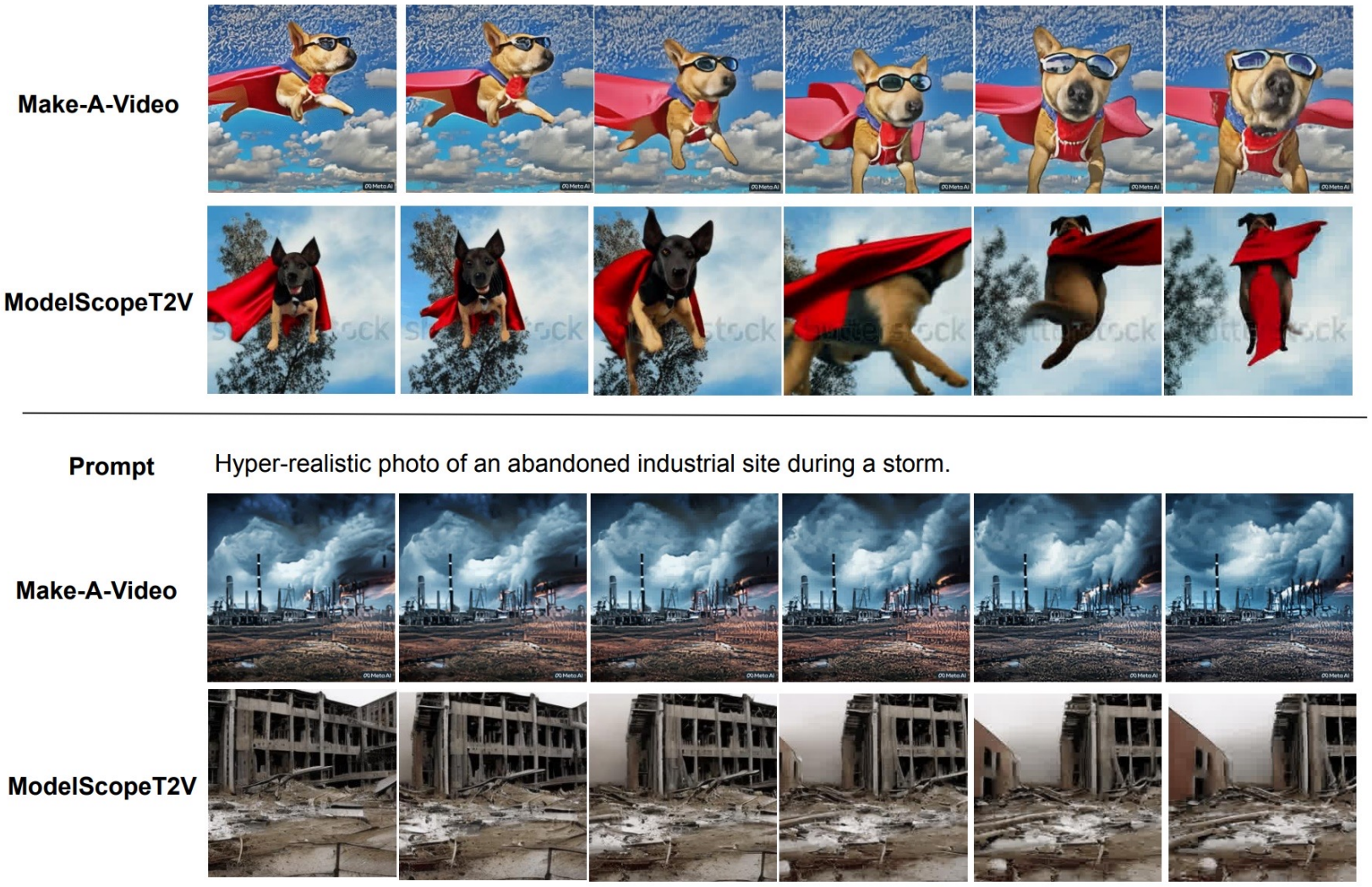
Table 1: **Quantitative comparison with state-of-the-art models on MSR-VTT.** We evaluate the models with three metrics (*i.e.*, FID-vid [15], FVD [57], and CLIPSIM [61]).

ZeroScope: finetunes ModelScope on a small set of high-quality videos, resulting into higher resolution at 1024 x 576, without the Shutterstock watermark

Wang et al., "ModelScope Text-to-Video Technical Report," arXiv 2023.

ModelScopeT2V

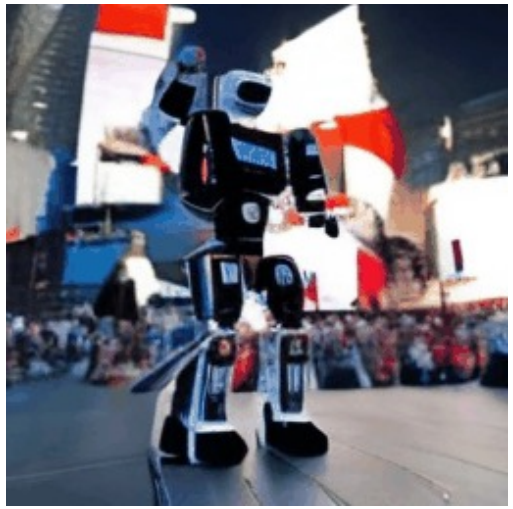
Leverage pretrained T2I models for video generation



ModelScopeT2V

Leverage pretrained T2I models for video generation

" Robot dancing in times square," arXiv 2023.



" Clown fish swimming through the coral reef," arXiv 2023.

" Melting ice cream dripping down the cone," arXiv 2023.

" Hyper-realistic photo of an abandoned industrial site during a storm," arXiv 2023.

Wang et al., "ModelScope Text-to-Video Technical Report," arXiv 2023.

Show-1

Better text-video alignment? Generation in both pixel- and latent-domain

A **panda** besides the waterfall is holding a sign that says ***"Show Lab".***

A ***blue tiger*** in the grass in the sunset, ***surrounded by butterflies.***

Show-1 Gen-2 Zeroscope Modelscope

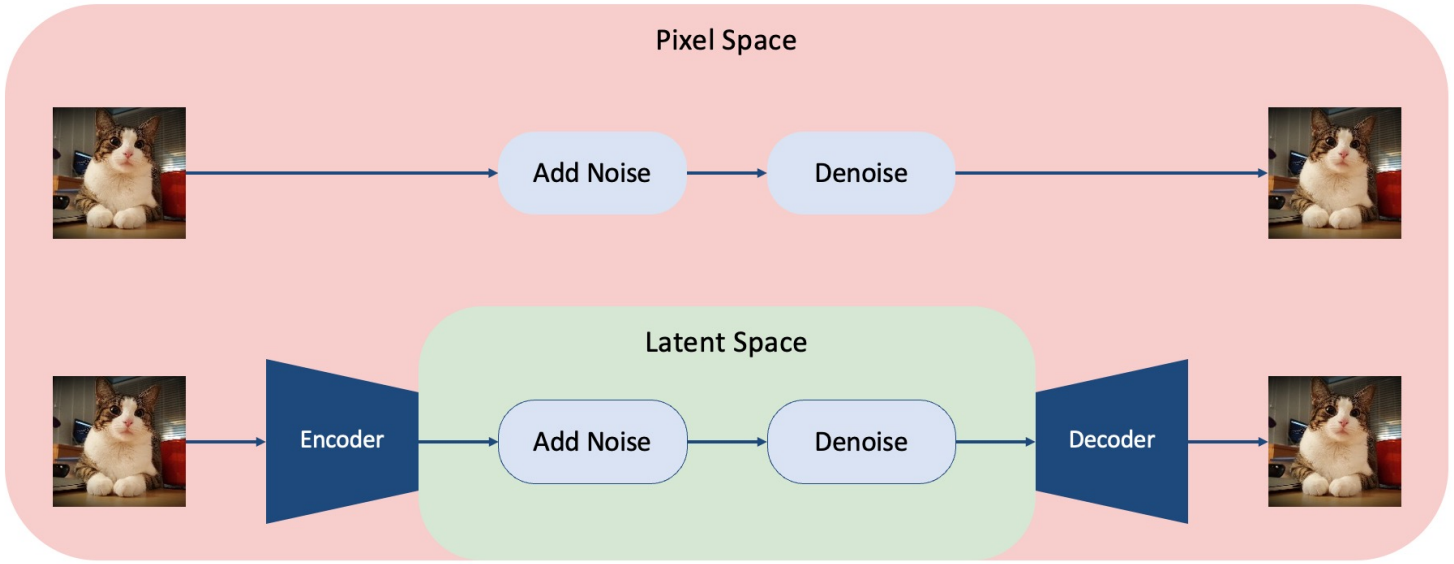
Zhang et al., “Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation,” arXiv 2023.

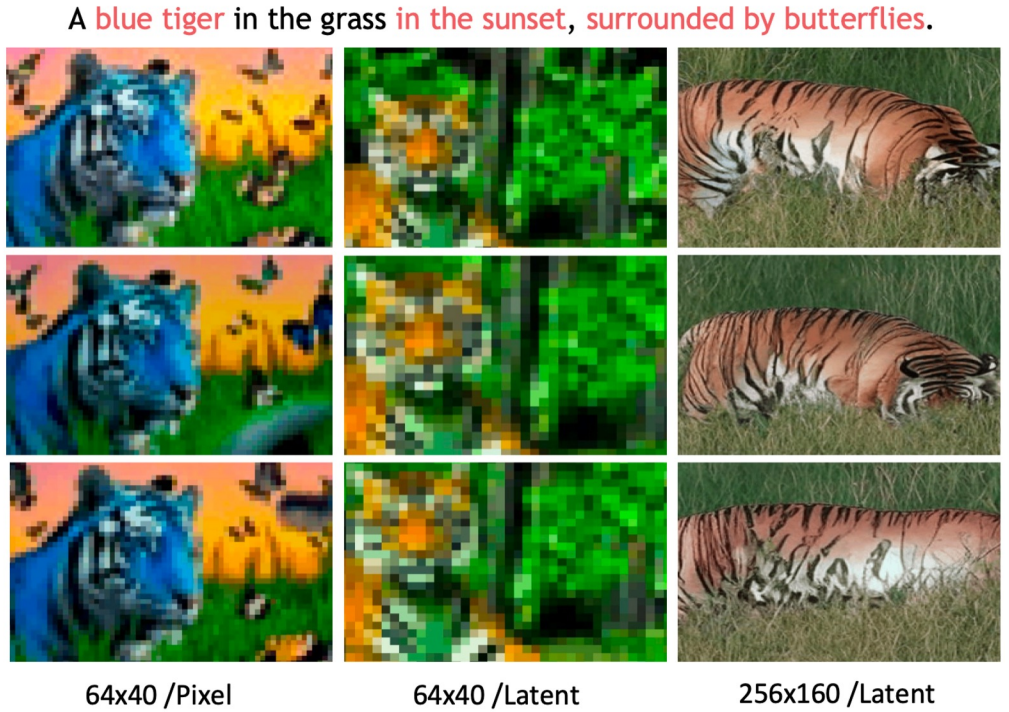
Show-1

Better text-video alignment? Generation in both pixel- and latent-domain

Motivation

- Pixel-based VDM achieves better text-video alignment than latent-based VDM





Zhang et al., "Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation," arXiv 2023.

Show-1

Generation in both pixel- and latent-domain

Motivation

- Pixel-based VDM achieves better text-video alignment than latent-based VDM
- Pixel-based VDM takes much larger memory than latent-based VDM

Keyframes Stage	Final Super-Res. Stage	CLIP-T Score	T-V	Max Mem.
64x40 or 256x160 /Pixel	576x320 /Pixel	--	--	72GB
64x40 /Pixel	576x320 /Latent	0.3026	71%	15GB
64x40 /Latent	576x320 /Latent	0.2441	3%	15GB
256x160 /Latent	576x320 /Latent	0.2874	26%	15GB

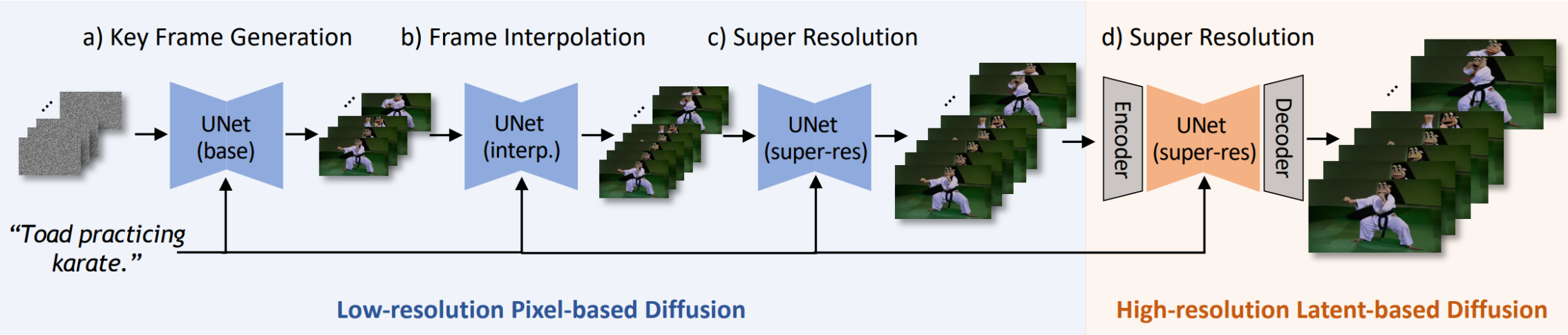
Zhang et al., "Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation," arXiv 2023.

Show-1

Generation in both pixel- and latent-domain

Motivation

- Use Pixel-based VDM in low-res stage
- Use latent-based VDM in high-res stage



Zhang et al., "Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation," arXiv 2023.

Show-1

Generation in both pixel- and latent-domain

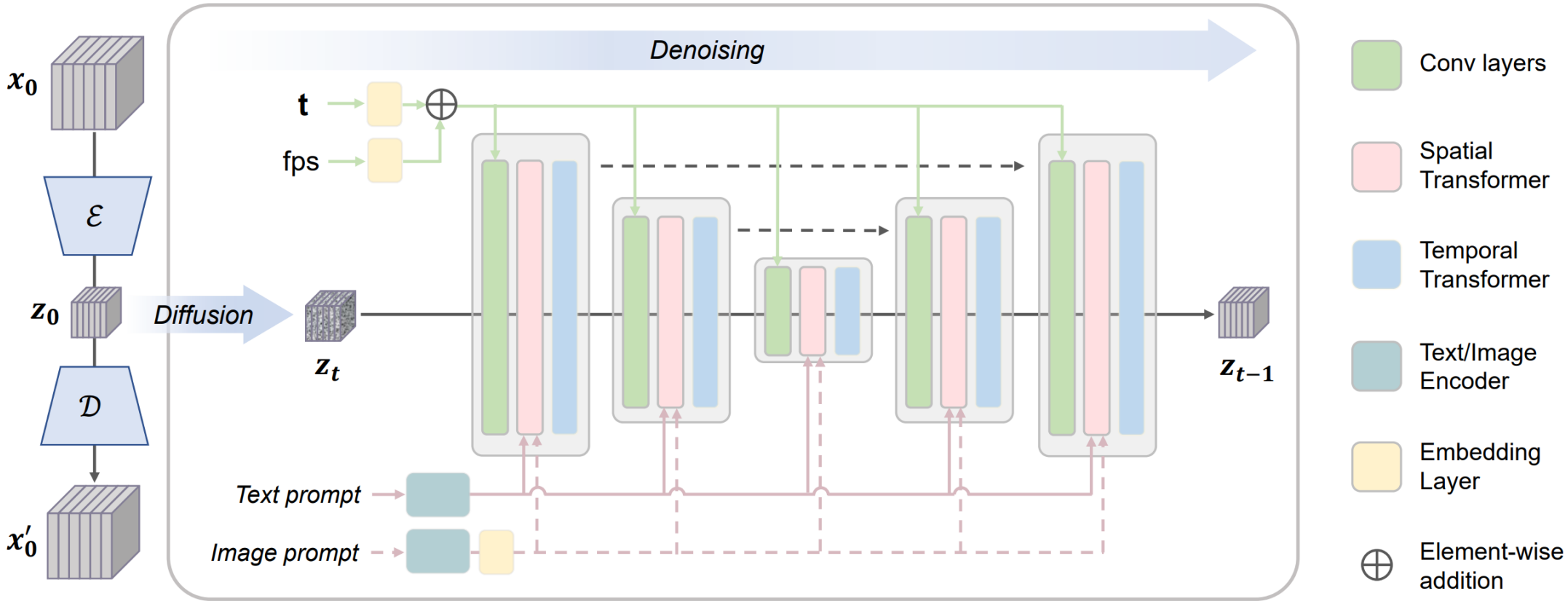
https://github.com/showlab/Show-1

- Better text-video alignment
- Can synthesize large motion
- Memory-efficient

Zhang et al., “Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation,” arXiv 2023.

VideoCrafter

- Latent diffusion inserted with temporal layers

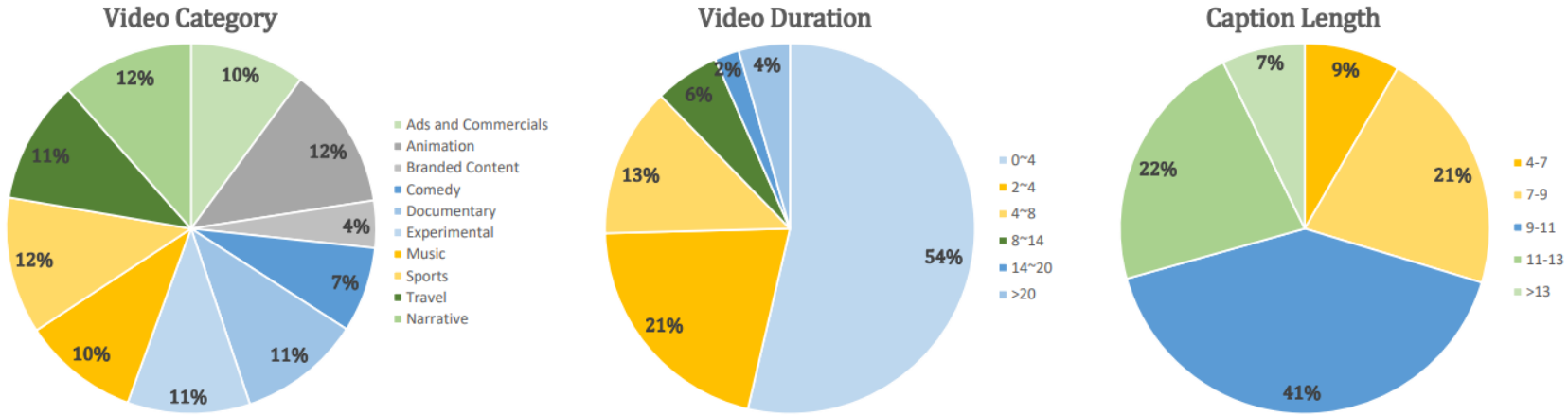


Chen et al., "VideoCrafter1: Open Diffusion Models for High-Quality Video Generation," arXiv 2023.

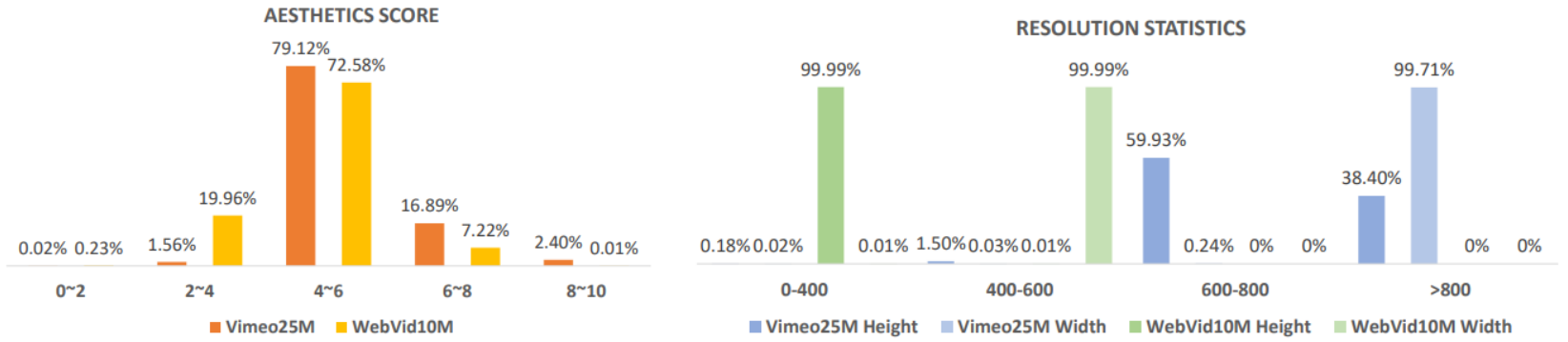
Joint image-video finetuning with curriculum learning

Proposed Dataset: Vimeo25M

General Information Statistics



Aesthetics Score and Video Resolution Statistics



(a) Aesthetics score statistics

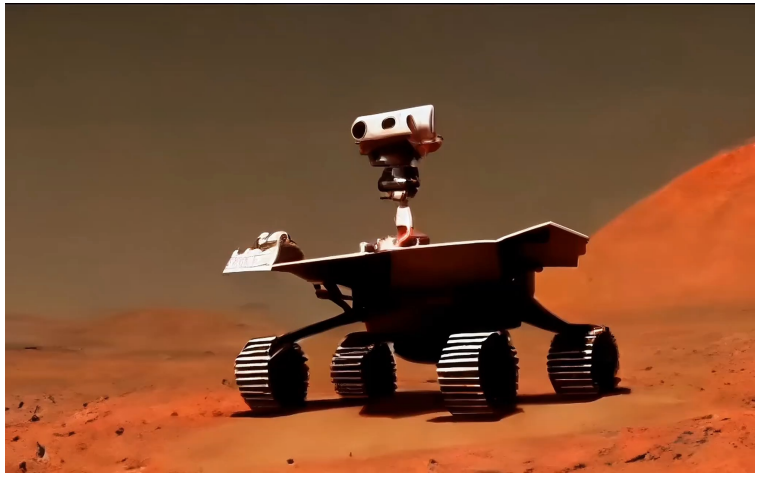
(b) Resolution statistics

Wang et al., "LAVIE: High-Quality Video Generation with Cascaded Latent Diffusion Models," arXiv 2023.

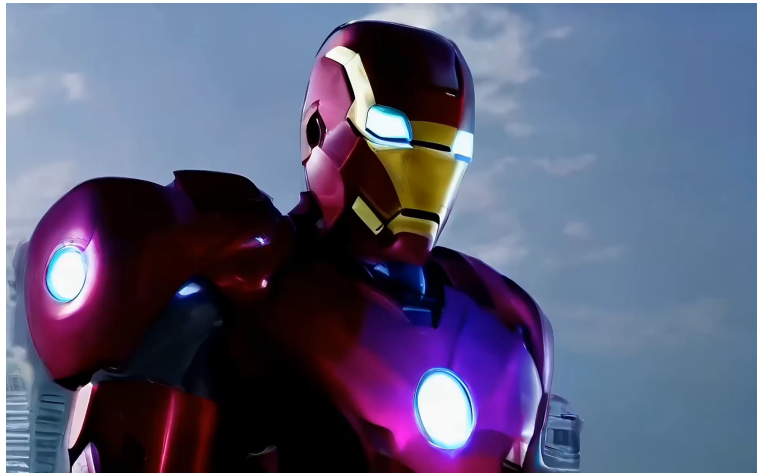
LaVie

Joint image-video finetuning with curriculum learning

A jellyfish floating through the ocean, with bioluminescent tentacles



A Mars rover moving on Mars



Iron Man flying in the sky

The bund Shanghai, oil painting

Wang et al., “LAVIE: High-Quality Video Generation with Cascaded Latent Diffusion Models,” arXiv 2023.

Stable Video Diffusion

Scaling latent video diffusion models to large datasets

Data Processing and Annotation

- Cut Detection and Clipping
  - Detect cuts/transitions at multiple FPS levels
  - Extract clips precisely using keyframe timestamps
- Synthetic Captioning
  - Use CoCa image captioner to caption the mid-frame of each clip
  - Use V-BLIP to obtain video-based caption
  - Use LLM to summarise the image- and video-based caption
  - Compute CLIP similarities and aesthetic scores
- Filter Static Scene
  - Use dense optical flow magnitudes to filter static scenes
- Text Detection
  - Use OCR to detect and remove clips with excess text

Blattmann et al., “Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets,” 2023.

Stable Video Diffusion

Scaling latent video diffusion models to large datasets

Data Processing and Annotation

Table 1. Comparison of our dataset before and after fitering with publicly available research datasets.

	LVD	*LVD-F*	*LVD-10M*	*LVD-10M-F*	*WebVid*	*InternVid*
#Clips	577M	152M	9.8M	2.3M	10.7M	234M
Clip Duration (s)	11.58	10.53	12.11	10.99	18.0	11.7
Total Duration (y)	212.09	50.64	3.76	0.78	5.94	86.80
Mean #Frames	325	301	335	320	-	-
Mean Clips/Video	11.09	4.76	1.2	1.1	1.0	32.96
Motion Annotations?	✓	✓	✓	✓	✗	✗

Blattmann et al., "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets," 2023.

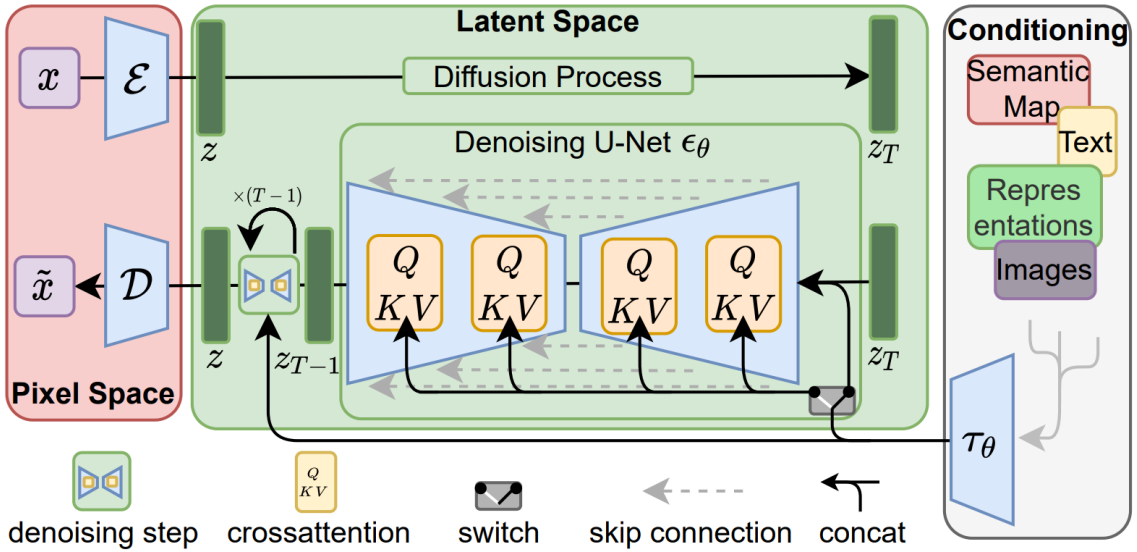
Stable Video Diffusion

Scaling latent video diffusion models to large datasets

Stage I: Image Pretraining

- Initialize weights from Stable Diffusion 2.1 (text-to-image model)

Stable Diffusion 2.1



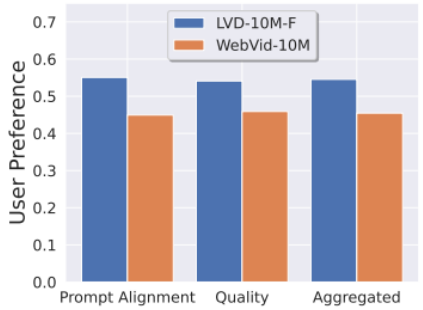
→ Inflate to 3D

Stable Video Diffusion

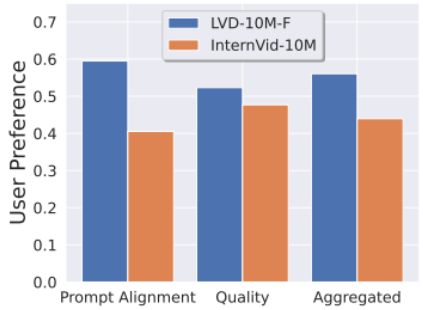
Scaling latent video diffusion models to large datasets

Stage II: Curating a Video Pretraining Dataset

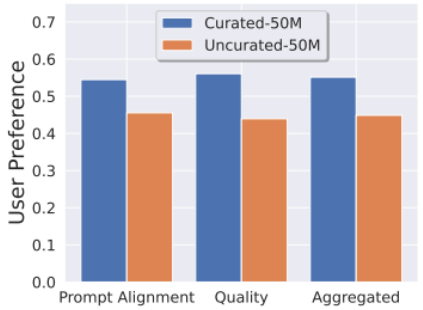
- Systematic Data Curation
  - Curate subsets filtered by various criteria (CLIP-, OCR-, optical flow-, aesthetic-scores…)
  - Assess human preferences on models trained on different subsets
  - Choose optimal filtering thresholds via Elo rankings for human preference votes
- Well-curated beats un-curated pretraining dataset



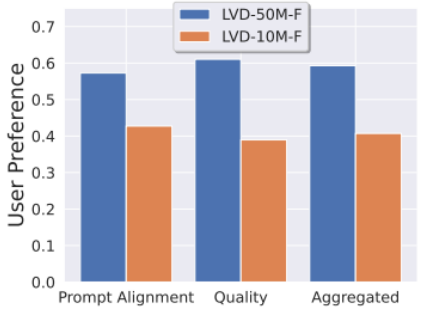
(a) User preference for *LVD-10M-F* and WebVid [6].



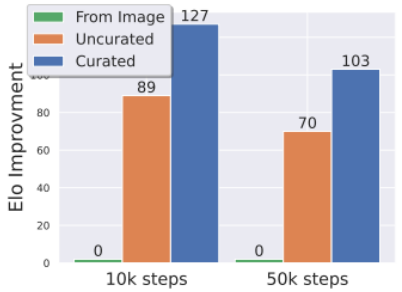
(b) User preference for *LVD-10M-F* and InternVid [96].



(c) User preference at 50M samples scales.



(d) User preference on scaling datasets.



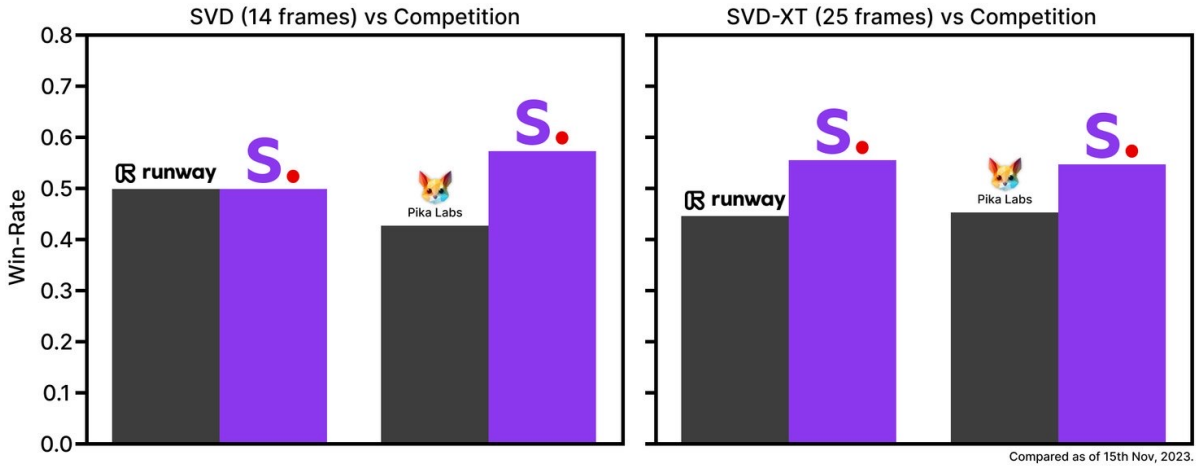
(e) Relative ELO progression over time during Stage III.

Stable Video Diffusion

Scaling latent video diffusion models to large datasets

Stage III: High-Quality Finetuning

- Finetune base model (pretrained from Stages I-II) on high-quality video data
  - High-Resolution Text-to-Video Generation
    - ~1M samples. Finetune for 50K iterations at 576x1024 (in contrast to 320x576 base resolution)
  - High Resolution Image-to-Video Generation
  - Frame Interpolation
  - Multi-View Generation
- Performance gains from curation persists after finetuning



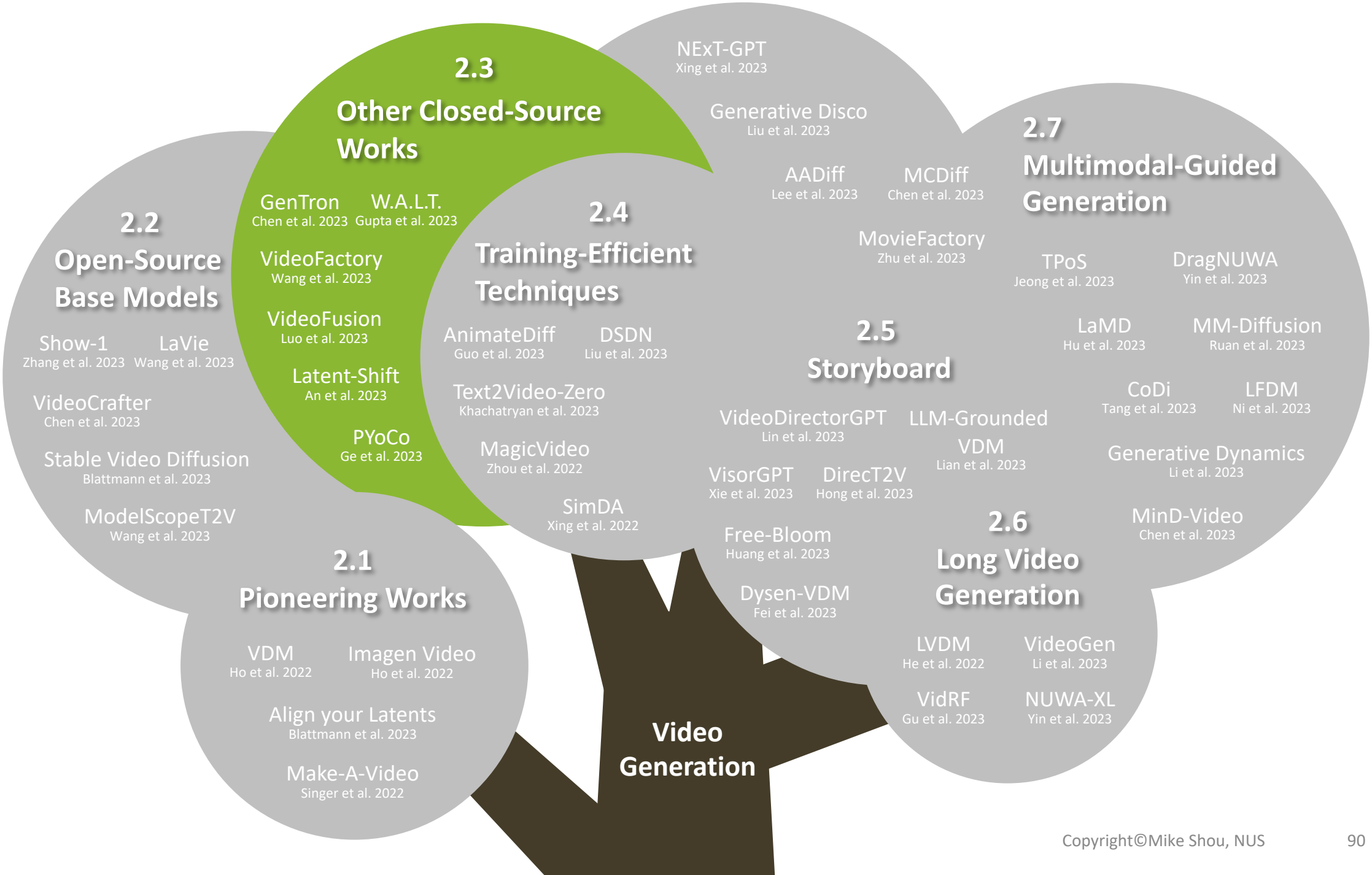
Stable Video Diffusion

Scaling latent video diffusion models to large datasets

Blattmann et al., “Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets,” 2023.

2 Video Generation

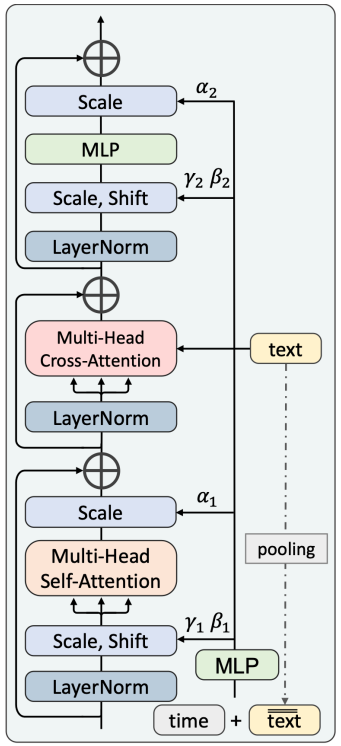
2.3 Other closed-source works



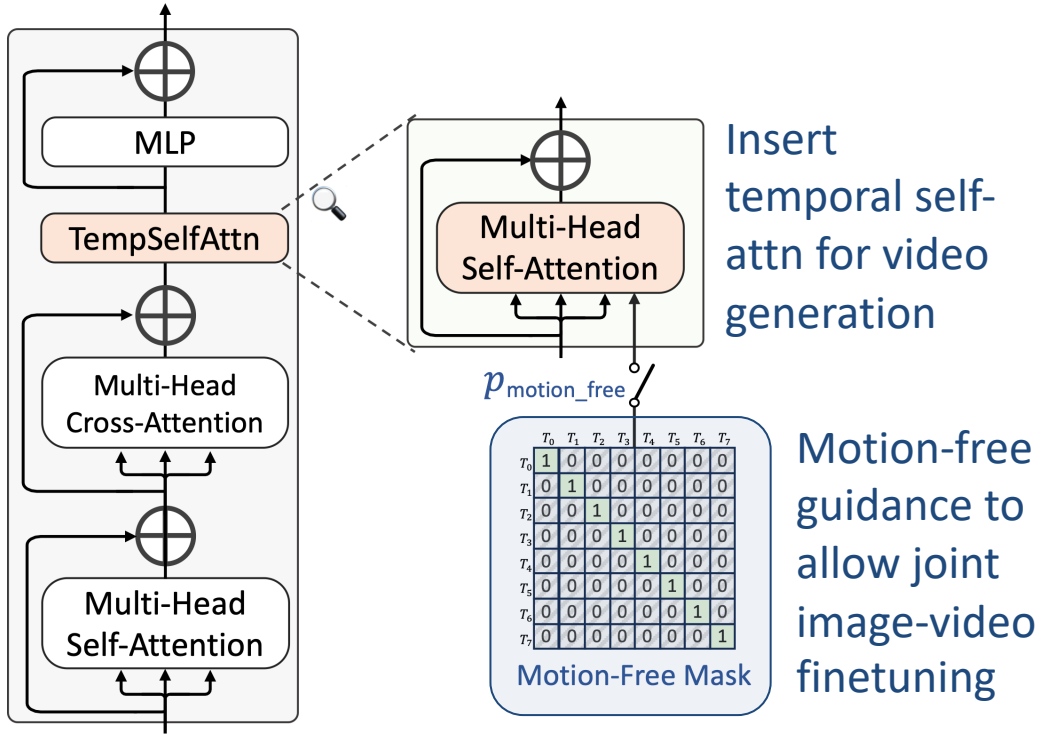
GenTron

Transformer-based diffusion for text-to-video generation

- Transformer-based architecture extended from DiT (class-conditioned transformer-based LDM)
- Train T2I → insert temporal self-attn → joint image-video finetuning (motion-free guidance)



Text-to-image architecture



Text-to-video architecture

GenTron

Transformer-based diffusion for text-to-video generation

"A fantasy landscape trending on Artstation, 4k"

"An astronaut flying in space, 4k high resolution"

"An astronaut riding a horse high definition, 4k"

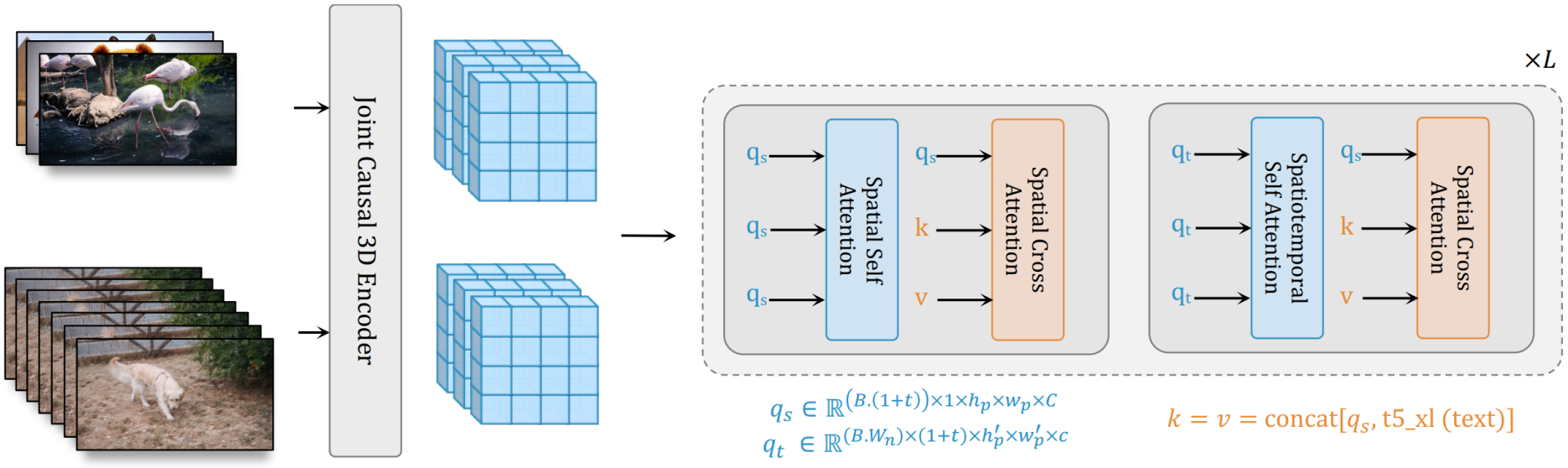
"Flying through a fantasy landscape, 4k high resolution"

Chen et al., "GenTron: Delving Deep into Diffusion Transformers for Image and Video Generation," arXiv 2023.

W.A.L.T.

Transformer-based diffusion for text-to-video generation

- Transformer-based denoising diffusion backbone
- Joint image-video training via unified image/video latent space (created by a joint 3D encoder with causal 3D conv layers, allowing the first frame of a video to be tokenized independently)
- Window attention to reduce computing/memory costs
- Cascaded pipeline for high-quality generation

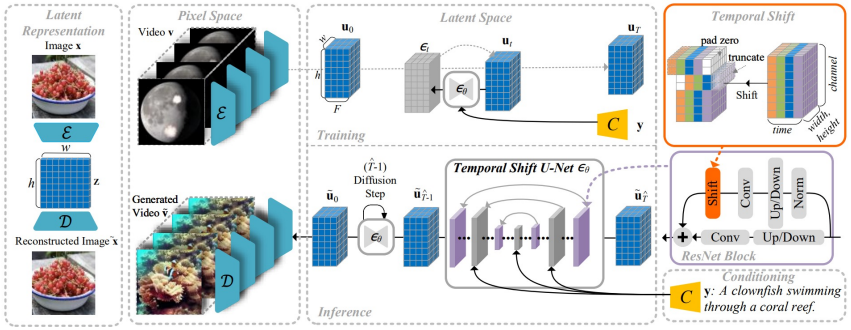


W.A.L.T.

Transformer-based diffusion for text-to-video generation

Gupta et al., “Photorealistic Video Generation with Diffusion Models,” arXiv 2023.

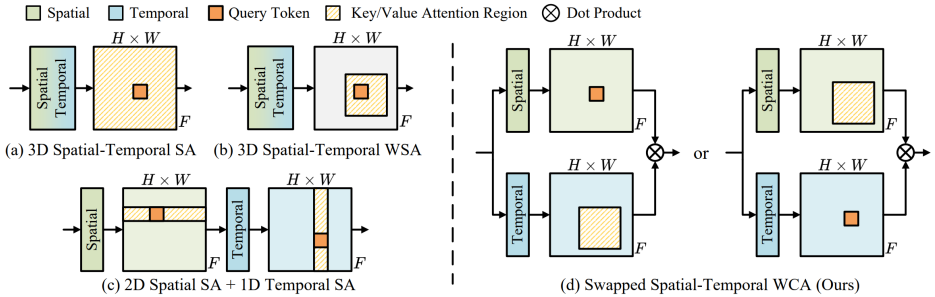
Other Closed-Source Works



Latent Shift (An et al.)

Shift latent features for better temporal coherence

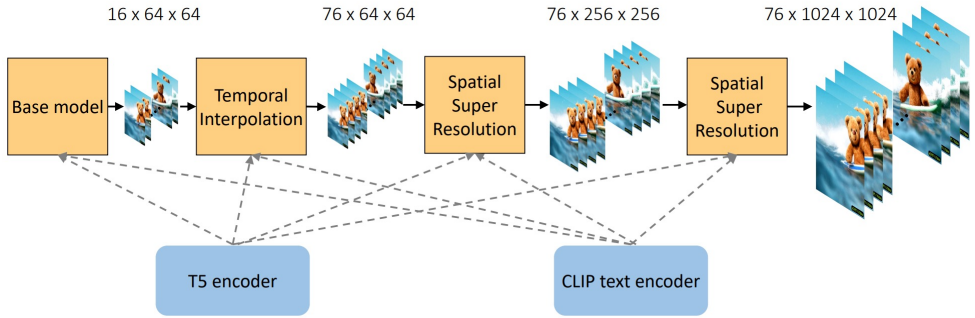
"Latent-Shift: Latent Diffusion with Temporal Shift for Efficient Text-to-Video Generation," arXiv 2023.



Video Factory (Wang et al.)

Modify attention mechanism for better temporal coherence

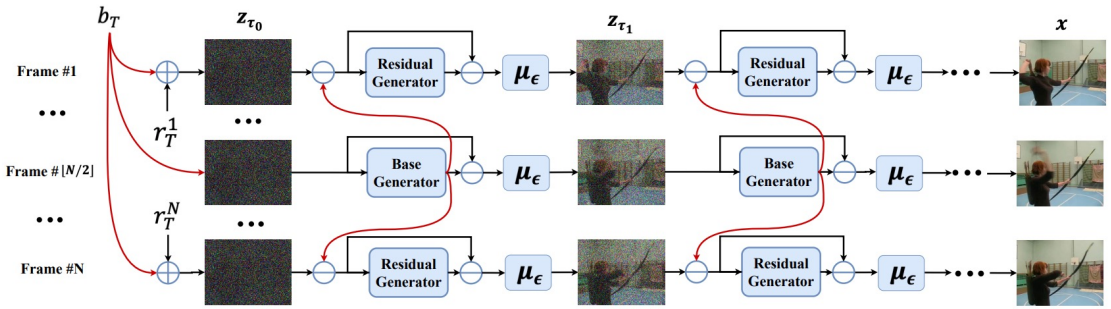
"VideoFactory: Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation," arXiv 2023.



PYoCo (Ge et al.)

Generate video frames starting from similar noise patterns

"Preserve Your Own Correlation: A Noise Prior for Video Diffusion Models," ICCV 2023.



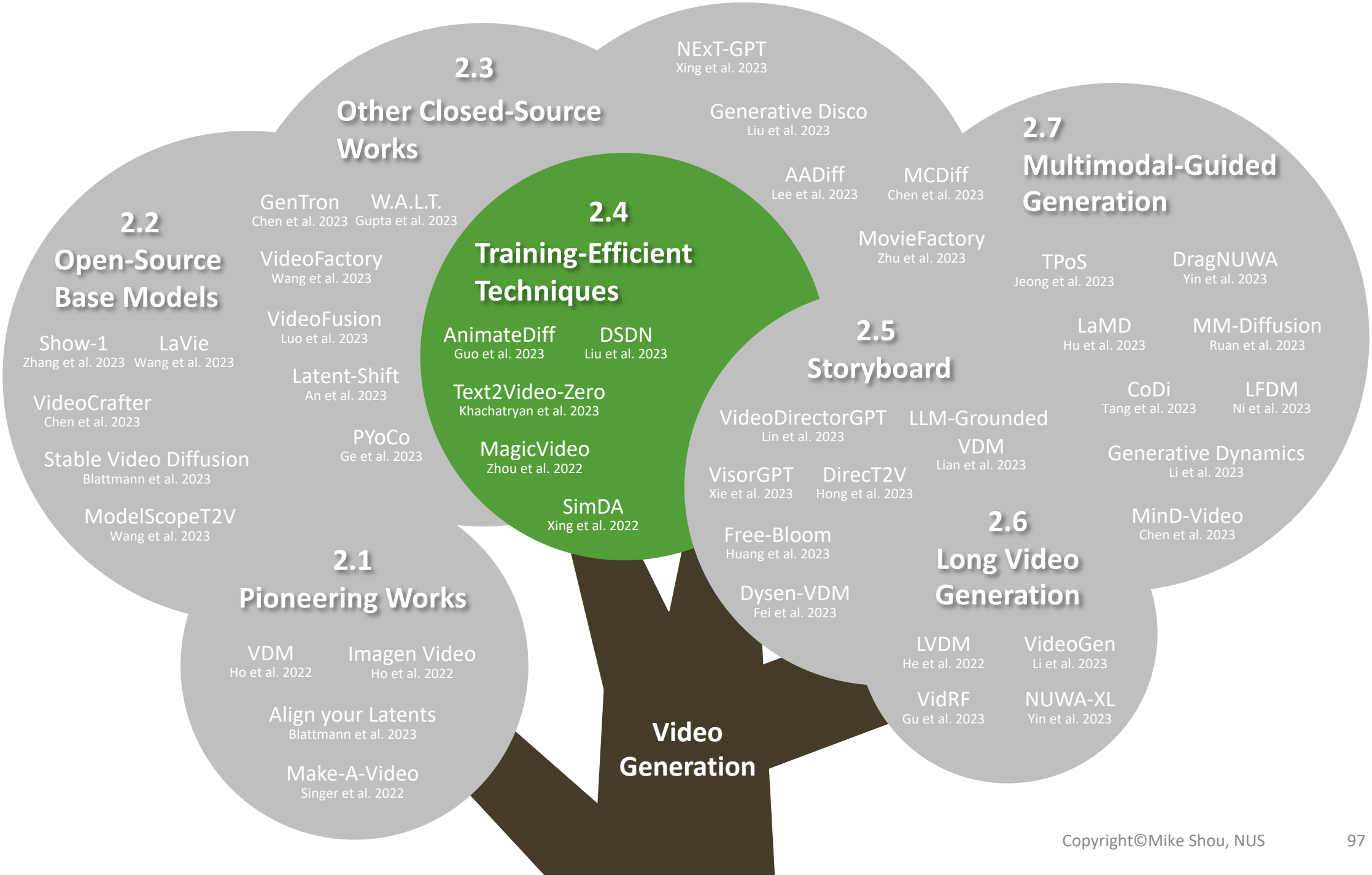
VideoFusion (Lorem et al.)

Decompose noise into shared "base" and individual "residuals"

"VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation," CVPR 2023.

2 Video Generation

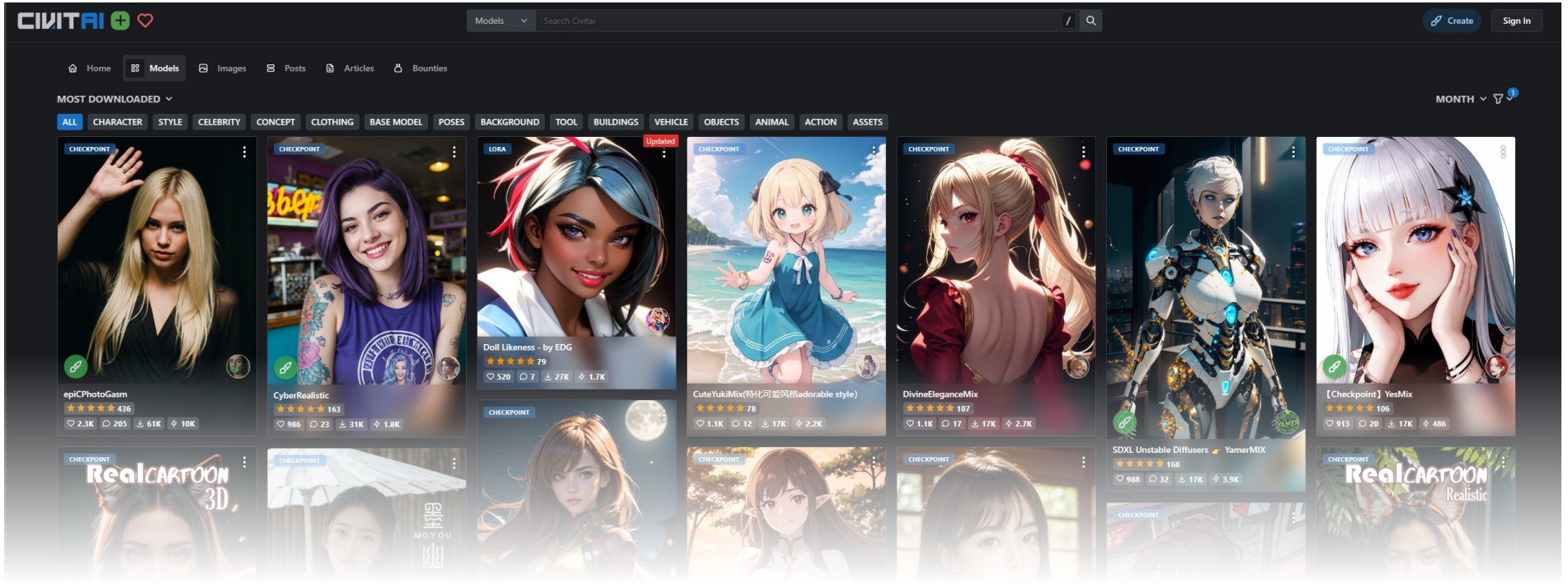
2.4 Training-efficient techniques



AnimateDiff

Transform domain-specific T2I models to T2V models

- Domain-specific (personalized) models are widely available for image
  - Domain-specific finetuning methodologies: LoRA, DreamBooth…
  - Communities: Hugging Face, CivitAI…
- **Task: turn these image models into T2V models, without specific finetuning**



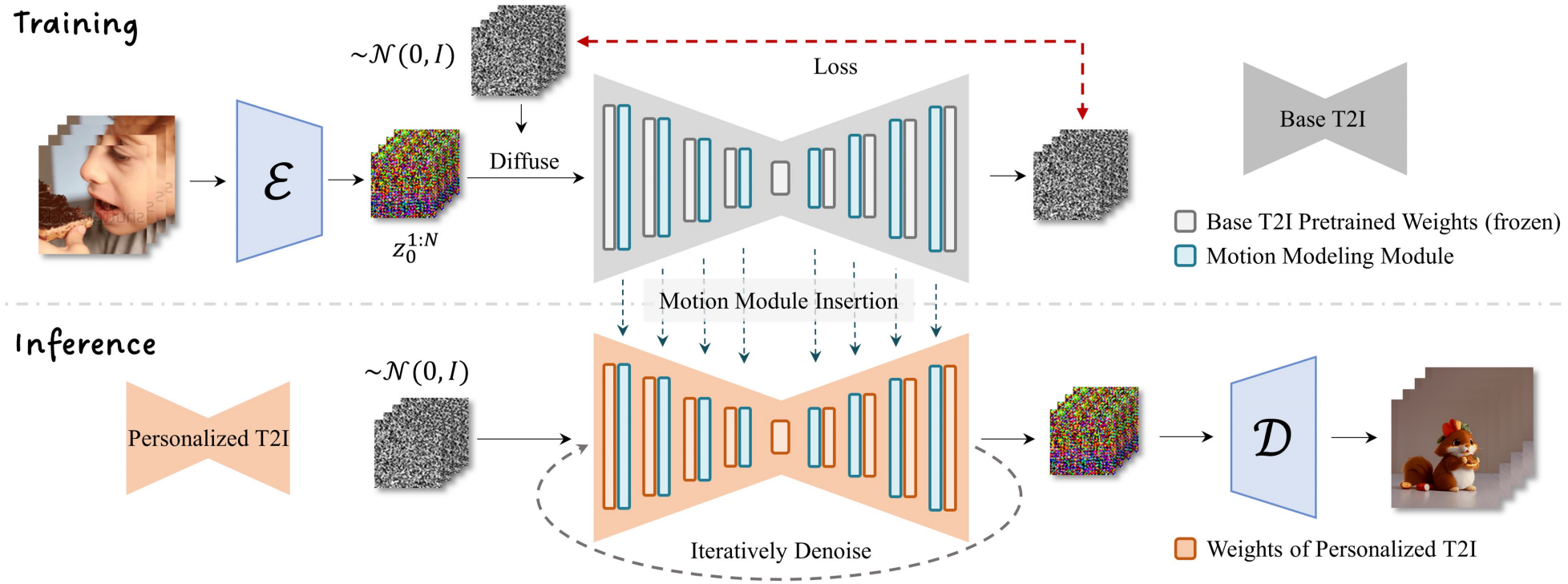
Guo et al., "AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning," arXiv 2023.

