CVPR

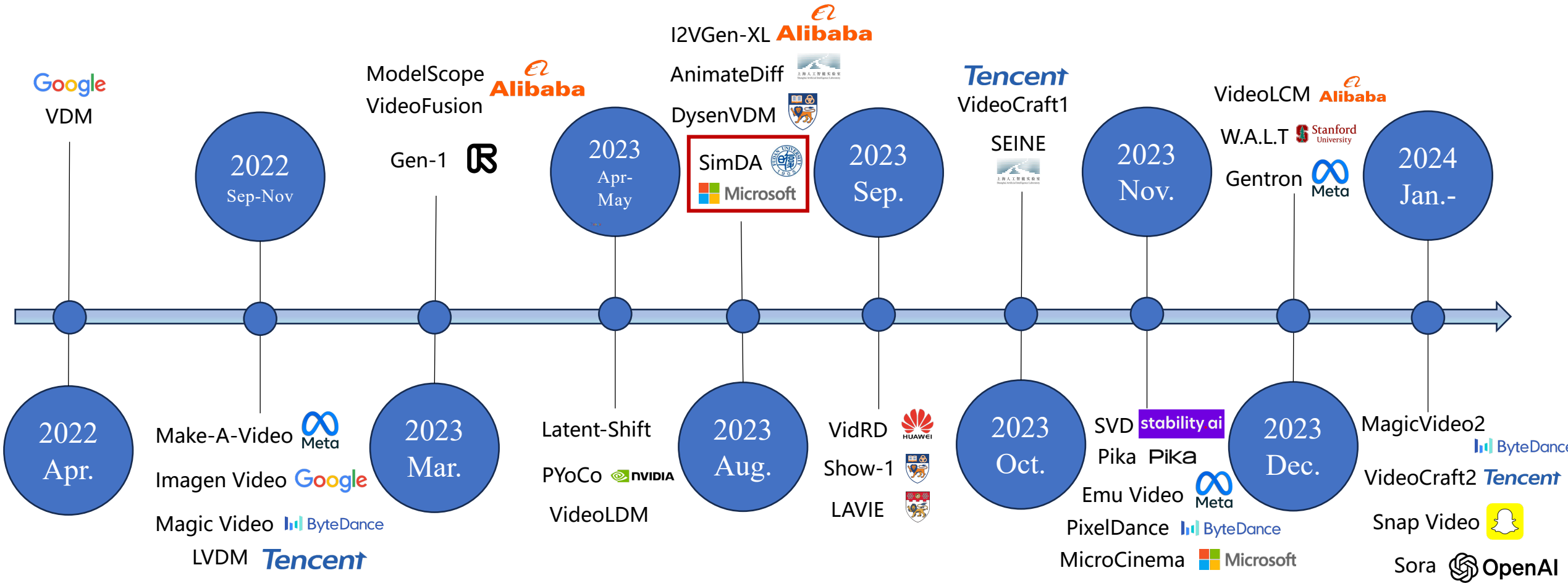
SimDA: Simple Diffusion Adapter for Efficient Video Generation

Zhen Xing [1], Qi Dai [2], Han Hu [2], Zuxuan Wu [1], Yu-Gang Jiang [1]