AnimateDiff

Transform domain-specific T2I models to T2V models

Methodology

- Train a motion modeling module (some temporal layers) together with frozen base T2I model
- Plug it into a domain-specific T2I model during inference



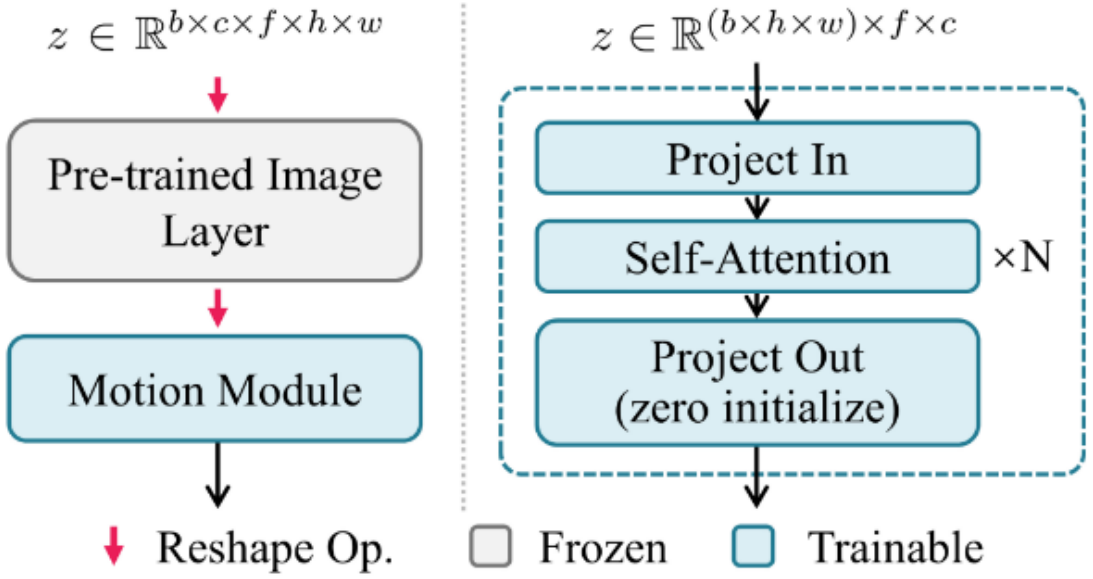
Guo et al., "AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning," arXiv 2023.

AnimateDiff

Transform domain-specific T2I models to T2V models

Methodology

- Train a motion modeling module (some temporal layers) together with frozen base T2I model
- Plug it into a domain-specific T2I model during inference



- Train on WebVid-10M, resized at 256x256 (experiments show can generalize to higher res.)

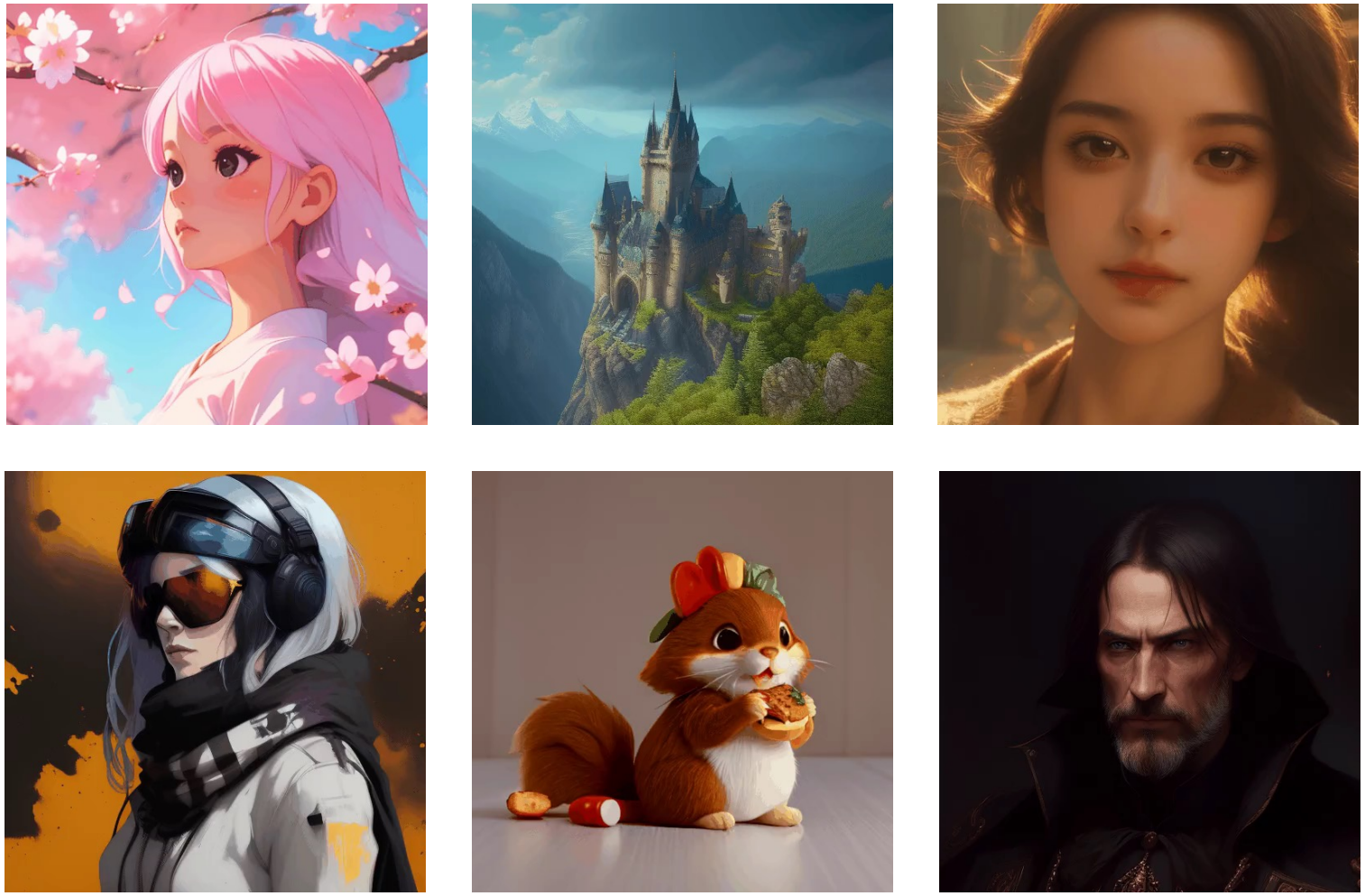
Guo et al., "AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning," arXiv 2023.

AnimateDiff

Transform domain-specific T2I models to T2V models

Model Name	Domain	Type
Counterfeit	Anime	DreamBooth
ToonYou	2D Cartoon	DreamBooth
RCNZ Cartoon	3D Cartoon	DreamBooth
Lyriel	Stylistic	DreamBooth
InkStyle	Stylistic	LoRA
GHIBLI Background	Stylistic	LoRA
majicMIX Realistic	Realistic	DreamBooth
Realistic Vision	Realistic	DreamBooth
FilmVelvia	Realistic	LoRA
TUSUN	Concept	LoRA

Table 1. Personalized models used for evaluation. We chose several representative personalized models contributed by artists from CivitAI [4] for our evaluation, covering a wide domain range from 2D animation to realistic photography.



Text2Video-Zero

Use Stable Diffusion to generate videos without any finetuning

Motivation: How to use Stable Diffusion for video generation without finetuning?

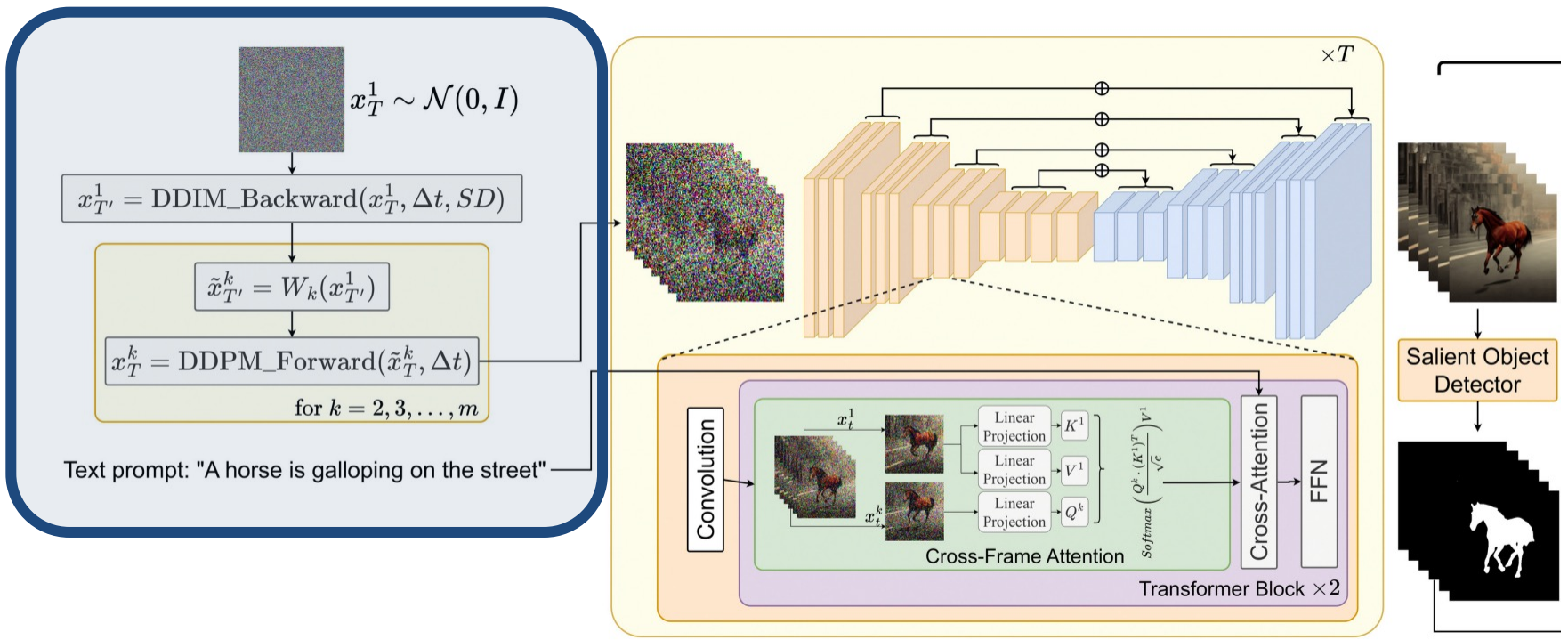
- Start from noises of similar pattern
- Make intermediate features of different frames to be similar

Khachatryan et al., “Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators,” arXiv 2023.

Text2Video-Zero

Use Stable Diffusion to generate videos without any finetuning

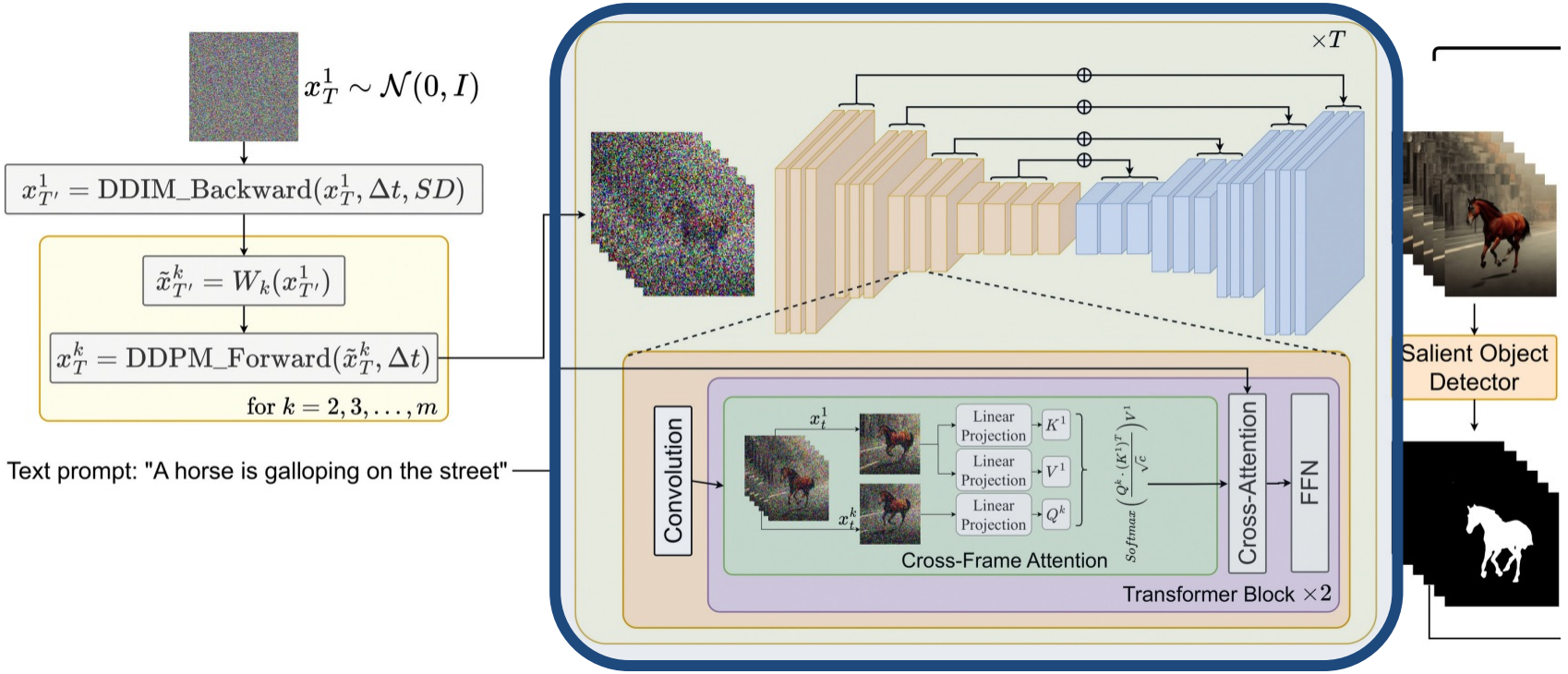
- Start from noises of similar pattern: given the first frame's noise, define a global scene motion, used to translate the first frame's noise to generate similar initial noise for other frames



Text2Video-Zero

Use Stable Diffusion to generate videos without any finetuning

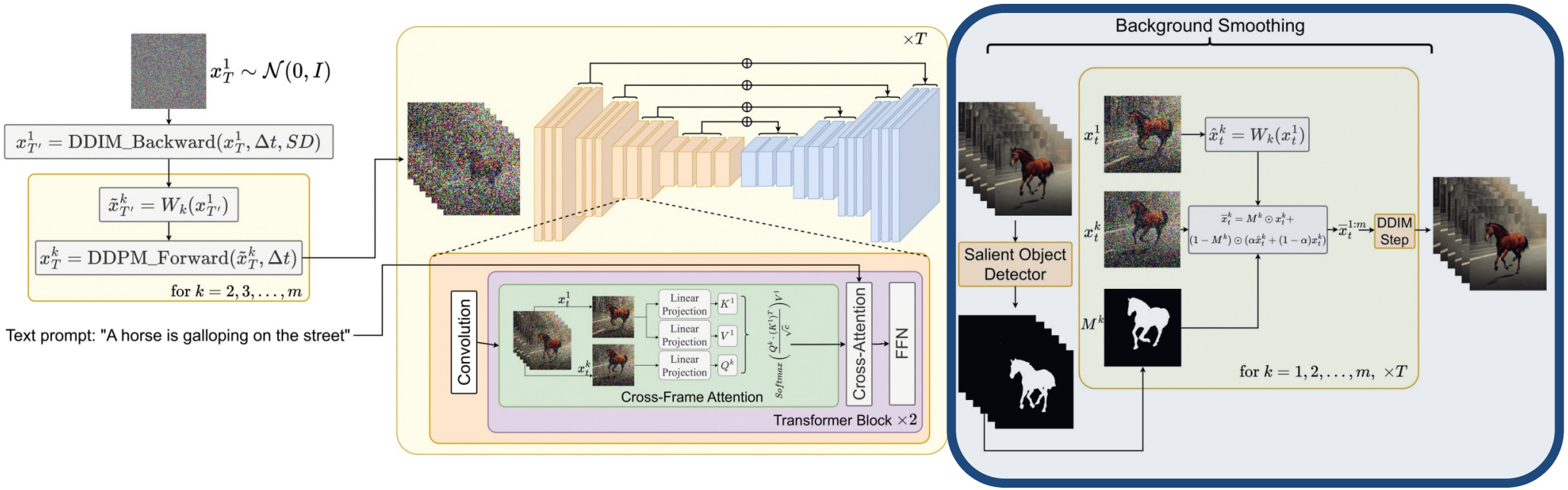
- Make intermediate features of different frames to be similar: always use K and V from the first frame in self-attention



Text2Video-Zero

Use Stable Diffusion to generate videos without any finetuning

- Optional background smoothing: regenerate the background, average with the first frame



Text2Video-Zero

Use Image Stable Diffusion to generate videos without any finetuning

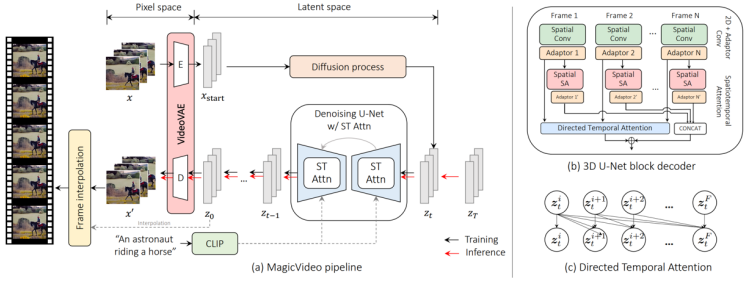
"A cat is running on the grass"

"A panda is playing guitar on times square"

"A man is running in the snow"

"An astronaut is skiing down the hill"

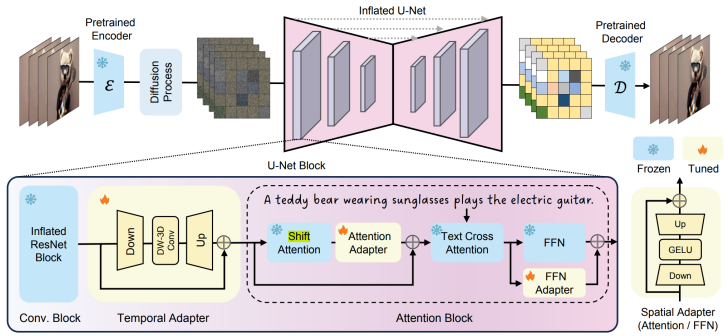
Training Efficient Techniques: More Works



MagicVideo (Zhou et al.)

Insert causal attention to Stable Diffusion for better temporal coherence

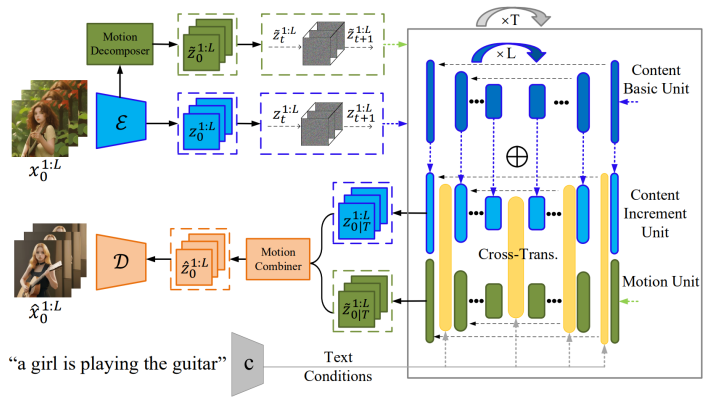
"MagicVideo: Efficient Video Generation With Latent Diffusion Models," arXiv 2022.



Simple Diffusion Adapter (Xing et al.)

Insert lightweight adapters to T2I models, shift latents, and finetune adapters on videos

"SimDA: Simple Diffusion Adapter for Efficient Video Generation," arXiv 2023.



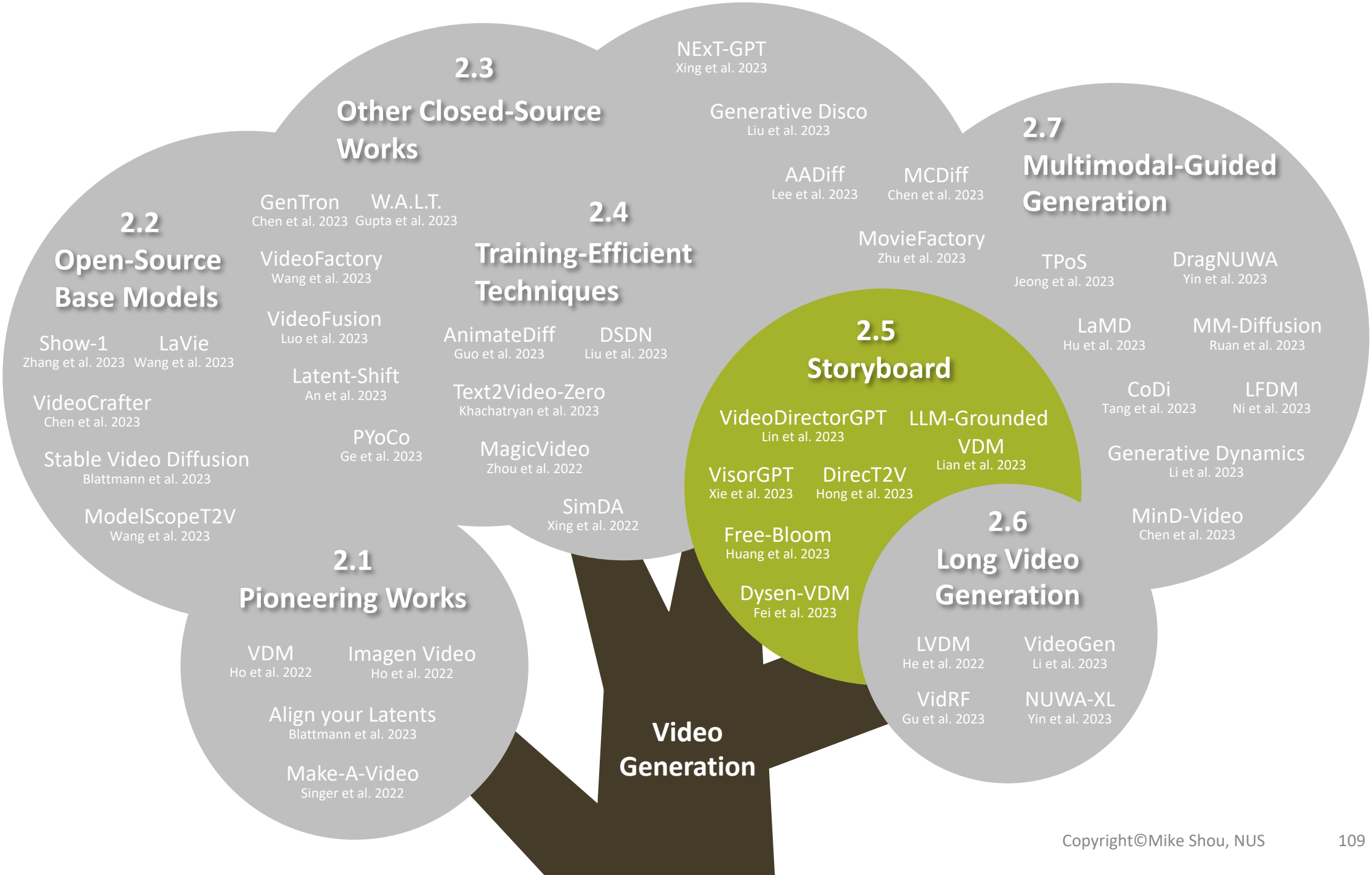
Dual-Stream Diffusion Net (Liu et al.)

Leverage multiple T2I networks for T2V

"Dual-Stream Diffusion Net for Text-to-Video Generation," arXiv 2023.

2 Video Generation

2.5 Storyboard



What is a storyboard?

Human can imagine what does the scene look like “roughly”

“Two men stand in the airport waiting room and stare at the airplane thru window”

What is in your mind now?

Storyboard image from deviantart.com.

What is a storyboard?

A concept in film production

- Rough sketches/drawings with notes
- Example: *Inception* by Christopher Nolan

Storyboard image from deviantart.com.

What is a storyboard?

A concept in film production

- How to generate such a storyboard?
- As humans, over the years, we have acquired such “visual prior” about object location, object shape, relation, etc.
- Can LLM model such visual prior?

Storyboard image from deviantart.com.

VisorGPT

Can we model such visual prior with LLM

Object Locations Object Shape Relations among various objects …

1. Object bounding boxes
2. Human keypoints
3. Semantic masks
4. …

How to discretize into tokens?

person, table, person, person; [xmin $377 ymin $250 ymax $406 ymax $288] [xmin $388 ymin $258 ymax $413 ymax $286] [xmin $271 ymin $129 ymax $395 ymax $370] [xmin $287 ymin $228 ymax $377 ymax $399] …

a sequence of **discrete tokens**

Represented by **continuous** coordinates

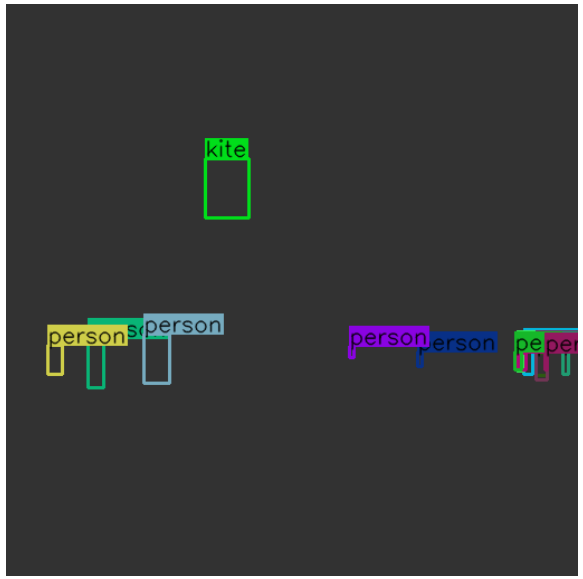
Xie et al., "VisorGPT: Learning Visual Prior via Generative Pre-Training," NeurIPS 2023.

VisorGPT

Prompt design

Table 1: Candidate choices of prompt template.

Annotation type	box; keypoint; mask
Data type	object centric; multiple instances
Size	small; medium; large
#Instances	1; 2; 3; · · ·
#Keypoints	14; 18
Category name	cup; person; dog; · · ·



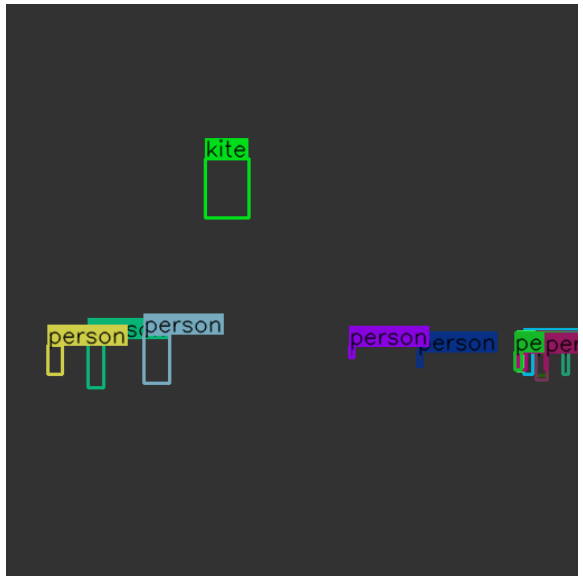
[CLS] box ; multiple instances ; small ; 14 ; 0 ; person , kite , person , person , person , person , person , person , person , person , person , person , person , person ; [xmin $ 73 ymin $ 299 ymax $ 87 ymax $ 343] [xmin $ 178 ymin $ 138 ymax $ 217 ymax $ 191] [xmin $ 463 ymin $ 308 ymax $ 471 ymax $ 331] [xmin $ 457 ymin $ 310 ymax $ 463 ymax $ 328] [xmin $ 474 ymin $ 312 ymax $ 484 ymax $ 332] [xmin $ 474 ymin $ 310 ymax $ 484 ymax $ 336] [xmin $ 458 ymin $ 312 ymax $ 465 ymax $ 327] [xmin $ 123 ymin $ 295 ymax $ 146 ymax $ 339] [xmin $ 37 ymin $ 305 ymax $ 50 ymax $ 331] [xmin $ 455 ymin $ 311 ymax $ 461 ymax $ 327] [xmin $ 368 ymin $ 311 ymax $ 372 ymax $ 324] [xmin $ 498 ymin $ 311 ymax $ 503 ymax $ 331] [xmin $ 482 ymin $ 312 ymax $ 484 ymax $ 327] [xmin $ 307 ymin $ 306 ymax $ 311 ymax $ 316] [SEP]

VisorGPT

Prompt design

Table 1: Candidate choices of prompt template.

Annotation type	box; keypoint; mask
Data type	object centric; multiple instances
Size	small; medium; large
#Instances	1; 2; 3; · · ·
#Keypoints	14; 18
Category name	cup; person; dog; · · ·



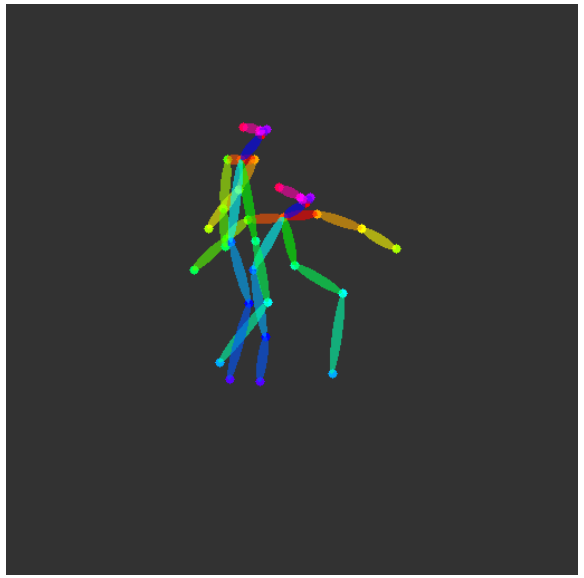
[CLS] box ; multiple instances ; small ; 14 ; 0 ; person , kite , person , person , person , person , person , person , person , person , person , person , person , person ; [xmin $ 73 ymin $ 299 ymax $ 87 ymax $ 343] [xmin $ 178 ymin $ 138 ymax $ 217 ymax $ 191] [xmin $ 463 ymin $ 308 ymax $ 471 ymax $ 331] [xmin $ 457 ymin $ 310 ymax $ 463 ymax $ 328] [xmin $ 474 ymin $ 312 ymax $ 484 ymax $ 332] [xmin $ 474 ymin $ 310 ymax $ 484 ymax $ 336] [xmin $ 458 ymin $ 312 ymax $ 465 ymax $ 327] [xmin $ 123 ymin $ 295 ymax $ 146 ymax $ 339] [xmin $ 37 ymin $ 305 ymax $ 50 ymax $ 331] [xmin $ 455 ymin $ 311 ymax $ 461 ymax $ 327] [xmin $ 368 ymin $ 311 ymax $ 372 ymax $ 324] [xmin $ 498 ymin $ 311 ymax $ 503 ymax $ 331] [xmin $ 482 ymin $ 312 ymax $ 484 ymax $ 327] [xmin $ 307 ymin $ 306 ymax $ 311 ymax $ 316] [SEP]

VisorGPT

Prompt design

Table 1: Candidate choices of prompt template.

Annotation type	box; keypoint; mask
Data type	object centric; multiple instances
Size	small; medium; large
#Instances	1; 2; 3; · · ·
#Keypoints	14; 18
Category name	cup; person; dog; · · ·



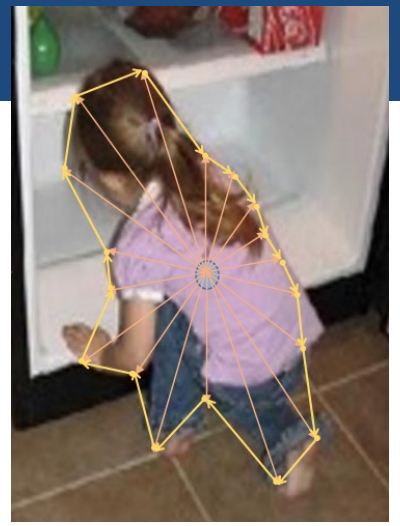
[CLS] key point ; multiple instances ; large ; 2 ; 18 ; person , person ; [a $ 268 $ 178 b $ 248 $ 191 c $ 278 $ 188 d $ 318 $ 201 e $ 349 $ 219 f $ 216 $ 193 g $ 197 $ 210 h $ 168 $ 238 i $ 258 $ 234 j $ 301 $ 259 k $ 292 $ 331 l $ 221 $ 238 m $ 232 $ 298 n $ 227 $ 338 o $ 272 $ 173 p $ 263 $ 173 q $ 0 $ 0 r $ 244 $ 164] [a $ 228 $ 117 b $ 210 $ 139 c $ 222 $ 139 d $ 208 $ 166 e $ 181 $ 201 f $ 198 $ 139 g $ 194 $ 183 h $ 196 $ 217 i $ 223 $ 212 j $ 234 $ 267 k $ 191 $ 321 l $ 201 $ 213 m $ 217 $ 268 n $ 200 $ 336 o $ 233 $ 112 p $ 227 $ 113 q $ 0 $ 0 r $ 212 $ 110] [SEP]

VisorGPT

Prompt design

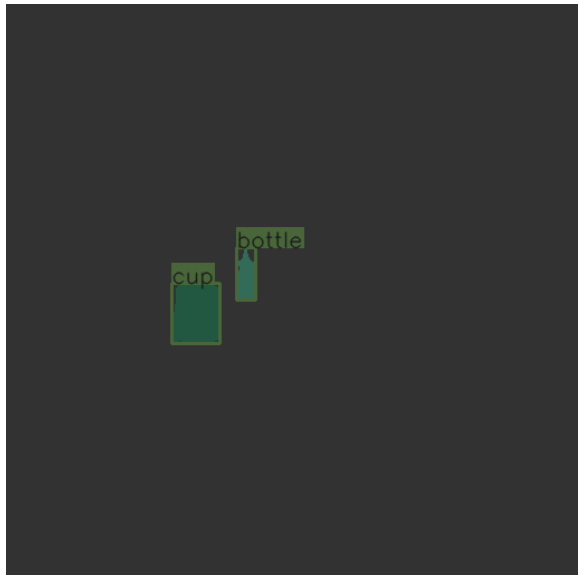
Table 1: Candidate choices of prompt template.

Annotation type	box; keypoint; mask
Data type	object centric; multiple instances
Size	small; medium; large
#Instances	1; 2; 3; · · ·
#Keypoints	14; 18
Category name	cup; person; dog; · · ·



"PolarMask: Single Shot Instance Segmentation with Polar Representation". CVPR 2020.

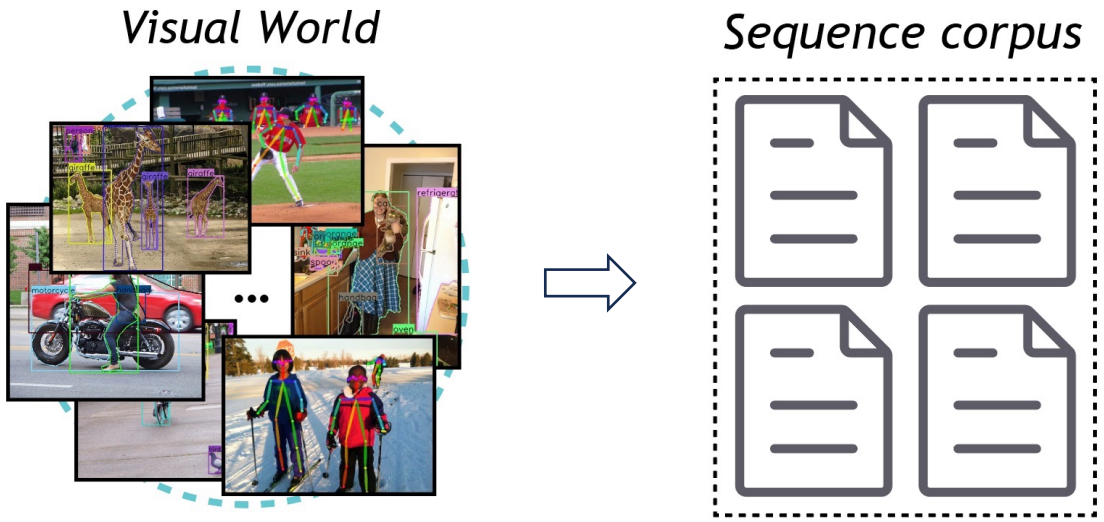
TL;DR. Represent a mask by 36 coordinates



[CLS] mask ; multiple instances ; medium ; 2 ; 0 ; bottle , cup ; [m0 272 248 m1 272 249 m2 272 251 m3 271 253 m4 271 256 m5 270 260 m6 266 272 m7 270 275 m8 276 271 m9 280 270 m10 284 269 m11 287 267 m12 291 266 m13 292 262 m14 292 258 m15 291 254 m16 291 252 m17 291 250 m18 291 248 m19 291 246 m20 291 244 m21 291 242 m22 292 238 m23 292 234 m24 291 230 m25 289 226 m26 285 221 m27 280 220 m28 275 219 m29 274 230 m30 273 236 m31 272 239 m32 272 241 m33 272 243 m34 272 245 m35 272 247] [m0 289 249 m1 289 250 m2 289 251 m3 289 253 m4 289 254 m5 289 256 m6 290 258 m7 290 263 m8 293 265 m9 295 265 m10 298 265 m11 301 265 m12 303 …] [SEP]

VisorGPT

Modeling Visual Prior via Generative Pre-Training



Pretext Objective. After processing the visual locations $\boldsymbol{x}$ as textual sequences $\mathbf{t}$ in § 3.2, we tokenize each sequence by byte-pair encoding (BPE) algorithm [37] to obtain a sequence with n tokens $\mathbf{u} = \{u_1, u_2, \cdots, u_n\}$ such that a standard language modeling objective can be directly employed to learn visual prior by maximizing the following likelihood:

$$\mathcal{L} = \sum_i \log p(u_i | u_{i-k}, \cdots, u_{i-1}; \Theta), \tag{1}$$

where k is the size of context window, and $p(\cdot|\cdot)$ indicates the conditional probability which is modeled by the neural network Θ. Stochastic gradient descent is used to train the neural network.

Xie et al., "VisorGPT: Learning Visual Prior via Generative Pre-Training," NeurIPS 2023.

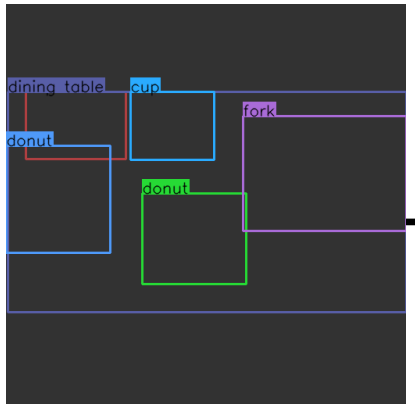
VisorGPT

Sample from the LLM which has learned visual prior

box; multiple instances; large; 6; 0; cup, dining table, donut,

box ; multiple instances ; large ; 6 ; 0 ; cup , dining table , donut , donut , fork , cup ; [xmin 25 ymin 112 xmax 153 ymax 198] [xmin 2 ymin 112 xmax 512 ymax 394] [xmin 174 ymin 242 xmax 307 ymax 358] [xmin 0 ymin 181 xmax 133 ymax 318] [xmin 303 ymin 143 xmax 512 ymax 290] [xmin 159 ymin 112 xmax 266 ymax 199]

Decode



GLIGEN

Xie et al., "VisorGPT: Learning Visual Prior via Generative Pre-Training," NeurIPS 2023.

VisorGPT

Sample from the LLM which has learned visual prior

(d) Object Centric Bounding-Box

box; object centric; large; 1; 0;

box; object centric; large; 1; 0; great white shark; [xmin 95 ymin 66 xmax 510 ymax 310]

(e) Continuous Generation

key point; multiple instances; medium;1; 18;

key point ; multiple instances ; large ; 1 ; 18 ; [person a 244 162 b 240 190 c 279 198 d 320 256 e 321 324 f 200 182 g 157 211 h 122 228 i 240 284 j 264 333 k 224 380 l 187 283 m 186 330 n 183 399 o 252 152 p 234 150 q 0 0 r 218 156]

Add 2 person

key point; multiple instances; large; 3; 18; [person a 244 162 b 240 190 c 279 198 d 320 256 e 321 324 f 200 182 g 157 211 h 122 228 i 240 284 j 264 333 k 224 380 l 187 283 m 186 330 n 183 399 o 252 152 p 234 150 q 0 0 r 218 156] [person ...] [person ...]

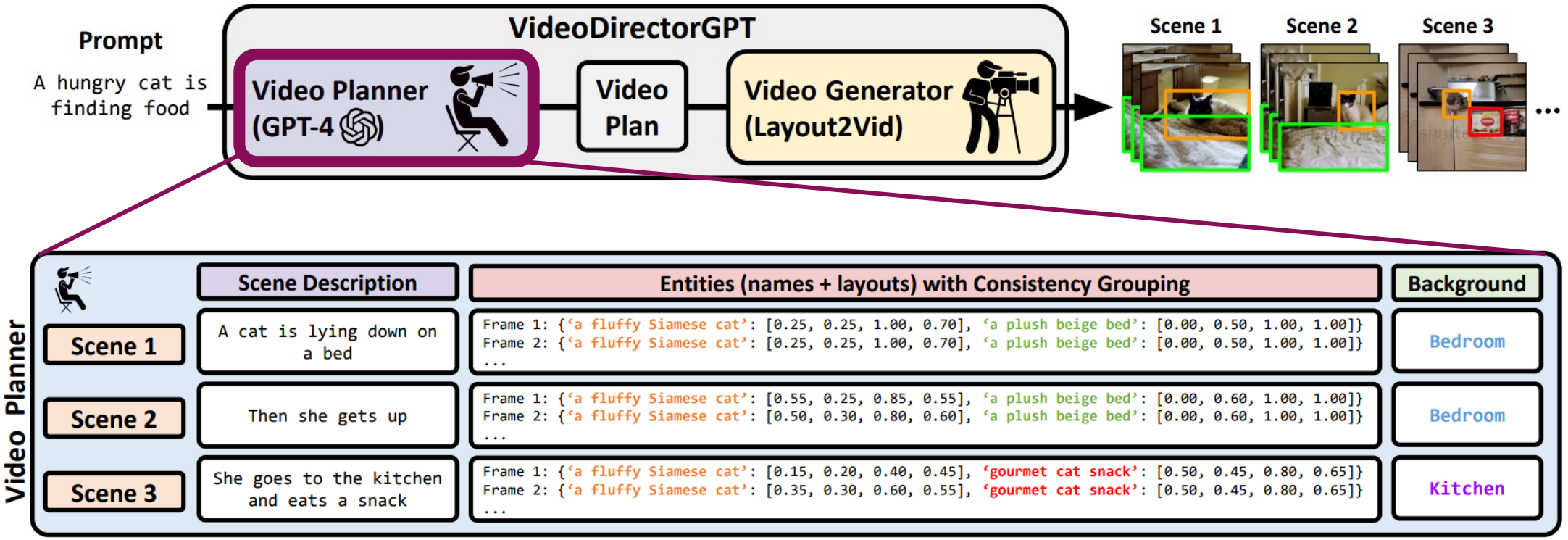
Add 5 person

key point; multiple instances; large; 8; 18; [person a 244 162 b 240 190 c 279 198 d 320 256 e 321 324 f 200 182 g 157 211 h 122 228 i 240 284 j 264 333 k 224 380 l 187 283 m 186 330 n 183 399 o 252 152 p 234 150 q 0 0 r 218 156] [person ...] [person ...] [person ...] ...

Xie et al., "VisorGPT: Learning Visual Prior via Generative Pre-Training," NeurIPS 2023.

VideoDirectorGPT

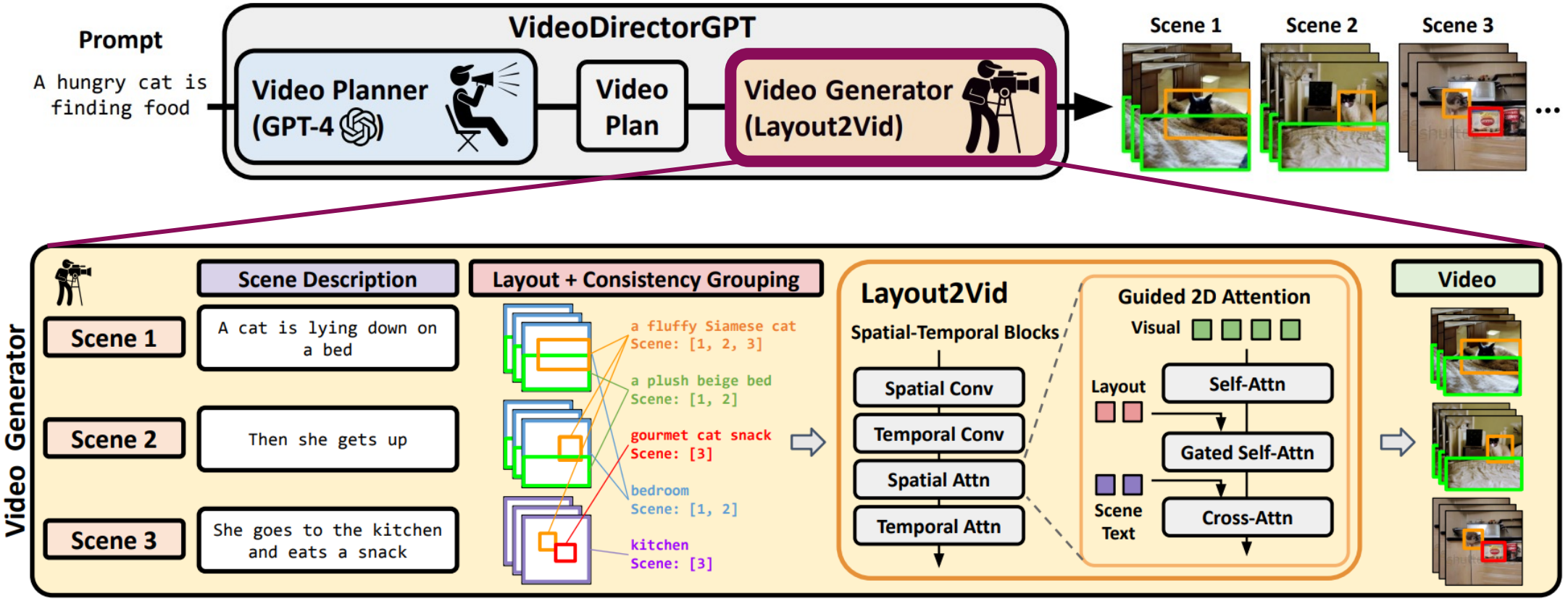
Use storyboard as condition to generate video



Lin et al., "VideoDirectorGPT: Consistent Multi-scene Video Generation via LLM-Guided Planning," arXiv 2023.

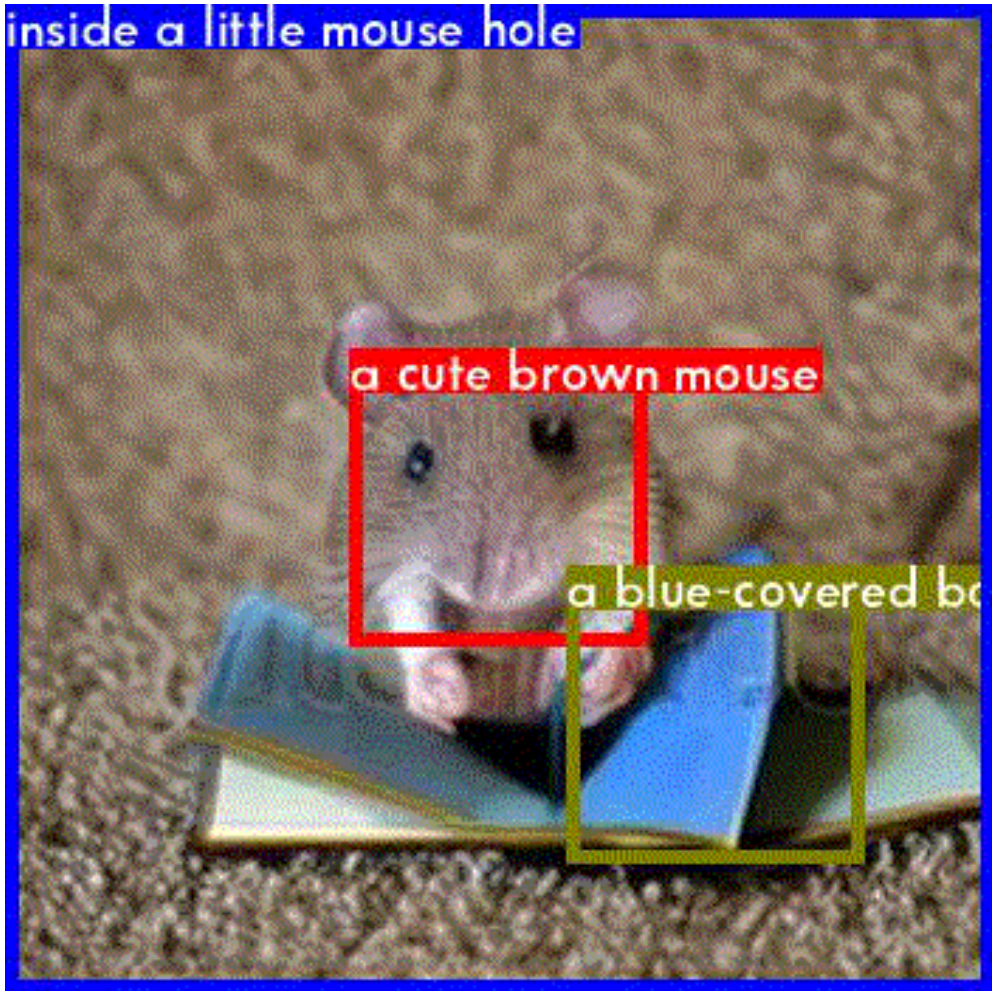
VideoDirectorGPT

Use storyboard as condition to generate video



VideoDirectorGPT

Use storyboard as condition to generate video



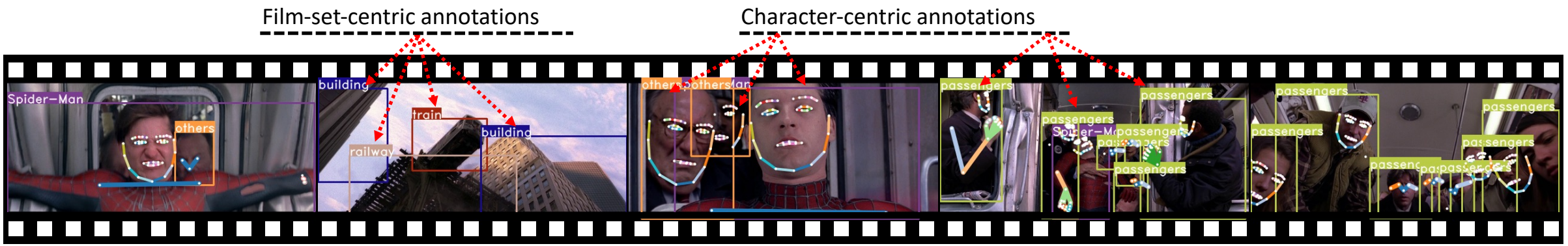
Scene 1: mouse is holding a book and makes a happy face.
Scene 2: he looks happy and talks.
Scene 3: he is pulling petals off the flower.
Scene 4: he is ripping a petal from the flower.
Scene 5: he is holding a flower by his right paw.
Scene 6: one paw pulls the last petal off the flower.
Scene 7: he is smiling and talking while holding a flower on his right paw.

Lin et al., "VideoDirectorGPT: Consistent Multi-scene Video Generation via LLM-Guided Planning," arXiv 2023.

Long-form Video Prior

GPT can be trained to learn better long-form video prior (e.g., object position, relative size, human interaction)

A new dataset - Storyboard20K

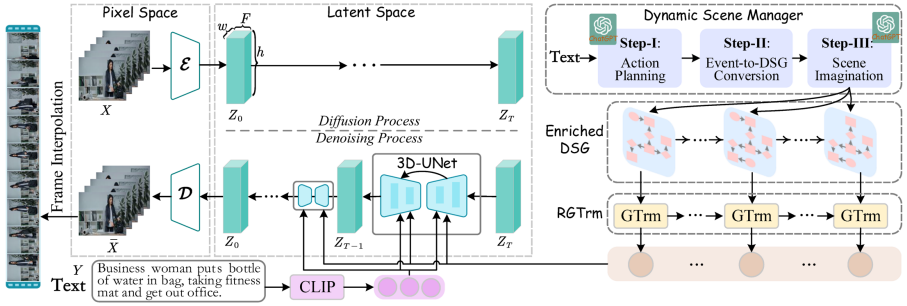


Scripts (shot by shot):
(1) Spider-Man stands as if glued to the front of the train.
(2) Spider-Man continues to hold the webs attached to the buildings behind them.
(3) Spider-Man's eyes close.
(4) The passengers lift the unconscious, unmasked Spider-Man into the carriage.
(5) His face visible, the passengers stare at him.

Summative annotations:
Title: "the heroic sacrifice: spider-man saves the train"
Genre: "action and dramatic rescue"
Emotion: "exhaustion, surprise, admiration"
Scene: "train"
Summary: "spider-man saves a train from disaster but collapses from exhaustion, receiving support and admiration from the grateful passengers"

Xie et al., "Learning Long-form Video Prior via Generative Pre-Training," to be released in 2024.
https://github.com/showlab/Long-form-Video-Prior

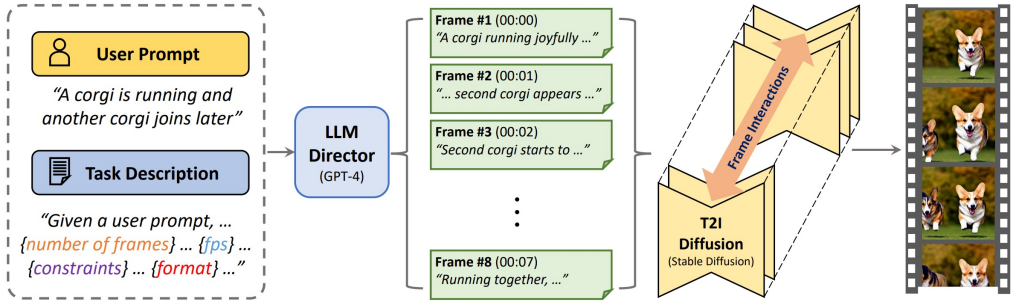
Storyboard: More Works



Dysen-VDM (Fei et al.)

Storyboard through scene graphs

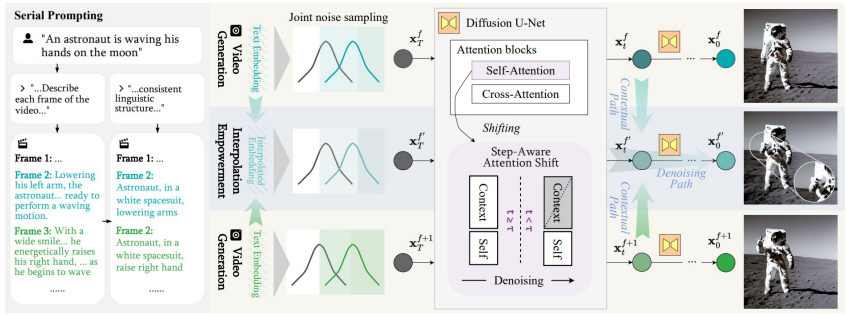
"Empowering Dynamics-aware Text-to-Video Diffusion with Large Language Models," arXiv 2023.



DirectT2V (Hong et al.)

Storyboard through bounding boxes

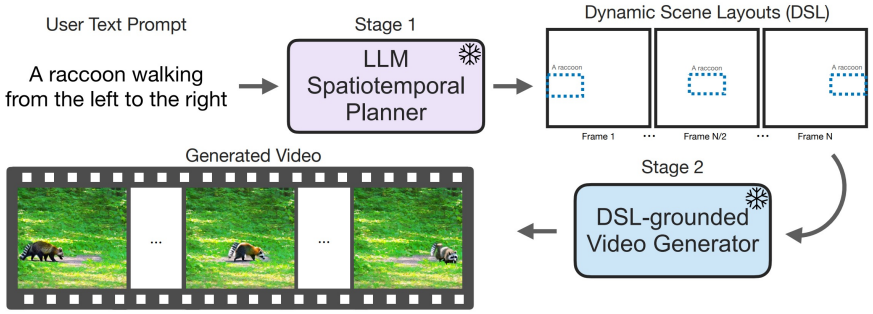
"Large Language Models are Frame-level Directors for Zero-shot Text-to-Video Generation," arXiv 2023.



Free-Bloom (Huang et al.)

Storyboard through detailed text prompts

"Free-Bloom: Zero-Shot Text-to-Video Generator with LLM Director and LDM Animator," NeurIPS 2023.



LLM-Grounded Video Diffusion Models

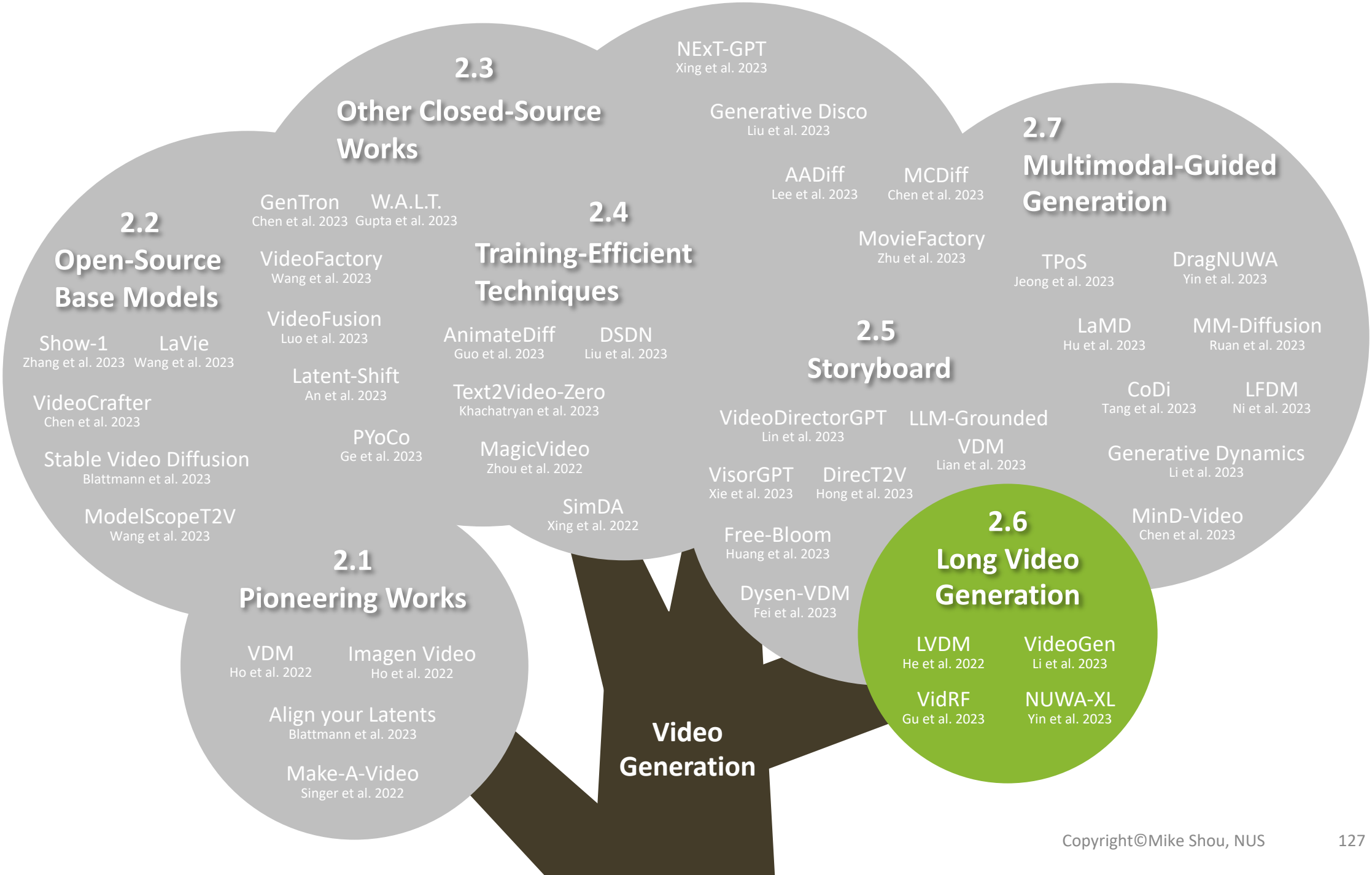
(Lian et al.)

Storyboard through foreground bounding boxes

"LLM-grounded Video Diffusion Models," arXiv 2023.

2 Video Generation

2.6 Long video generation



NUWA-XL

Recursive interpolations for generating very long videos

Method Proposed

- A “diffusion over diffusion” architecture for very long video generation

Key Idea

- Key idea: coarse-to-fine hierarchical generation

Other Highlights

- Trained on very long videos (3376 frames)
- Enables parallel inference
- Built FlintstonesHD: a new dataset for long video generation, contains 166 episodes with an average of 38000 frames of 1440 × 1080 resolution

Yin et al., “NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation,” arXiv 2023.

Recursive interpolations for generating very long videos

Generation Pipeline

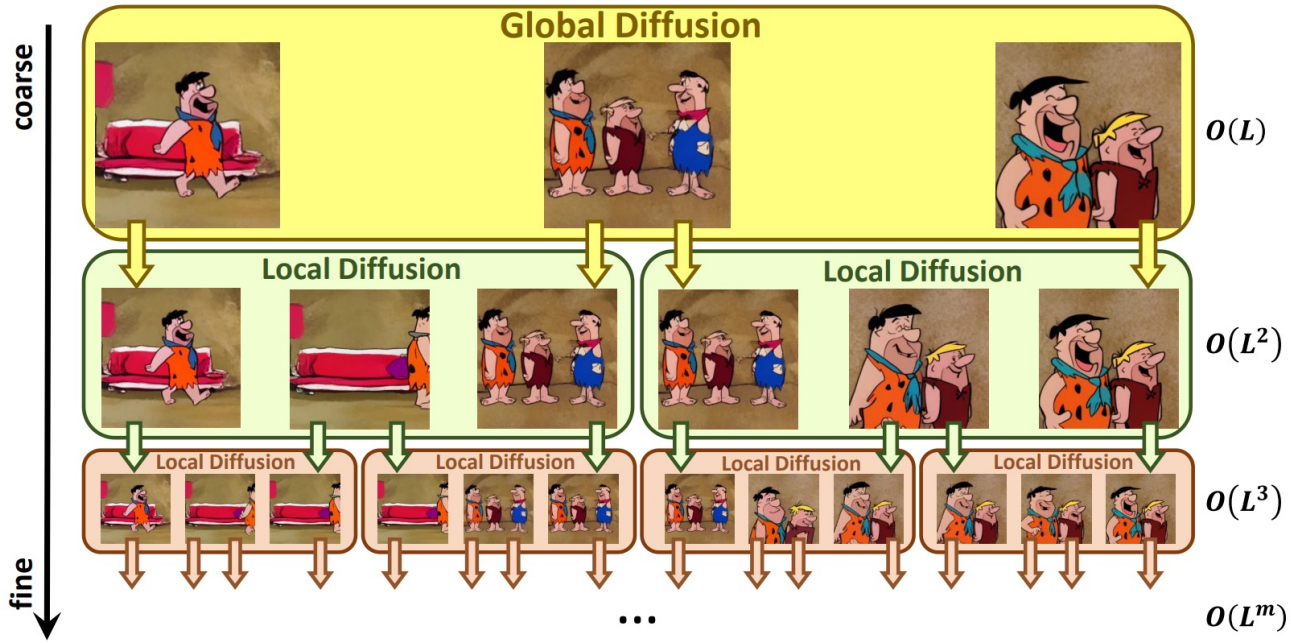
- Storyboard through multiple text prompts

NUWA-XL

Recursive interpolations for generating very long videos

Generation Pipeline

- Storyboard through multiple text prompts
- Global diffusion model: L text prompts → L keyframes
- Local diffusion model: 2 text prompts + 2 keyframes → L keyframes

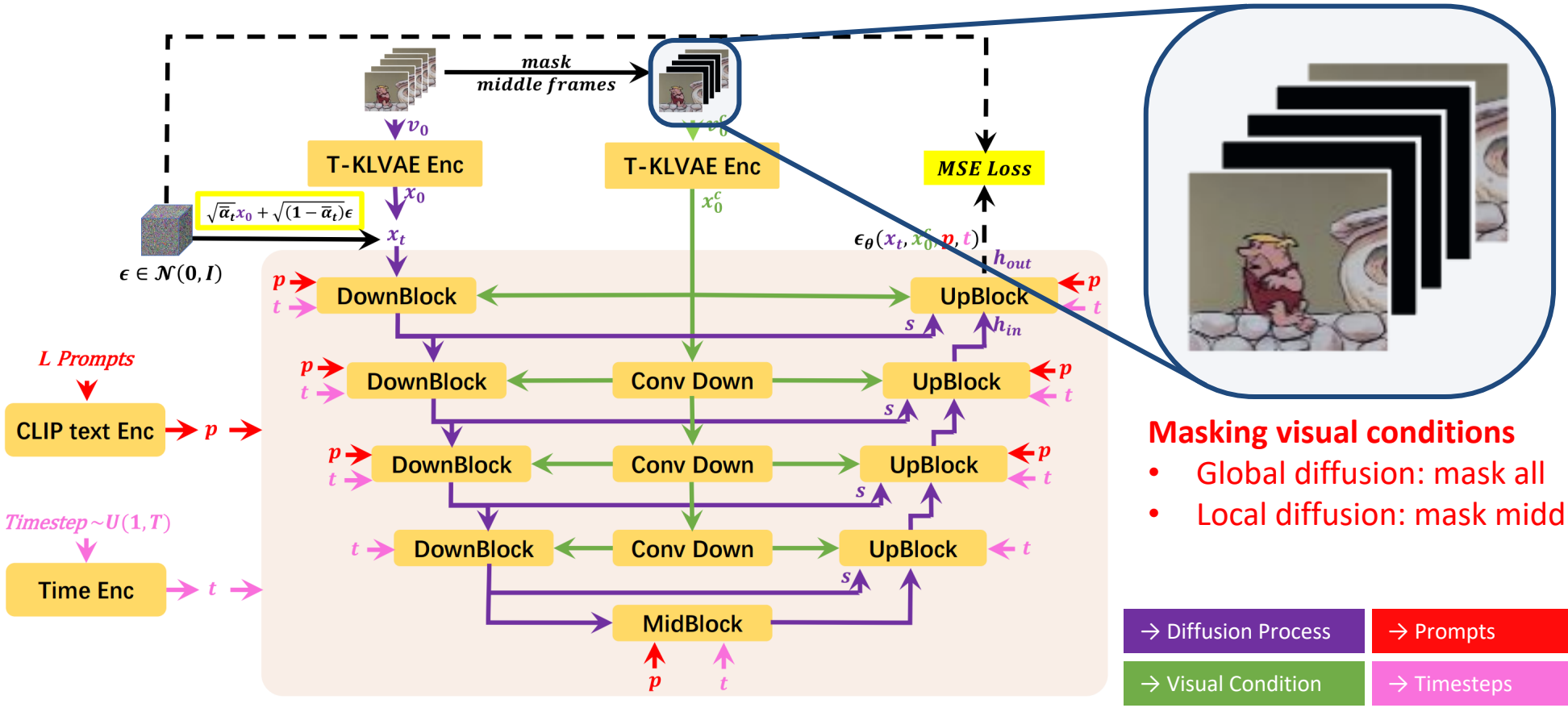


Yin et al., "NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation," arXiv 2023.

Recursive interpolations for generating very long videos

Mask Temporal Diffusion (MTD)

- A basic diffusion model for global & local diffusion models



Masking visual conditions

- Global diffusion: mask all
- Local diffusion: mask middle frames

→ Diffusion Process | → Prompts | → Visual Condition | → Timesteps

NUWA-XL

Recursive interpolations for generating very long videos

0:00.000

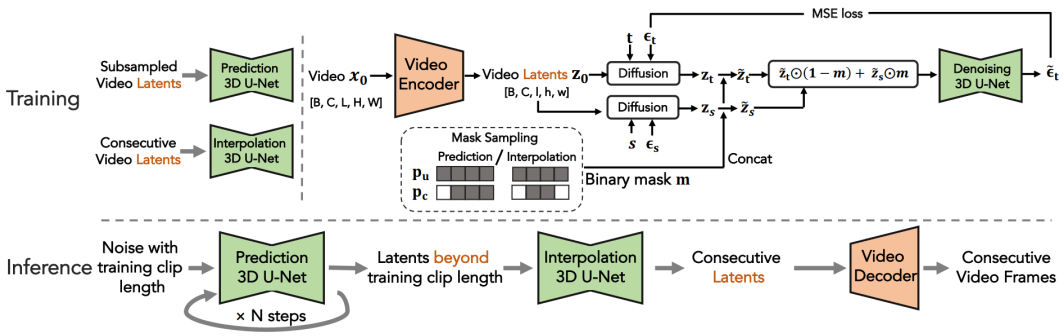
5:00.000 onwards

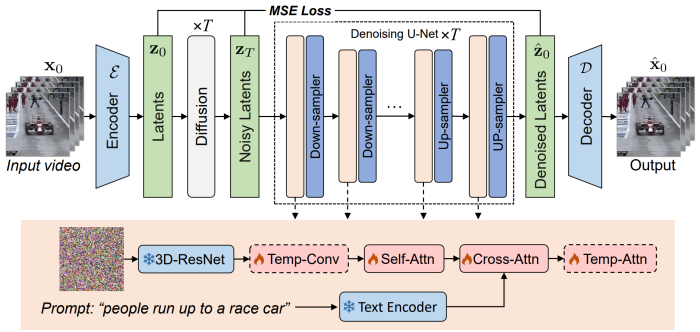
10:00.000 onwards

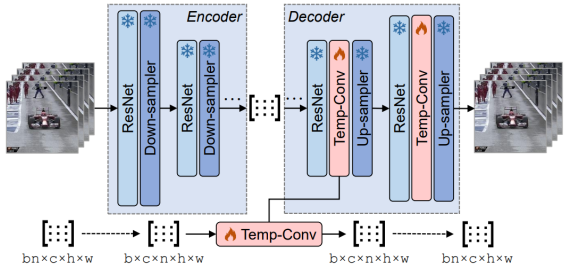
Total 11:15.221

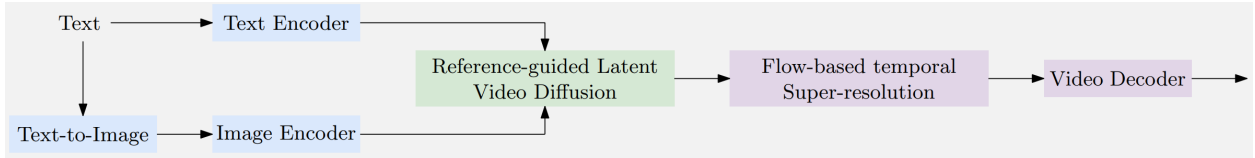
Yin et al., “NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation,” arXiv 2023.

Long Video Generation: More Works









Latent Video Diffusion Models for High-Fidelity Long Video Generation

(He et al.)

Generate long videos via autoregressive generation & interpolation

"Latent Video Diffusion Models for High-Fidelity Long Video Generation," arXiv 2022.

VidRD (Gu et al.)

Autoregressive long video generation

"Reuse and Diffuse: Iterative Denoising for Text-to-Video Generation," arXiv 2023.

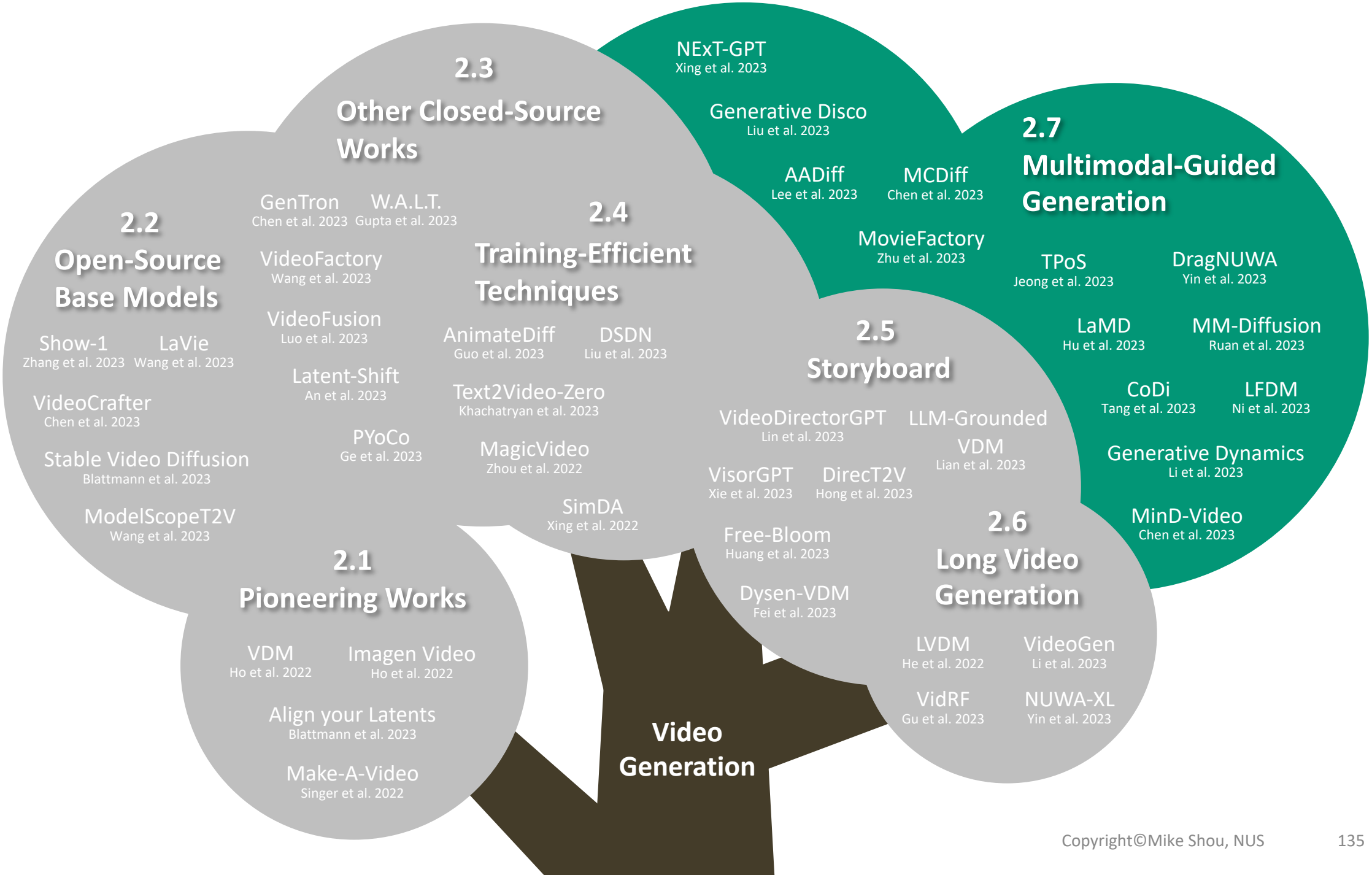
VideoGen (Li et al.)

Cascaded pipeline for long video generation

"VideoGen: A Reference-Guided Latent Diffusion Approach for High Definition Text-to-Video Generation," arXiv 2023.

2 Video Generation

2.7 Multimodal-guided generation



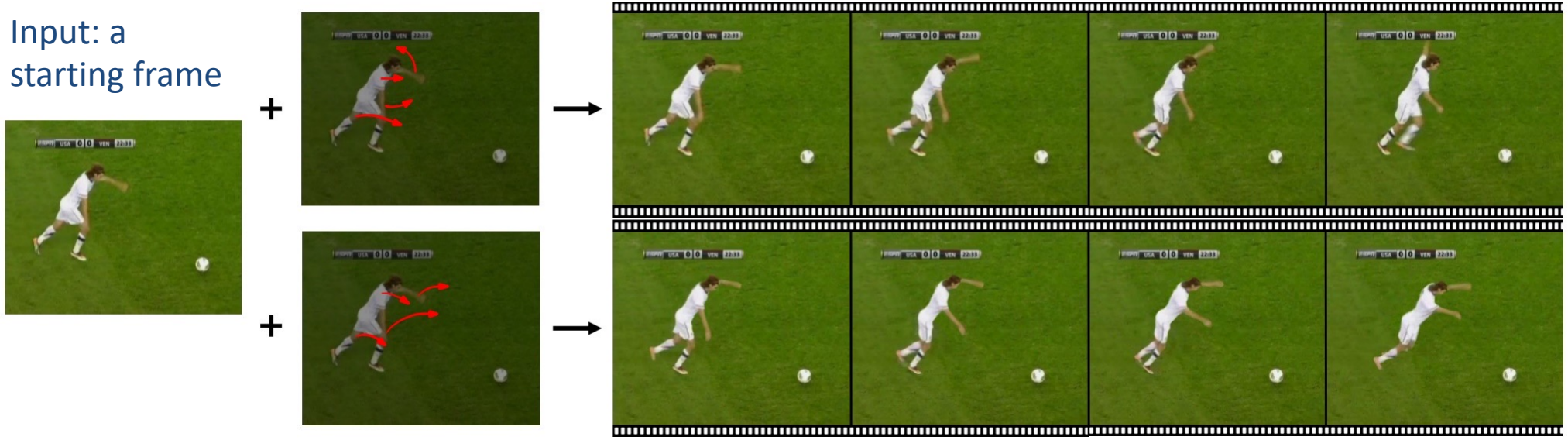
MCDiff

Motion-guided video generation

Input: a starting frame

Input: many "flow" frames

Output: consecutive video frames



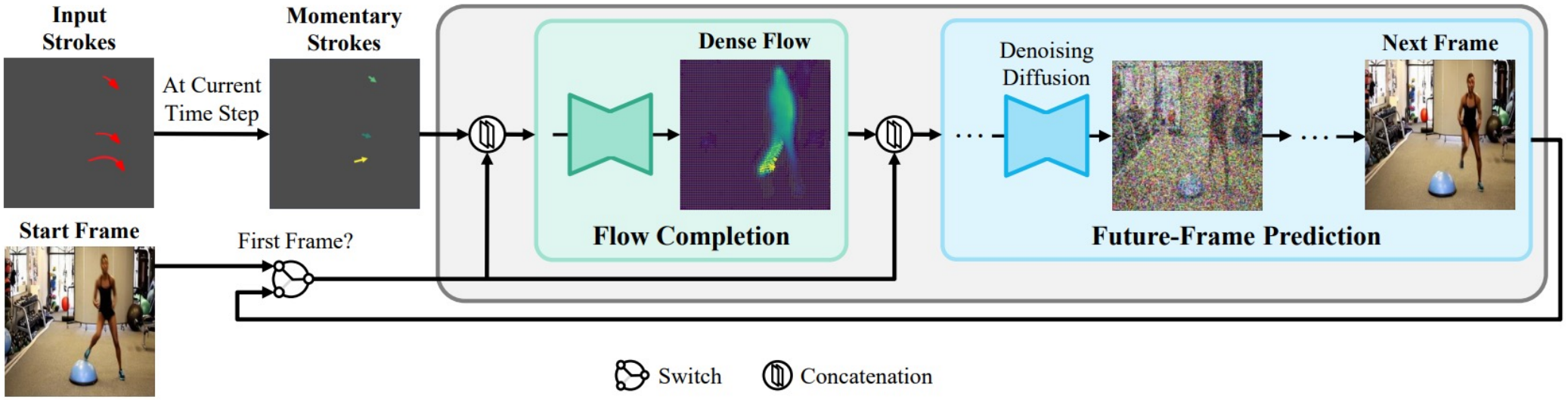
Flow = pixel-wise motion vectors between consecutive video frames

Chen et al., "Motion-Conditioned Diffusion Model for Controllable Video Synthesis," arXiv 2023.

MCDiff

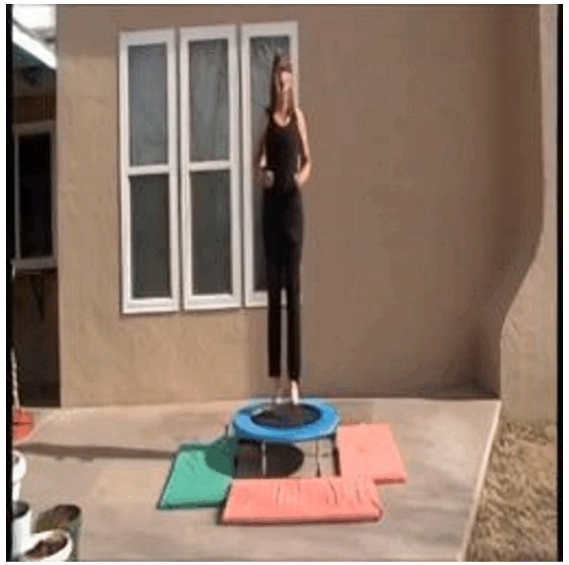
Motion-guided video generation

- Two-stage autoregressive generation



MCDiff

Motion-guided video generation



Start Frame

Start Frame

Start Frame

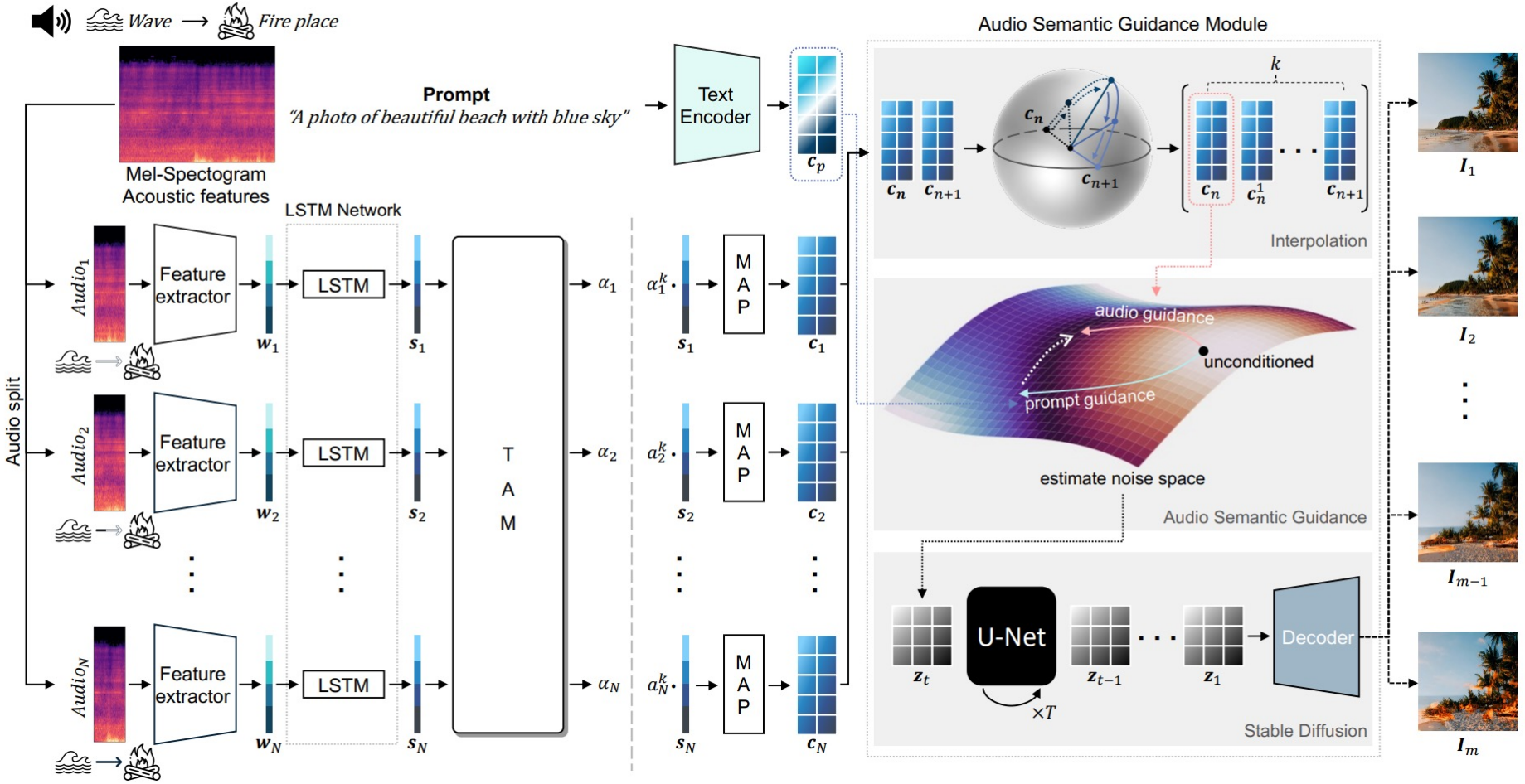
Start Frame

Chen et al., “Motion-Conditioned Diffusion Model for Controllable Video Synthesis,” arXiv 2023.

The Power of Sound (TPoS)

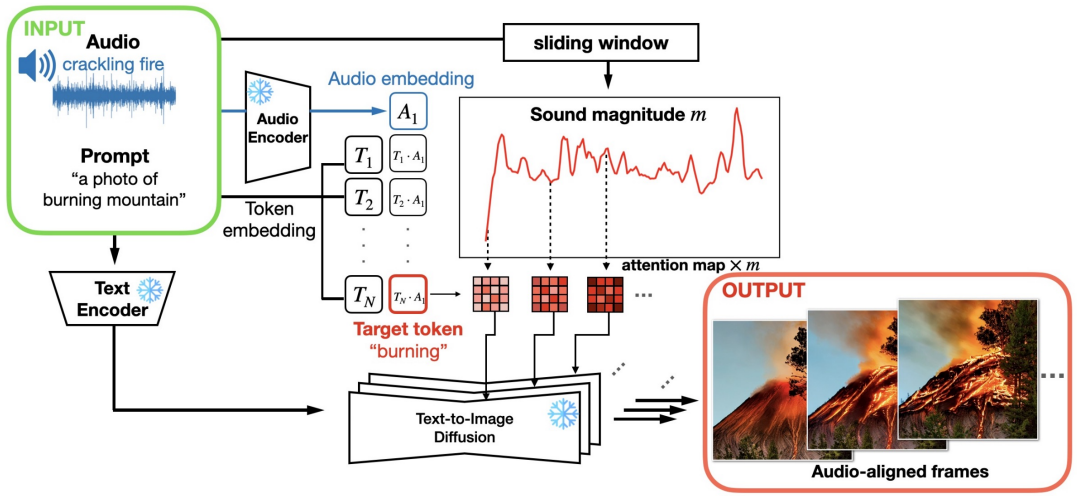
Sound- and text-guided video generation

- Input/output: a text prompt + an audio segment → a video



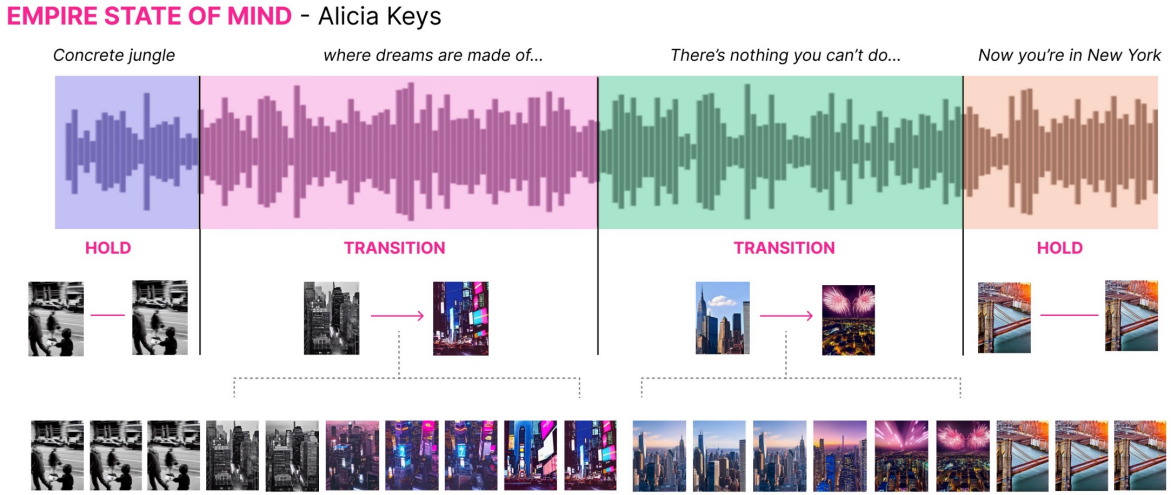
Jeong et al., "The Power of Sound (TPoS): Audio Reactive Video Generation with Stable Diffusion," ICCV 2023.

Sound-Guided Video Generation: More Works



AADiff (Lee et al.)

"AADiff: Audio-Aligned Video Synthesis with Text-to-Image Diffusion," CVPRW 2023.

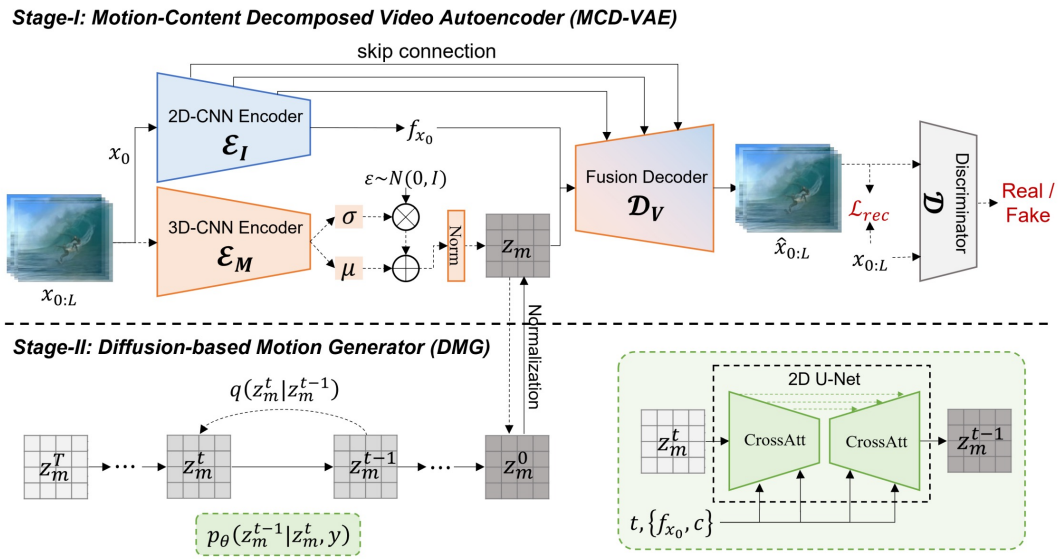


Generative Disco (Liu et al.)

"Generative Disco: Text-to-Video Generation for Music Visualization," arXiv 2023.

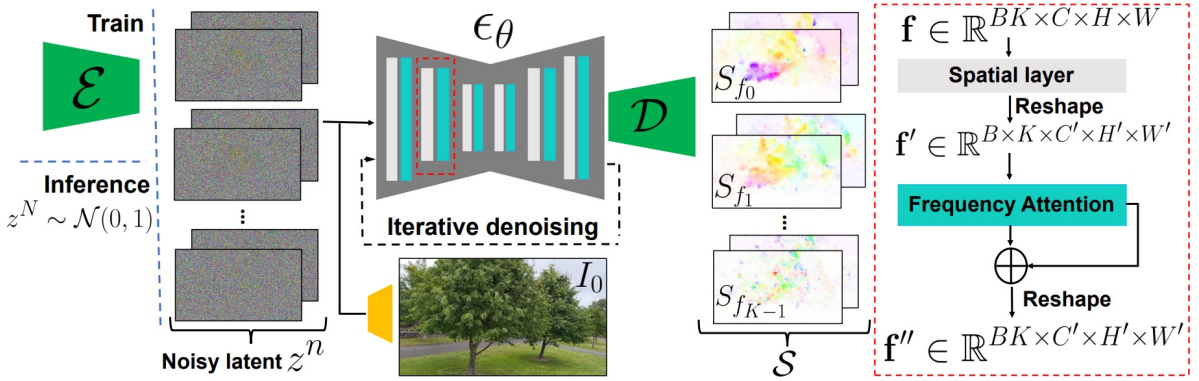
Xing et al., "A Survey on Video Diffusion Models," arXiv 2023.

Image-Guided Video Generation: More Works



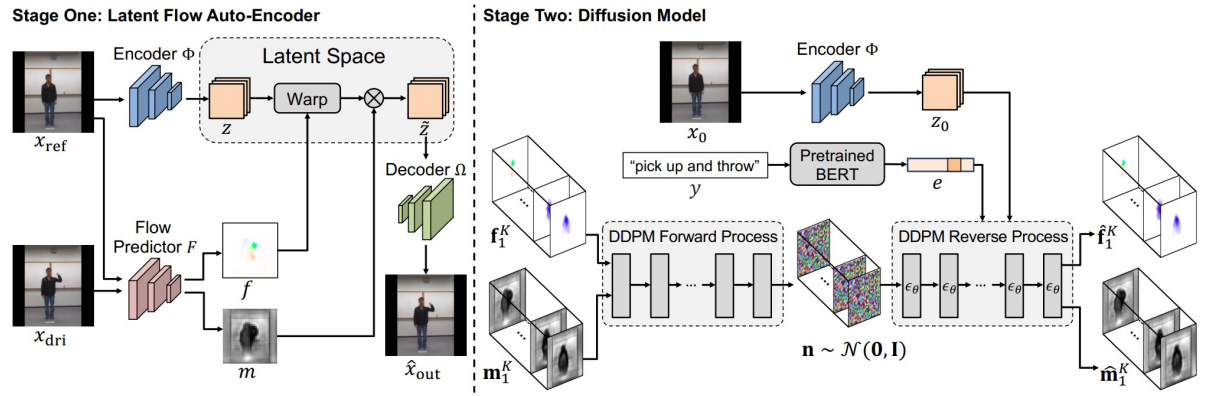
Generative Image Dynamics (Li et al.)

"Generative Image Dynamics," arXiv 2023.



LaMD (Hu et al.)

"LaMD: Latent Motion Diffusion for Video Generation," arXiv 2023.



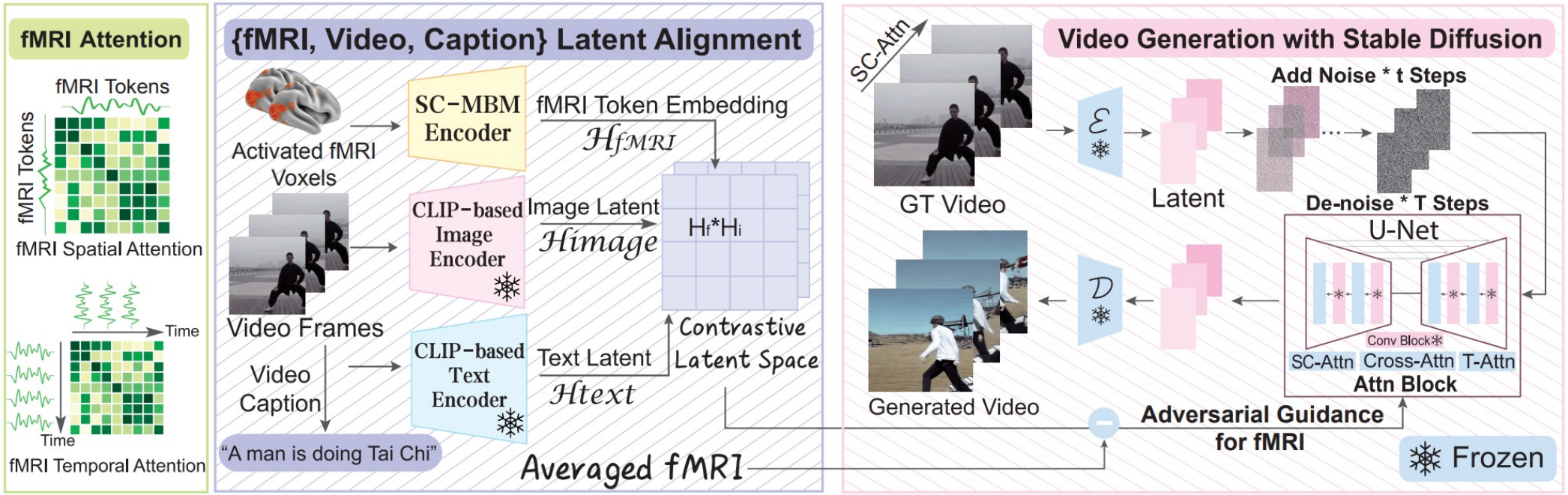
LFDM (Ni et al.)

"Conditional Image-to-Video Generation with Latent Flow Diffusion Models," CVPR 2023.

Cinematic Mindscapes

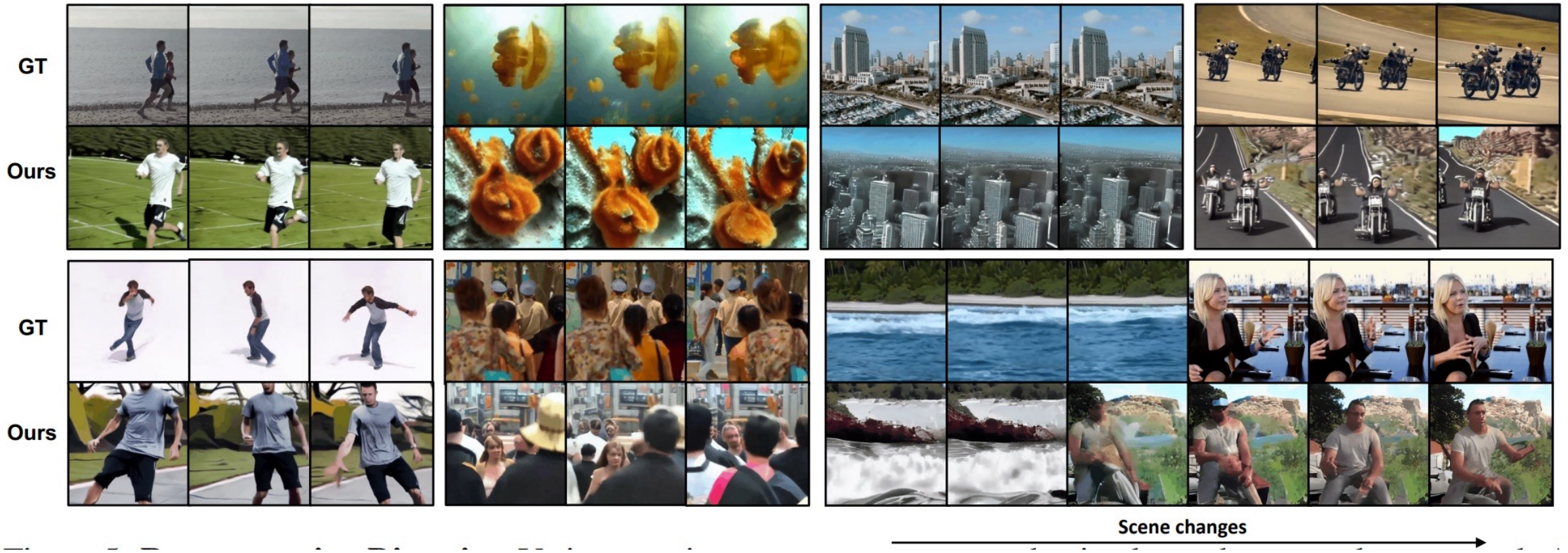
Brain activity-guided video generation

- Task: human vision reconstruction via fMRI signal-guided video generation



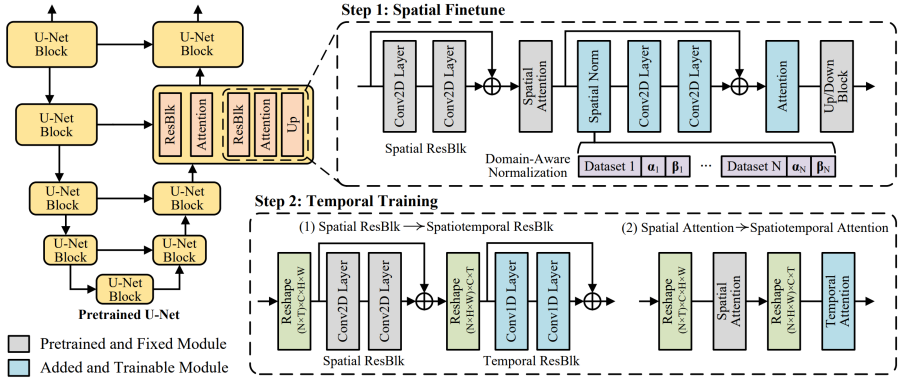
Cinematic Mindscapes

Brain activity-guided video generation



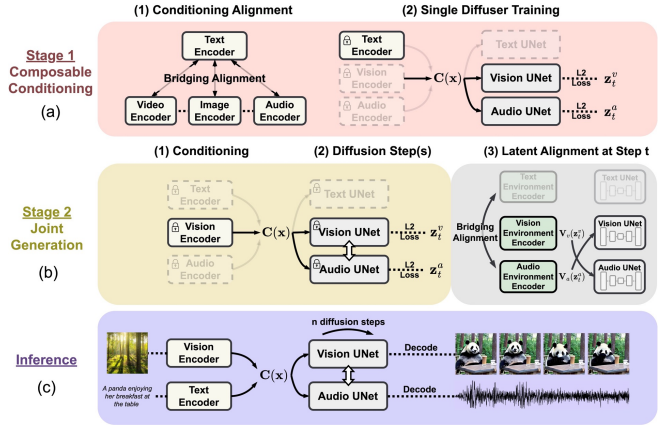
Chen et al., “Cinematic Mindscapes: High-quality Video Reconstruction from Brain Activity,” arXiv 2023.

Multimodal-Guided Video Generation: More Works



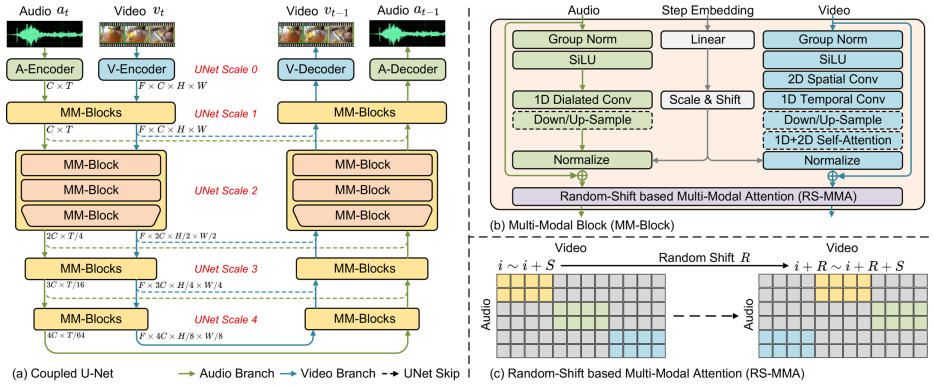
MovieFactory (Zhu et al.)

"MovieFactory: Automatic Movie Creation from Text using Large Generative Models for Language and Images," arXiv 2023.



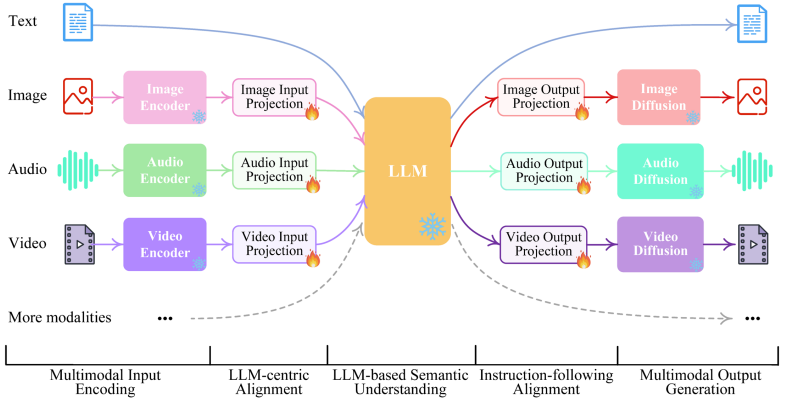
CoDi (Tang et al.)

"Any-to-Any Generation via Composable Diffusion," NeurIPS 2023.



MM-Diffusion (Ruan et al.)

"MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation," CVPR 2023.

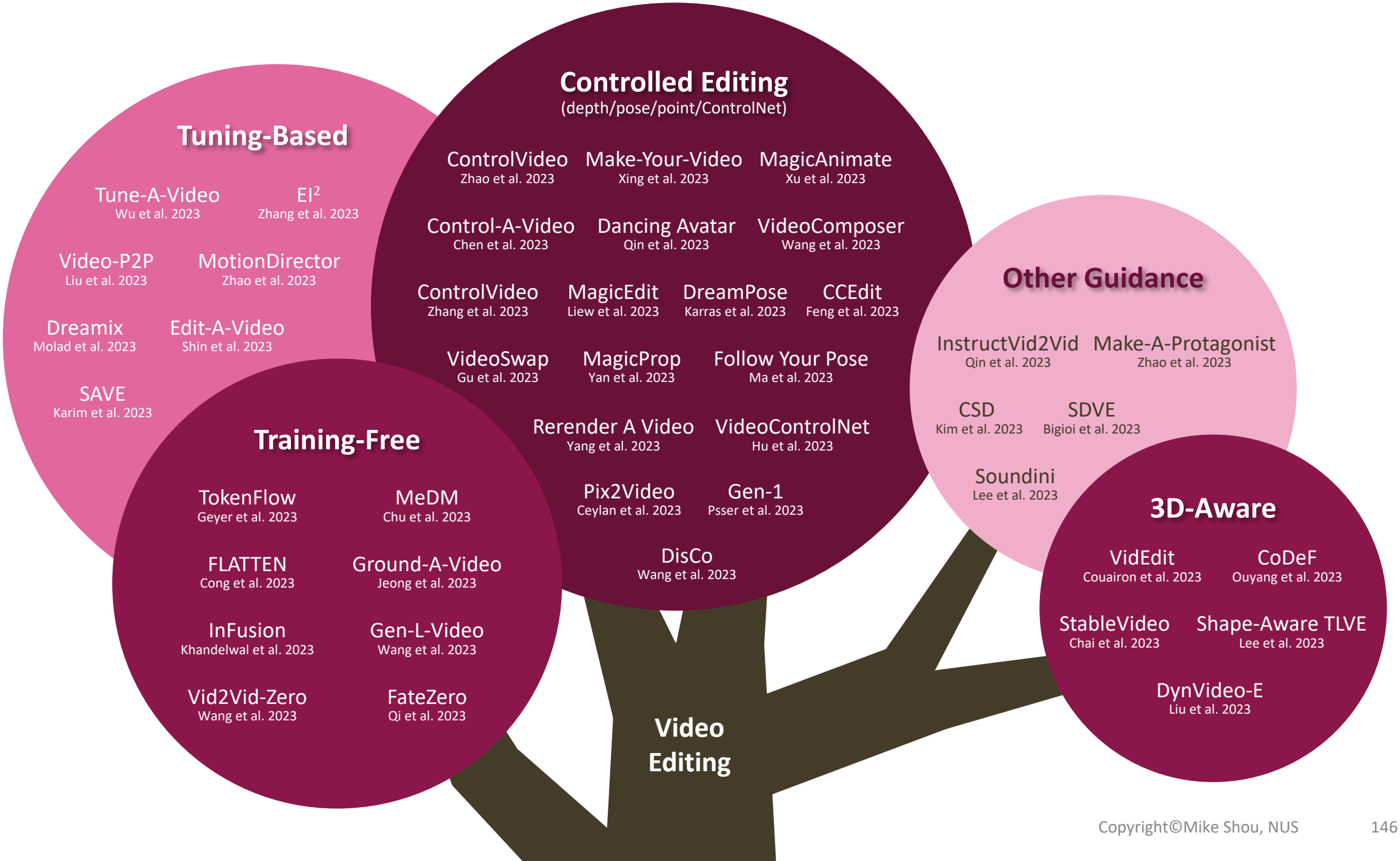


NExT-GPT (Wu et al.)