1.Fudan University 2. Microsoft Research Asia

Zhen Xing
2024.04.24

Video Generation Models



Video Generation Models

Video Diffusion Models（Google）
2022-04

A red Cardinal on a tree branch stands out when the snow is falling

Sea waves with foam on white tropical sandy beach

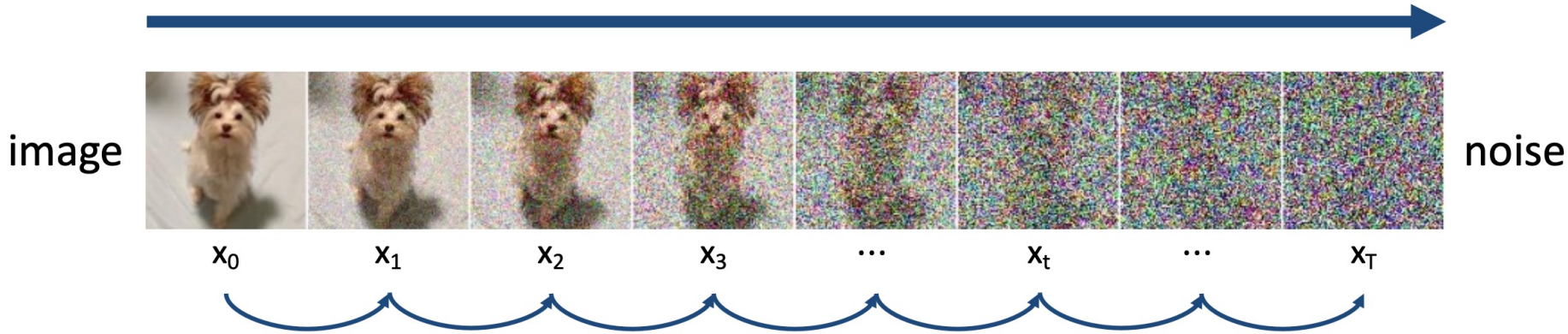
SimDA（Ours）
2023-08

Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.

Sora (OpenAI)
2024-02

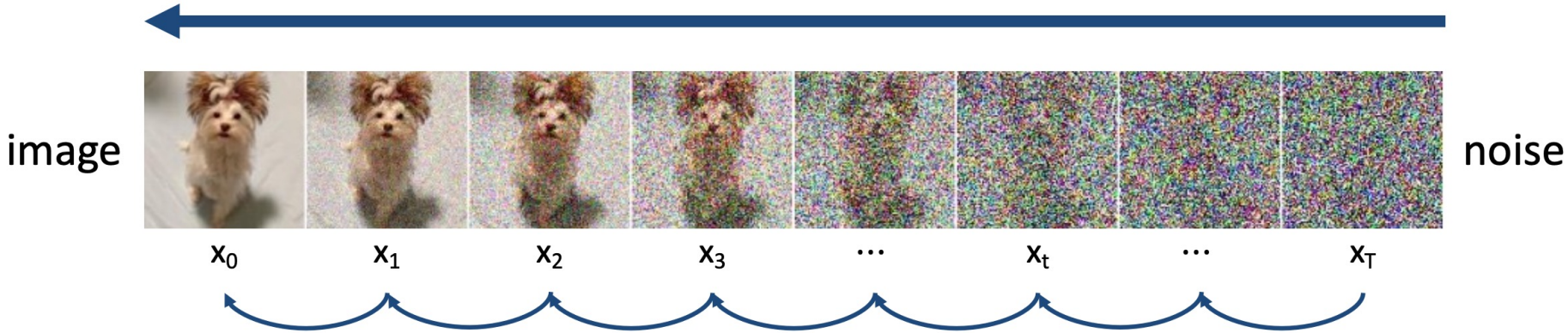
DDPM

Forward process/diffusion process: add noise



$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathrm{I})$$

Reverse process/denoise process: remove noise

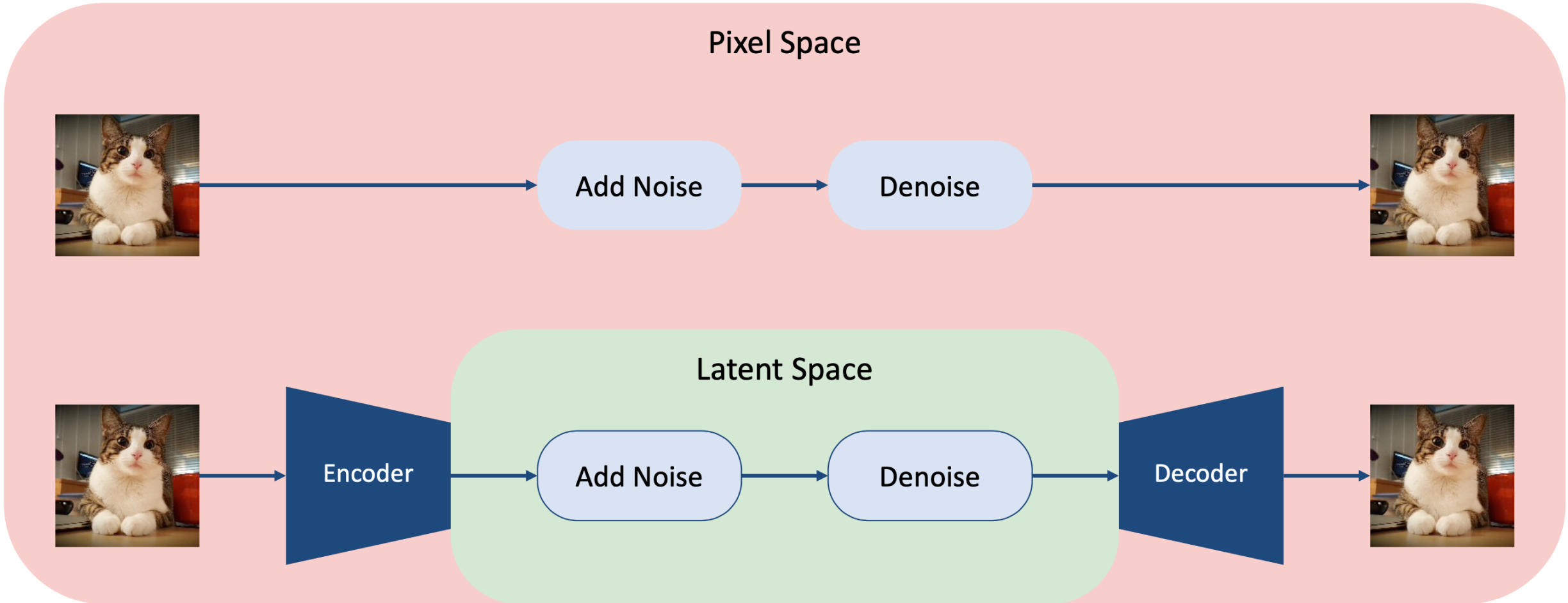


$$\mathrm{p}_\theta(x_{t-1}|x_t)$$

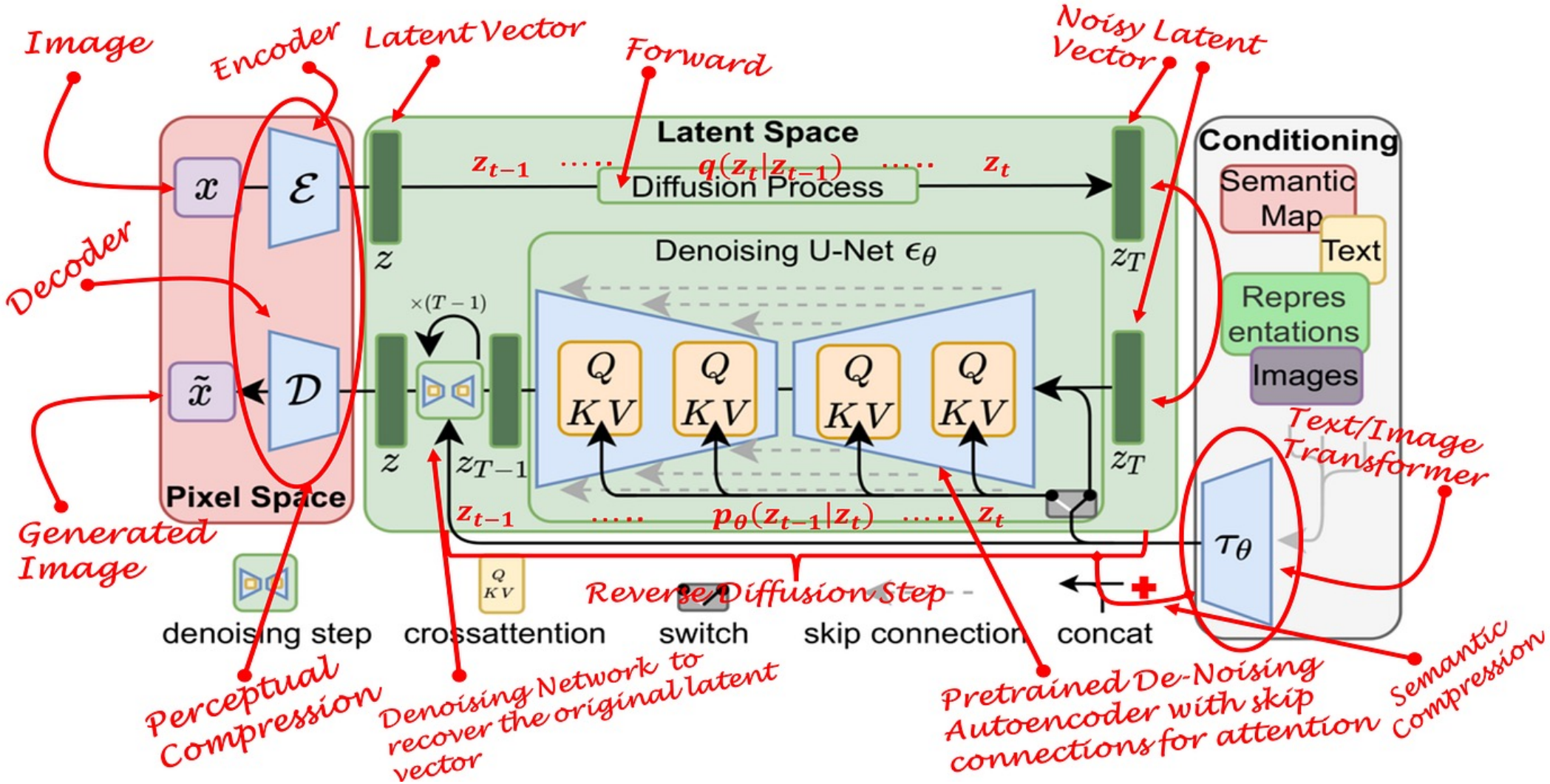
Train a neural network model to remove noise gradually