"NExT-GPT: Any-to-Any Multimodal LLM," arXiv 2023.

Xing et al., "A Survey on Video Diffusion Models," arXiv 2023.

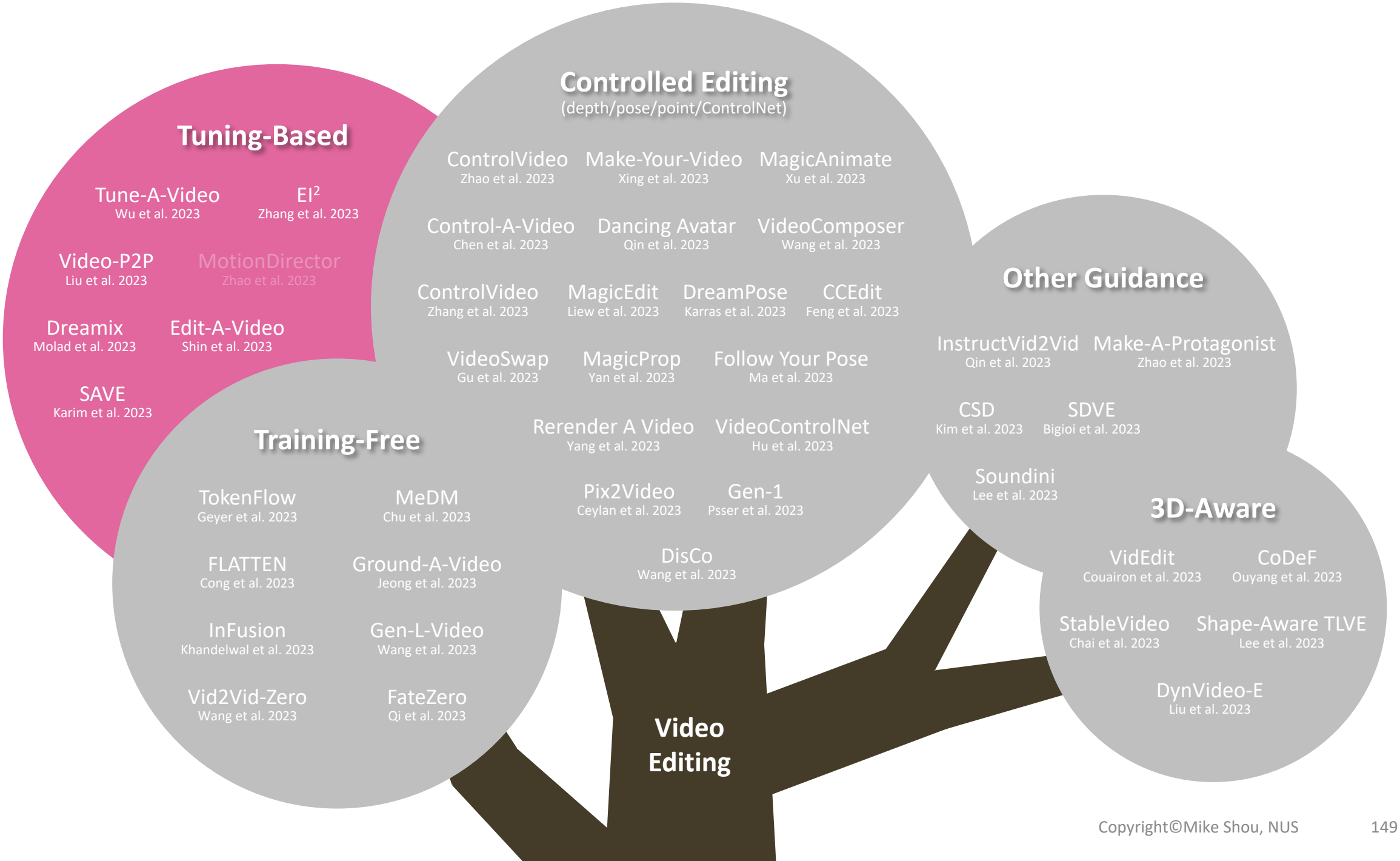
3 Video Editing



3 Video Editing

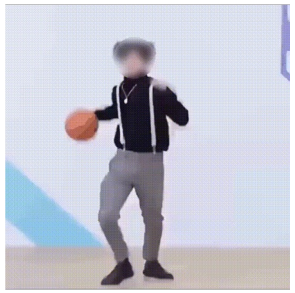
3.1 Tuning-based

One-Shot Tuned



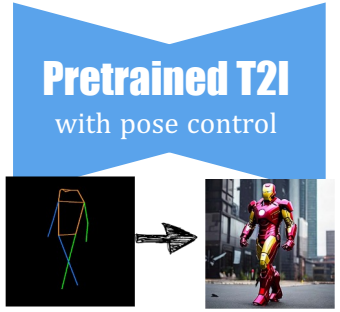
Tune-A-Video

One-shot tuning of T2I models for T2V generation/editing



"A man is dribbling a basketball"

"James Bond" "astronaut"



"Iron Man"

Tune-A-Video

"James Bond is ... on the beach"

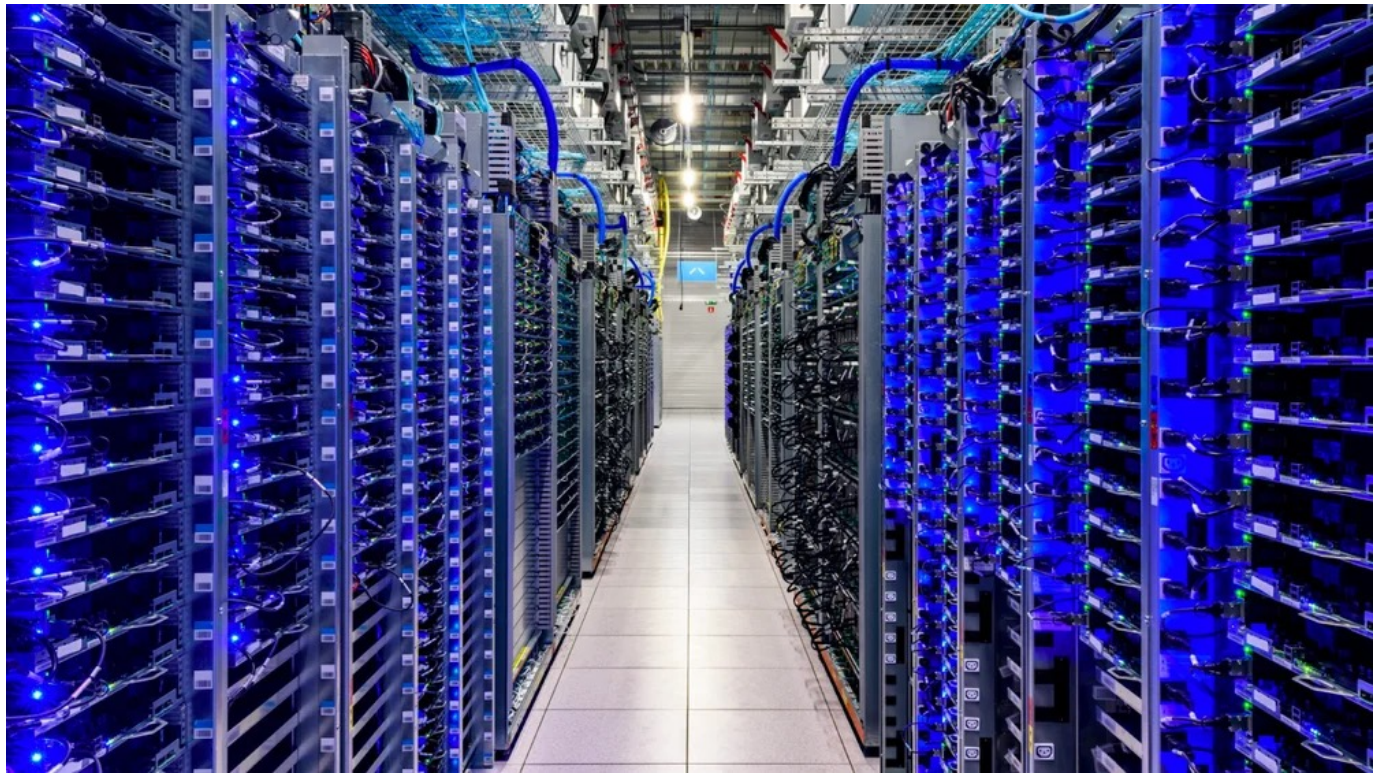
"An astronaut is ..., cartoon style"

"A prince with a crown is ..., modern disney style"

"Iron Man is dancing on the beach"

Computing Resources

Big Company



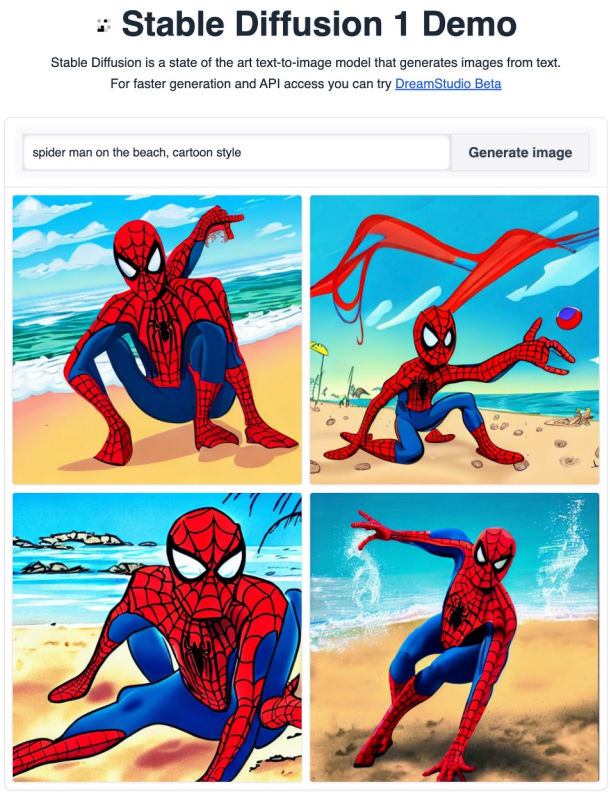
University Lab

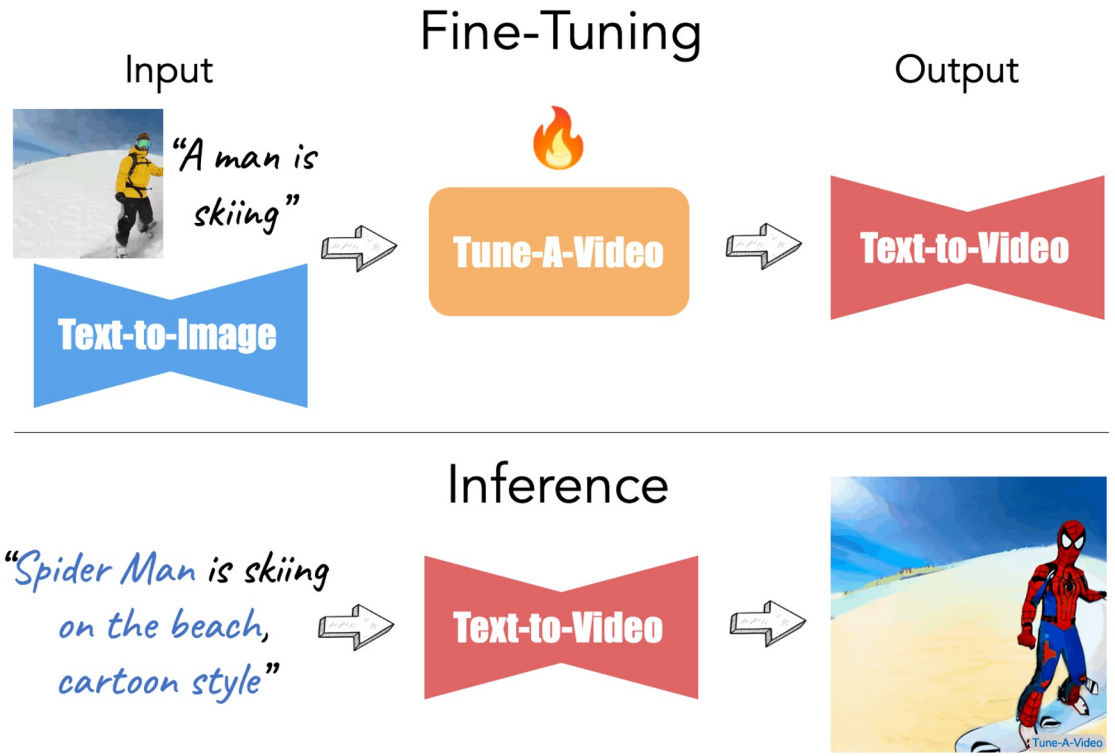
Tune-A-Video

One-shot tuning of T2I models for T2V generation/editing

https://github.com/showlab/Tune-A-Video

Motivation: appearance from pretrained T2I models, dynamics from a reference video



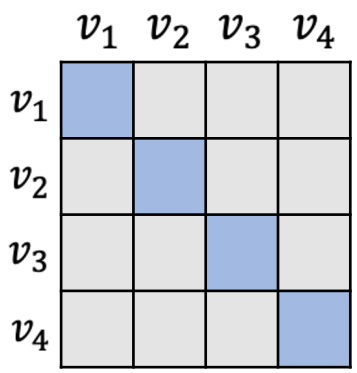


Wu et al., "Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation," ICCV 2023.

Tune-A-Video

One-shot tuning of T2I models for T2V generation/editing

Obs #1: Still images that accurately represent the verb terms

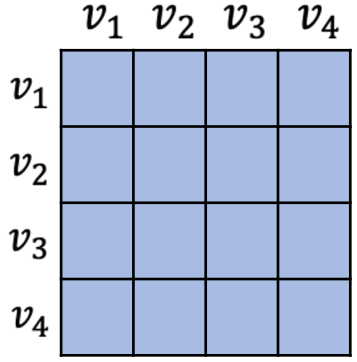


spatial self-attention

"A man is running on the beach"

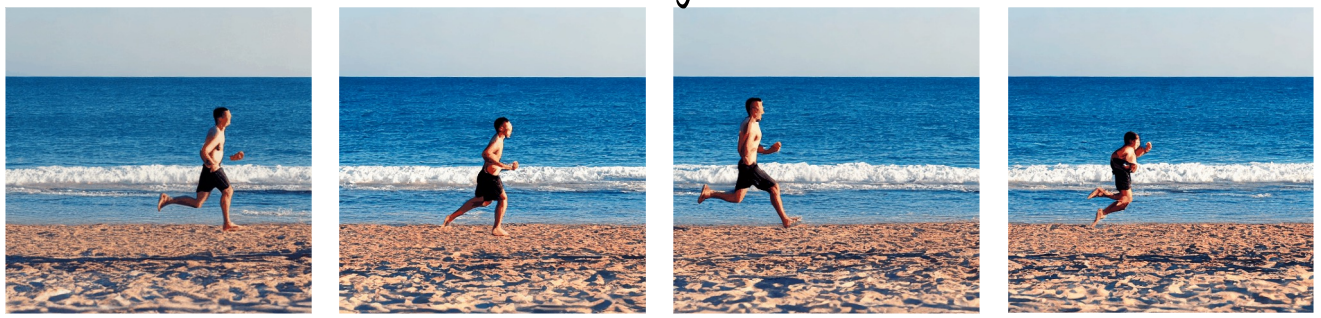
v_1 v_2 v_3 v_4

Obs #2: Extending attention to spatio-temporal yields consistent content



spatio-temporal attention

"A man is running on the beach"



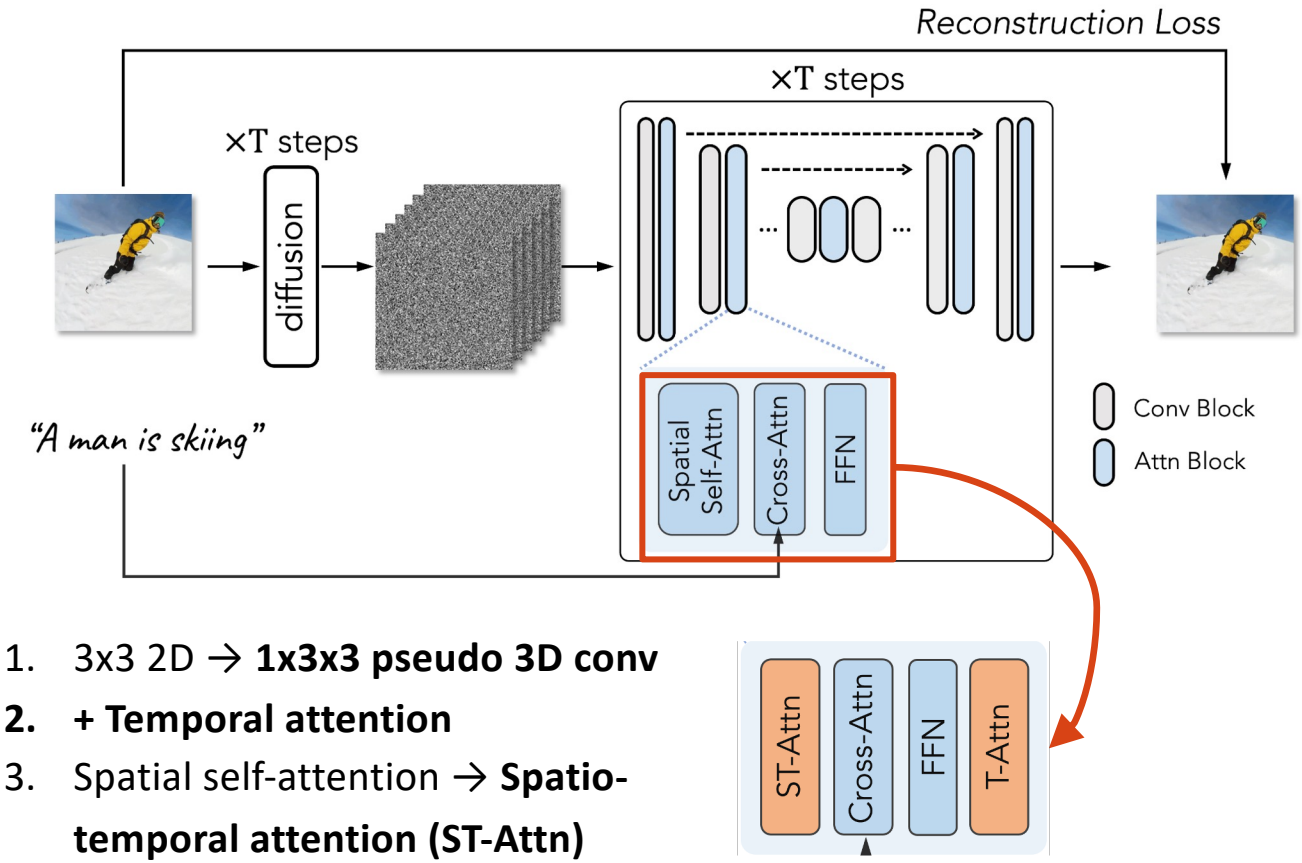
v_1 v_2 v_3 v_4

Wu et al., "Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation," ICCV 2023.

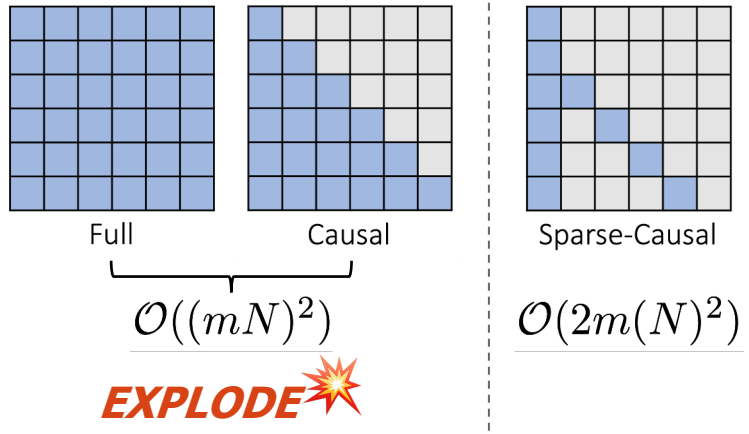
Tune-A-Video

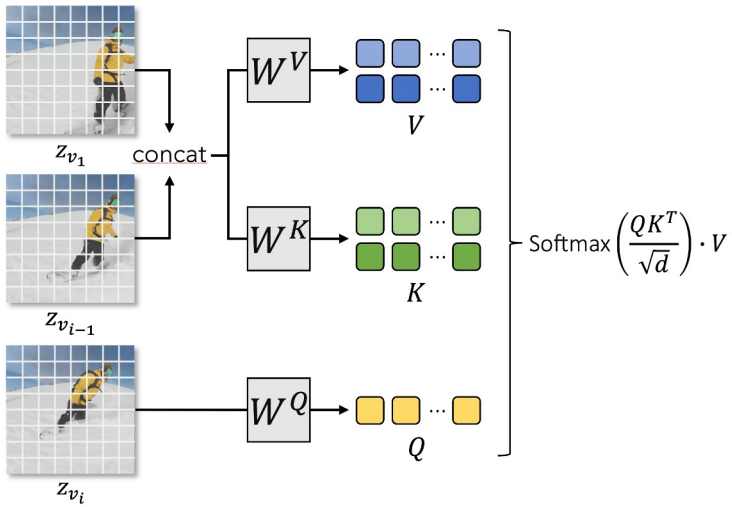
One-shot tuning of T2I models for T2V generation/editing

Network Inflation



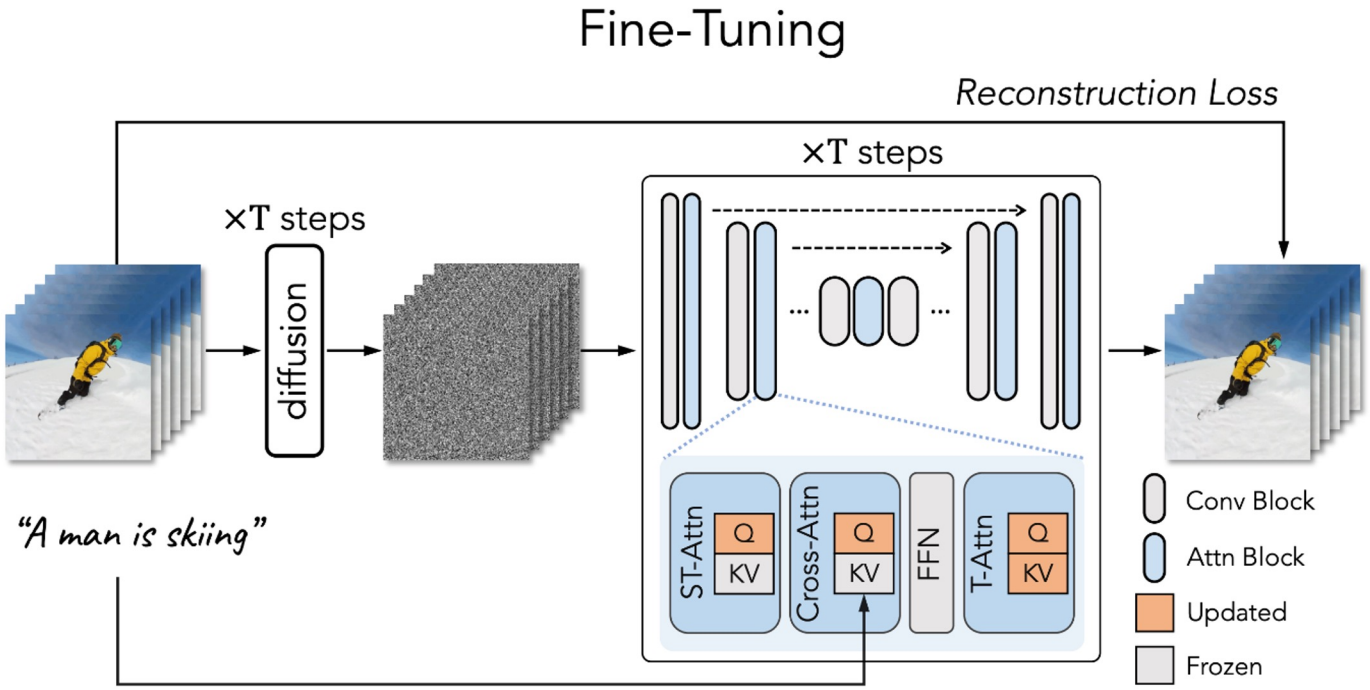
1. 3x3 2D → **1x3x3 pseudo 3D conv**
2. **+ Temporal attention**
3. Spatial self-attention → **Spatio-temporal attention (ST-Attn)**

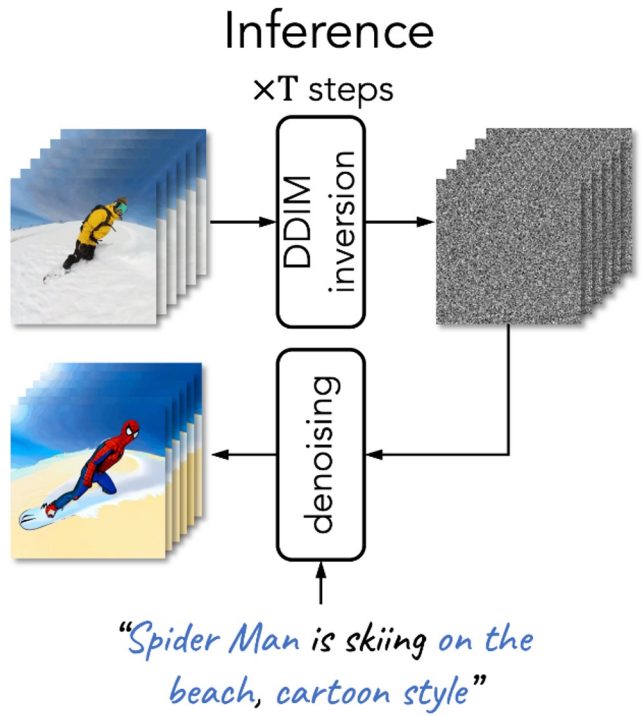




Tune-A-Video

One-shot tuning of T2I models for T2V generation/editing





Full finetuning: finetunes the entire network

- inefficient, especially when #frames increases;
- prone to overfitting → poor editing ability.

Our tuning strategy: update the specific projection matrices

- parameter efficient and fast (~10 min);
- retains the original property of pre-trained T2I diffusion models.

$$\mathcal{V}^* = \mathcal{D}(\text{DDIM-samp}(\text{DDIM-inv}(\mathcal{E}(\mathcal{V})), \mathcal{T}^*)).$$

Structure guidance via DDIM inversion

- preserves the structural information
- improves temporal consistency

Wu et al., "Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation," ICCV 2023.

Tune-A-Video

Pretrained T2I (Stable Diffusion)

"A man is skiing"

"Spider Man is ... on the beach, cartoon style"

"Wonder Woman, wearing a cowboy hat, is ..."

"A man, wearing pink clothes, is ..., at sunset"

"A man is dribbling a basketball"

"James Bond is dribbling a basketball on the beach"

"A lego man in a black suit is dribbling a basketball"

"An astronaut is dribbling a basketball, cartoon style"

Tune-A-Video

Pretrained T2I (Stable Diffusion)

"A rabbit is eating a watermelon on the table"

"A rabbit is ~~eating a watermelon~~ on the table"

"A cat with sunglasses is ... on the beach"

"A puppy is eating a cheeseburger ..., comic style"

"A lion is roaring"

"A tiger is roaring"

"A wolf is roaring in New York City"

"A lion is roaring, Van Gogh style"

Wu et al., "Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation," ICCV 2023.

Tune-A-Video

Pretrained T2I ( **)**

"Modern Disney Style"

"A bear is playing guitar"

"A rabbit is playing guitar, modern disney style"

"A prince is playing guitar, modern disney style"

"A princess wearing sunglasses is ..., modern disney style"

"A man is dribbling a basketball"

"Mickey Mouse is ..., modern disney style"

"A prince with a crown is ..., modern disney style"

"A panda is ... on the beach, modern disney style"

Wu et al., "Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation," ICCV 2023.

Tune-A-Video

Pretrained T2I ( **)**

anything-v4.0

"A bear is playing guitar"

"1girl is ..., white hair, cat ears, scarf, jacket, streets"

"1boy is ..., bishounen, coffee shop, bokeh"

"1girl is ..., red hair, long hair, dress, beach, sea"

Pretrained T2I ( **)**

"Mr. Potato Head"

"A bear is playing guitar"

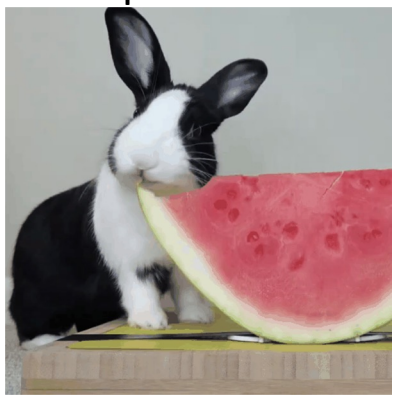
"Mr Potato Head, made of lego, is ... on the snow"

"Mr Potato Head wearing sunglasses, is ..., beach"

"Mr Potato Head is ... in the starry night, Van Gogh style"

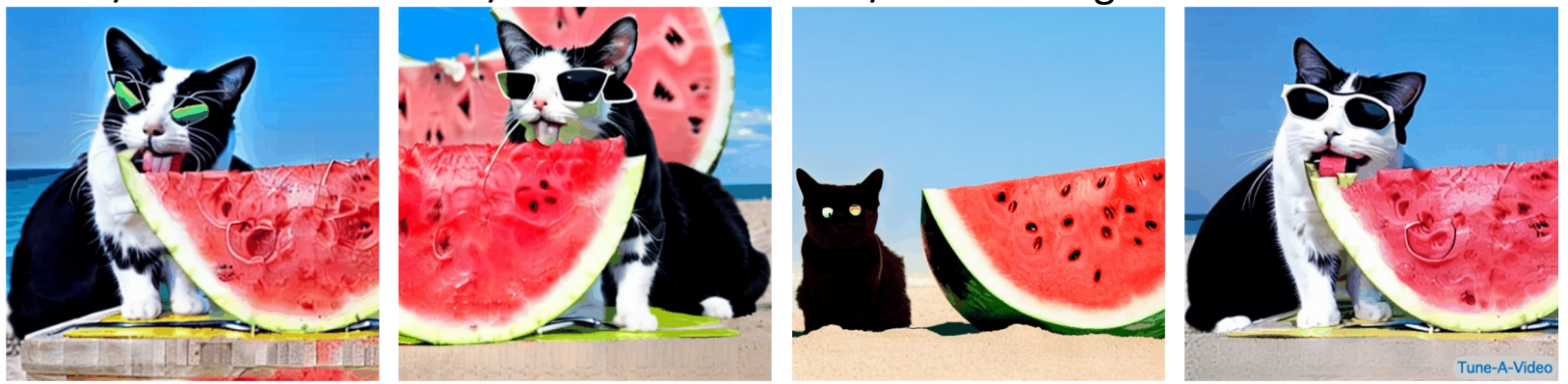
Ablation Study

Input Video



"A rabbit is eating a watermelon on the table"

w/o ST-Attn	w/o inversion	w/o finetuning	Tune-A-Video



"A cat with sunglasses is eating a watermelon on the beach"

"A puppy is eating a cheeseburger on the table, comic style"

Tune-A-Video

One-shot tuning of T2I models for T2V generation/editing

Method	Frame Consisitency		Textual alignment	
	CLIP Score	User Preference	CLIP Score	User Preference
CogVideo	90.64	12.14	23.91	15.00
Plug-and-Play	88.89	37.86	27.56	23.57
Tune-A-Video	**92.40**	**87.86* / 62.14****	**27.58**	**85.00* / 76.43****

* indicates Tune-A-Video vs. CogVideo, ** indicates Tune-A-Video vs. Plug-and-Play.

Automatic metrics – CLIP Score

- *Frame Consistency*: the average cosine similarity between all pairs of video frames
- *Textual Alignment*: average CLIP score between all frames of output videos and corresponding edited prompts

User study

Compare two videos generated by our method and a baseline (shown in random order):

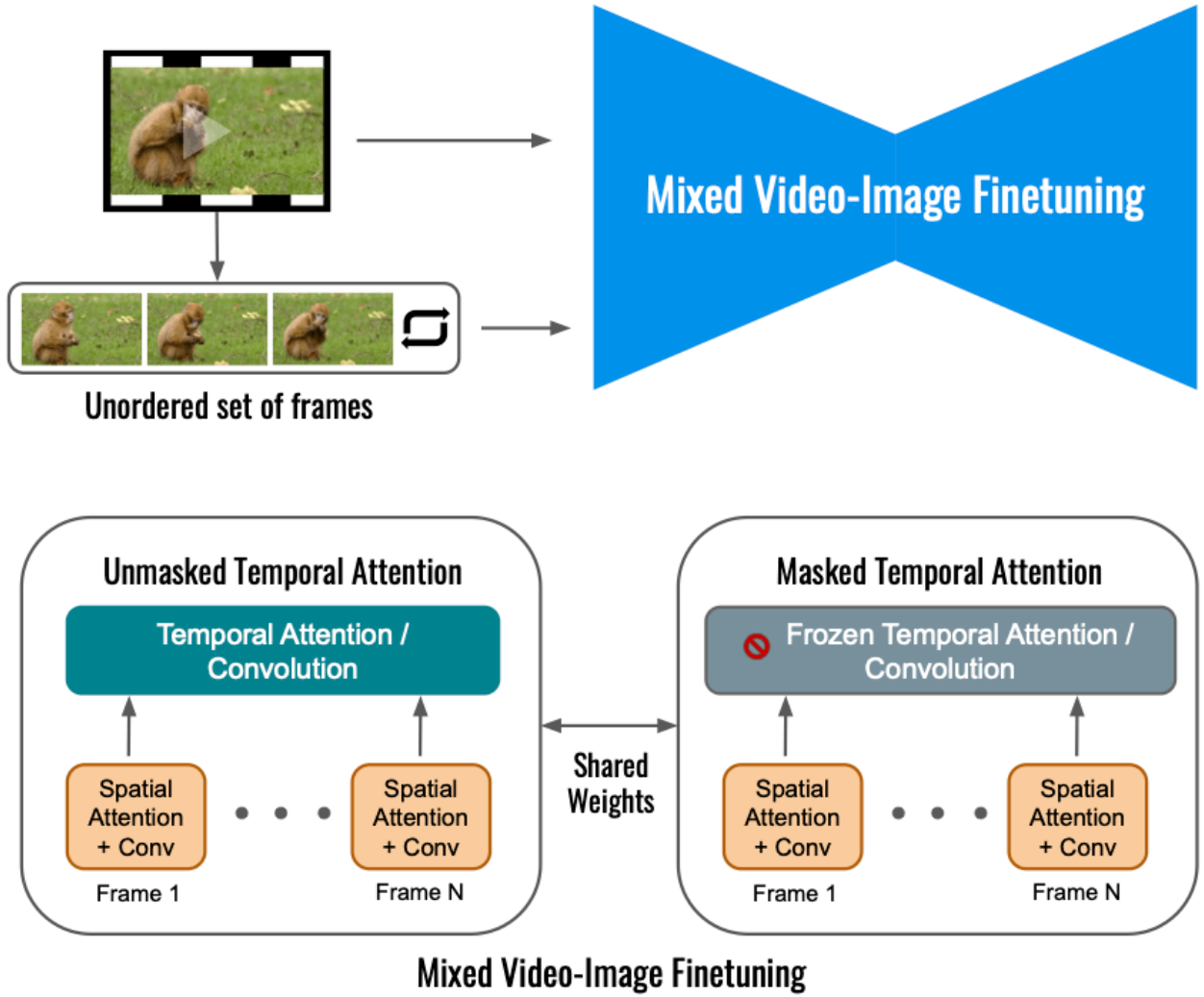
- *Which video has better temporal consistency?*
- *Which video better aligns with the textual description?*

Dreamix

Few-shot finetuning for personalized video editing

Main idea: Mixed Video-Image Finetuning

- Finetune Imagen Video (Ho et al., 2022) which is a strong video foundation model
- Finetuned to generate individual frames (bypassing temporal attentions) & video

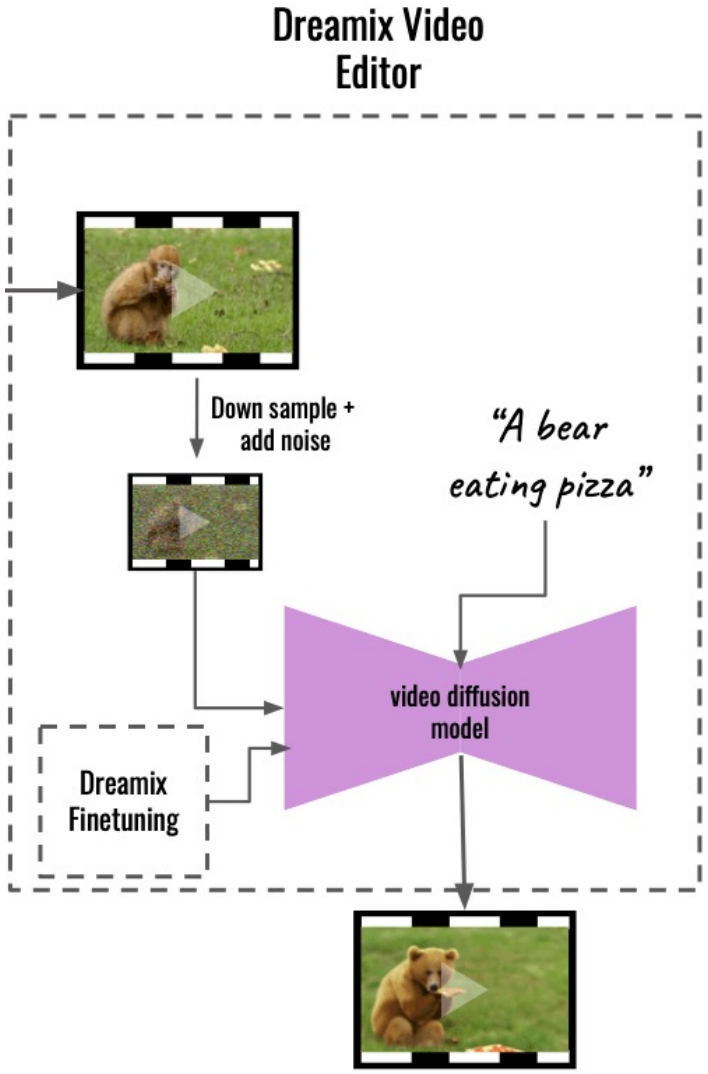


Dreamix

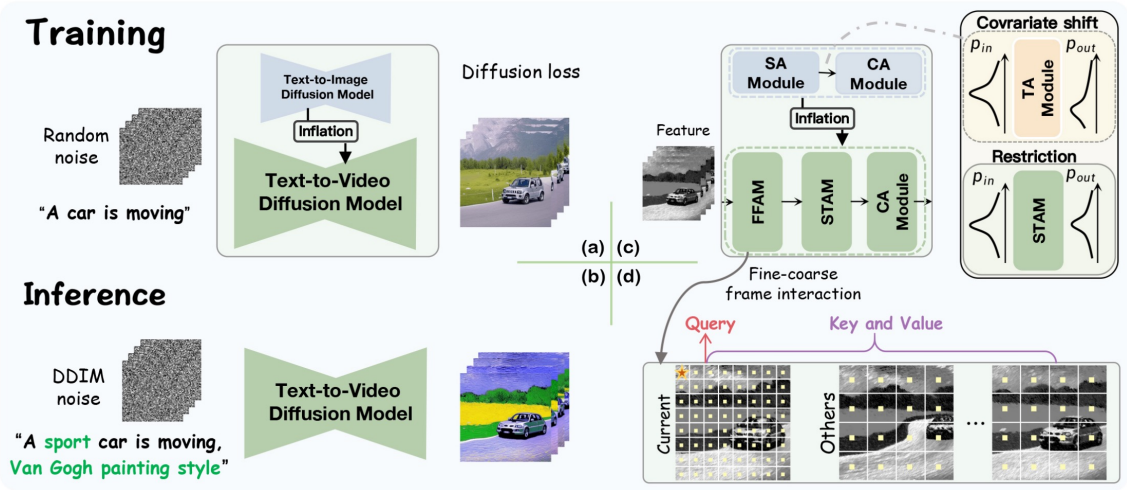
Few-shot finetuning for personalized video editing

Inference Overview

- Corrupt the input video by down-sampling and add noise
- Apply the finetuned video diffusion model to denoise and upscale



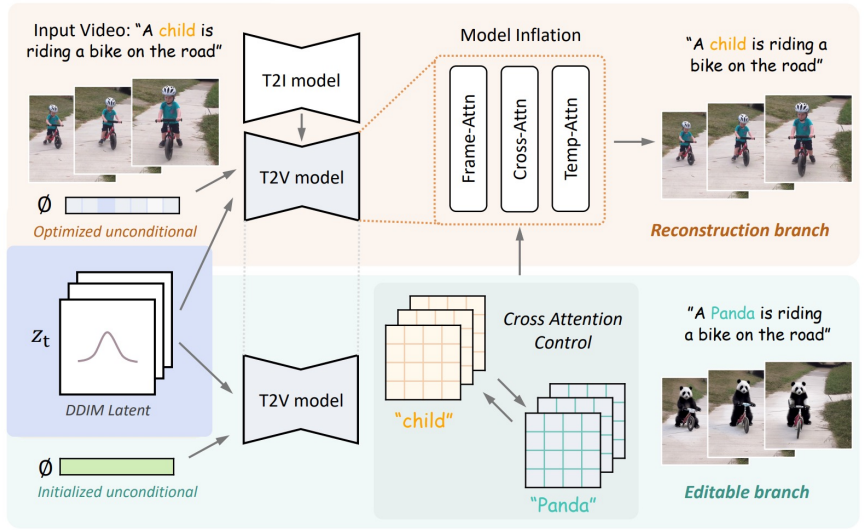
One-Shot Tuned Video Editing: More Works



EI^2 (Zhang et al.)

Modify self-attention for better temporal consistency

"Towards Consistent Video Editing with Text-to-Image Diffusion Models," arXiv 2023.



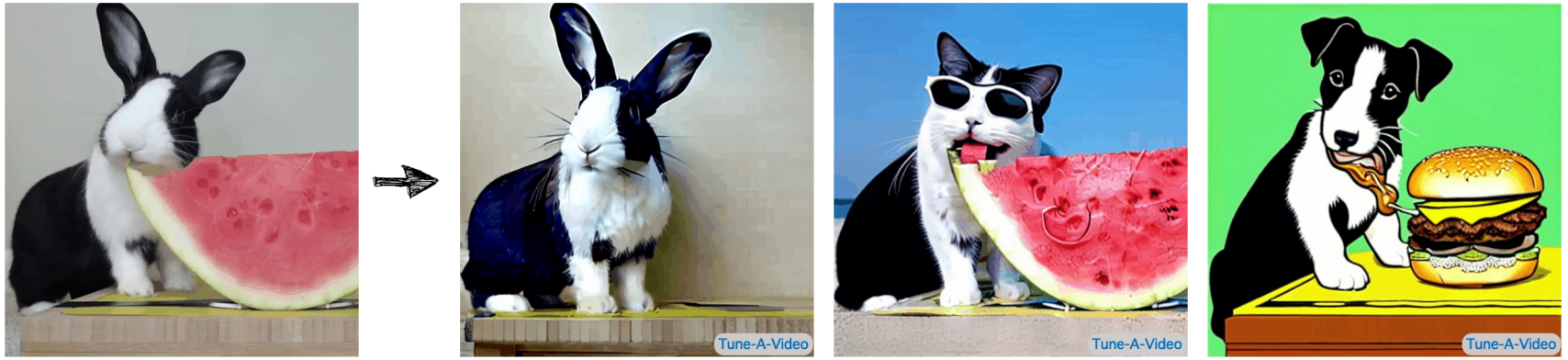
Video-P2P (Liu et al.)

Improve input-output semantic consistency of video editing via shared embedding optimization and cross-attention control

"Video-P2P: Video Editing with Cross-attention Control," arXiv 2023.

One-Shot Tuned Video Editing: More Works

Pretrained T2I (Stable Diffusion)



"A rabbit is eating a watermelon on the table"

"A rabbit is ~~eating a watermelon~~ on the table"

"A cat with sunglasses is ... on the beach"

"A puppy is eating a cheeseburger ..., comic style"

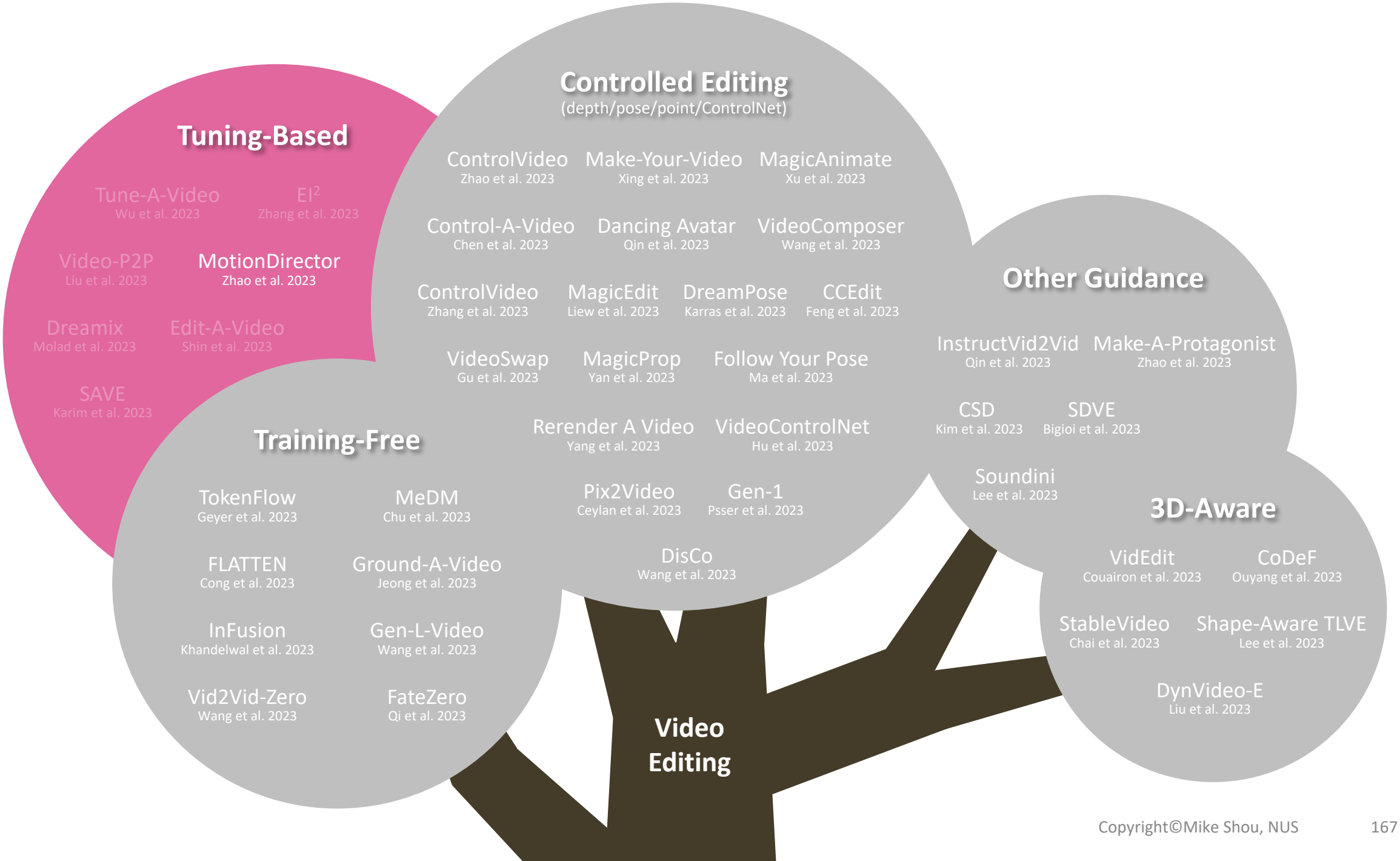
Compared to training-free editing methods:

- Cons: still need 1 video for training
- Pros: supports significant shape change

Wu et al., "Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation," ICCV 2023.

Multiple-Shot Tuned

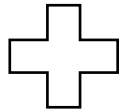
Video Editing: Text Conditioned

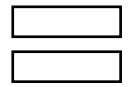


MotionDirector

Tune on multiple videos of a motion to be customised

Imagine what a monkey playing golf looks like?



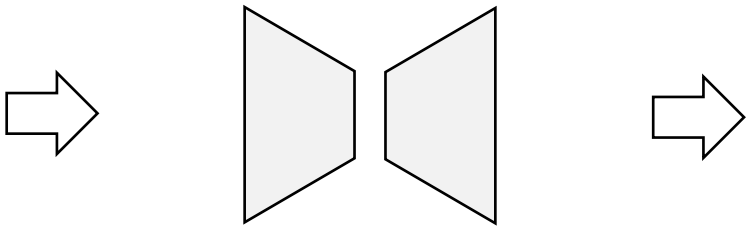


MotionDirector

Tune on multiple videos of a motion to be customised

Foundation models have no idea

" A monkey is playing golf, side view. "



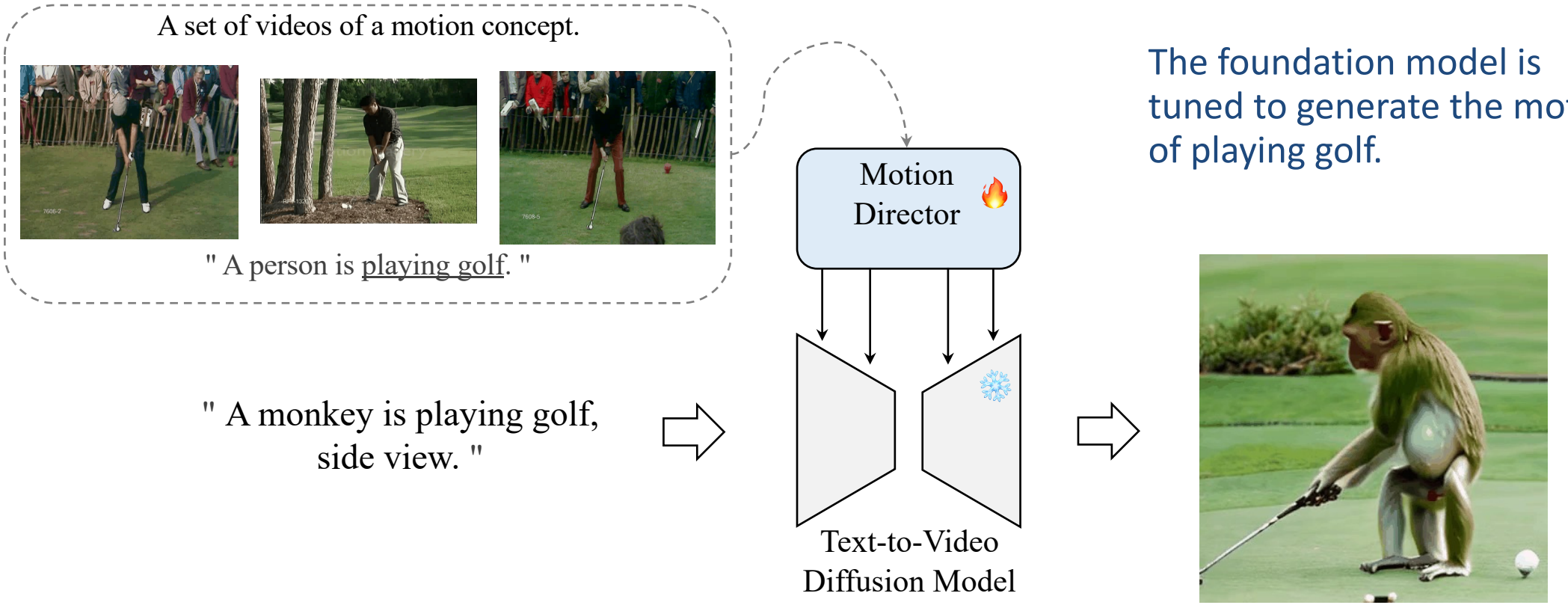
Foundational Text-to-Video Diffusion Model

Zhao et al., “MotionDirector: Motion Customization of Text-to-Video Diffusion Models,” arXiv 2023.

MotionDirector

Tune on multiple videos of a motion to be customised

Let’s give the foundation model some hints.