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.

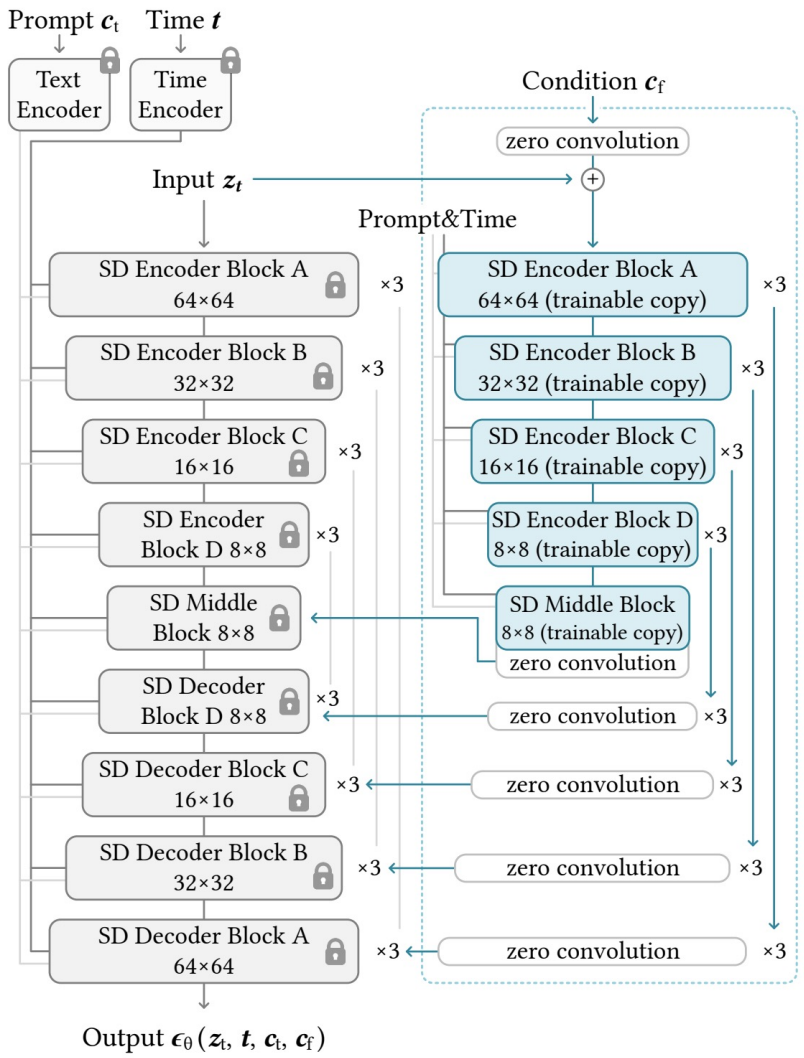
Pixel Diffusion Model v.s Latent Diffusion Model