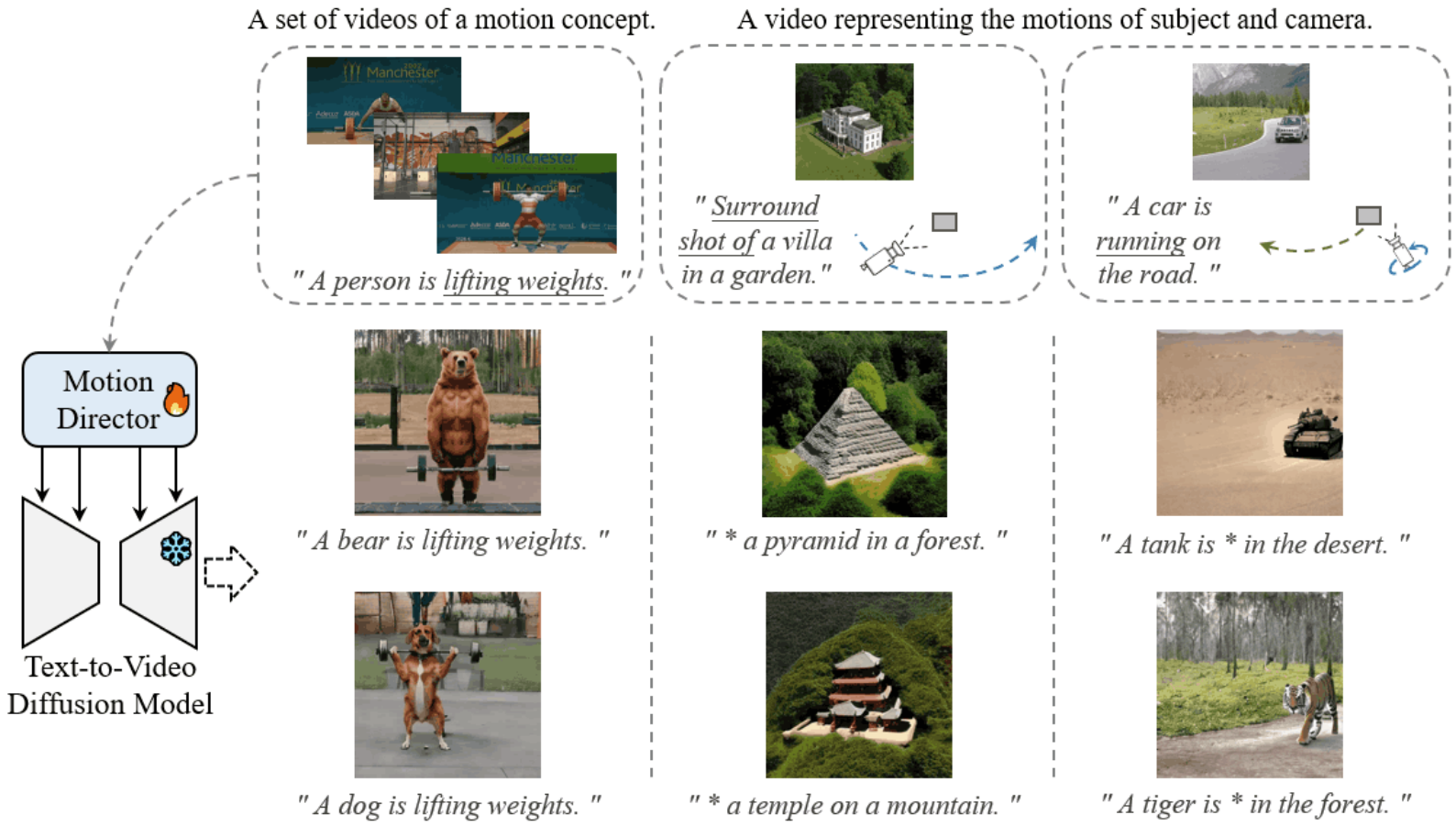
The foundation model is tuned to generate the motion of playing golf.

Zhao et al., “MotionDirector: Motion Customization of Text-to-Video Diffusion Models,” arXiv 2023.

MotionDirector

Tune on multiple videos of a motion to be customised

- MotionDirector can customize foundation models to generate videos with desired motions.

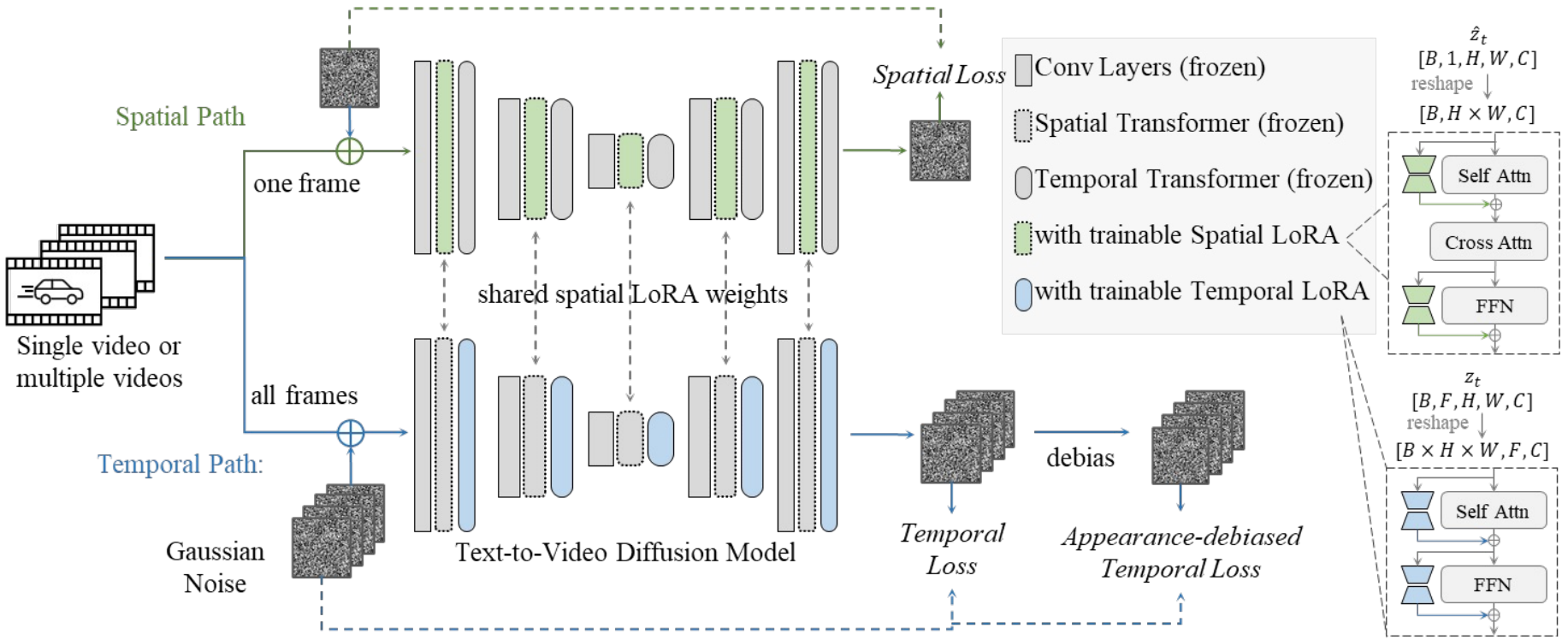


Zhao et al., "MotionDirector: Motion Customization of Text-to-Video Diffusion Models," arXiv 2023.

MotionDirector

Tune on multiple videos of a motion to be customised

- The challenge is generalizing the learned motions to diverse appearance.
- MotionDirector learns the appearances and motions in reference videos in a decoupled way, to avoid overfitting on the limited appearances.

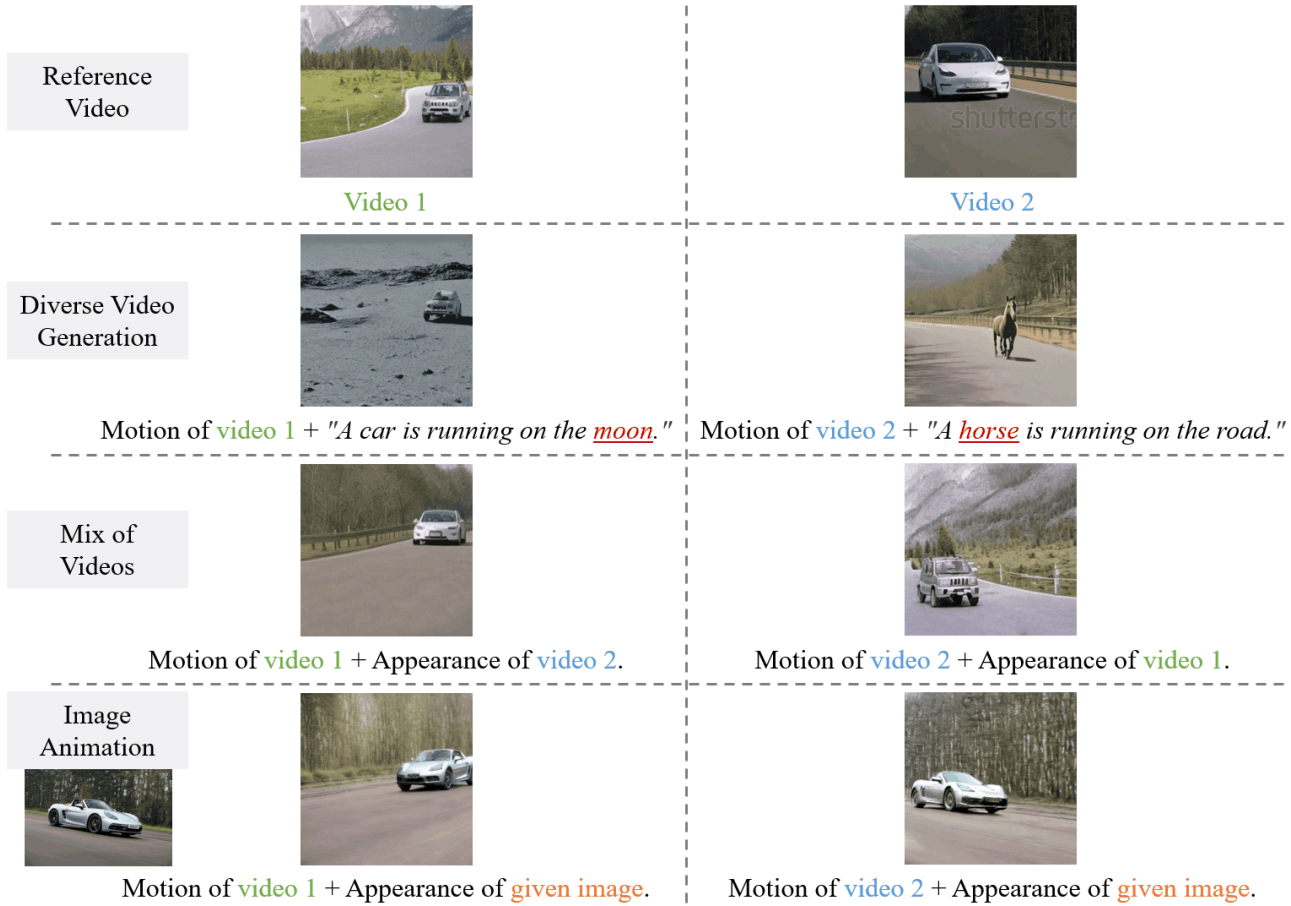


Zhao et al., "MotionDirector: Motion Customization of Text-to-Video Diffusion Models," arXiv 2023.

MotionDirector

Tune on multiple videos of a motion to be customised

- Decouple appearance and motion.



Zhao et al., "MotionDirector: Motion Customization of Text-to-Video Diffusion Models," arXiv 2023.

MotionDirector

Tune on multiple videos of a motion to be customised

- Comparing with other methods.

MotionDirector

Tune on multiple videos of a motion to be customised

- Comparing with other methods.

Automatic Evaluations					Human Evaluations			
		Appearance Diversity (↑)	Temporal Consistency (↑)	Pick Score (↑)		Appearance Diversity	Temporal Consistency	Motion Fidelity
Tune-A-Video		28.22	92.45	20.20	v.s. Base Model (ModelScope) v.s. Base Model (ZeroScope)	25.00 v.s. 75.00 44.00 v.s. 56.00	25.00 v.s. 75.00 16.67 v.s. 83.33	40.00 v.s. 60.00 53.33 v.s. 46.67
ModelScope	Base Model	28.55	92.54	20.33				
	Coupled Tuned	**25.66 (-2.89)**	90.66	19.85	v.s. Base Model (ModelScope)	**23.08 v.s. 76.92**	40.00 v.s. 60.00	52.00 v.s. 48.00
	w/o AD-Loss	28.32 (-0.23)	91.17	20.34	v.s. Base Model (ModelScope)	53.12 v.s. 46.88	49.84 v.s. 50.16	62.45 v.s. 37.55
	ours	**28.66 (+0.11)**	92.36	20.59	v.s. Base Model (ModelScope)	**54.84 v.s. 45.16**	56.00 v.s. 44.00	**75.00 v.s. 25.00**
ZeroScope	Base Model	28.40	92.94	20.76				
	Coupled Tuned	**25.52 (-2.88)**	90.67	19.99	v.s. Base Model (ZeroScope)	**37.81 v.s. 62.19**	41.67 v.s. 58.33	54.55 v.s. 45.45
	w/o AD-Loss	28.61 (+0.21)	91.37	20.56	v.s. Base Model (ZeroScope)	50.10 v.s. 49.90	48.00 v.s. 52.00	58.33 v.s. 41.67
	ours	**28.94 (+0.54)**	92.67	20.80	v.s. Base Model (ZeroScope)	**52.94 v.s. 47.06**	55.00 v.s. 45.00	**76.47 v.s. 23.53**

Zhao et al., "MotionDirector: Motion Customization of Text-to-Video Diffusion Models," arXiv 2023.

MotionDirector

Tune on multiple videos of a motion to be customised

Reference Videos:
" A person is skateboarding. "

Generated Video:
" A *bear* is skateboarding. "

Reference Videos:
" A person is skateboarding. "

Generated Video:
" A *man* is skateboarding *on the moon*."

Reference Video:
" A women is eating a pizza with various toppings. "

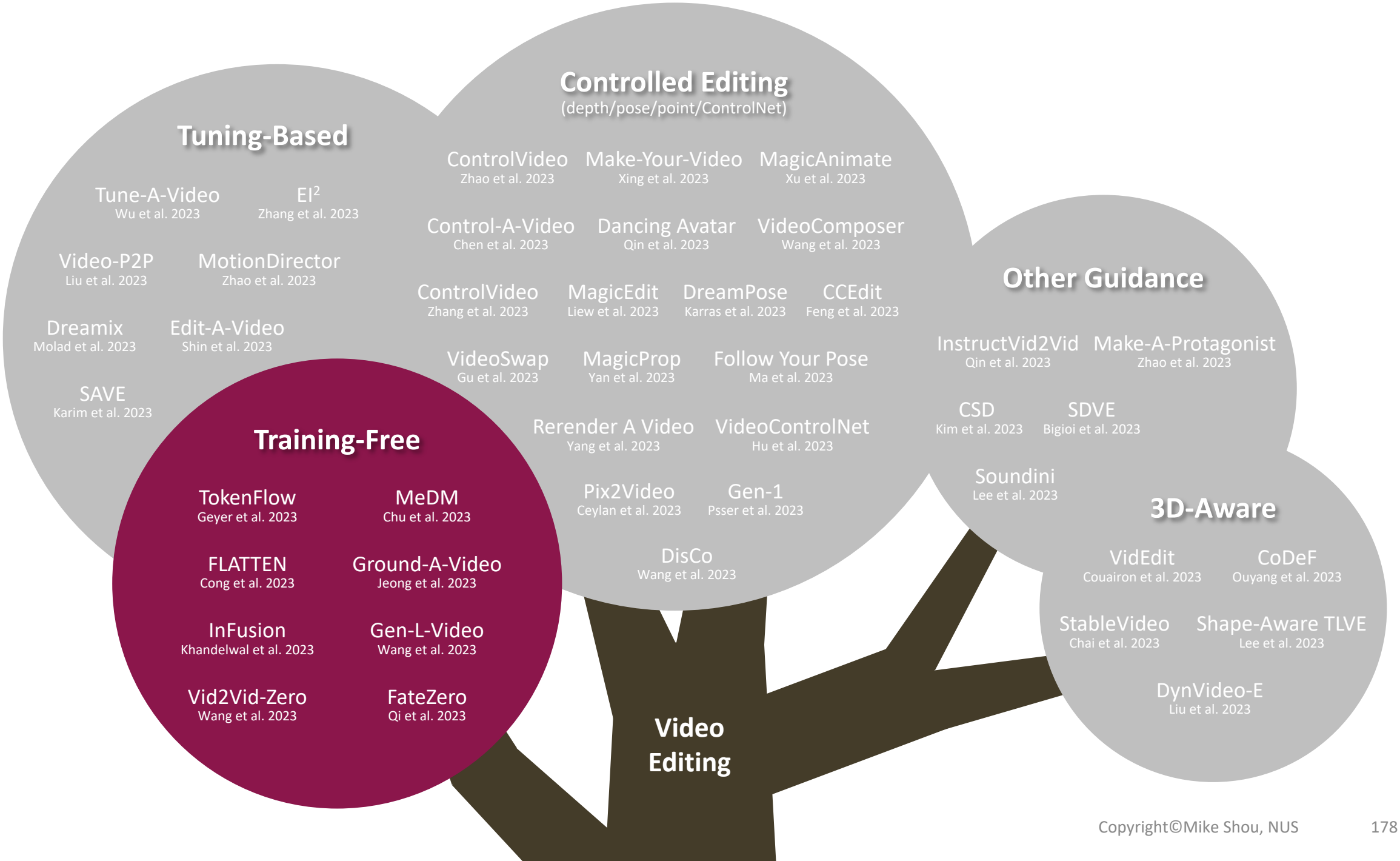
Generated Video:
" A *panda* is eating a pizza with various toppings. "

Reference Video:
" Surround shot of a villa in a garden. "

Generated Video:
" Surround shot of *an alien base on Mars*. "

3 Video Editing

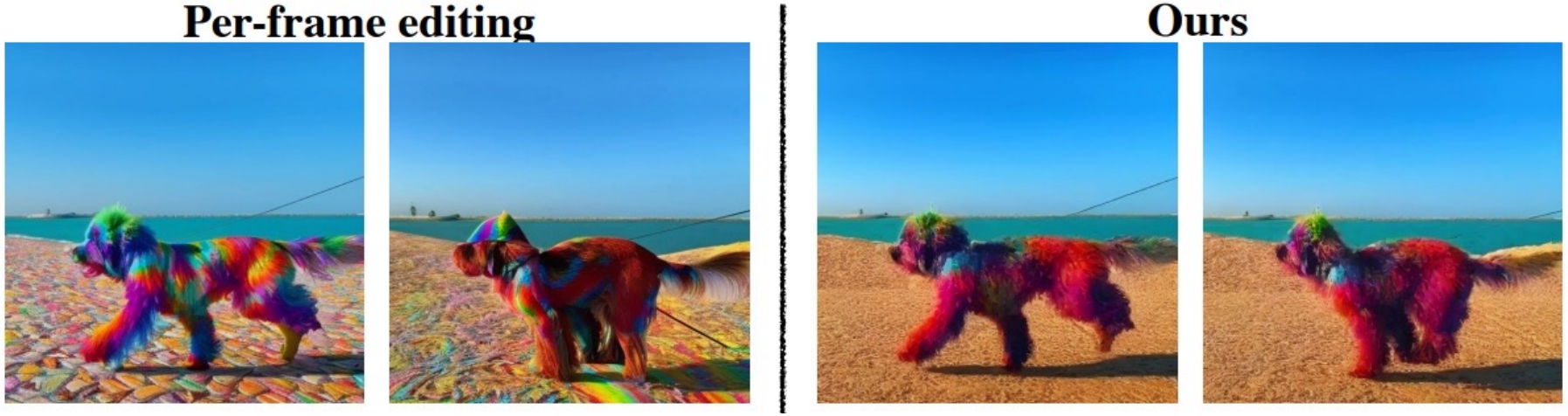
3.2 Training-free



TokenFlow

Consistent high-quality semantic edits

Main challenge using T2I to edit videos without finetuning: temporal consistency



Inconsistent patterns

Good consistency

TokenFlow

Consistent high-quality semantic edits

Key Idea

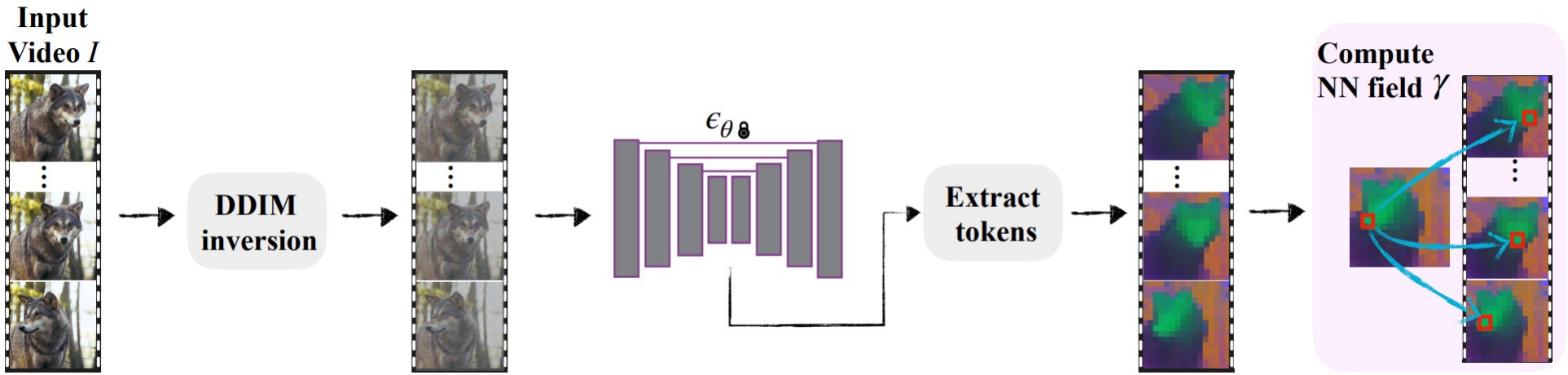
- Achieve consistency by enforcing the inter-frame correspondences in the original video

Geyer et al., “TokenFlow: Consistent Diffusion Features for Consistent Video Editing,” arXiv 2023.

TokenFlow

Consistent high-quality semantic edits

Main idea



Find similar features across frame
Record the correspondence

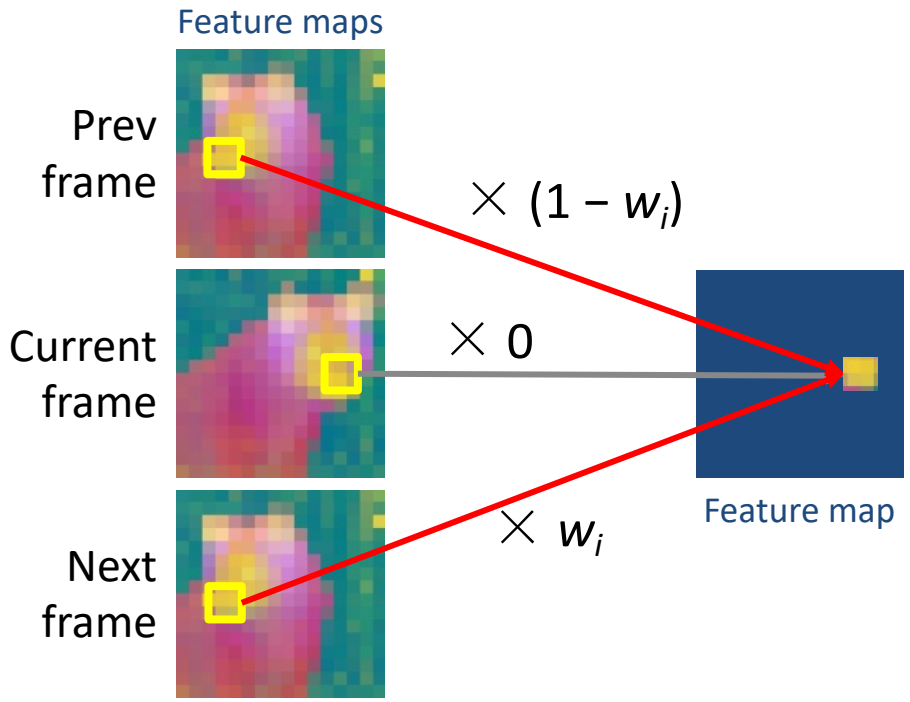
$$\gamma^{i\pm}[p] = \arg\min_{q} \mathcal{D}\left(\phi(\boldsymbol{x}^{i})[p], \phi(\boldsymbol{x}^{i\pm})[q]\right)$$

TokenFlow

Consistent high-quality semantic edits

Main idea

During conditional denoising, use features from corresponding positions in preceding and following frames instead of the pixel's own feature at output of extended-attention



$$\mathcal{F}_\gamma(\mathbf{T}_{base}, i, p) = w_i \cdot \phi(\boldsymbol{J}^{i+})[\gamma^{i+}[p]] + (1 - w_i) \cdot \phi(\boldsymbol{J}^{i-})[\gamma^{i-}[p]]$$

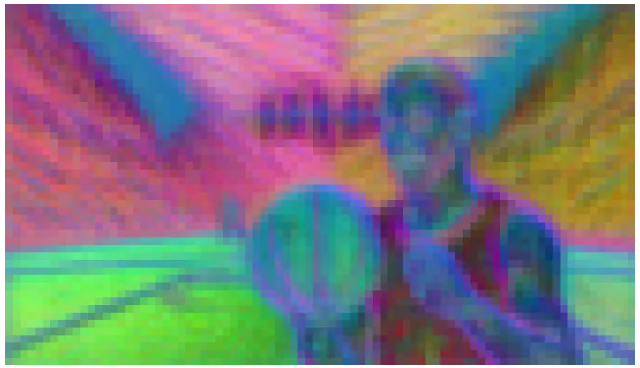
Geyer et al., "TokenFlow: Consistent Diffusion Features for Consistent Video Editing," arXiv 2023.

TokenFlow

Consistent high-quality semantic edits

No temporal finetuning, good temporal consistency

Original



Per-frame editing

TokenFlow

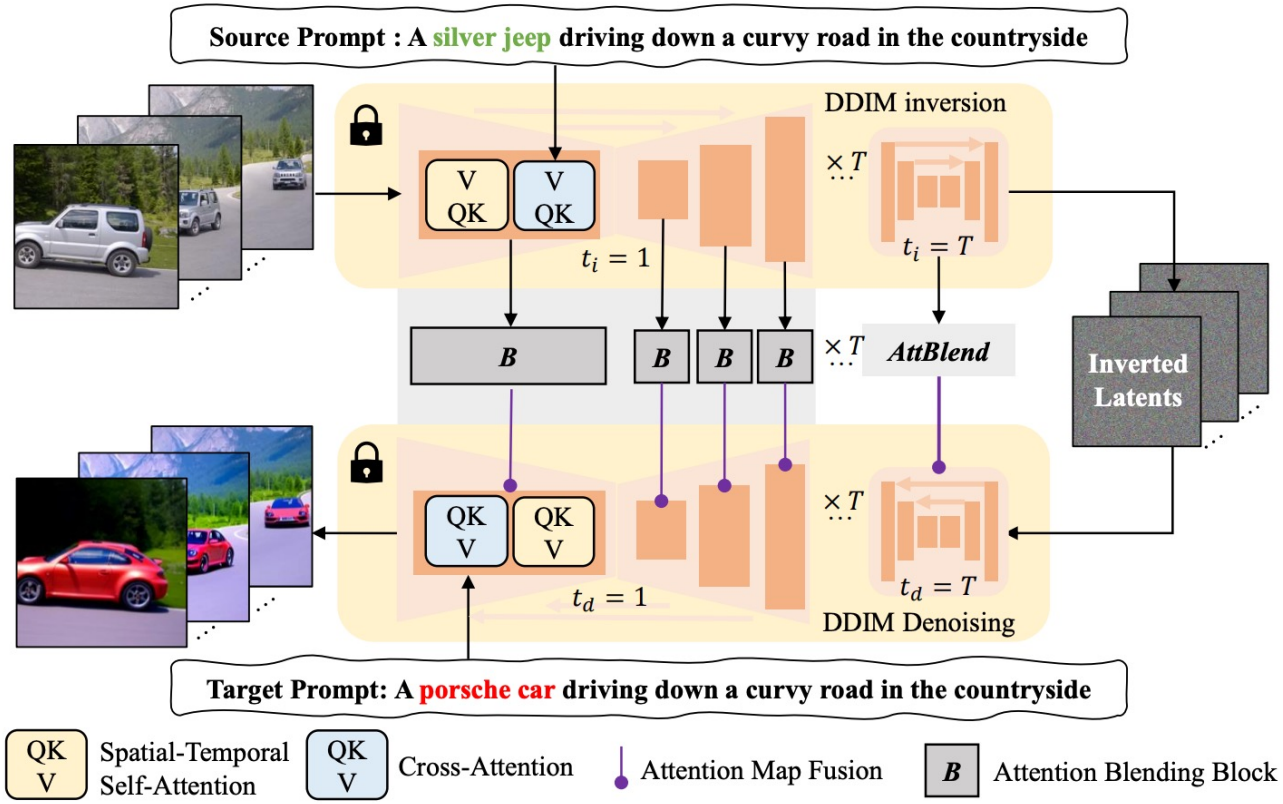
Geyer et al., “TokenFlow: Consistent Diffusion Features for Consistent Video Editing,” arXiv 2023.

FateZero

Attention map fusing for better temporal consistency

Methodology

- During DDIM inversion, save inverted self-/cross-attention maps
- During editing, use some algorithms to blend inverted maps and generated maps



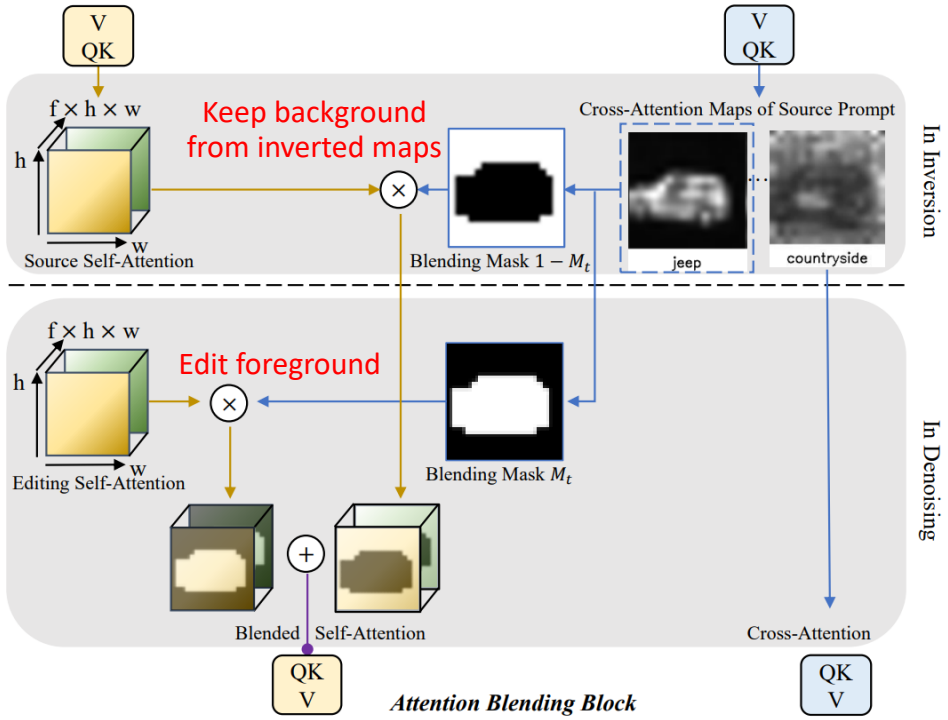
Qi et al., "FateZero: Fusing Attentions for Zero-shot Text-based Video Editing," ICCV 2023.

FateZero

Attention map fusing for better temporal consistency

Methodology

- During DDIM inversion, save inverted self-/cross-attention maps
- During editing, use some algorithms to blend inverted maps and generated maps

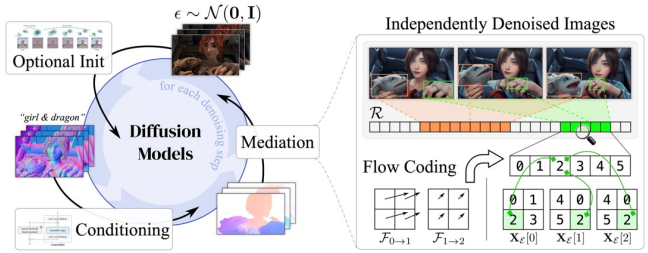


FateZero

Attention map fusing for better temporal consistency

Qi et al., "FateZero: Fusing Attentions for Zero-shot Text-based Video Editing," ICCV 2023.

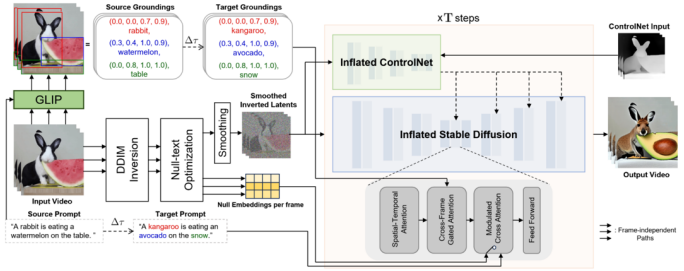
Training-Free Video Editing: More Works



MeDM (Chu et al.)

Optical flow-based guidance for temporal consistency

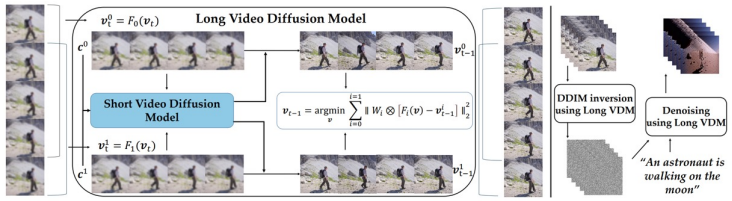
"MeDM: Mediating Image Diffusion Models for Video-to-Video Translation with Temporal Correspondence Guidance," arXiv 2023.



Ground-A-Video (Jeong et al.)

Improve temporal consistency via modified attention and optical flow

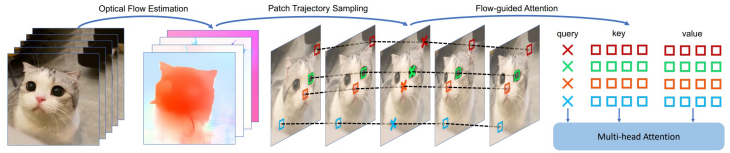
"Ground-A-Video: Zero-shot Grounded Video Editing using Text-to-image Diffusion Models," arXiv 2023.



Gen-L-Video (Lorem et al.)

Edit very long videos using existing generators

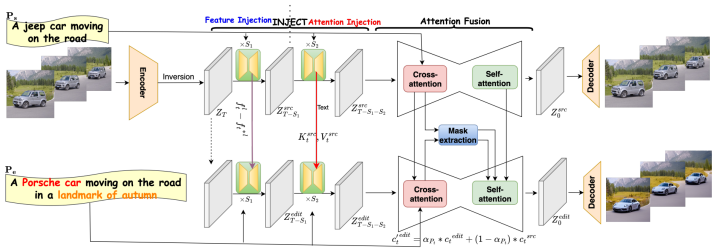
"Gen-L-Video: Multi-Text to Long Video Generation via Temporal Co-Denoising," arXiv 2023.



FLATTEN (Cong et al.)

Optical flow-guided attention for temporal consistency

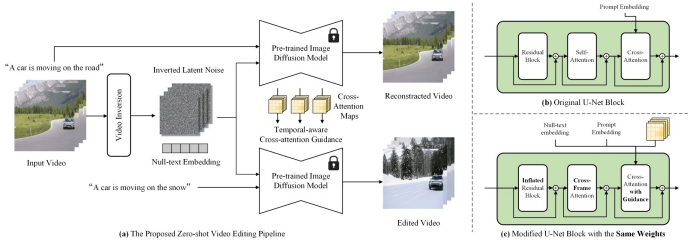
"Flatten: optical flow-guided attention for consistent text-to-video editing," arXiv 2023.



InFusion (Khandelwal et al.)

Improve temporal consistency via fusing latents

"InFusion: Inject and Attention Fusion for Multi Concept Zero-Shot Text-based Video Editing," ICCVW 2023.



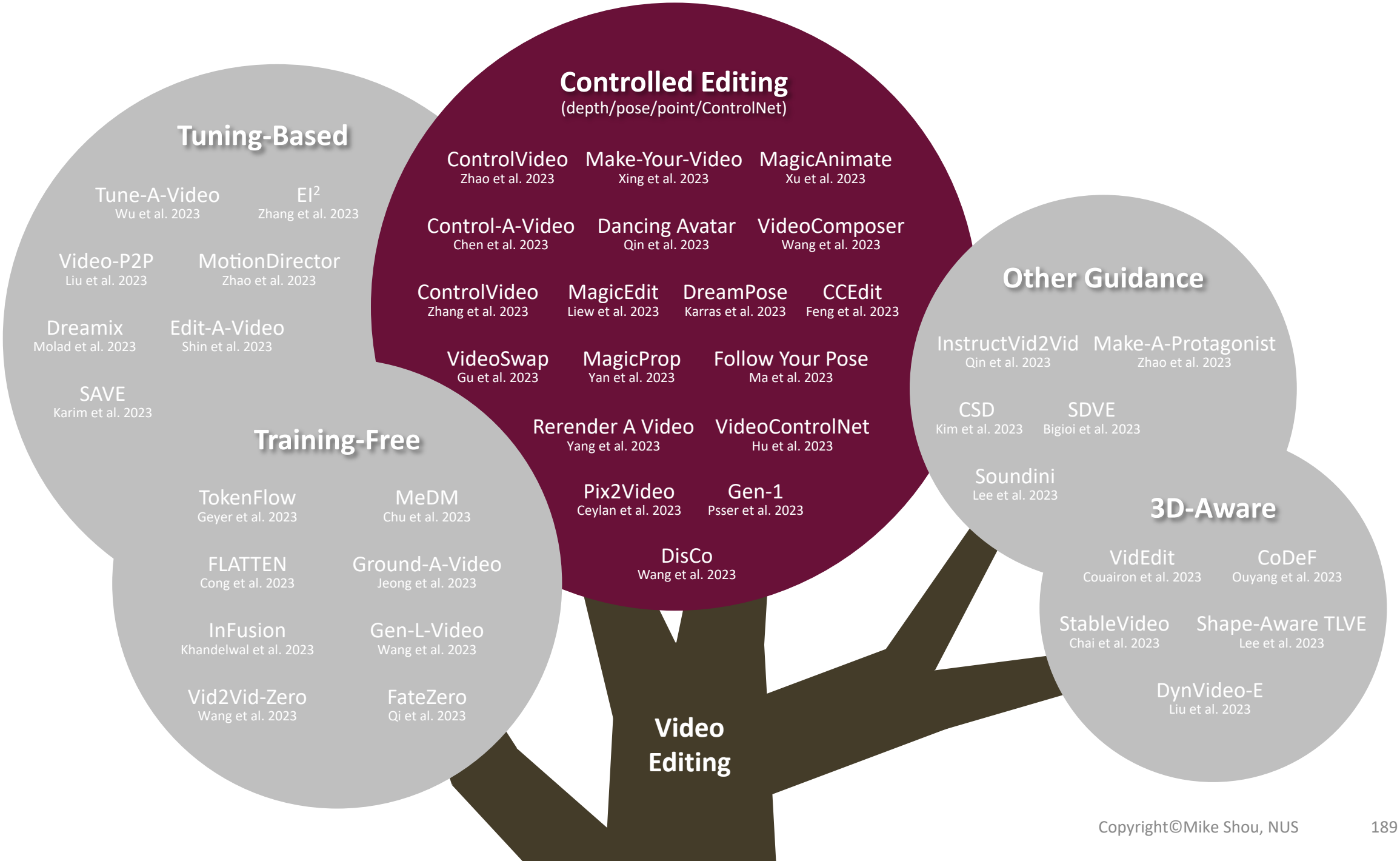
Vid2Vid-Zero (Wang et al.)

Improve temporal consistency via cross-attention guidance and null-text inversion

"Zero-Shot Video Editing Using Off-The-Shelf Image Diffusion Models," arXiv 2023.

3 Video Editing

3.3 Controlled Editing (depth/pose/point/ControlNet)



Depth Control

Gen-1

Video editing tool

RunwayML 4+

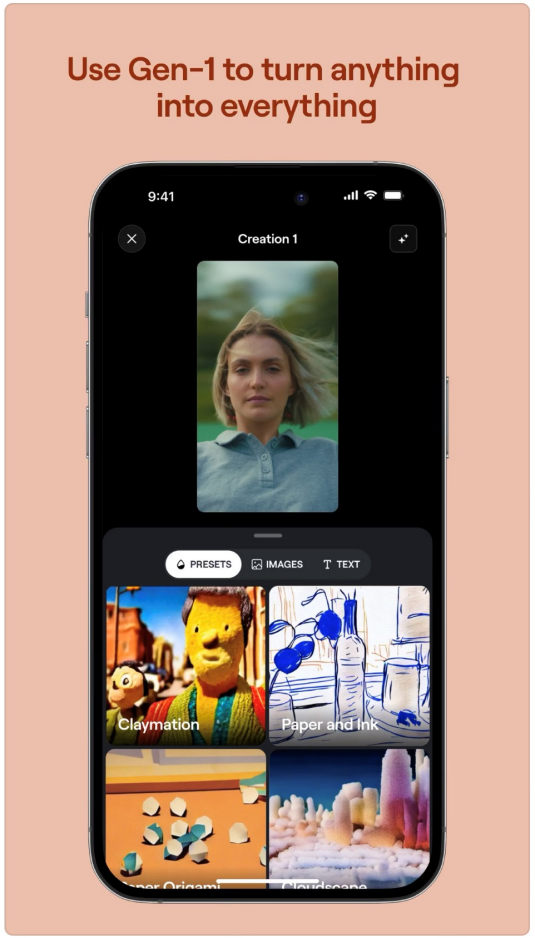
Magic AI video creation

runwayml

#156 in Photo & Video

★★★★½ 4.4 • 957 Ratings

Free · Offers In-App Purchases



Runaway Research, 2023.

Use MiDaS to offer depth condition

Depth estimating network

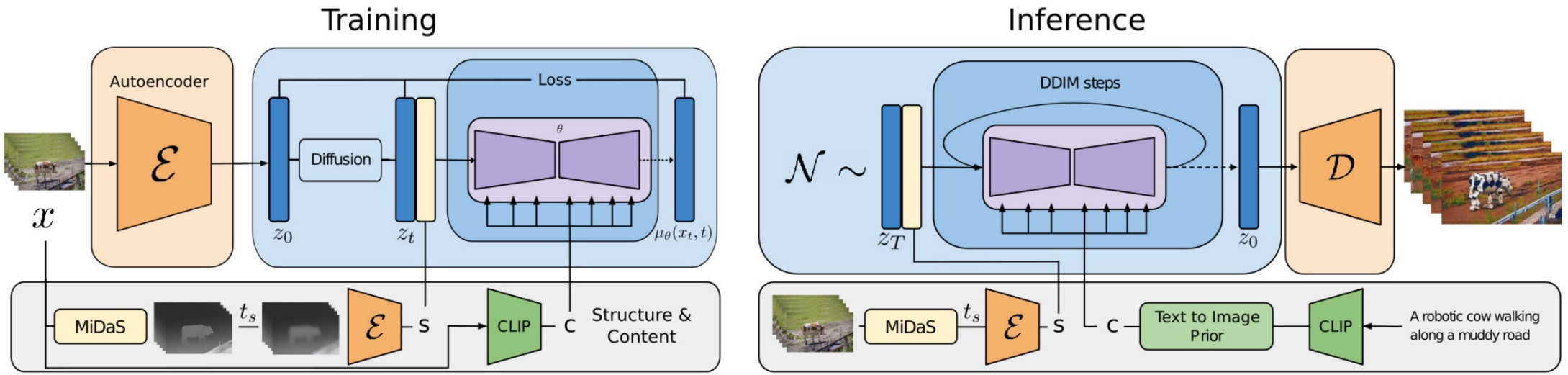
Image → depth

Ranftl et al., “Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer,” TPAMI 2022.

Gen-1

Framewise depth-guided video editing

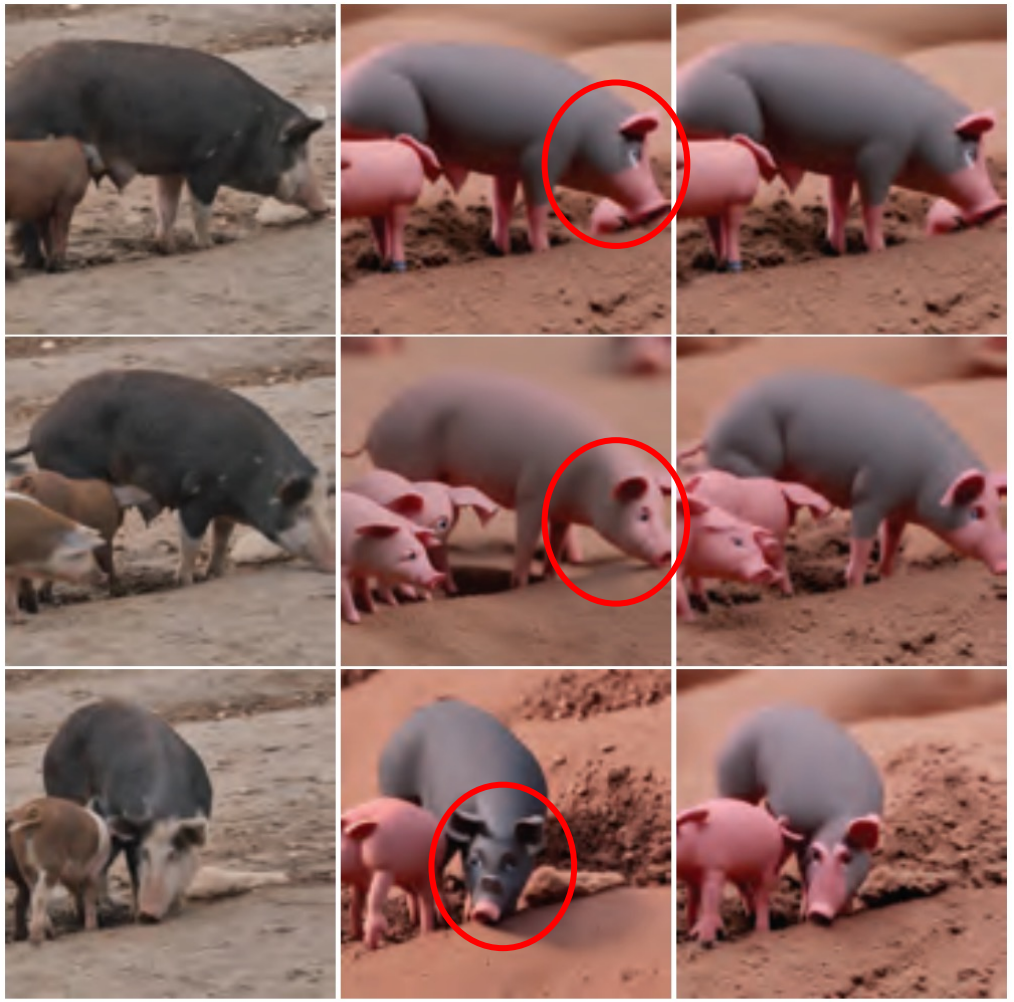
- Inflate Stable Diffusion to a 3D model, finetune on pretrained weights
- Insert temporal convolution/attention layers
- Finetune to take per-frame depth as conditions



Pix2Video

Framewise depth-guided video editing

- Leverage a pretrained per-frame depth-conditioned Stable Diffusion model to edit frame by frame, to maintain motion consistency between source video and edited video
- No need for training/finetuning
- Challenge is how to ensure temporal consistency?

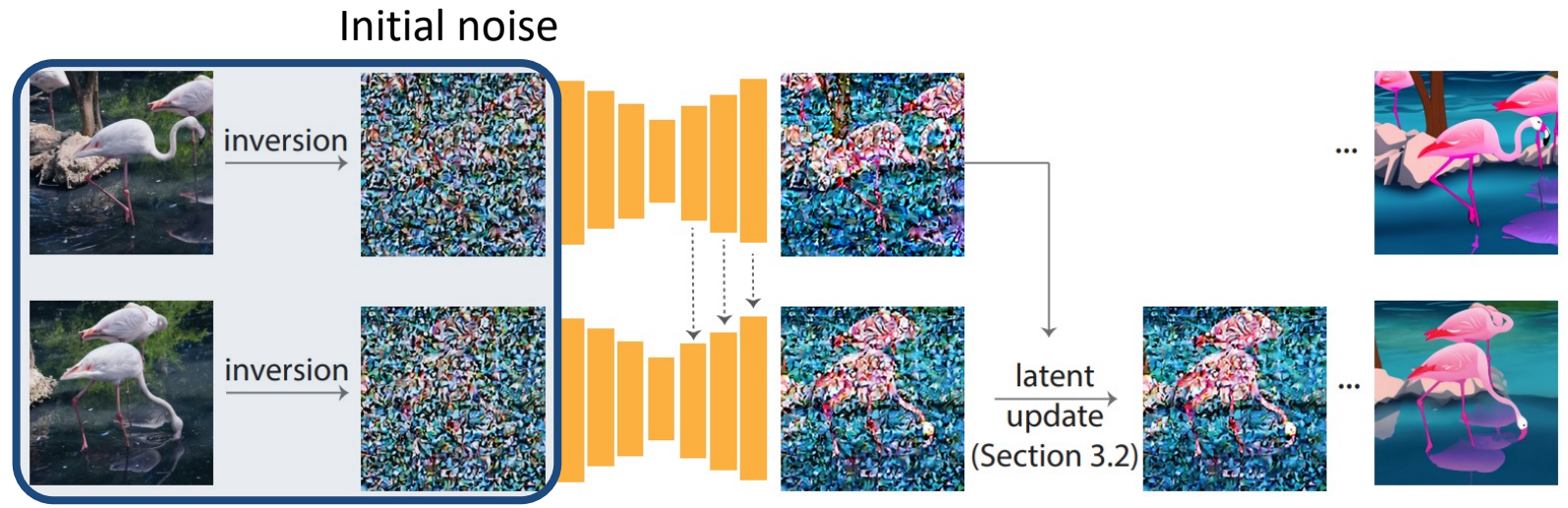


Ceylan et al., “Pix2Video: Video Editing using Image Diffusion,” ICCV 2023.

Pix2Video

Framewise depth-guided video editing

- How to ensure temporal consistency?
  - Obtain initial noise from DDIM inversion

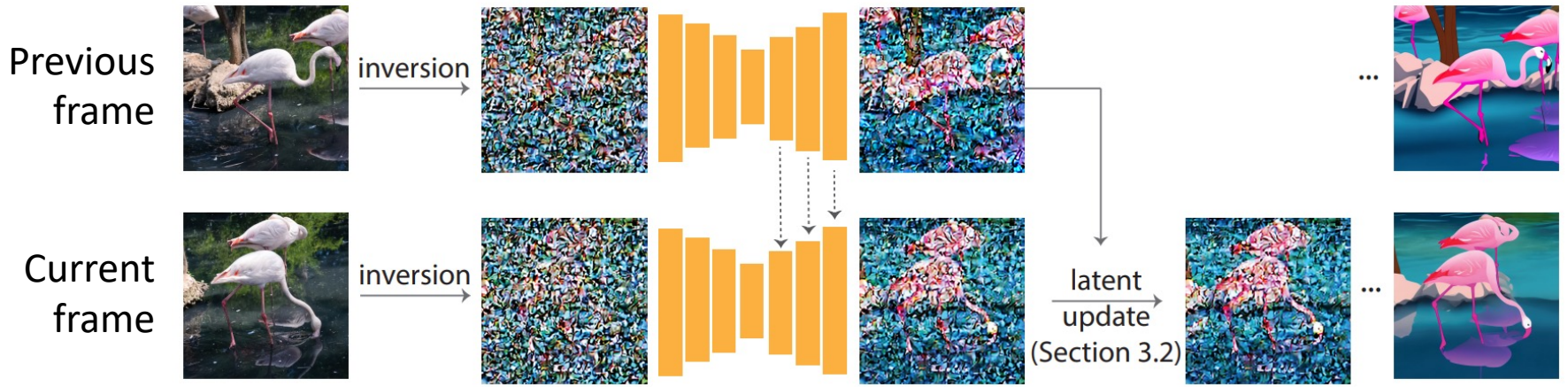


Initial noise obtained through DDIM inversion

Pix2Video

Framewise depth-guided video editing

- How to ensure temporal consistency?
  - Inject self-attention features from the previous frame in U-Net for generating the current frame
  - Use the latent of the previous frame to guide latent update of the current frame



Ceylan et al., "Pix2Video: Video Editing using Image Diffusion," ICCV 2023.

Pix2Video

Framewise depth-guided video editing

a jeep car is moving on the beach

a jeep car is moving on the snow

a jeep car is moving on the road

Ceylan et al., “Pix2Video: Video Editing using Image Diffusion,” ICCV 2023.

Pix2Video

Framewise depth-guided video editing

Iron Man is skiing on the snow

A man is surfing on the sea

A man wearing red is skiing on the snow

Ceylan et al., “Pix2Video: Video Editing using Image Diffusion,” ICCV 2023.