Latent Diffusion Model (Stable Diffusion)



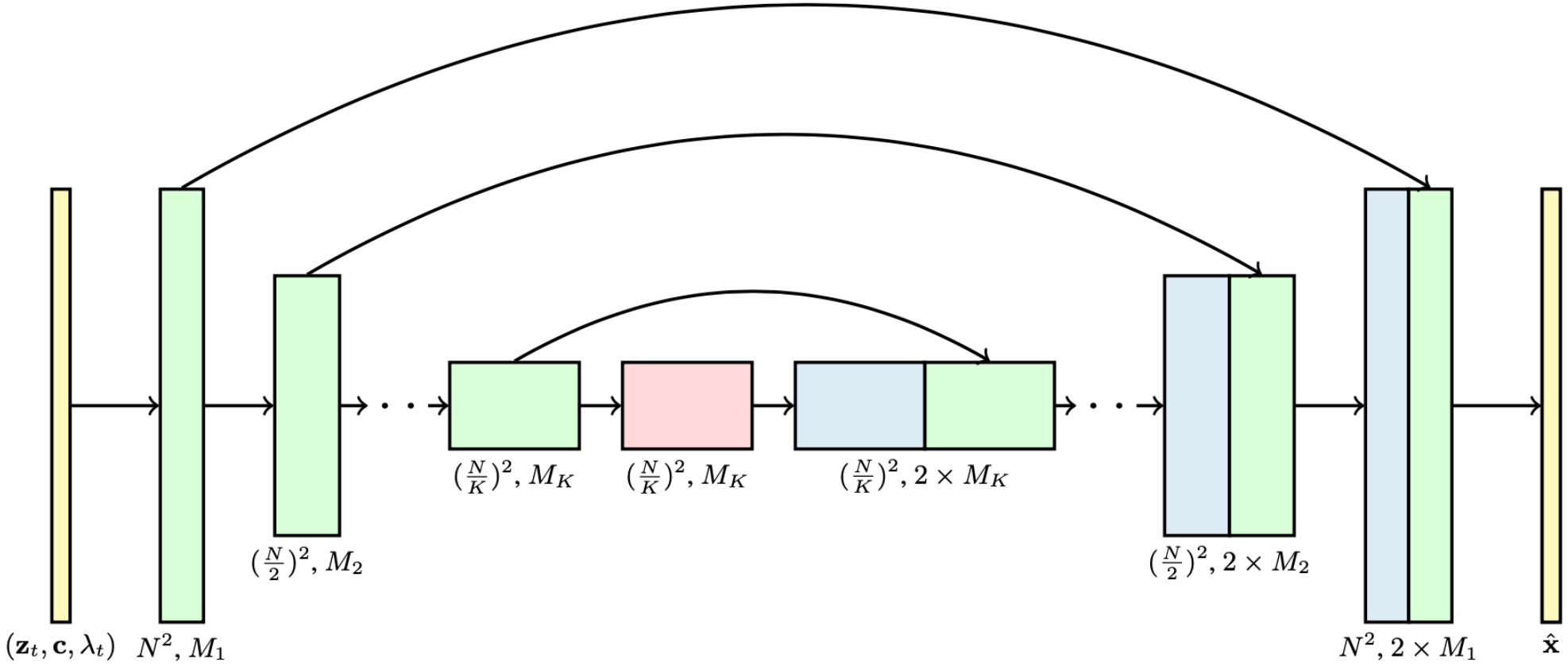
ControlNet



(a) Stable Diffusion (b) ControlNet

Conditional Control to Text-to-Image Diffusion Models （ICCV 2023）

Background (First Video Diffusion Model)