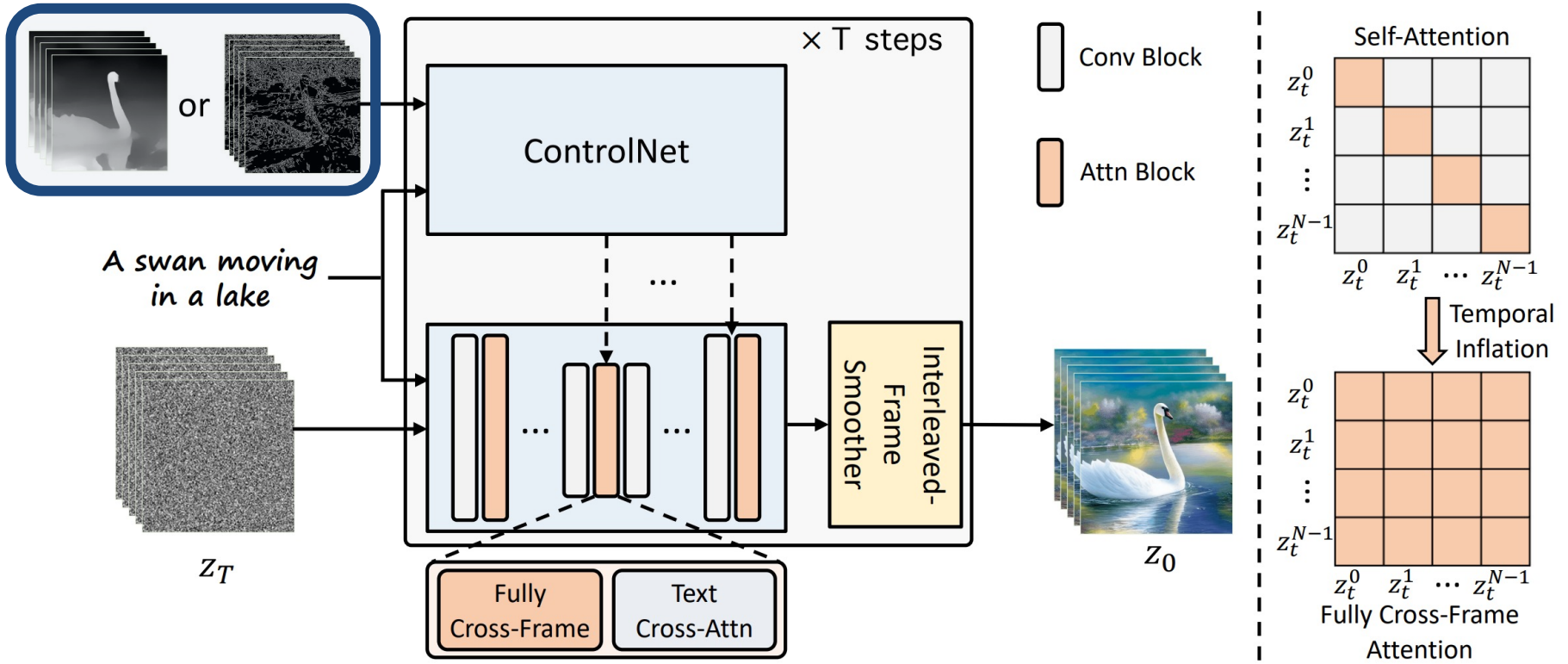
ControlNet / Multiple Control

ControlVideo (Zhang et al. 2023)

ControlNet-like video editing

- Input structural conditions through ControlNet

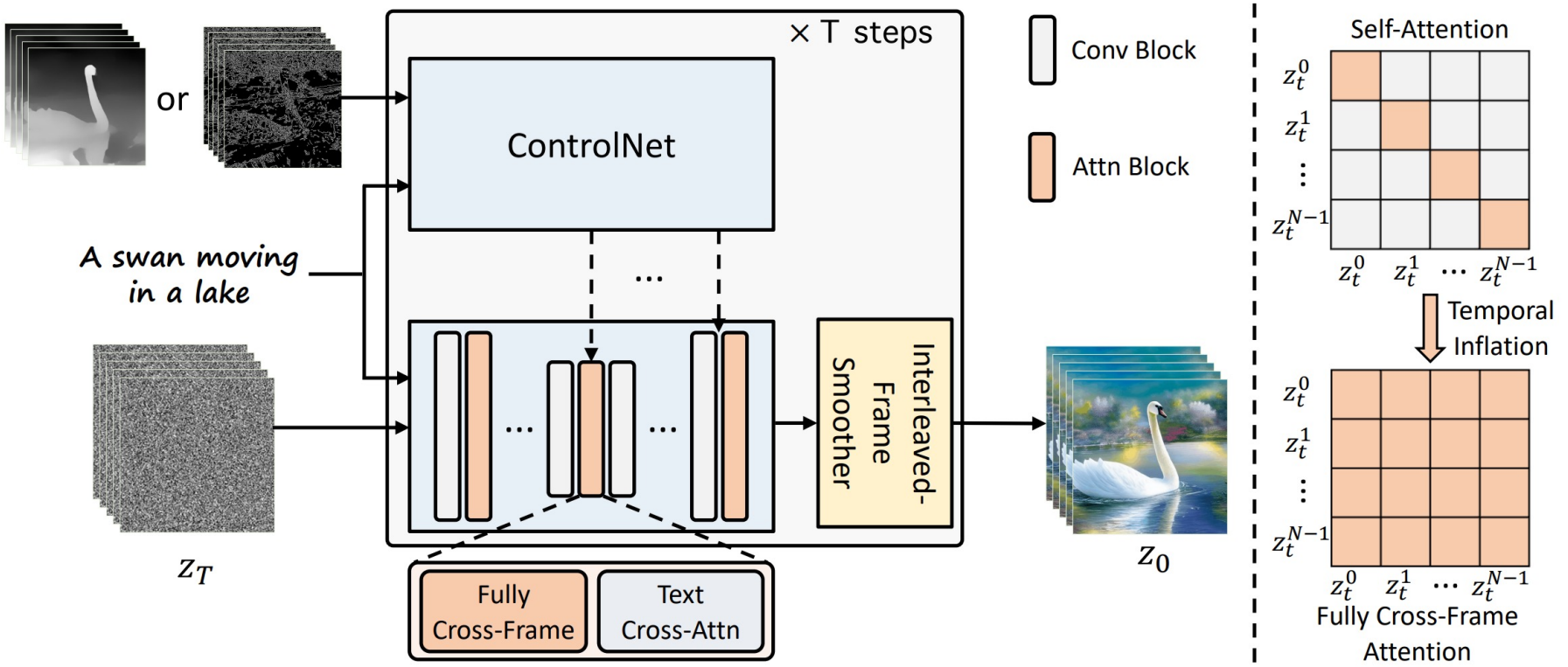
Structural conditions: depth/edges



ControlVideo (Zhang et al. 2023)

ControlNet-like video editing

- Use pretrained weights for Stable Diffusion & ControlNet, no training/finetuning
- Inflate Stable Diffusion and ControlNet along the temporal dimension
- Interleaved-frame smoothing during DDIM sampling for better temporal consistency

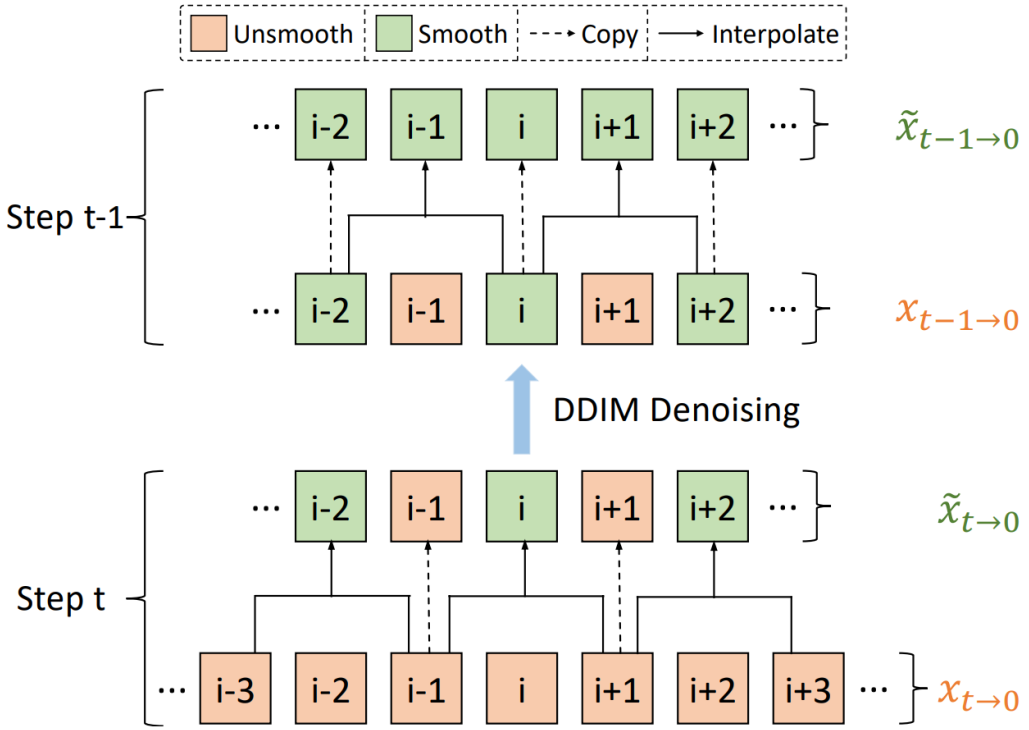


Zhang et al., "ControlVideo: Training-free Controllable Text-to-Video Generation," arXiv 2023.

ControlVideo (Zhang et al. 2023)

ControlNet-like video editing

- Use pretrained weights for Stable Diffusion & ControlNet, no training/finetuning
- Inflate Stable Diffusion and ControlNet along the temporal dimension
- Interleaved-frame smoothing during denoising for better temporal consistency



Zhang et al., "ControlVideo: Training-free Controllable Text-to-Video Generation," arXiv 2023.

ControlVideo (Zhang et al. 2023)

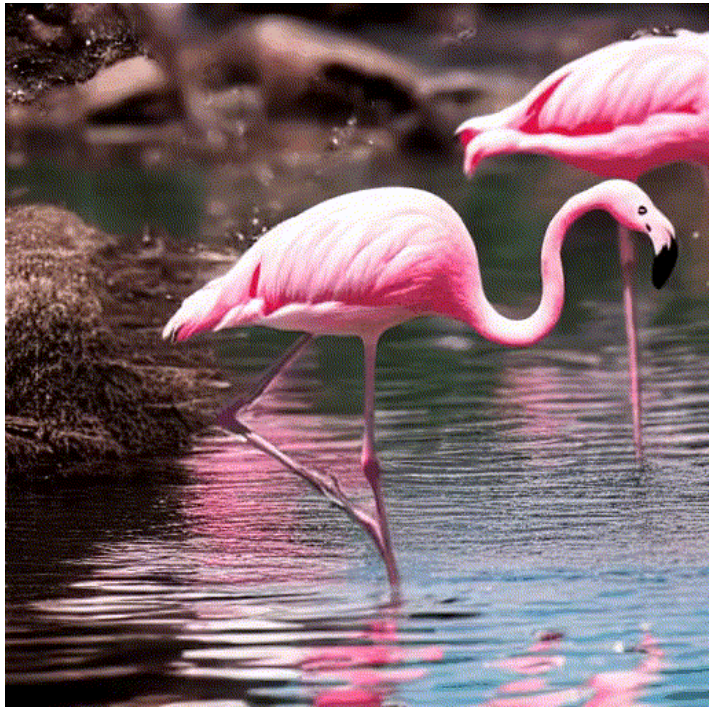
ControlNet-like video editing

Zhang et al., "ControlVideo: Training-free Controllable Text-to-Video Generation," arXiv 2023.

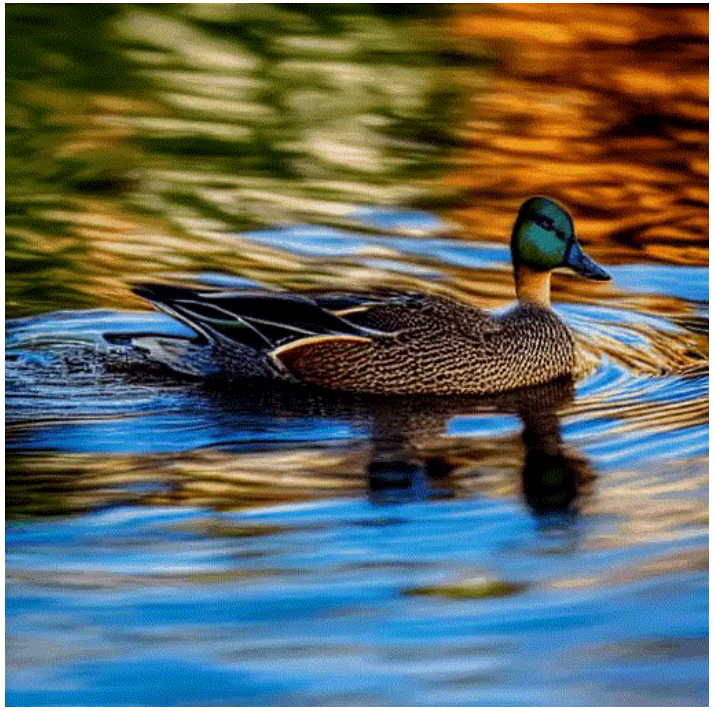
ControlVideo (Zhang et al. 2023)

ControlNet-like video editing

ControlVideo on depth maps



A charming flamingo gracefully wanders in the calm and serene water, its delicate neck curving into an elegant shape.



A striking mallard floats effortlessly on the sparkling pond.

A gigantic yellow jeep slowly turns on a wide, smooth road in the city.

Zhang et al., “ControlVideo: Training-free Controllable Text-to-Video Generation,” arXiv 2023.

ControlVideo (Zhang et al. 2023)

ControlNet-like video editing

ControlVideo on canny edges

A young man riding a sleek, black motorbike through the winding mountain roads.

A white swan movingon the lake, cartoon style.

A dusty old jeep was making its way down the winding forest road, creaking and groaning with each bump and turn.

Zhang et al., “ControlVideo: Training-free Controllable Text-to-Video Generation,” arXiv 2023.

ControlVideo (Zhang et al. 2023)

ControlNet-guided video editing

ControlVideo on human poses

James bond moonwalk on the beach, animation style.

Goku in a mountain range, surreal style.

Hulk is jumping on the street, cartoon style.

Zhang et al., "ControlVideo: Training-free Controllable Text-to-Video Generation," arXiv 2023.

ControlVideo (Zhao et al. 2023)

ControlNet-guided video editing

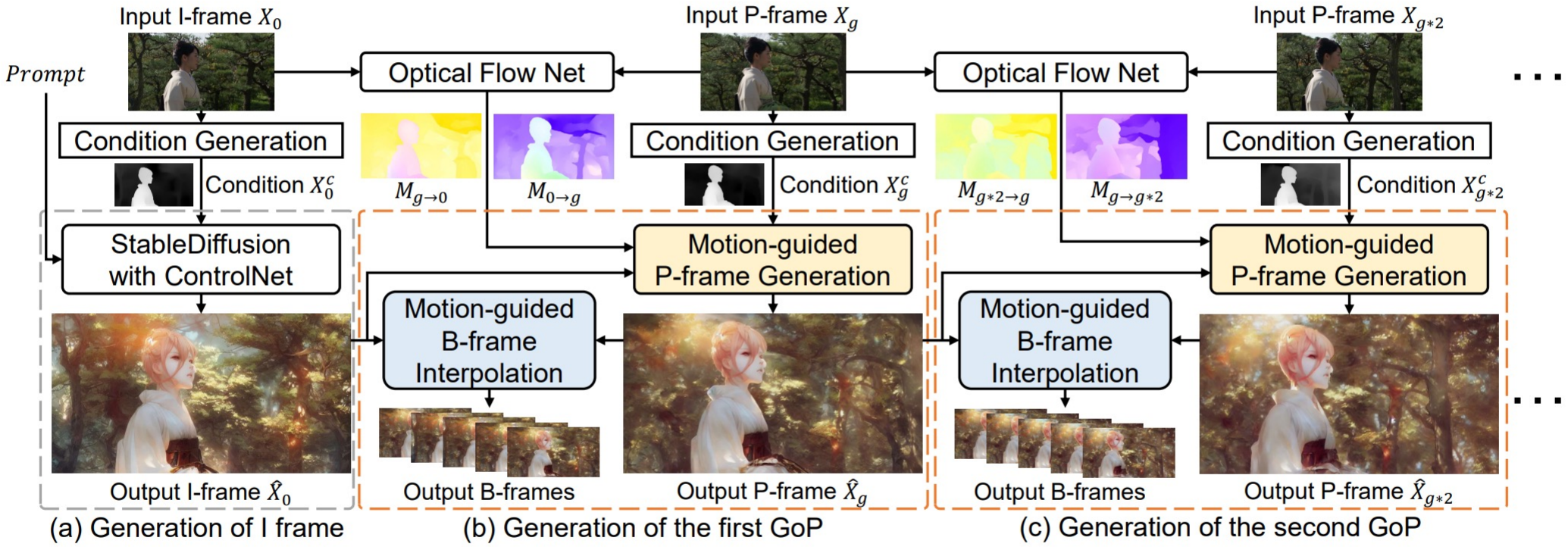
a car, autumn

a car, Vincent van Gough style

Zhao et al., “ControlVideo: Adding Conditional Control for One Shot Text-to-Video Editing,” arXiv 2023.

VideoControlNet

Optical flow-guided video editing; I, P, B frames in video compression



Hu et al., "VideoControlNet: A Motion-Guided Video-to-Video Translation Framework by Using Diffusion Model with ControlNet," arXiv 2023.

Rerender A Video

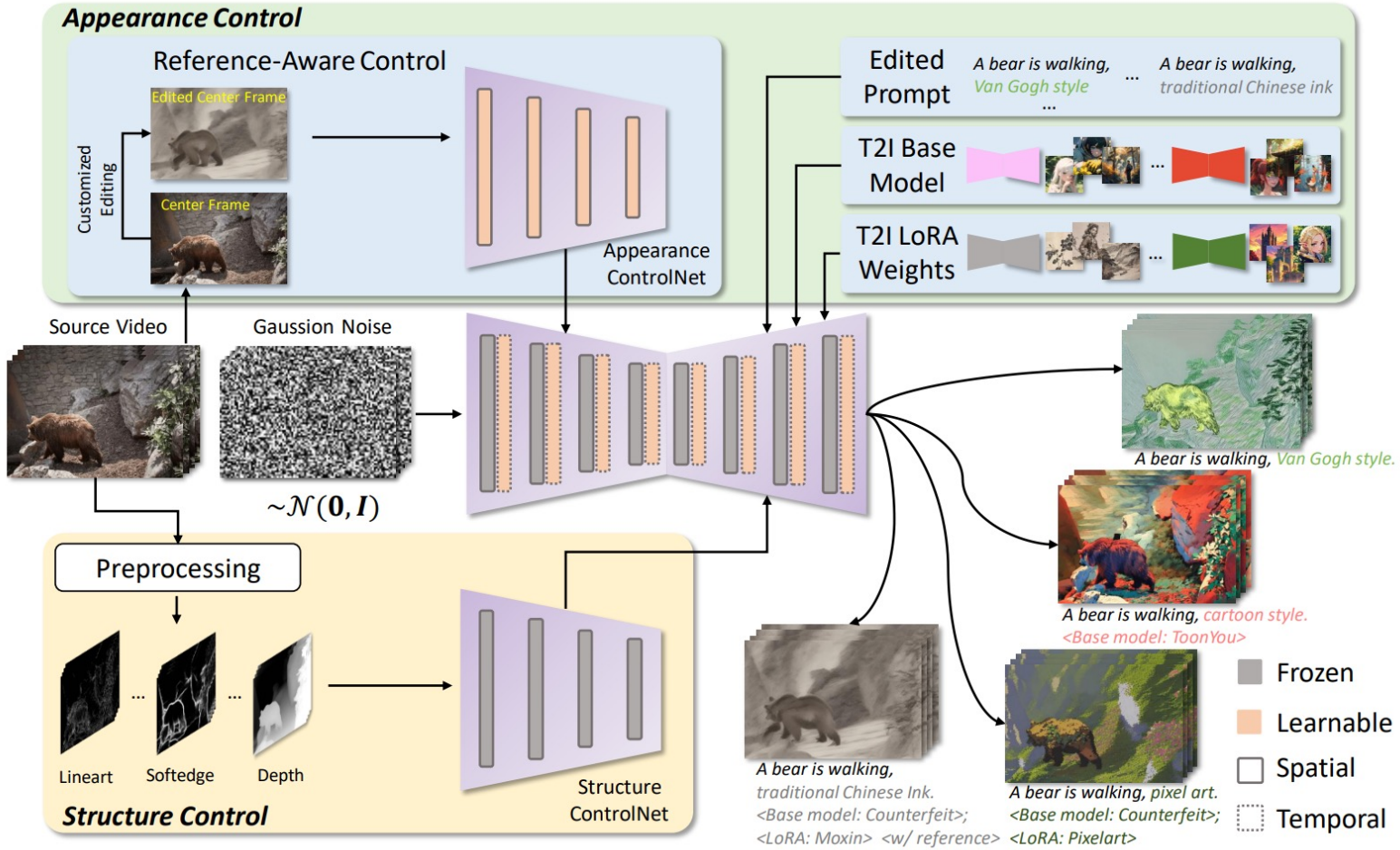
Various techniques for maintaining consistency

Input

Yang et al., “Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation,” SIGGRAPH Asia 2023.

CCEdit

Multimodal-guided video editing

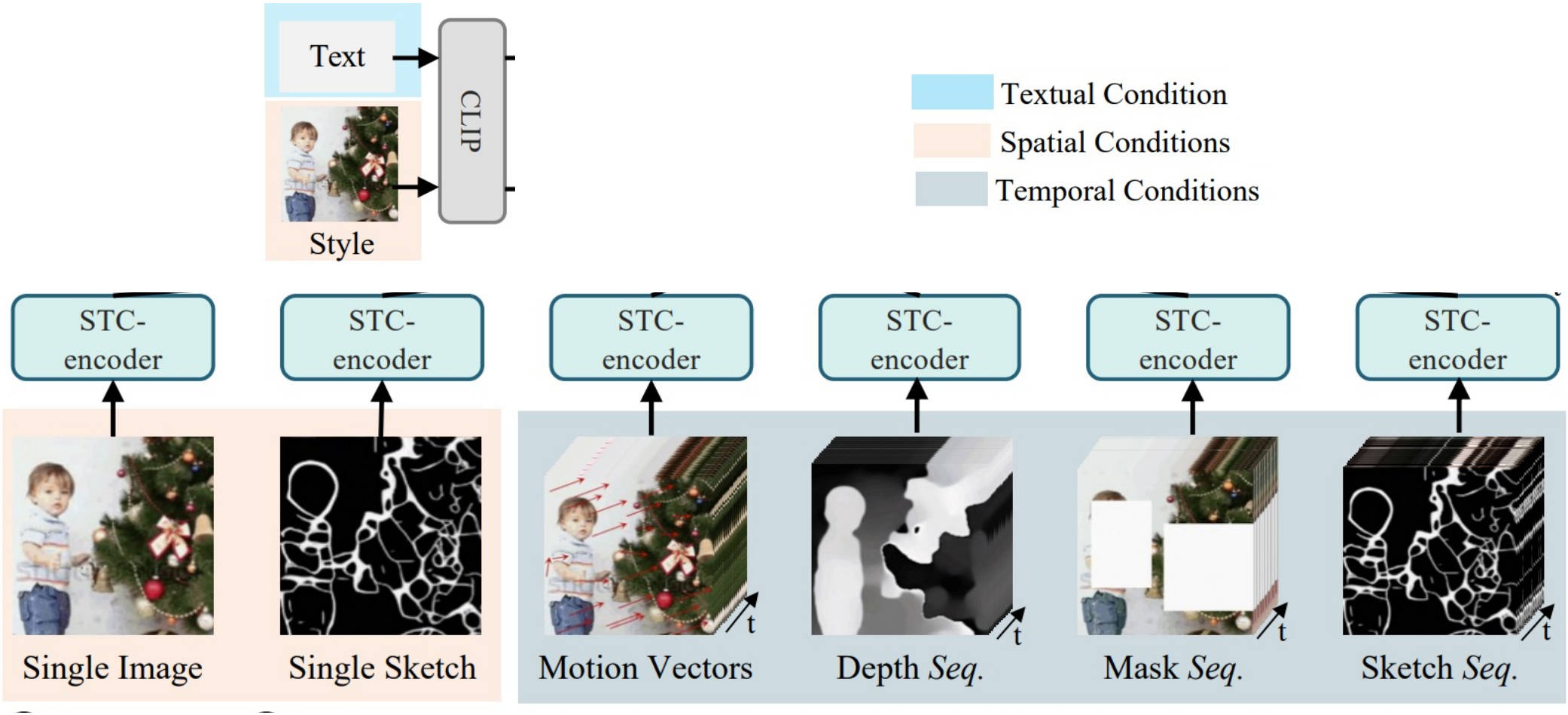


Feng et al., "CCEdit: Creative and Controllable Video Editing via Diffusion Models," arXiv 2023.

VideoComposer

Image-, sketch-, motion-, depth-, mask-controlled video editing

Video Editing based on Various Conditions

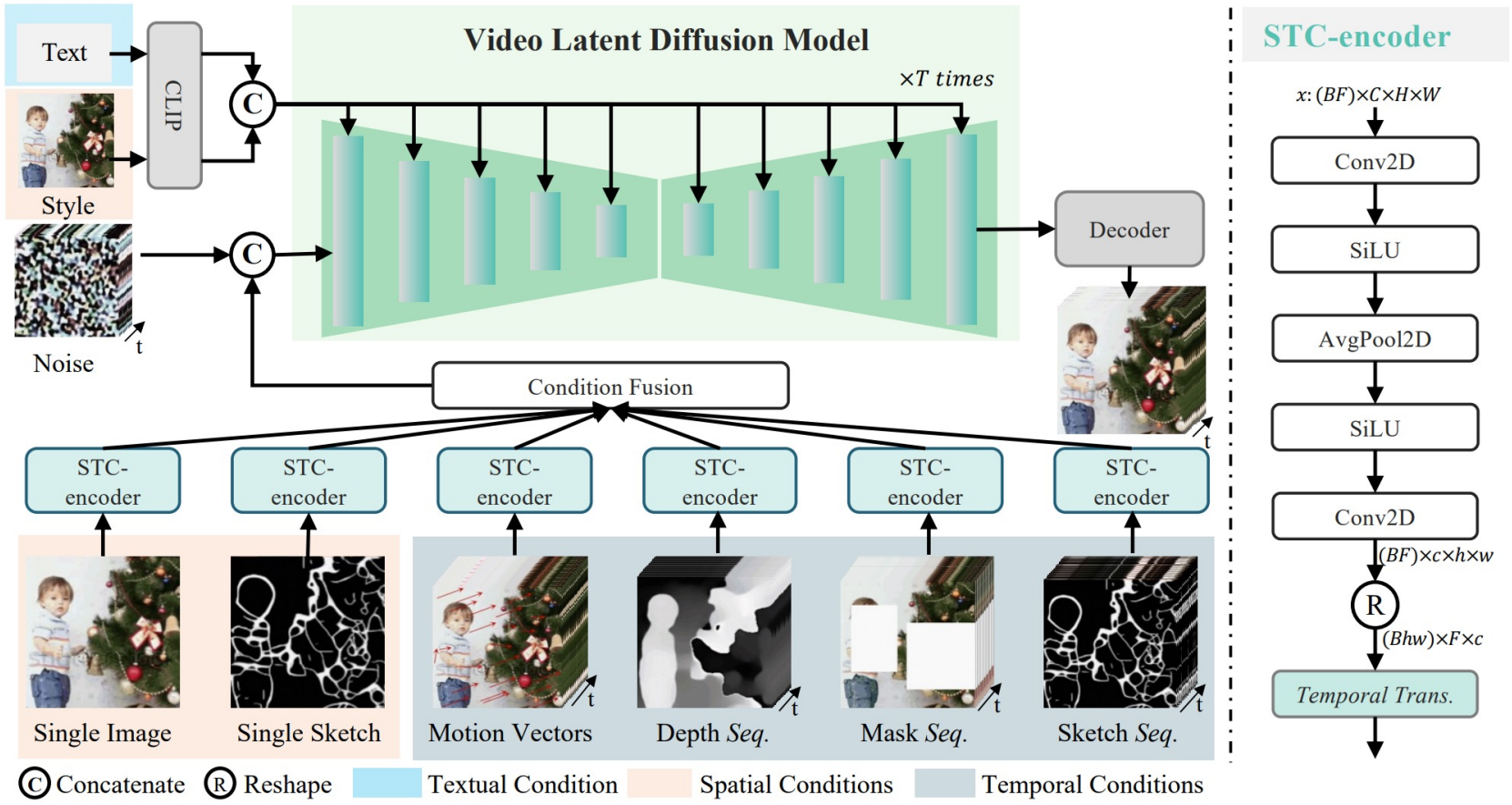


Wang et al., "VideoComposer: Compositional Video Synthesis with Motion Controllability," arXiv 2023.

VideoComposer

Image-, sketch-, motion-, depth-, mask-controlled video editing

- Spatio-Temporal Condition encoder (STC-encoder): a unified input interface for conditions



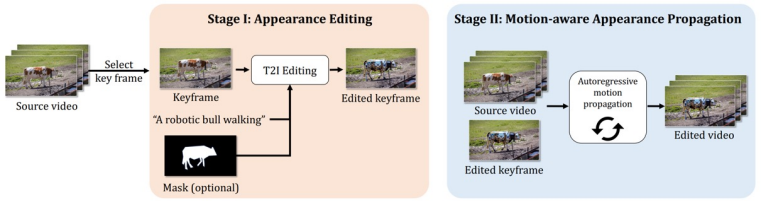
Wang et al., "VideoComposer: Compositional Video Synthesis with Motion Controllability," arXiv 2023.

VideoComposer

Image-, sketch-, motion-, depth-, mask-controlled video editing

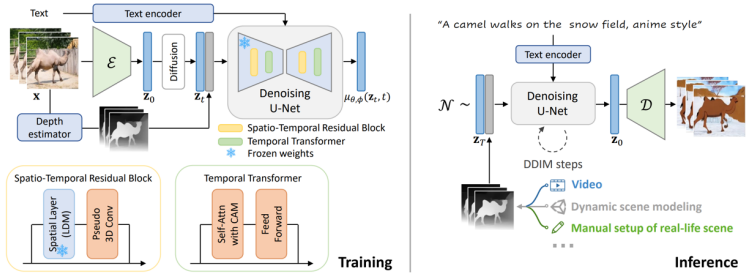
Wang et al., “VideoComposer: Compositional Video Synthesis with Motion Controllability,” arXiv 2023.

ControlNet- and Depth-Controlled Video Editing: More Works



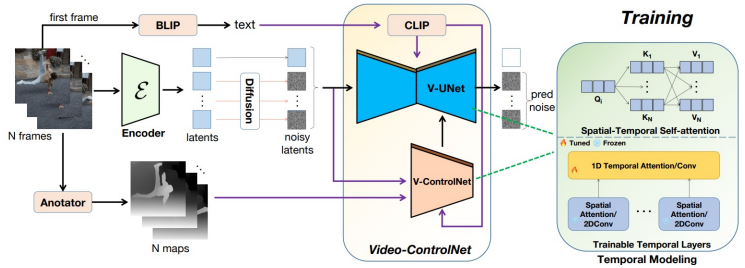
MagicProp (Yan et al.)

"MagicProp: Diffusion-based Video Editing via Motion-aware Appearance Propagation," arXiv 2023.



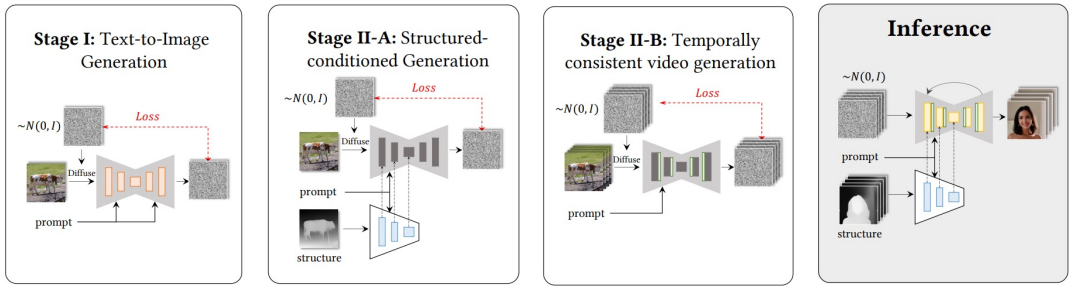
Make-Your-Video (Xing et al.)

"Make-Your-Video: Customized Video Generation Using Textual and Structural Guidance," arXiv 2023.



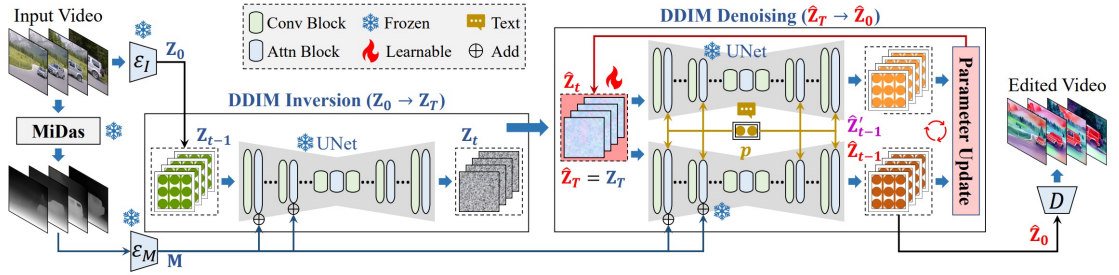
Control-A-Video (Lorem et al.)

"Control-A-Video: Controllable Text-to-Video Generation with Diffusion Models," arXiv 2023.



MagicEdit (Liew et al.)

"MagicEdit: High-Fidelity and Temporally Coherent Video Editing," arXiv 2023.



EVE (Chen et al.)

"EVE: Efficient zero-shot text-based Video Editing with Depth Map Guidance and Temporal Consistency Constraints," arXiv 2023.

Pose Control

DreamPose

Pose- and image-guided video generation

Input: image

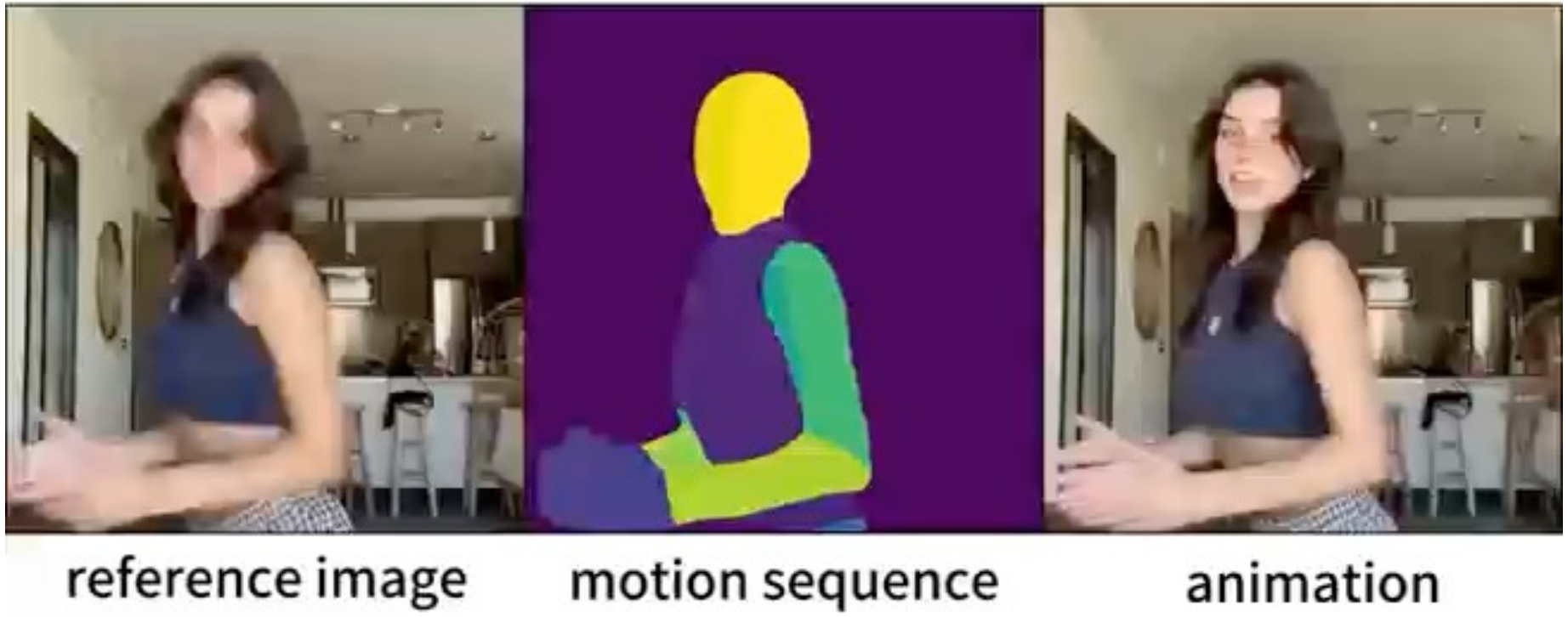
Input: pose sequence

Output: Video

Karras et al., “DreamPose: Fashion Image-to-Video Synthesis via Stable Diffusion,” arXiv 2023.

MagicAnimate

Pose- and image-guided video generation



Xu et al., “MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model,” arXiv 2023.

MagicAnimate

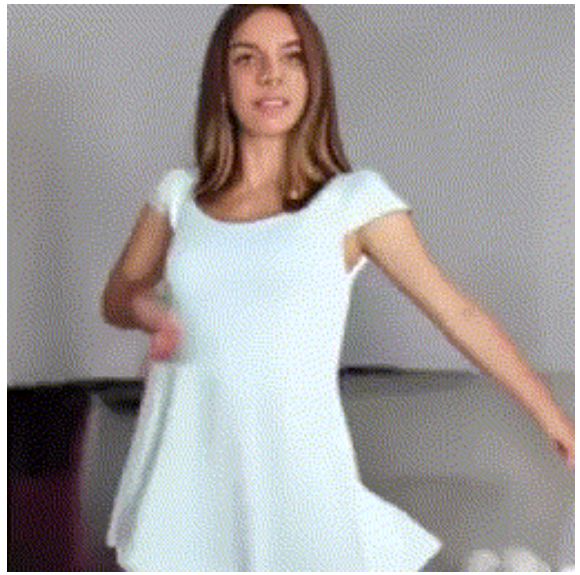
Pose- and image-guided video generation

Challenges

- Flickering video
- Cannot maintain background
- Short video animation results

Possible Cause

- Weak appearance preservation due to lack of temporal modeling

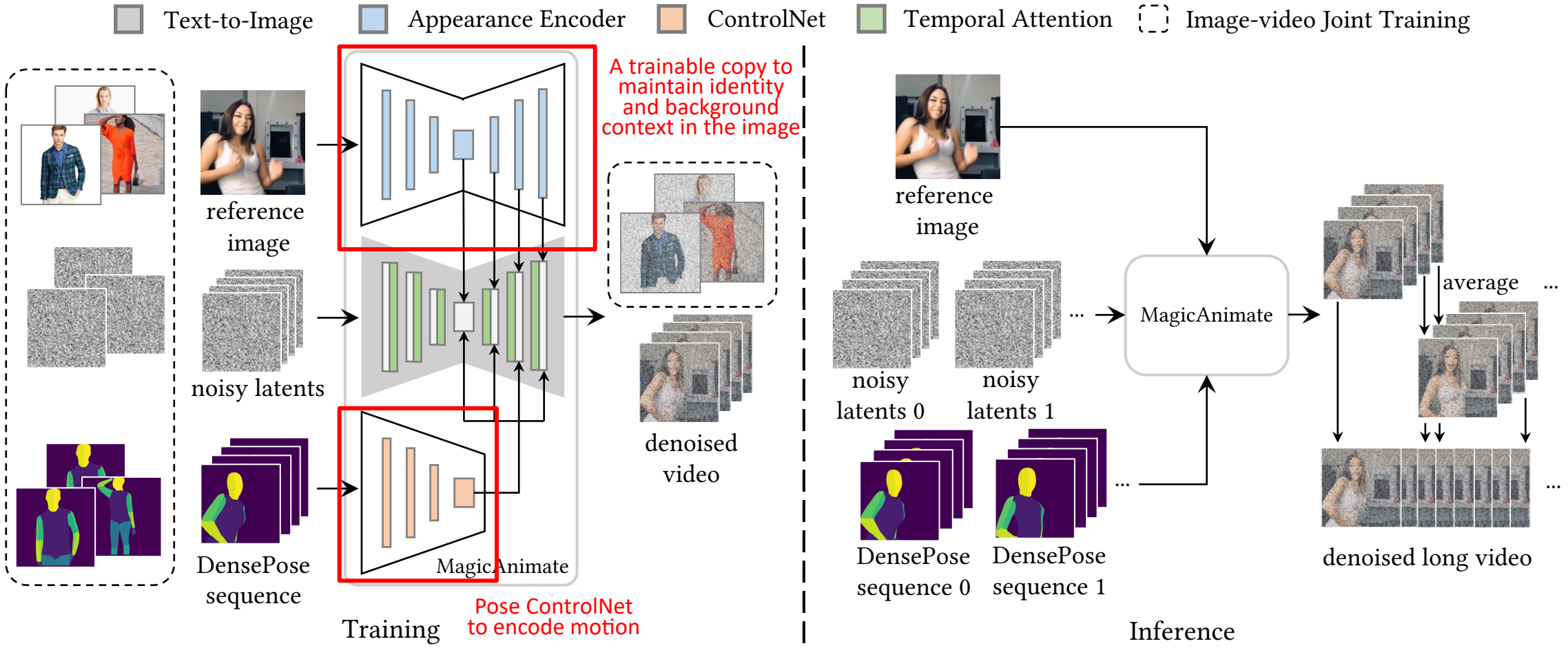


Xu et al., “MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model,” arXiv 2023.

MagicAnimate

Pose- and image-guided video generation

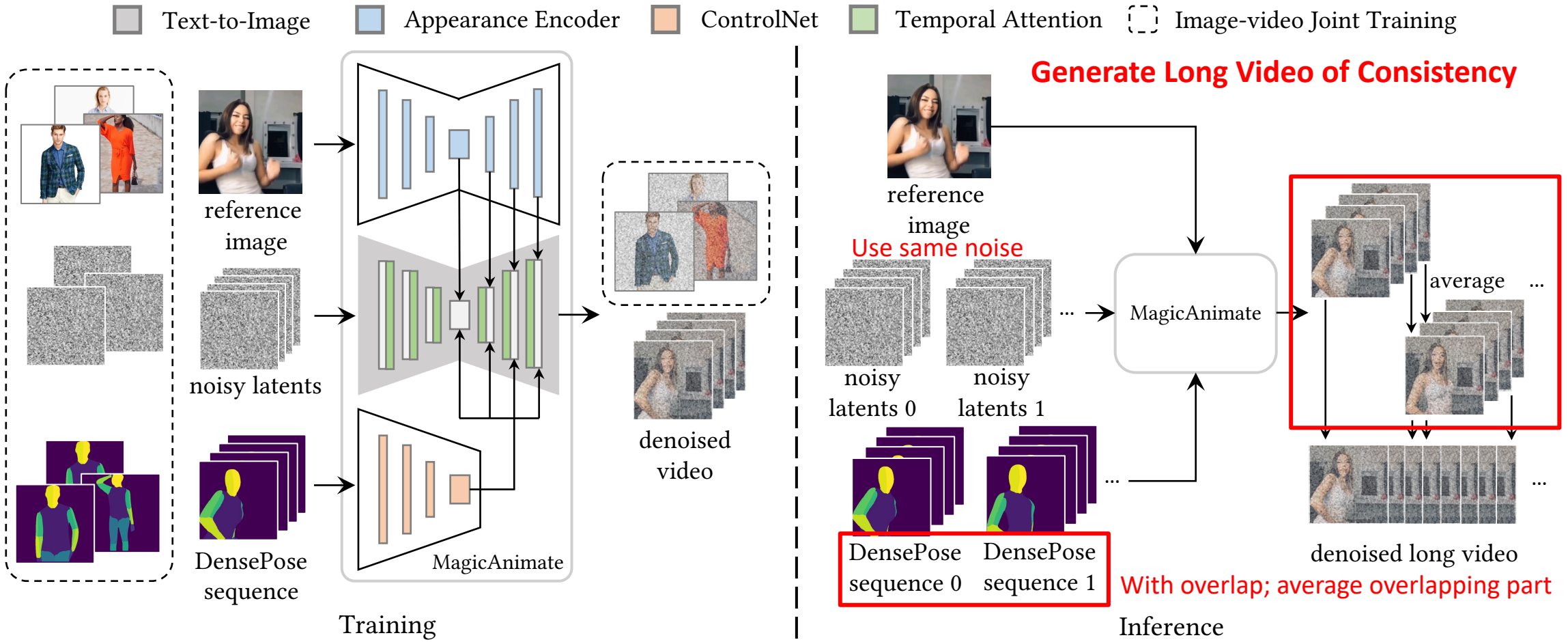
- Training data: TikTok dataset of 350 dancing videos; TED-talks dataset of 1,203 video clips



MagicAnimate

Pose- and image-guided video generation

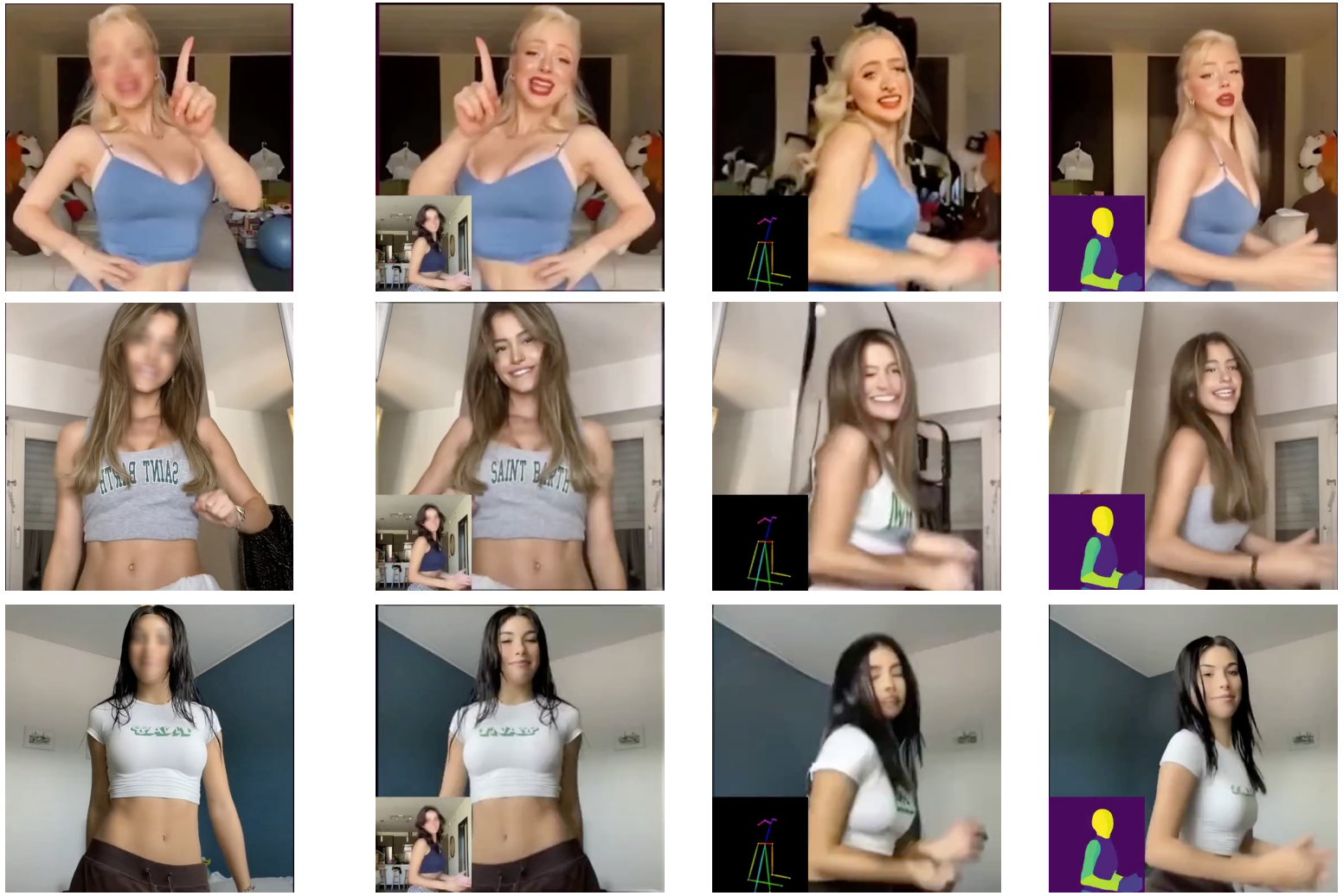
- Once trained, can be used to animate any image



Xu et al., "MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model," arXiv 2023.

MagicAnimate

Pose- and image-guided video generation

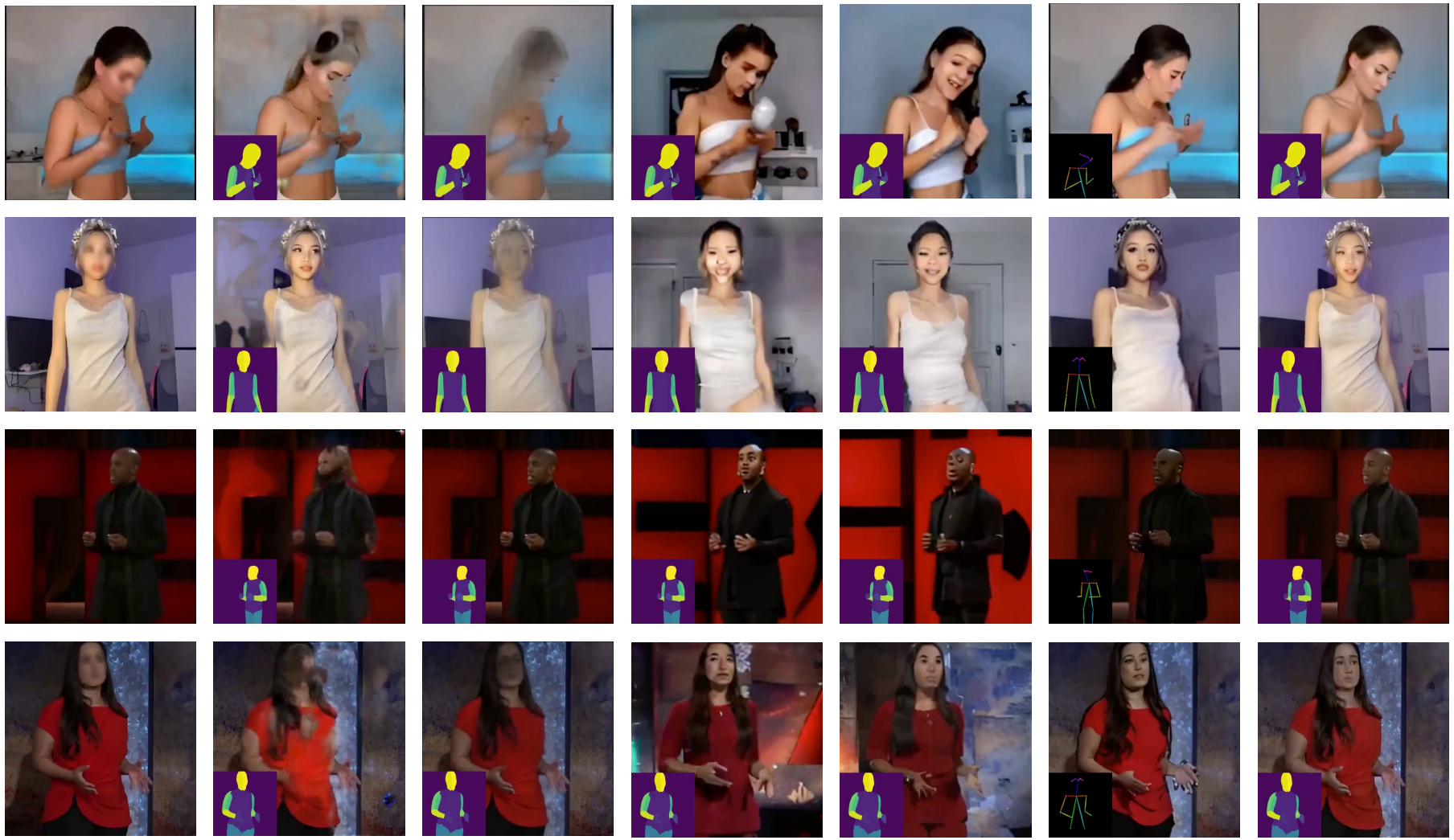


reference | MRAA | DisCo | MagicAnimate (ours)

Xu et al., “MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model,” arXiv 2023.

MagicAnimate

Pose- and image-guided video generation

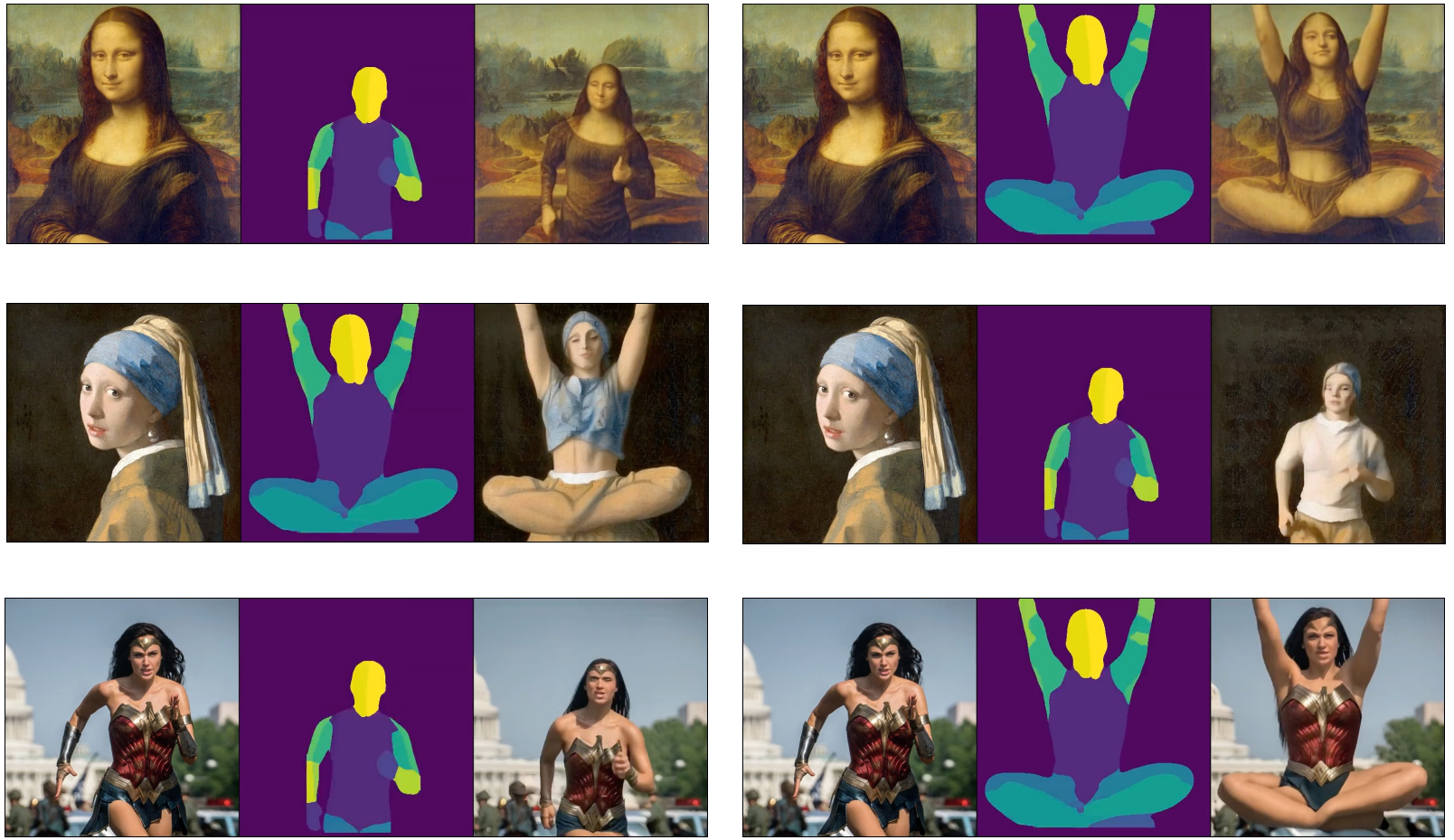


Xu et al., “MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model,” arXiv 2023.

MagicAnimate

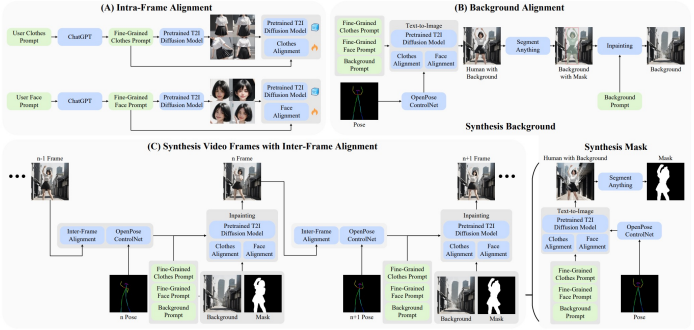
Pose-guided video generation

Animating Unseen Domain Image



Xu et al., “MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model,” arXiv 2023.

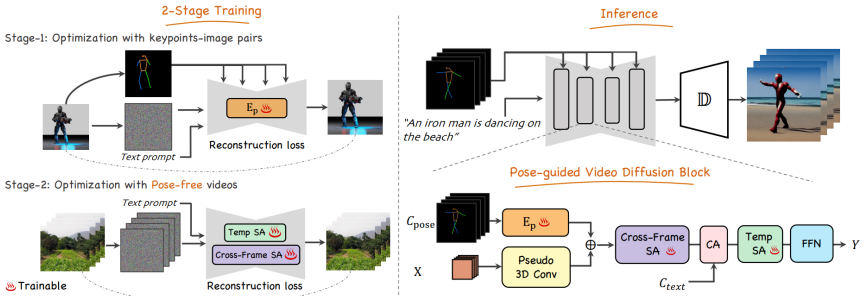
Video Editing Under Pose Guidance: More Works



Dancing Avatar (Qin et al.)

Pose-guided video editing

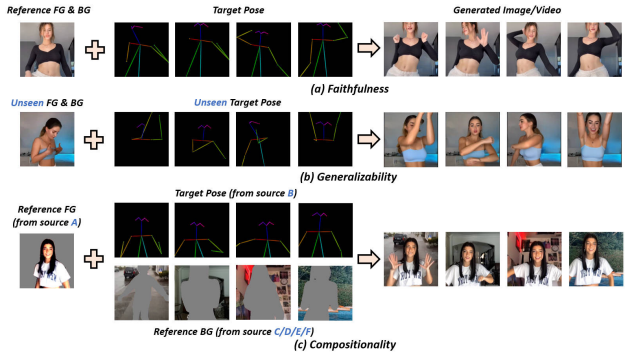
"Dancing avatar: Pose and text-guided human motion videos synthesis with image diffusion model," arXiv 2023.



Follow Your Pose (Ma et al.)

Pose-guided video editing

"Follow Your Pose: Pose-Guided Text-to-Video Generation using Pose-Free Videos," arXiv 2023.



DisCo (Wang et al.)

Pose-guided video editing

"Disco: Disentangled control for referring human dance generation in real world," arXiv 2023.

Point-Control

VideoSwap

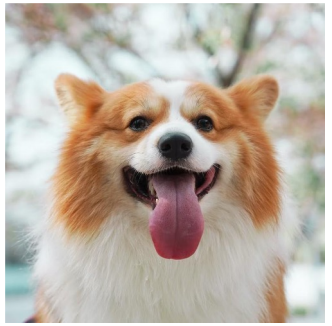
Customized video subject swapping via point control

Problem Formulation

- Subject replacement: change video subject to a **customized** subject
- Background preservation: preserve the unedited background same as the source video

Source Video

Target Subject



Swapped Result by VideoSwap

Gu et al., “VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence,” 2023.

VideoSwap

Customized video subject swapping via point control

Motivation

- Existing methods are promising but still often motion not well aligned
- Need ensure precise correspondence of **semantic points** between the source and target

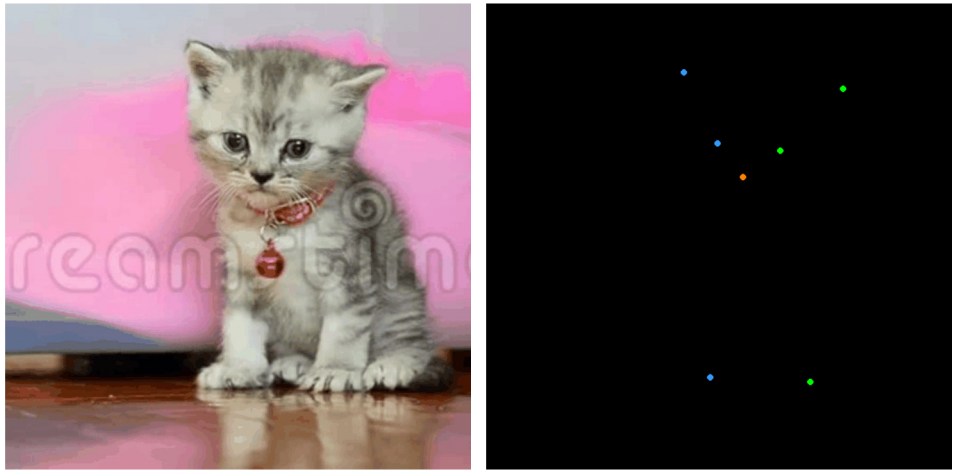
Gu et al., “VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence,” 2023.

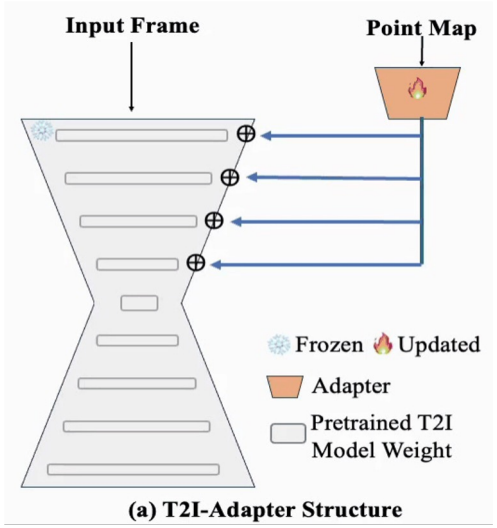
VideoSwap

Customized video subject swapping via point control

Empirical Observations

- **Question**: Can we learn semantic point control for a specific source video subject using only a small number of source video frames
- **Toy Experiment**: Manually define and annotate a set of semantic points on 8 frame; use such point maps as condition for training a control net, i.e., T2I-Adapter.





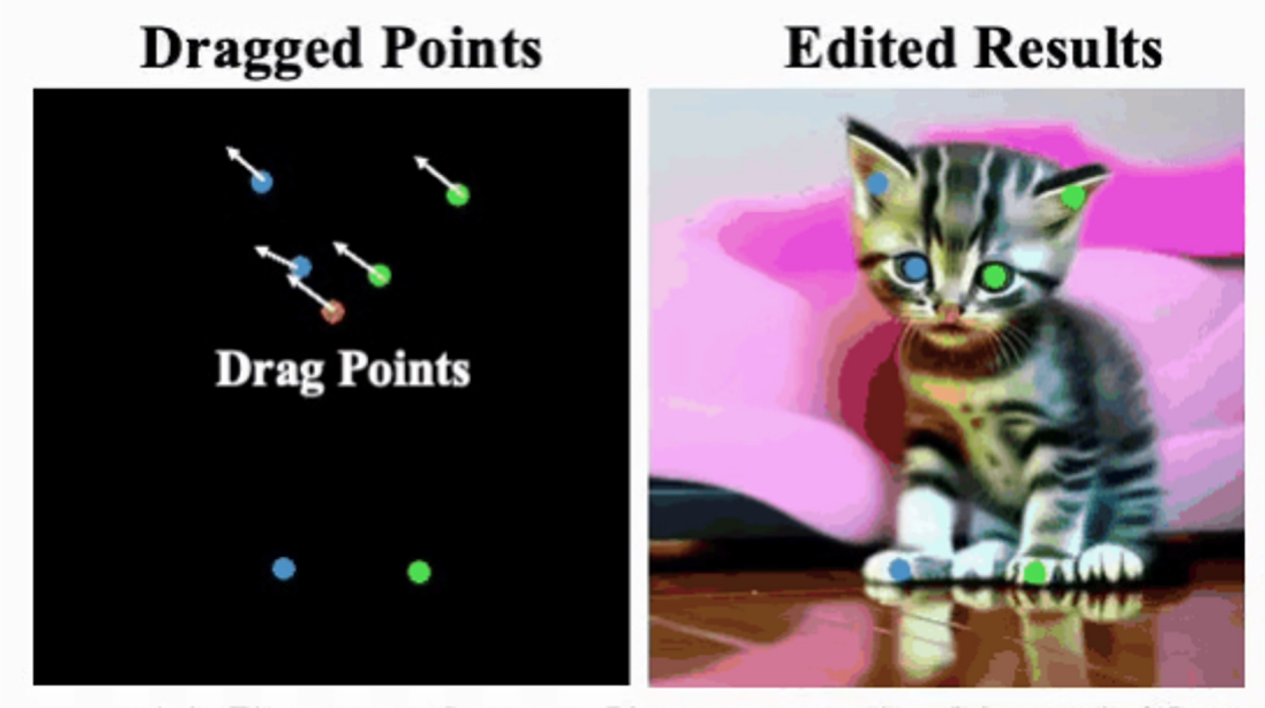
Gu et al., "VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence," 2023.

VideoSwap

Customized video subject swapping via point control

Empirical Observations

- **Observation 1:** If we can drag the points, the trained T2I-Aapter can generate new contents based on such dragged new points (new condition) → feasible to use semantic points as condition to control and maintain the source motion trajectory.



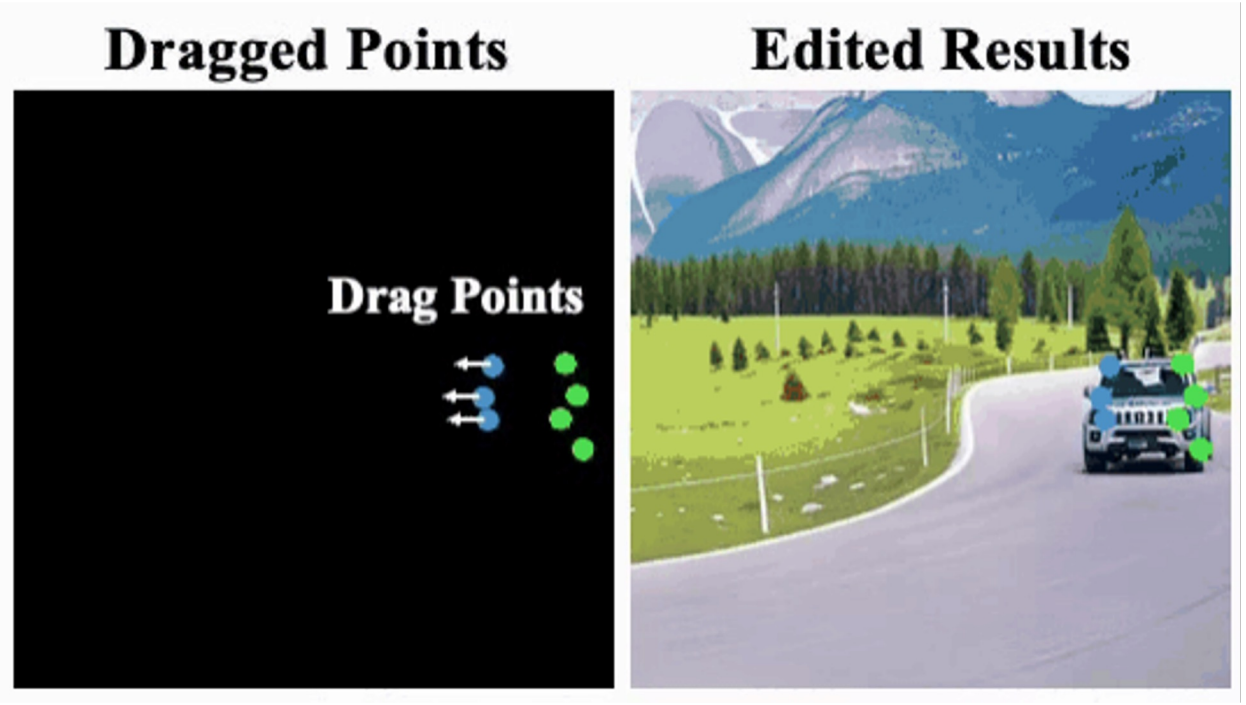
Gu et al., "VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence," 2023.

VideoSwap

Customized video subject swapping via point control

Empirical Observations

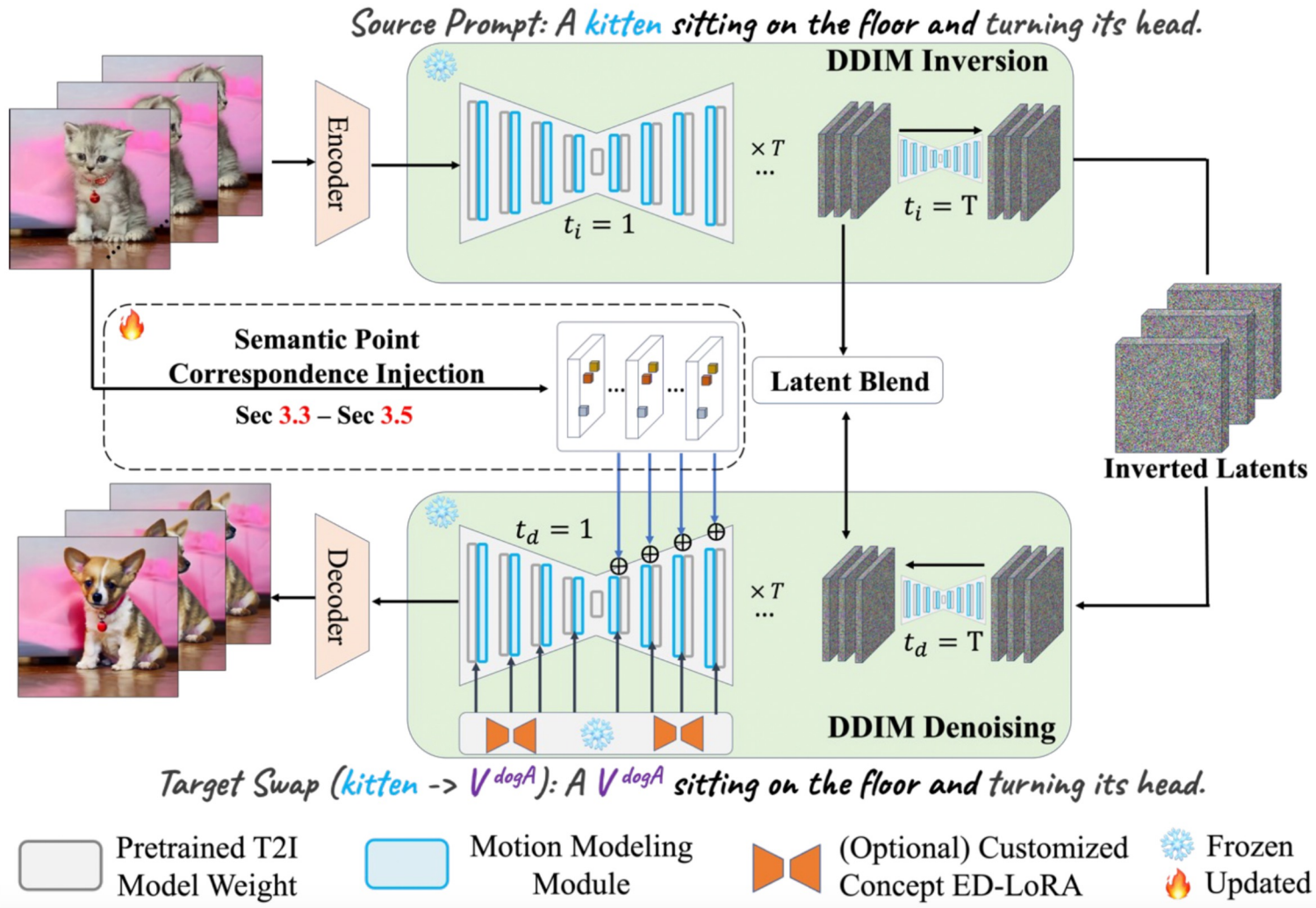
- **Observation 2:** Further, we can drag the semantic points to control the subject's shape



Gu et al., "VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence," 2023.

VideoSwap

Customized video subject swapping via point control



Framework

- **Motion layer**: use pretrained and fixed AnimateDiff to ensure essential temporal consistency
- **ED-LoRA** (Mix-of-Show): learn the concept to be customized

- **Key design aims:**
  - Introduce semantic point correspondences to guide motion trajectory
  - Reduce human efforts of annotating points

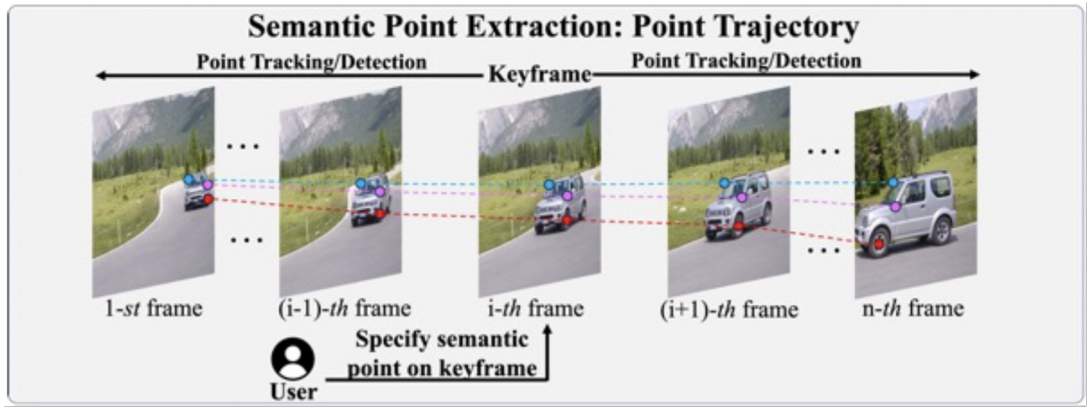
Gu et al. "VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence," 2023.
Gu et al. "Mix-of-Show: Decentralized Low-Rank Adaptation for Multi-Concept Customization of Diffusion Models." NeurIPS, 2023.

VideoSwap

Customized video subject swapping via point control

Step 1: Semantic Point Extraction

- Reduce human efforts in annotating points
  - User define point at one keyframe
  - Propagate to other frames by point tracking/detector
- Embedding



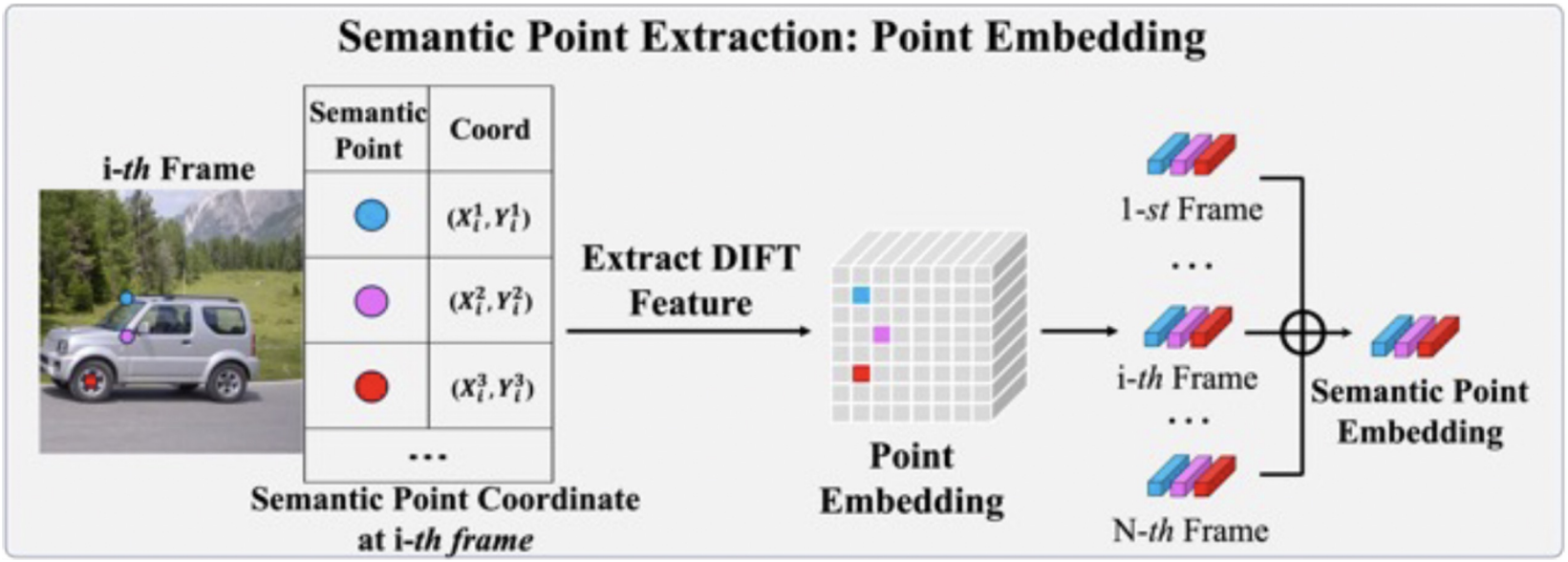
Gu et al., "VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence," 2023.

VideoSwap

Customized video subject swapping via point control

Methodology – Step 1: Semantic Point Extraction on the source video

- Reduce human efforts in annotating points
- Embedding
  - Extract DIFT embedding (intermediate U-Net feature) for each semantic point
  - Aggregate over all frames



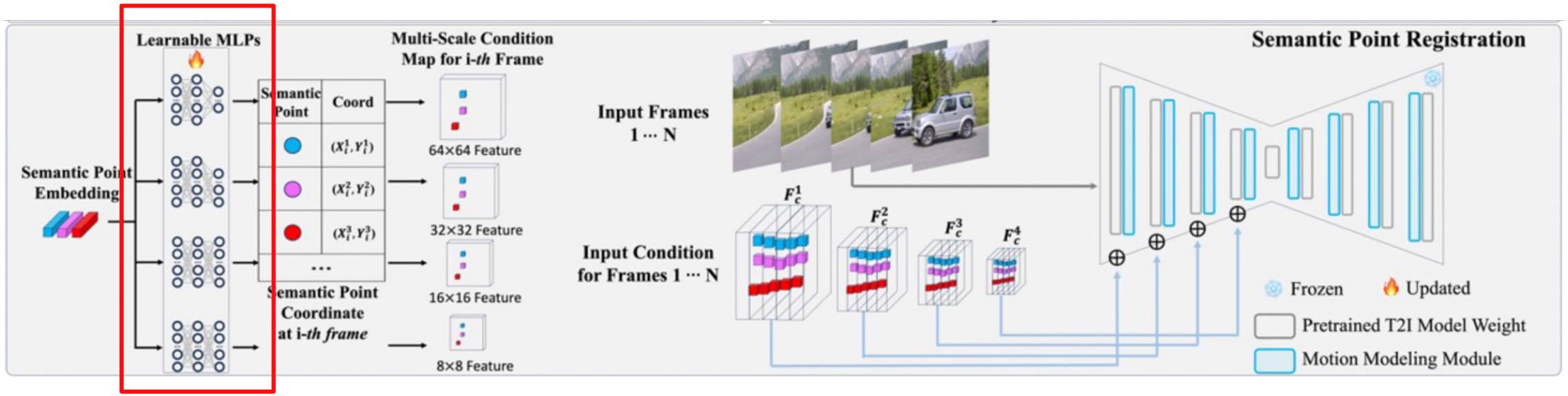
Gu et al., "VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence," 2023.

VideoSwap

Customized video subject swapping via point control

Methodology – Step 2: Semantic Point Registration on the source video

- Introduce several learnable MLPs, corresponding to different scales
- Optimize the MLPs
  - Point Patch Loss: restrict diffusion loss to reconstruct local patch around the point
  - Semantic-Enhanced Schedule: only sample higher timestep (0.5T, T), which prevents overfitting to low-level details



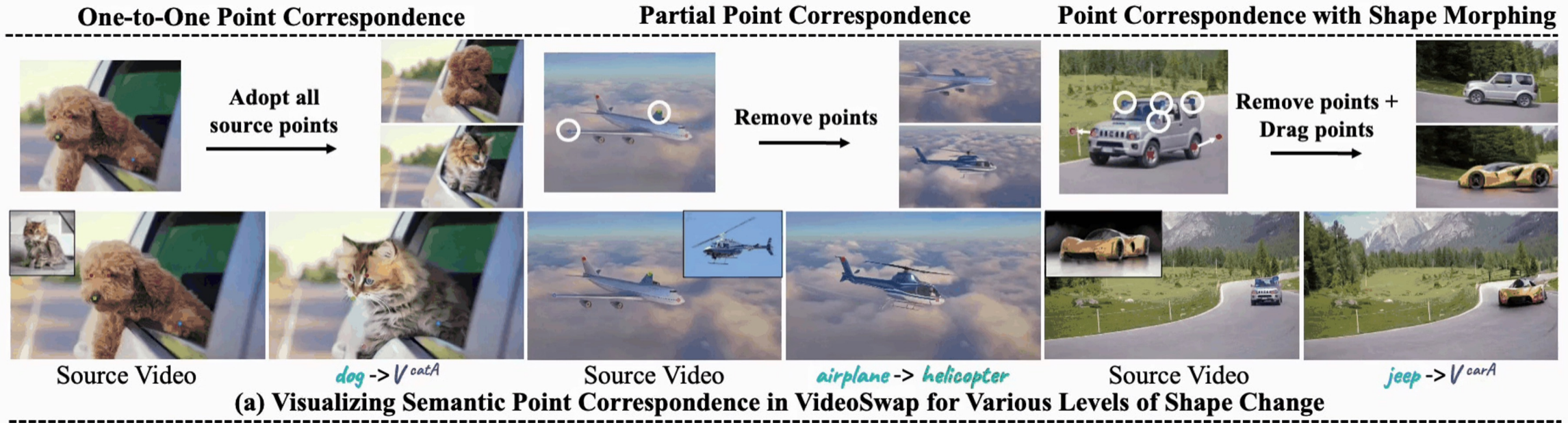
Gu et al., "VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence," 2023.

VideoSwap

Customized video subject swapping via point control

Methodology

- After Step1 (Semantic Point Extraction) and Step2 (Semantic Point Registration), those semantic points can be used to guide motion
- User-point interaction for various applications



(a) Visualizing Semantic Point Correspondence in VideoSwap for Various Levels of Shape Change

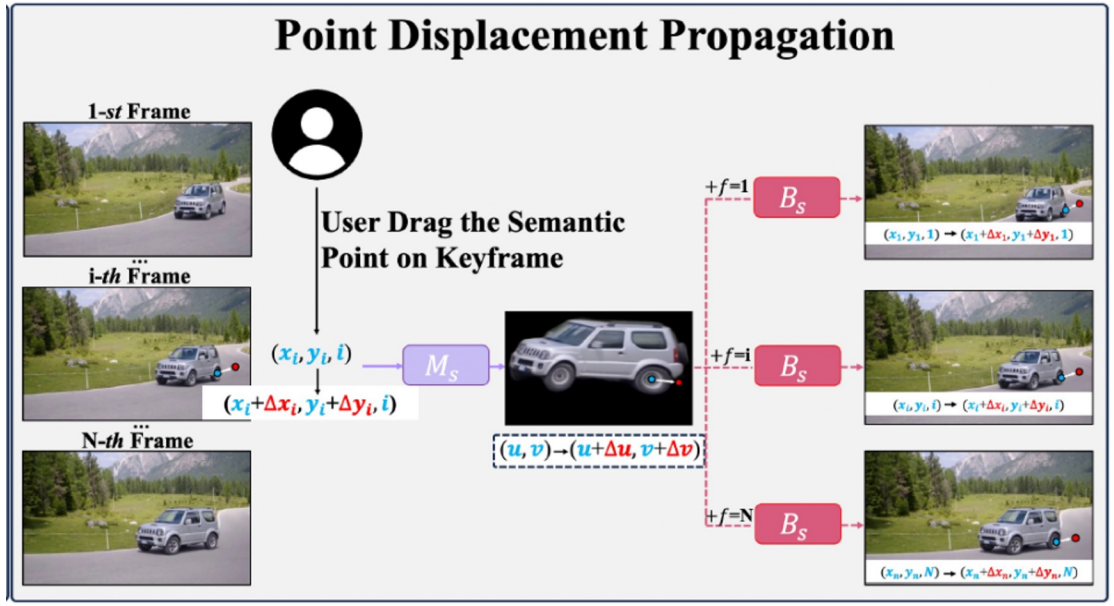
Gu et al., "VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence," 2023.

VideoSwap

Customized video subject swapping via point control

Methodology

- How to drag point for shape change?
  - Dragging at one frame is straightforward, propagating drag displacement over time is non-trivial, because of complex camera motion and subject motion in video.
  - Resort to canonical space (i.e., Layered Neural Atlas) to propagate displacement.



Gu et al., "VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence," 2023.

VideoSwap

Customized video subject swapping via point control

Methodology

- How to drag point for shape change?
  - Dragging at one frame is straightforward, propagating drag displacement over time is non-trivial because of complex camera motion and subject motion in video.
  - Resort to canonical space (i.e., Layered Neural Atlas) to propagate displacement.

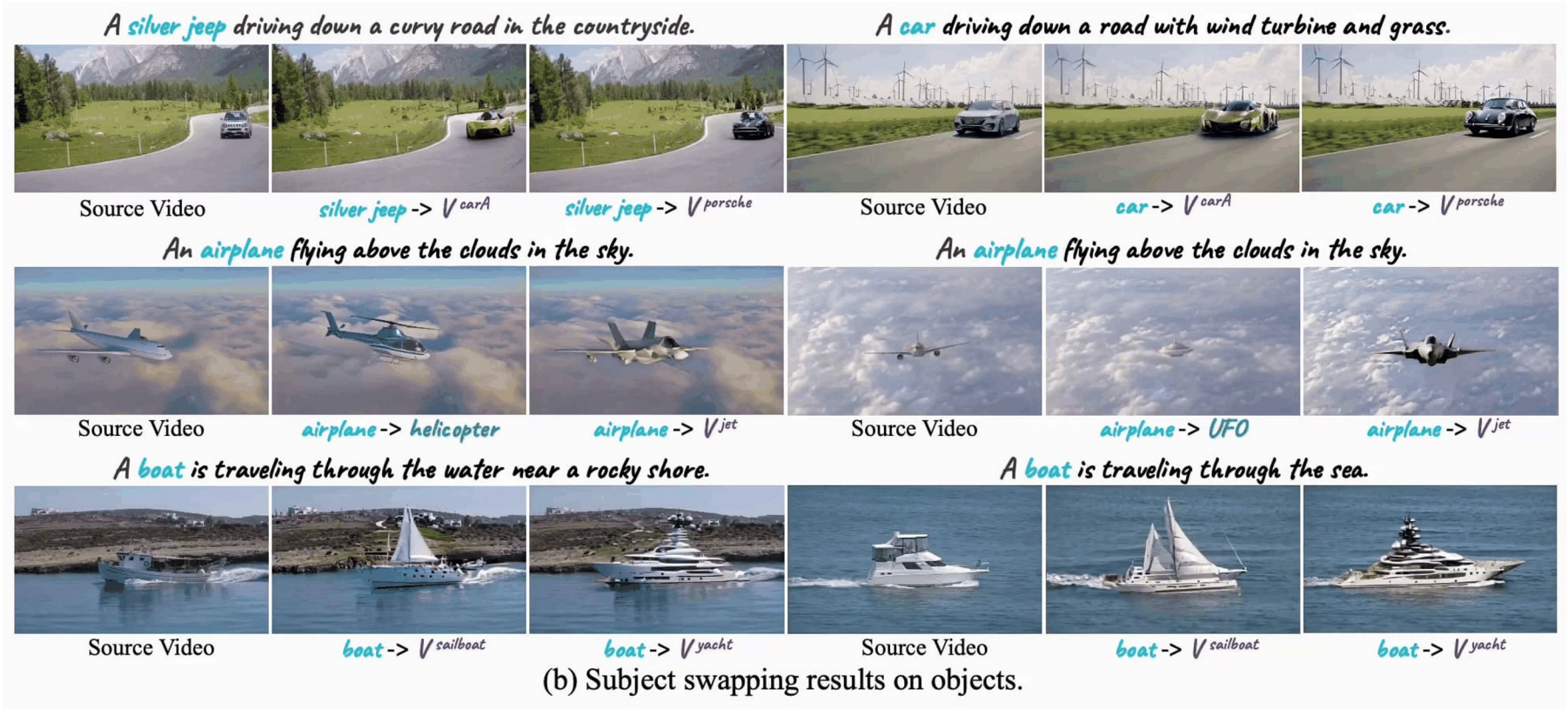
(d) Ablation Study of Drag-based Point Control

Gu et al., "VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence," 2023.

VideoSwap

Customized video subject swapping via point control

Experiments in swapping objects



(b) Subject swapping results on objects.

Gu et al., "VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence," 2023.

VideoSwap

Customized video subject swapping via point control

Experiments in swapping animals

(a) Subject swapping results on animals.

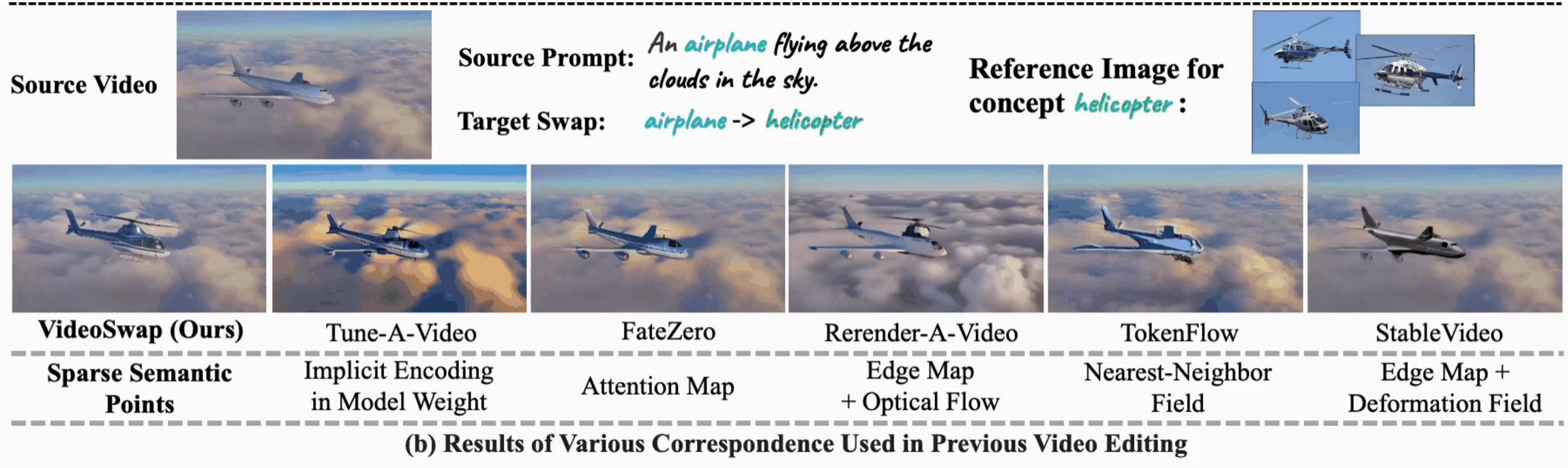
Gu et al., "VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence," 2023.

VideoSwap

Customized video subject swapping via point control

Qualitative Comparisons to previous works

- VideoSwap can **support shape change** in the target swap results, leading to the correct identity of target concept.



Gu et al., "VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence," 2023.

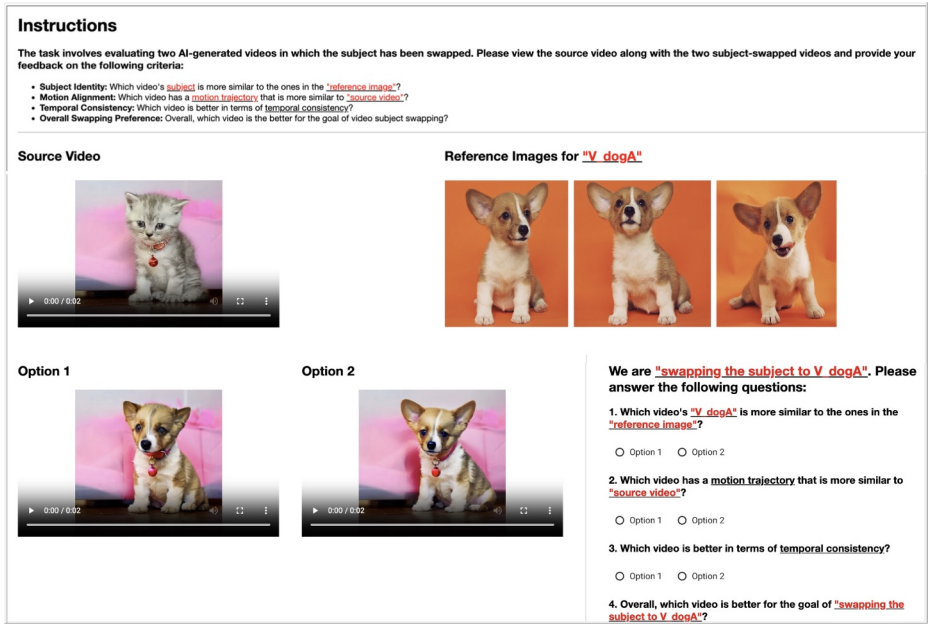
VideoSwap

Customized video subject swapping via point control

Human Evaluation

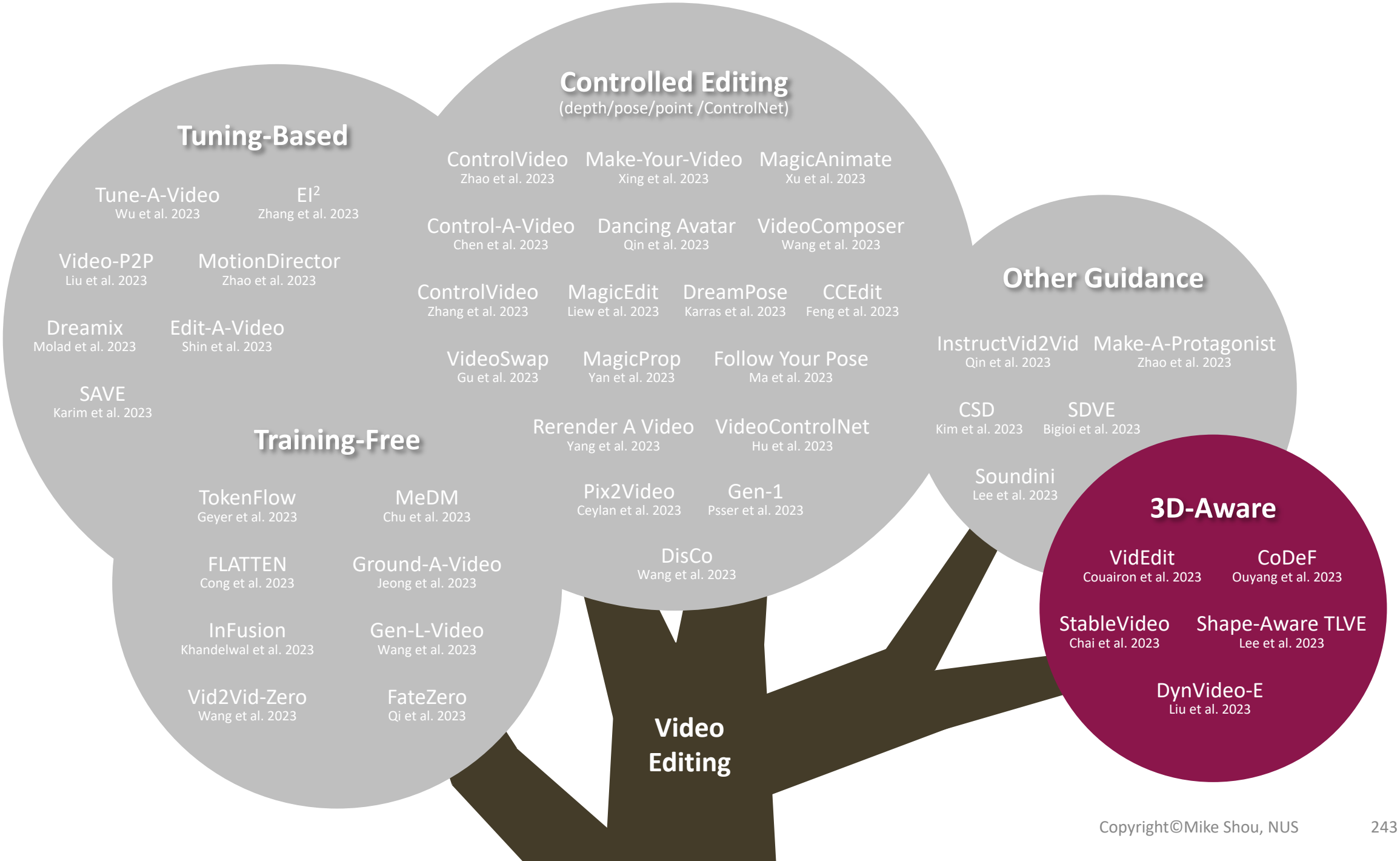
	Methods/Metrics	Subject Identity	Motion Alignment	Temporal Consistency	Overall Preference
	Compare to Previous Video Editing Methods				
VideoSwap v.s.	Tune-A-Video	**80%** v.s. 20%	**72%** v.s. 28%	**80%** v.s. 20%	**78%** v.s. 22%
	FateZero	**74%** v.s. 26%	**67%** v.s. 33%	**73%** v.s. 27%	**70%** v.s. 30%
	Text2Video-Zero	**76%** v.s. 24%	**71%** v.s. 29%	**80%** v.s. 20%	**80%** v.s. 20%
	Rerender-A-Video	**81%** v.s. 19%	**72%** v.s. 28%	**84%** v.s. 16%	**84%** v.s. 16%
	Compare to Baselines on AnimateDiff				
	w/ DDIM	**70%** v.s. 30%	**74%** v.s. 26%	**69%** v.s. 31%	**73%** v.s. 27%
	w/ DDIM+TAV	**68%** v.s. 32%	**60%** v.s. 40%	**66%** v.s. 34%	**69%** v.s. 31%
	w/ DDIM+T2I-Adapter	**66%** v.s. 34%	**59%** v.s. 41%	**65%** v.s. 35%	**66%** v.s. 34%

Table 1. Human Evaluation on Video Subject Swapping Results.



3 Video Editing

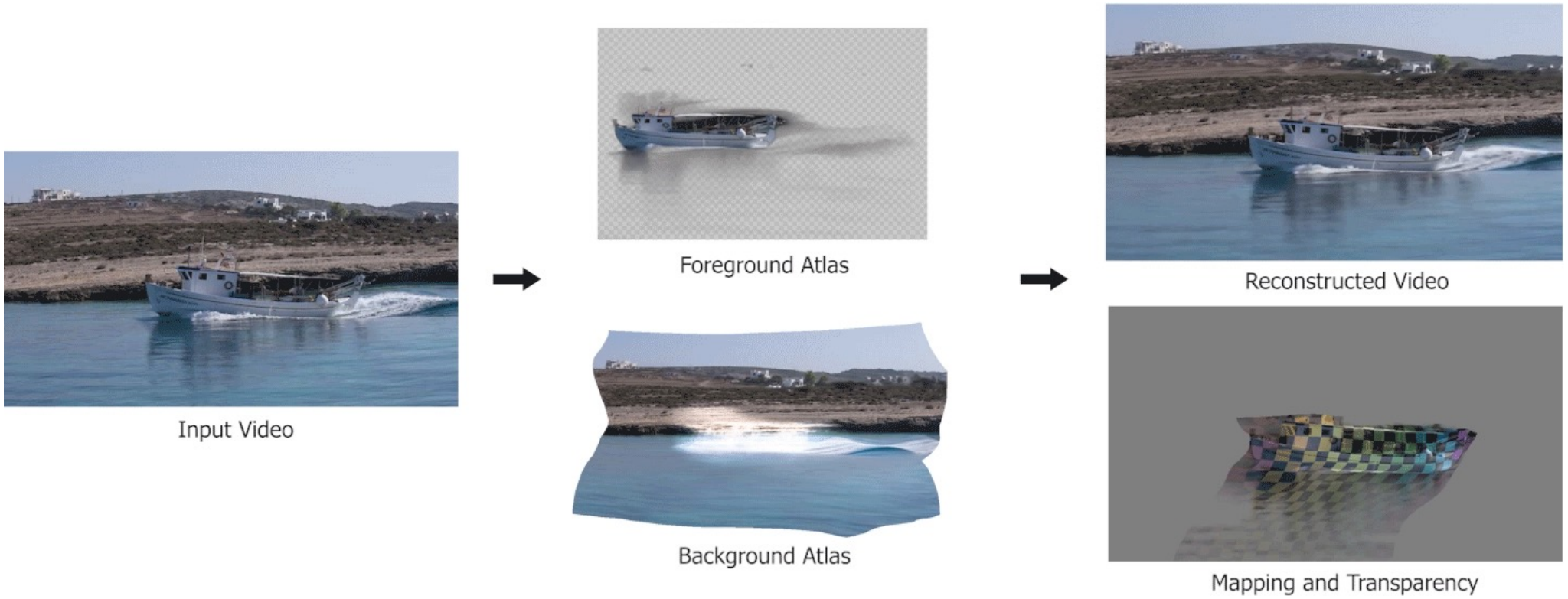
3.4 3D-Aware



Layered Neural Atlases

Decompose a video into two images

- Decompose a video into a foreground image + a background image

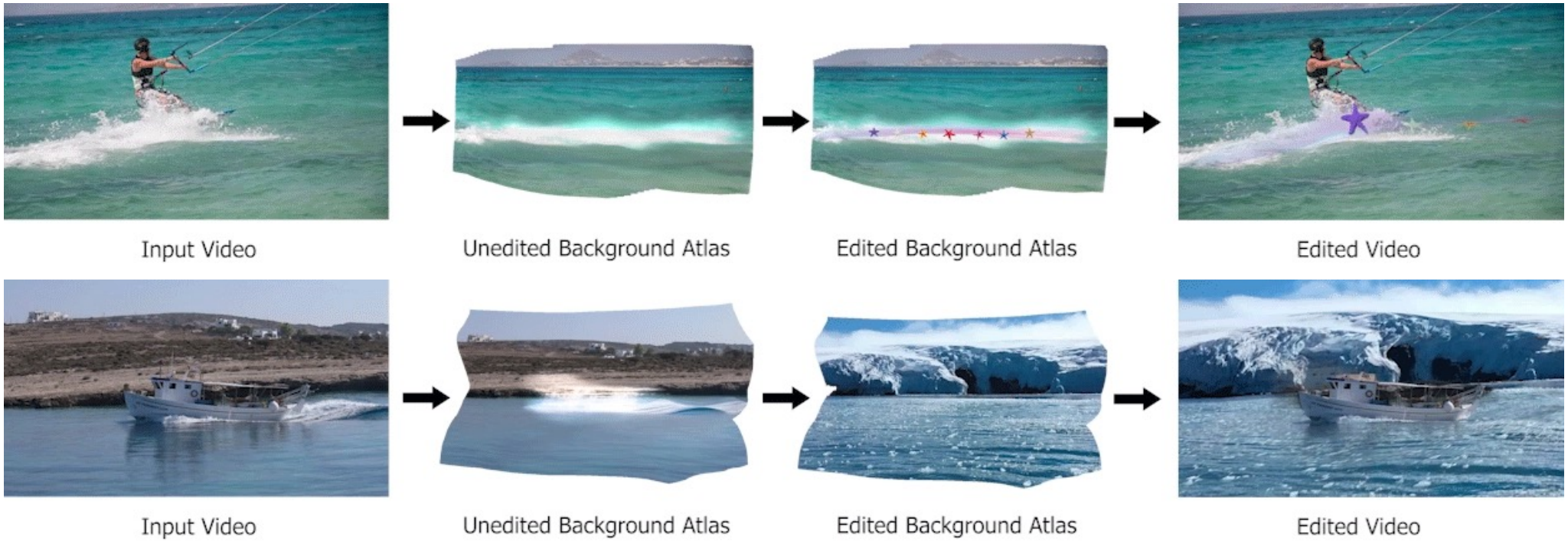


Kasten et al., "Layered Neural Atlases for Consistent Video Editing," arXiv 2023.

Layered Neural Atlases

Decompose a video into two images

- Decompose a video into a foreground image + a background image
- Edit the foreground/background image = edit the video

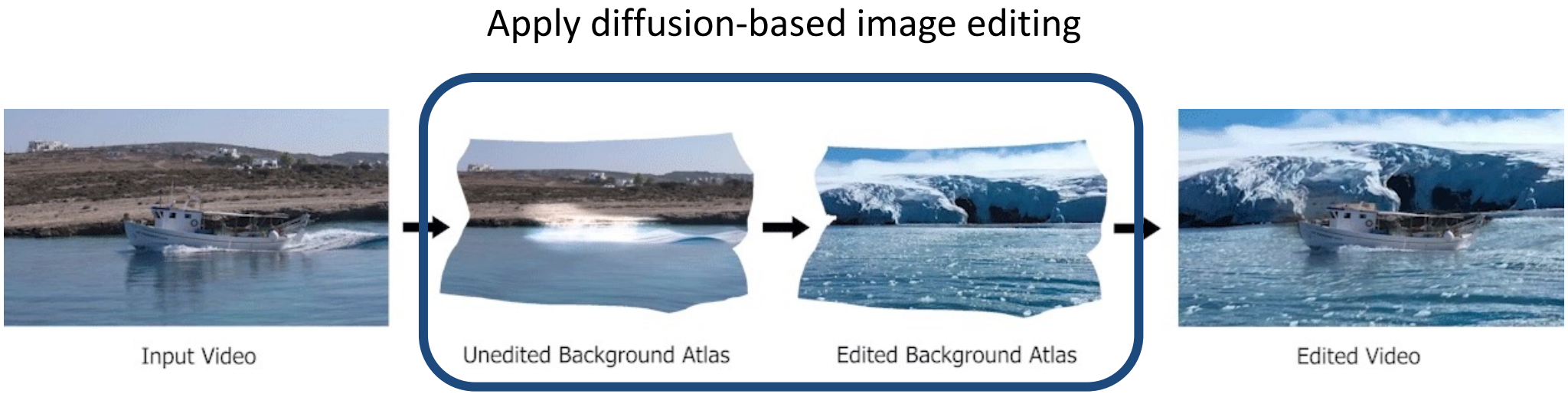


Kasten et al., "Layered Neural Atlases for Consistent Video Editing," arXiv 2023.

VidEdit

Atlas-based video editing

- Decompose a video into a foreground image + a background image
- Edit the foreground/background image = edit the video
- Use diffusion to edit foreground/background atlas

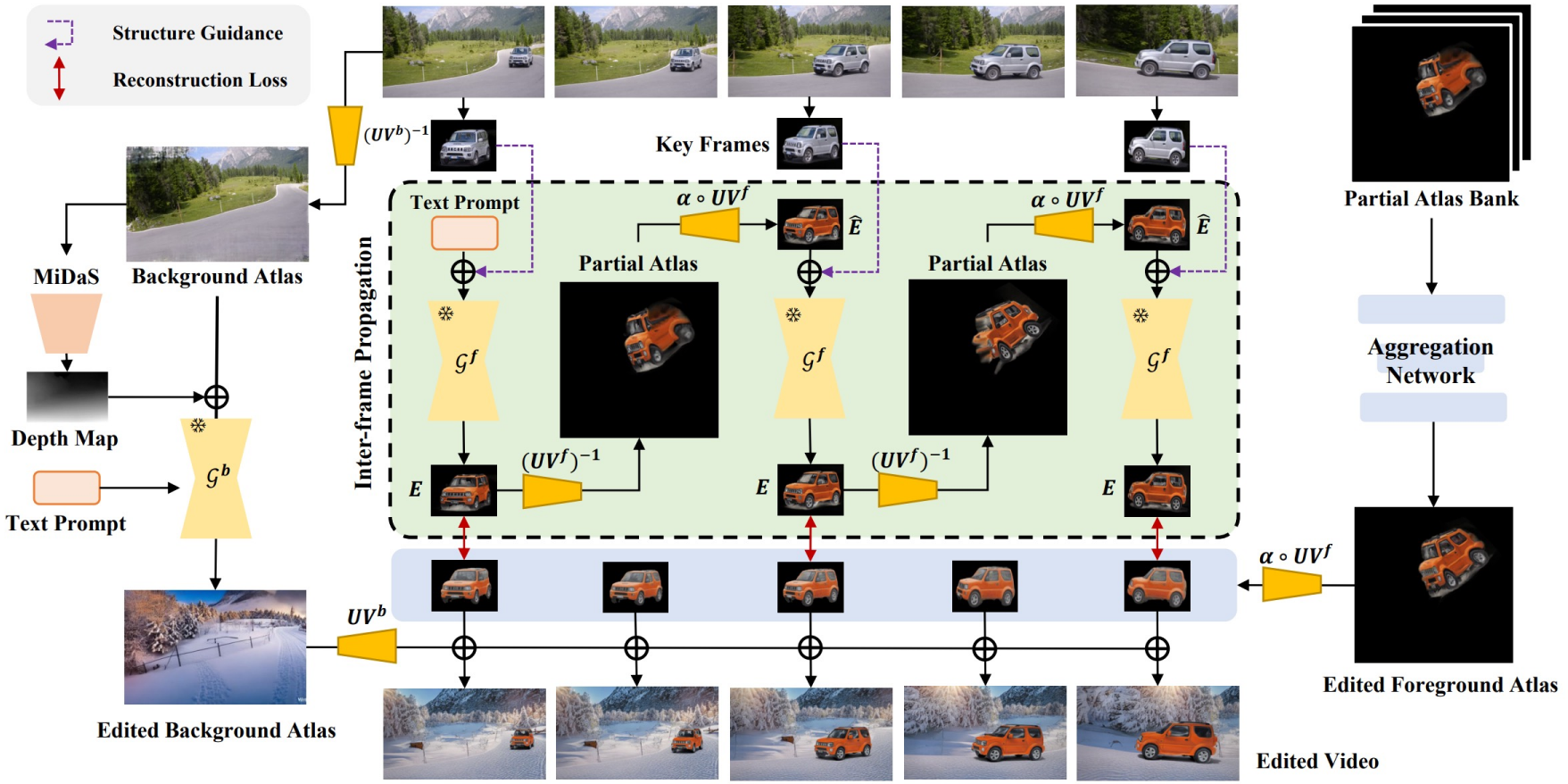


Video from Kasten et al., "Layered Neural Atlases for Consistent Video Editing," arXiv 2023.
Couairon et al., "VidEdit: Zero-Shot and Spatially Aware Text-Driven Video Editing," arXiv 2023.

StableVideo & Shape-aware Text-drive Layered Video Editing

Atlas-based video editing

Edit keyframes using Stable Diffusion → Aggregate the edits in decomposed atlases → Convert edited atlases back to edited video



Lee et al., "Shape-aware Text-driven Layered Video Editing," CVPR 2023.
Chai et al., "StableVideo: Text-driven Consistency-aware Diffusion Video Editing," ICCV 2023.

StableVideo

Atlas-based video editing

Chai et al., "StableVideo: Text-driven Consistency-aware Diffusion Video Editing," ICCV 2023.

Content Deformation Field (CoDeF)

Edit a video = edit a canonical image + learned deformation field

- Limitations of Neural Layered Atlases
  - Limited capacity for faithfully reconstructing intricate video details, missing subtle motion features like blinking eyes and slight smiles
  - Distorted nature of the estimated atlas leads to impaired semantic information
- Content Deformation Field: inspired by dynamic NeRF works, a new way of representing video, as a 2d canonical image + 3D deformation field over time
- Edit a video = edit a canonical image + learned deformation field

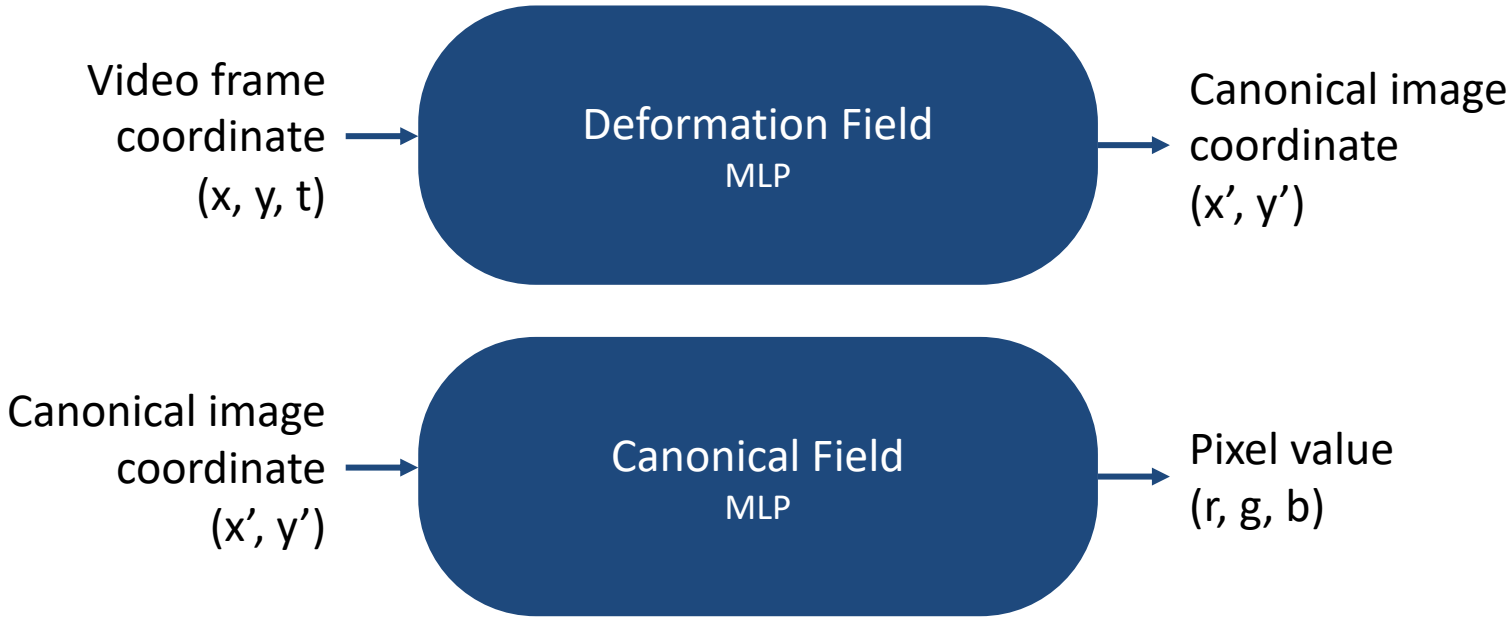
Ouyang et al., “CoDeF: Content Deformation Fields for Temporally Consistent Video Processing,” arXiv 2023.

Content Deformation Field (CoDeF)

Edit a video = edit a canonical image + learned deformation field

Problem Formulation

- Decode a video into a 2D canonical field and a 3D temporal deformation field
- Deformation Field: video (x, y, t) → canonical image coordinate (x', y')
- Canonical Field: (x', y') → (r, g, b), like a "2D image"



Ouyang et al., "CoDeF: Content Deformation Fields for Temporally Consistent Video Processing," arXiv 2023.

Content Deformation Field (CoDeF)

Edit a video = edit a canonical image + learned deformation field

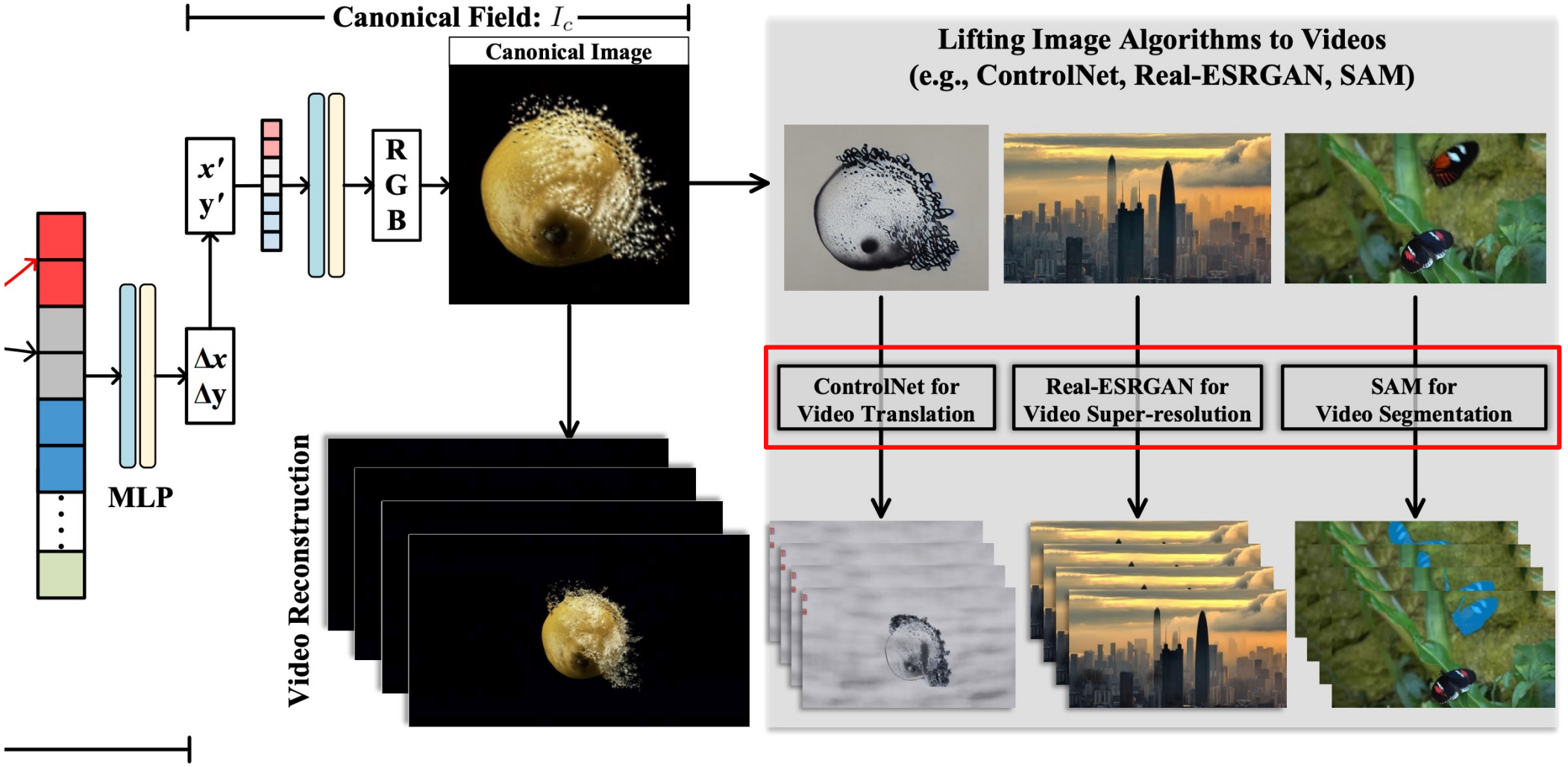
CoDeF compared to Atlas

- Superior robustness to non-rigid motion
- Effective reconstruction of subtle movements (e.g. eyes blinking)
- More accurate reconstruction: 4.4dB higher PSNR

Ouyang et al., "CoDeF: Content Deformation Fields for Temporally Consistent Video Processing," arXiv 2023.

Content Deformation Field (CoDeF)

Edit a video = edit a canonical image + learned deformation field

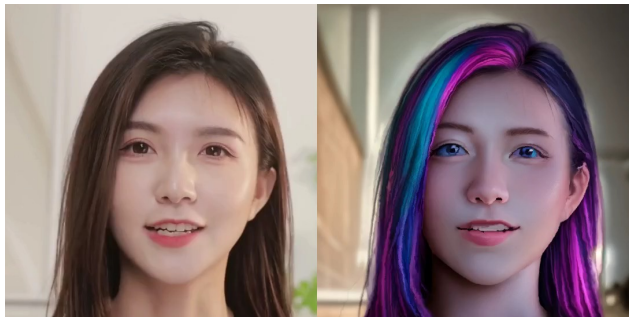


Ouyang et al., "CoDeF: Content Deformation Fields for Temporally Consistent Video Processing," arXiv 2023.

Content Deformation Field (CoDeF)

Edit a video = edit a canonical image + learned deformation field

Edit Canonical Image with ControlNet



Ouyang et al., "CoDeF: Content Deformation Fields for Temporally Consistent Video Processing," arXiv 2023.

DynVideo-E

Edit a video = edit a canonical ~~image~~ 3D NeRF

Canonical image in CoDeF is still 2D

Can we represent the video in a truly 3D space?

DynVideo-E

Edit a video = edit a canonical ~~image~~ 3D NeRF

Original Video

Background Style

Hulk

Editing Reference

DynVideo-E (Ours)

Renreder-A-Video

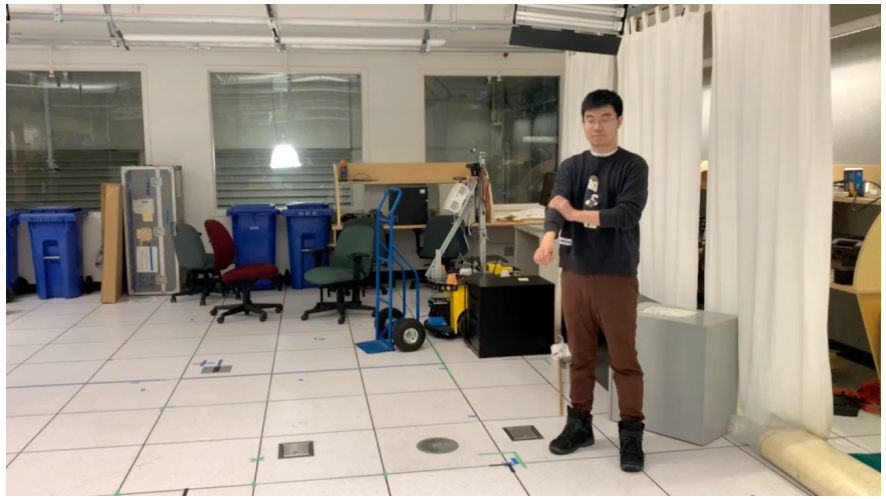
Stable Video

T2V Zero

Text2Live

DynVideo-E

Edit a video = edit a canonical ~~image~~ 3D NeRF



Original Video

Background Style

Luffy

Editing Reference

DynVideo-E (Ours)

Renreder-A-Video

Stable Video

T2V Zero

Text2Live

Liu et al., "DynVideo-E: Harnessing Dynamic NeRF for Large-Scale Motion- and View-Change Human-Centric Video Editing," arXiv 2023.

DynVideo-E

Edit a video = edit a canonical ~~image~~ 3D NeRF

Main idea

- For the first time introduce the dynamic NeRF as an innovative video representation for large-scale motion- and view-change human-centric video editing.

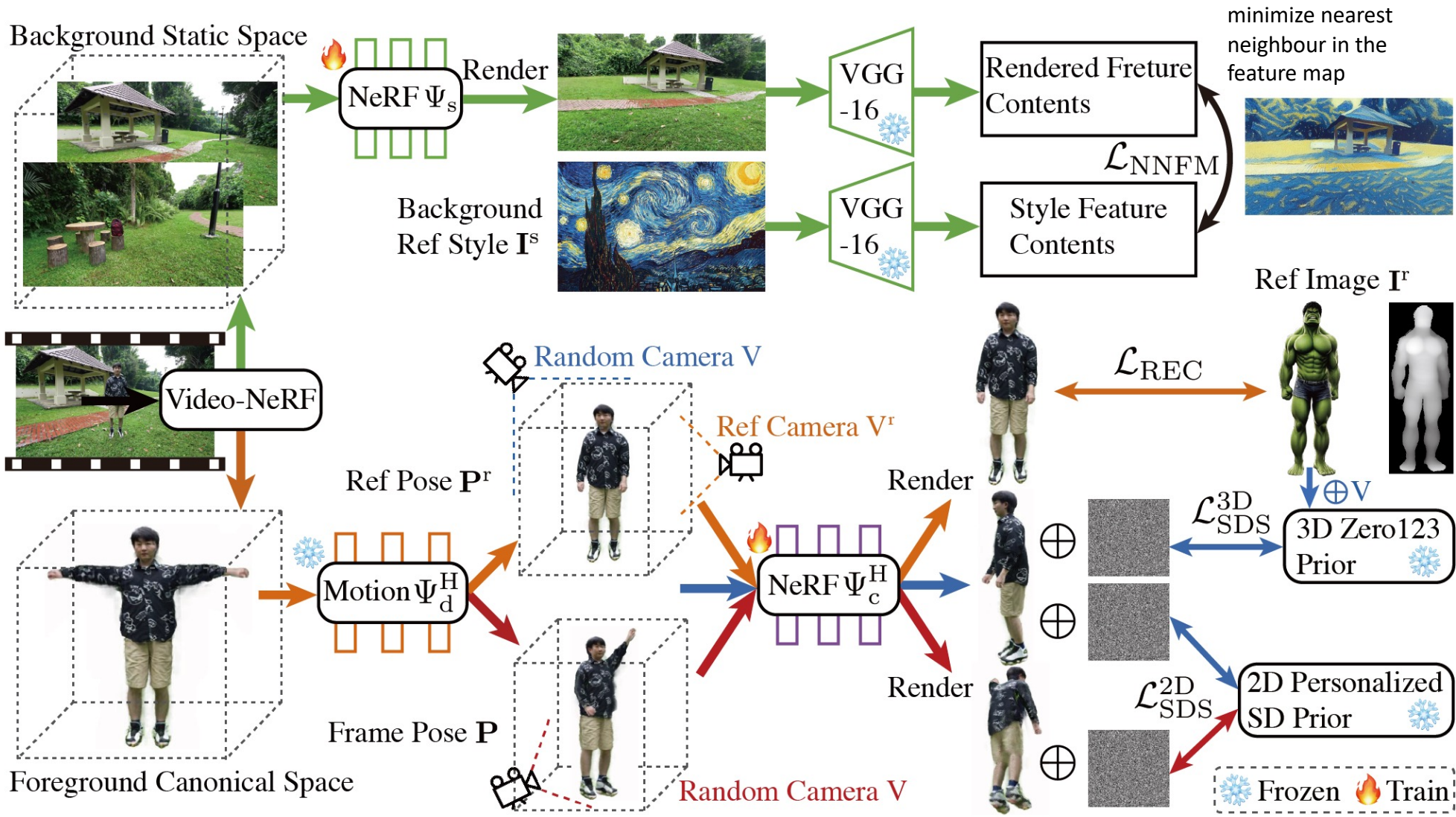
DynVideo-E

Edit a video = edit a canonical ~~image~~ 3D NeRF

Follow HOSNeRF, represent the video as:

- Background NeRF
- Human NeRF
- Deformation Field

Edit background NeRF and human NeRF respectively



Liu et al., "HOSNeRF: Dynamic Human-Object-Scene Neural Radiance Fields from a Single Video," ICCV 2023.
Liu et al., "DynVideo-E: Harnessing Dynamic NeRF for Large-Scale Motion- and View-Change Human-Centric Video Editing," arXiv 2023.

DynVideo-E

Edit a video = edit a canonical ~~image~~ 3D NeRF

DynVideo-E significantly outperforms SOTA approaches on two challenging datasets by a large margin of 50% ~ 95% in terms of human preference

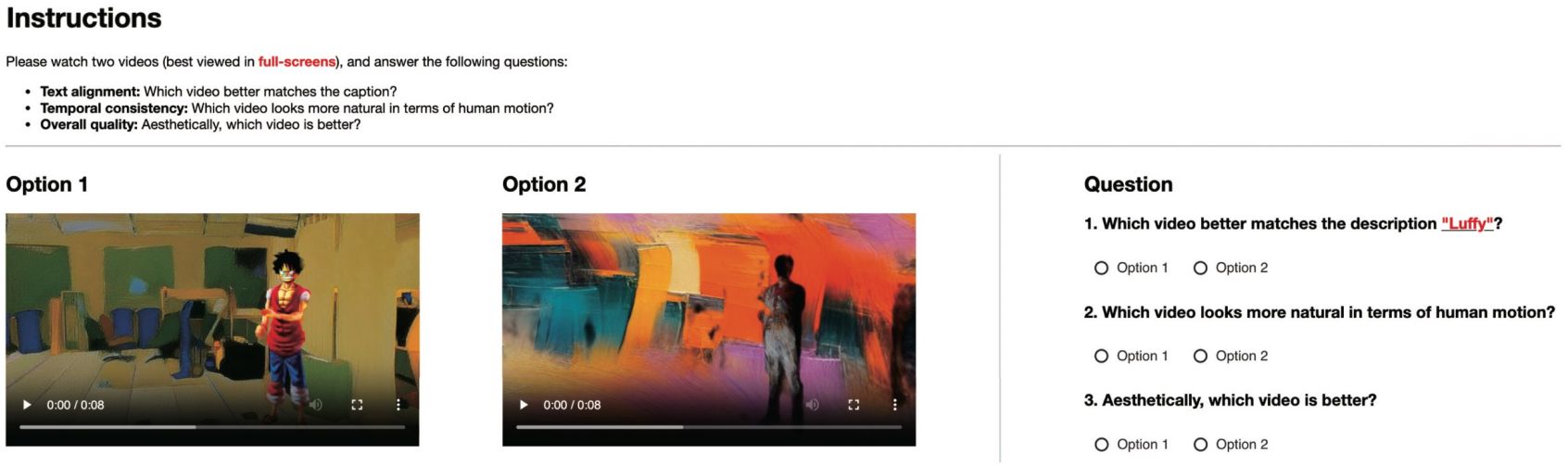


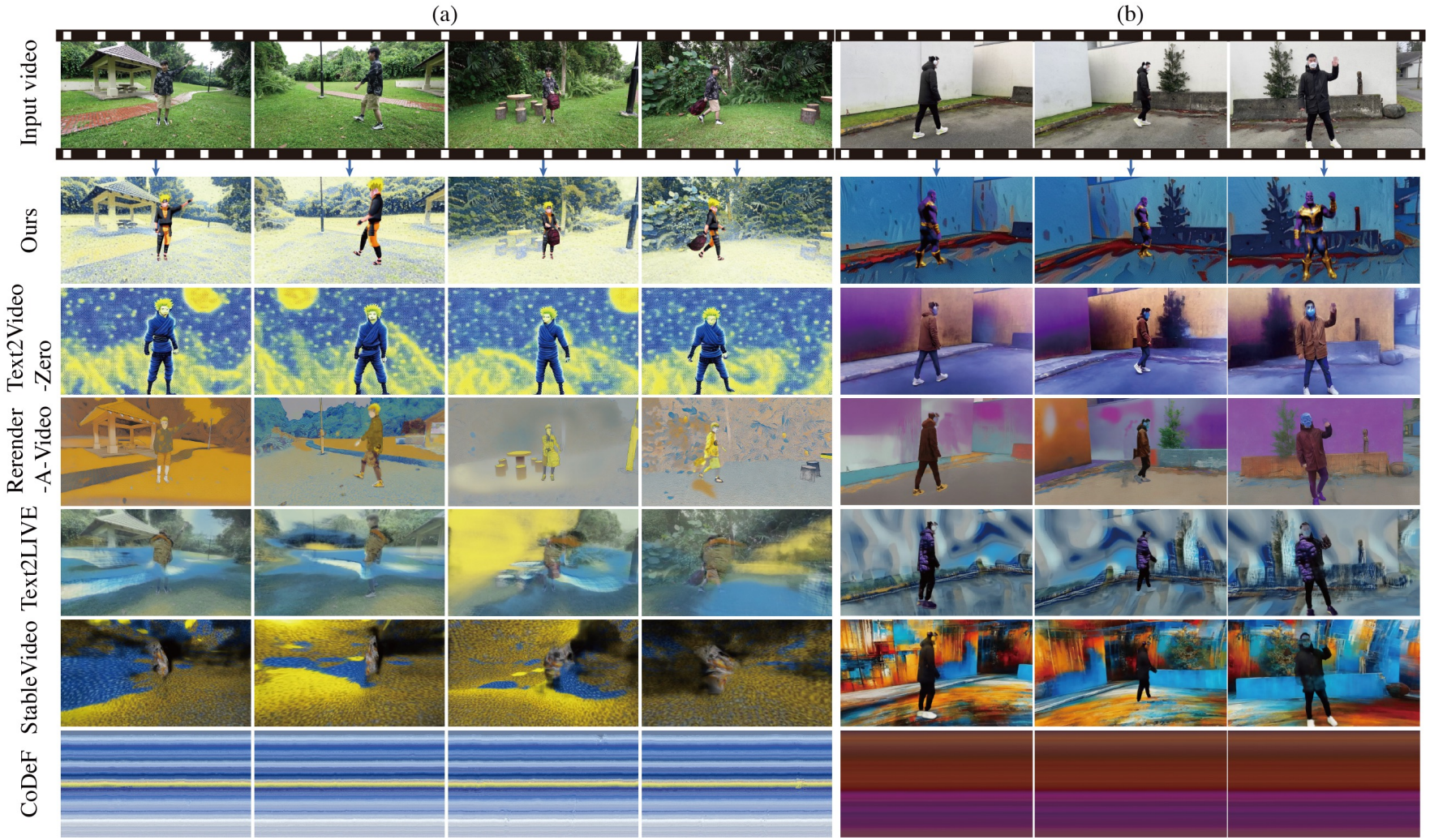
Figure 8: One comparison example from our questionnaires.

	METRICS	HUMAN PREFERENCE		
	CLIPScore (↑)	Textual Faithfulness (↑)	Temporal Consistency (↑)	Overall Quality (↑)
Text2Video-Zero [20]	26.70	9.17 v.s. **90.83 (Ours)**	21.25 v.s. **78.75 (Ours)**	12.08 v.s. **87.92 (Ours)**
Rerender-A-Video [55]	26.11	6.67 v.s. **93.33 (Ours)**	25.00 v.s. **75.00 (Ours)**	9.58 v.s. **90.42 (Ours)**
Text2LIVE [2]	22.77	3.81 v.s. **96.19 (Ours)**	26.67 v.s. **73.33 (Ours)**	9.05 v.s. **90.95 (Ours)**
StableVideo [6]	22.02	4.29 v.s. **95.71 (Ours)**	24.29 v.s. **75.71 (Ours)**	6.19 v.s. **93.81 (Ours)**
CoDeF [32]	16.77	1.25 v.s. **98.75 (Ours)**	3.75 v.s. **96.25 (Ours)**	1.25 v.s. **98.75 (Ours)**
DynVideo-E (Ours)	**31.31**	–	–	–

Table 1. Quantitative comparisons of our DynVideo-E against SOTA approaches on HOSNeRF dataset [27] and NeuMan dataset [18].

DynVideo-E

Edit a video = edit a canonical ~~image~~ 3D NeRF



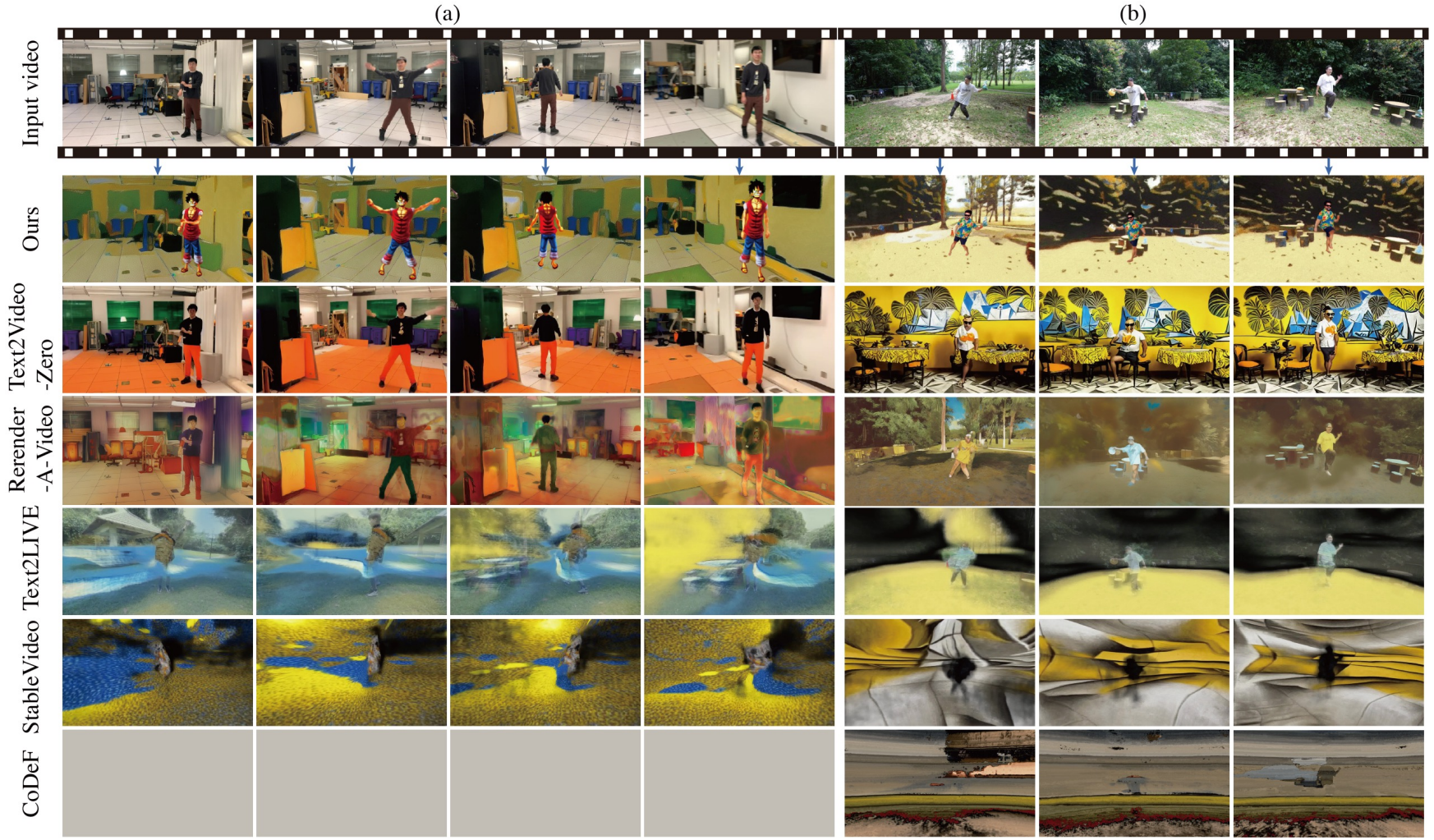
Liu et al., "DynVideo-E: Harnessing Dynamic NeRF for Large-Scale Motion- and View-Change Human-Centric Video Editing," arXiv 2023.

DynVideo-E

Edit a video = edit a canonical ~~image~~ 3D NeRF

Fail to edit the person

Fail to form correct 2D altas



DynVideo-E

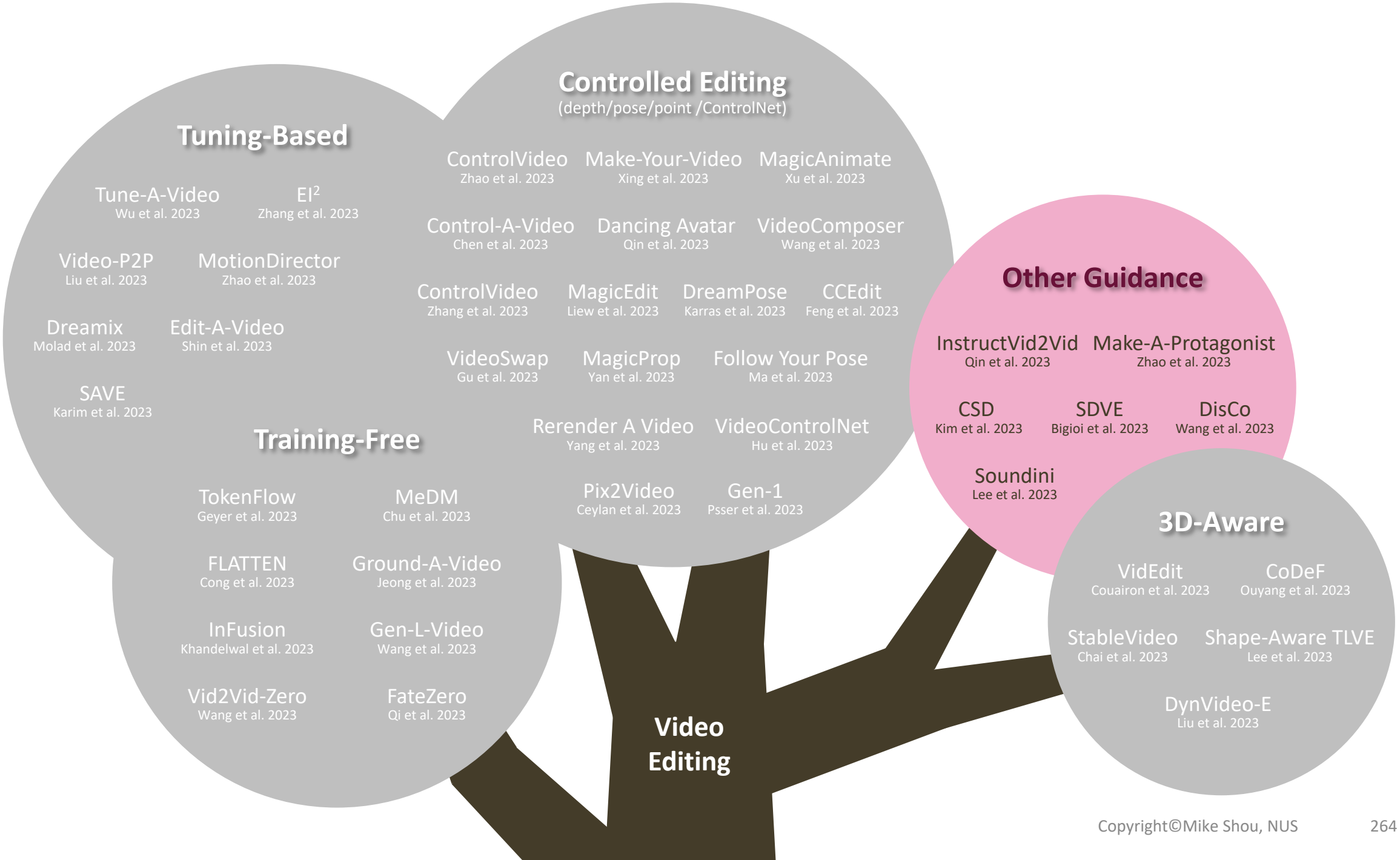
Edit a video = edit a canonical ~~image~~ 3D NeRF

Free Viewpoints for the edited scene

Liu et al., “DynVideo-E: Harnessing Dynamic NeRF for Large-Scale Motion- and View-Change Human-Centric Video Editing,” arXiv 2023.

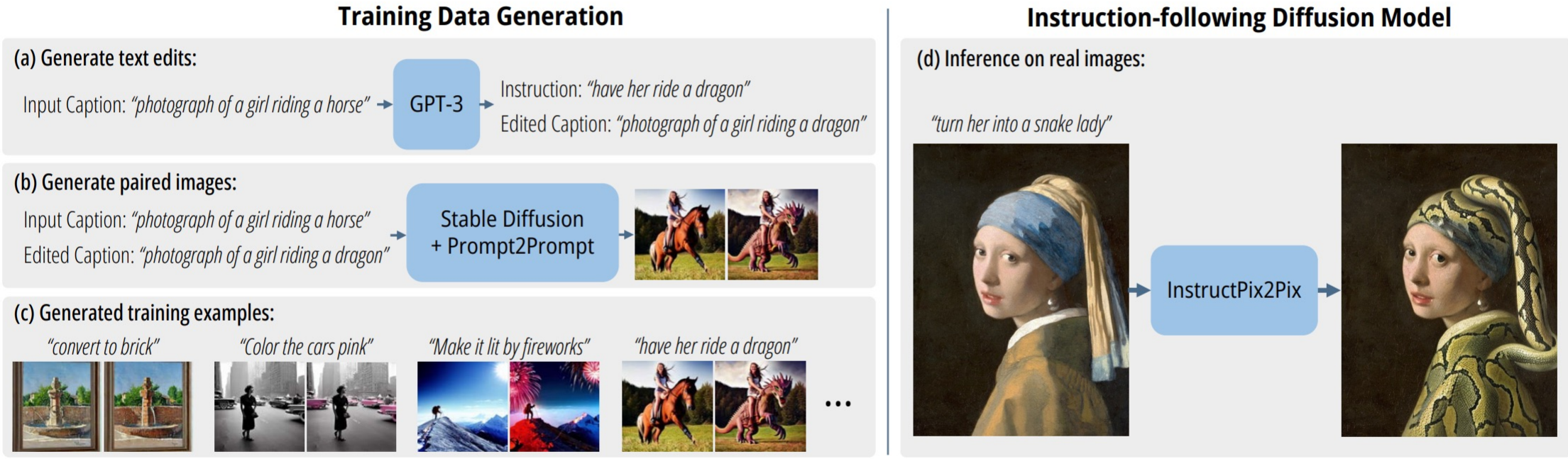
3 Video Editing

3.5 Other Guidance



InstructPix2Pix

Instruction-guided image editing

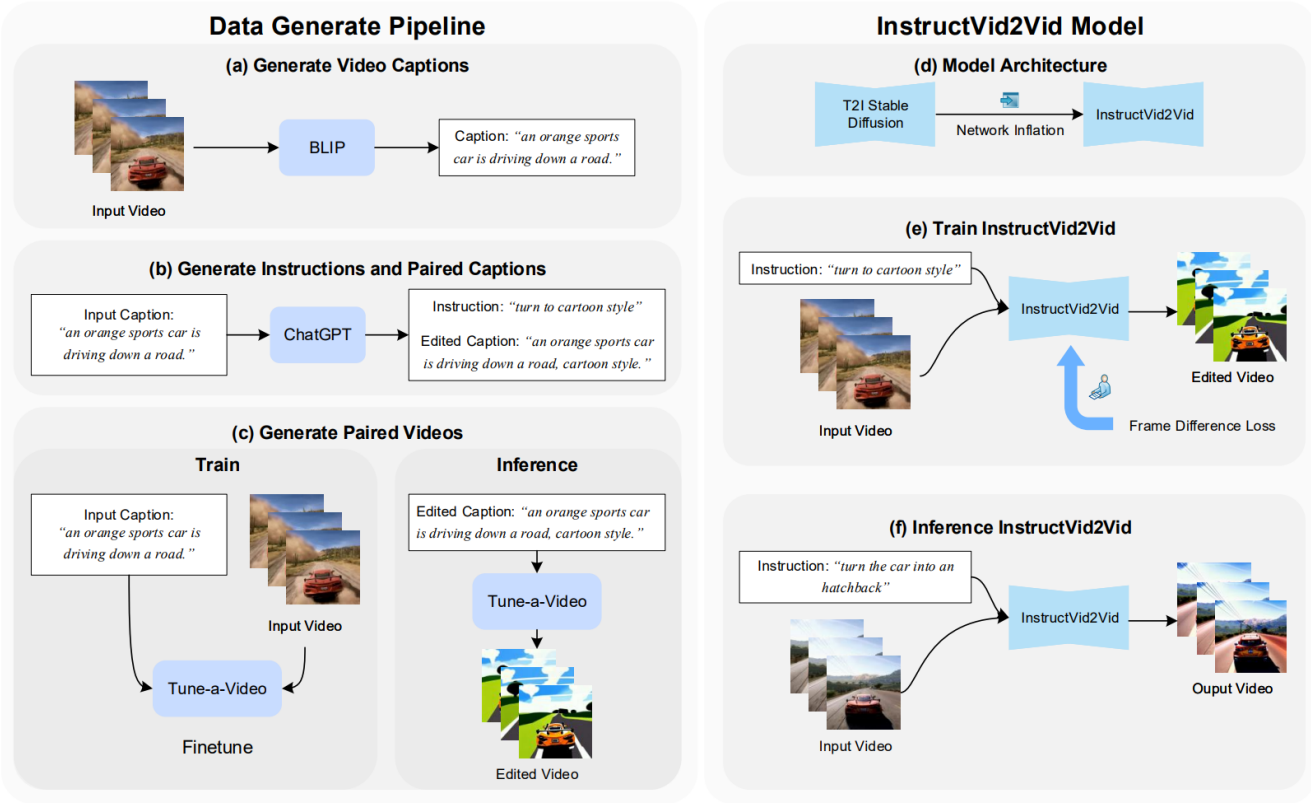


Brooks et al., "InstructPix2Pix: Learning to Follow Image Editing Instructions," CVPR 2023.

InstructVid2Vid

Instruction-guided Video Editing

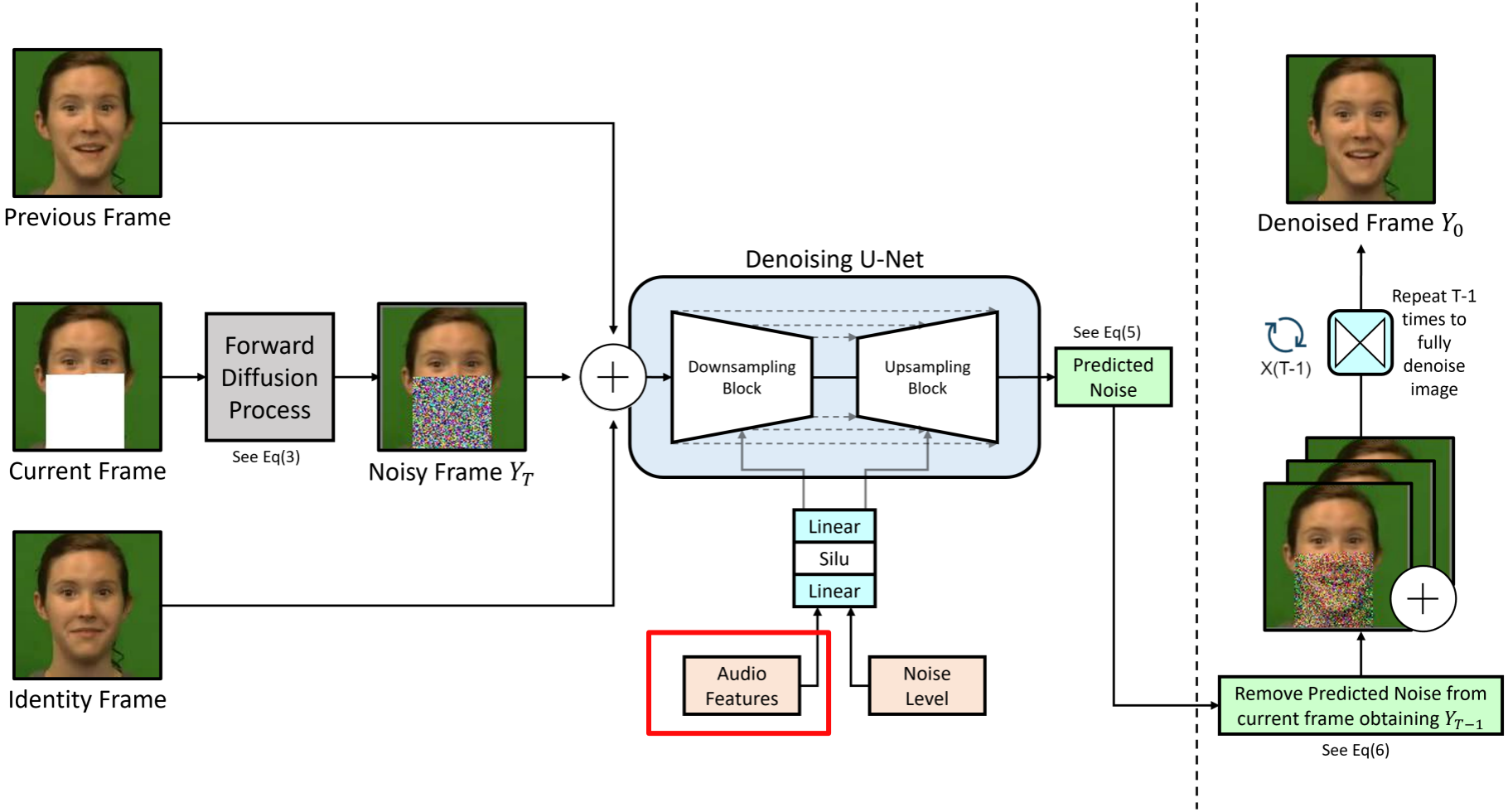
- Generate ⟨instruction, video⟩ dataset using ChatGPT, BLIP and Tune-A-Video
- Train inflated Stable Diffusion for instruction-guided video editing



Qin et al., "InstructVid2Vid: Controllable Video Editing with Natural Language Instructions," arXiv 2023.

Speech Driven Video Editing via an Audio-Conditioned Diffusion Model

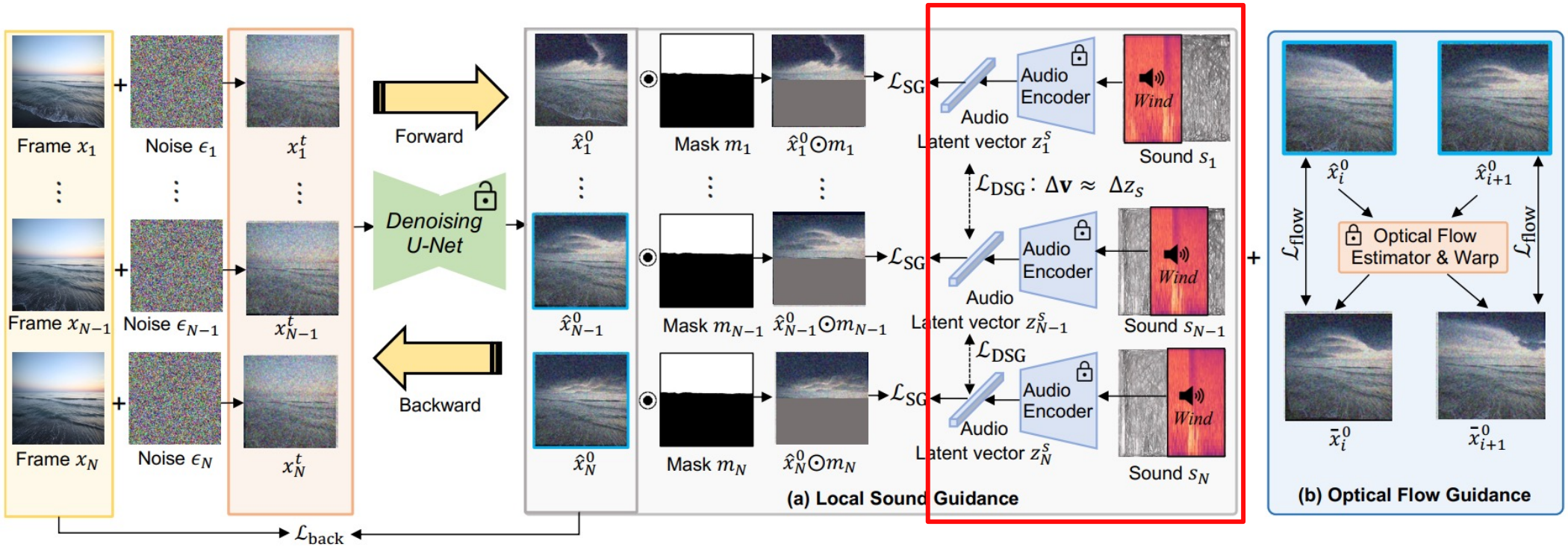
Speech-driven video editing



Bigioi et al., "Speech Driven Video Editing via an Audio-Conditioned Diffusion Model," arXiv 2023.

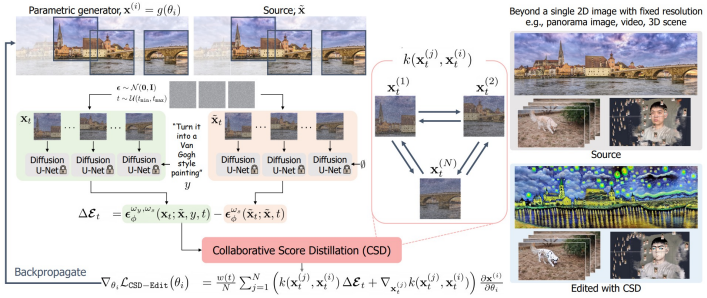
Soundini

Sound-guided video editing



Lee et al., "Soundini: Sound-Guided Diffusion for Natural Video Editing," arXiv 2023.

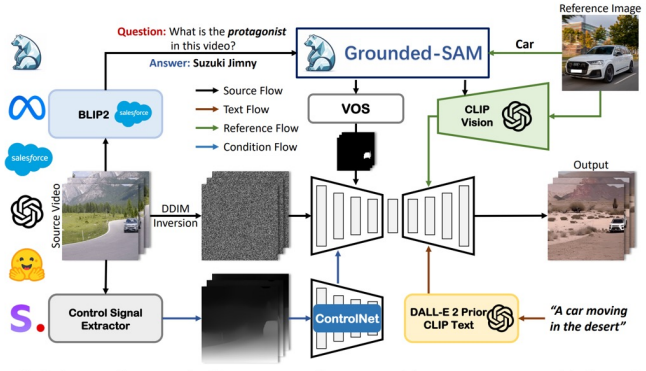
Video Editing Under Various Guidance: More Works



Collaborative Score Distillation (Kim et al.)

Instruction-guide video editing

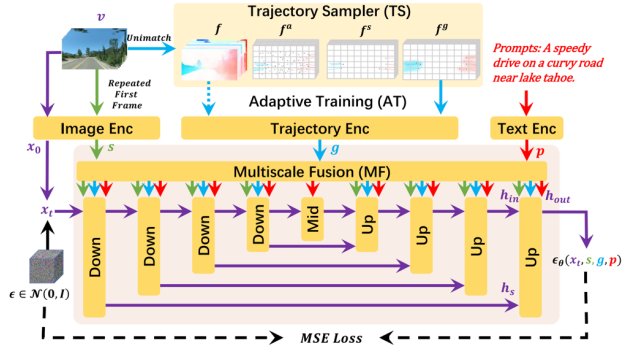
"Collaborative Score Distillation for Consistent Visual Synthesis," NeurIPS 2023.



Make-A-Protagonist (Zhao et al.)

Video editing with an ensemble of experts

"Make-A-Protagonist: Generic Video Editing with An Ensemble of Experts," arXiv 2023.

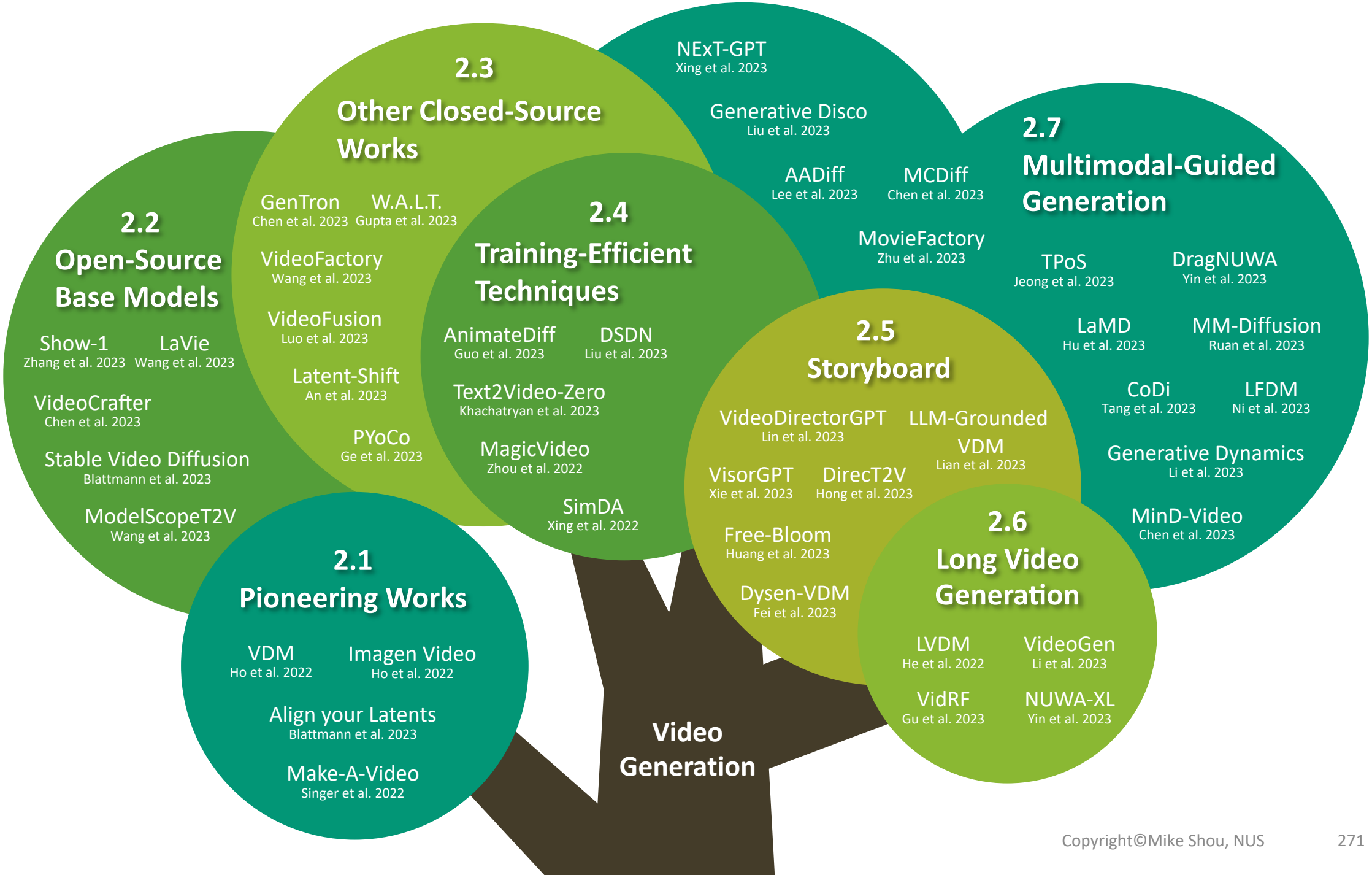


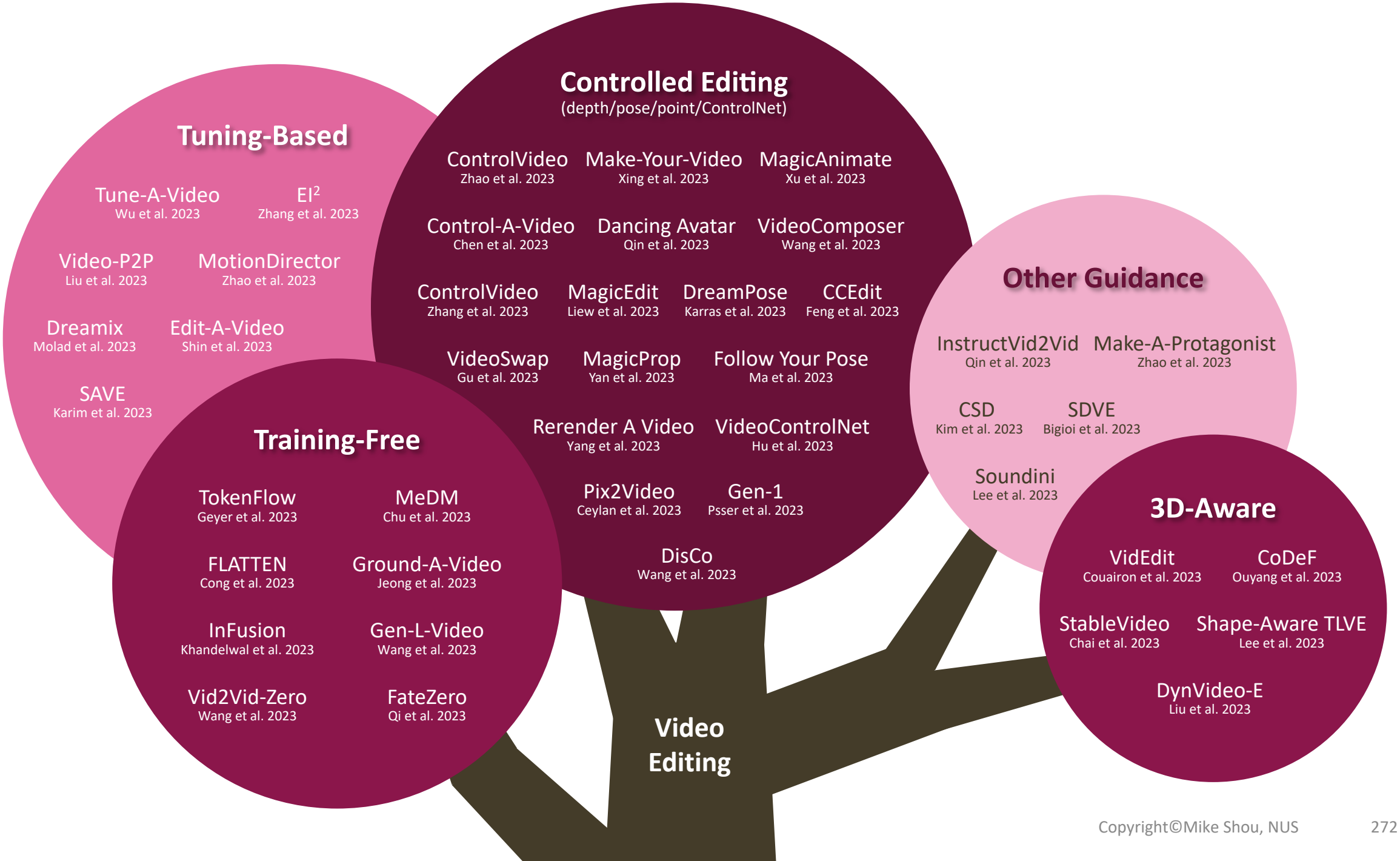
DragNUWA (Yin et al.)

Multimodal-guided video editing

"DragNUWA: Fine-grained Control in Video Generation by Integrating Text, Image, and Trajectory," arXiv 2023.

4 Summary





A list of video diffusion papers -- will keep adding latest papers!

showlab / **Awesome-Video-Diffusion**

Type / to search

Code | Issues 1 | Pull requests 1 | Actions | Projects | Security | Insights | Settings

Awesome-Video-Diffusion Public

Edit Pins | Unwatch 75 | Fork 65 | Star 1.3k

main | 1 branch | 0 tags | Go to file | Add file | Code

zhangjiewu update e7701d6 2 days ago **123** commits

README.md update 2 days ago

README.md

Awesome Video Diffusion awesome

About

A curated list of recent diffusion models for video generation, editing, restoration, understanding, etc.

awesome | video-editing | video-understanding | video-generation | diffusion-models | text-to-video | video-restoration | text-to-motion

Readme

Thank You!

Mike Shou

Asst Prof, National U. of Singapore

Joint work with Pei Yang & Jay Wu

Slides: https://sites.google.com/view/showlab/tutorial