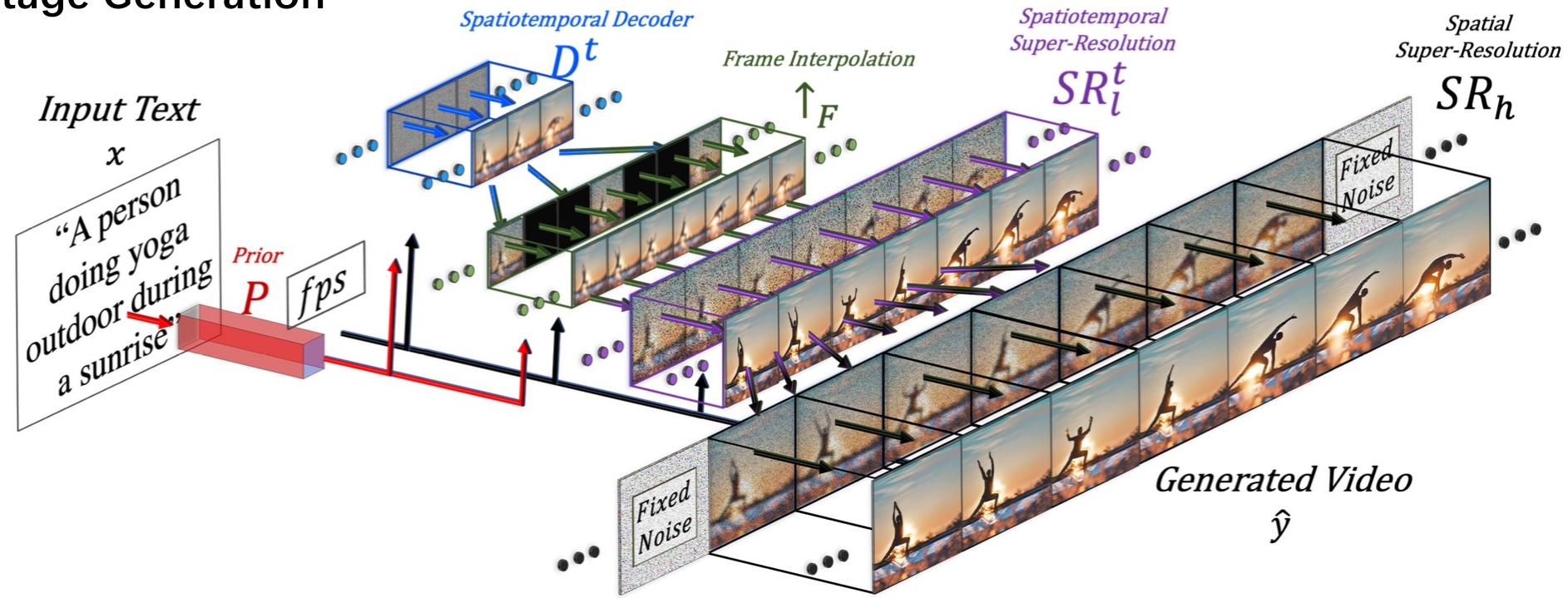


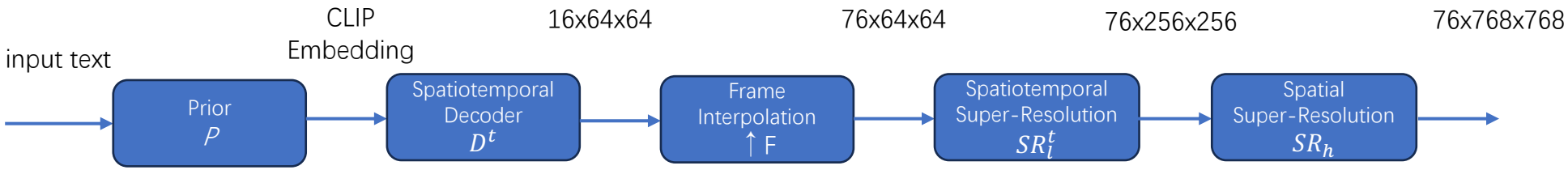
- Conv2D → 3D （3x3 -> 1x3x3）
- Space Attention →Divided Space-Temporal Attention
- Joint training on video and image modeling

Video Diffusion Models （NeurIPS 2022）

Background (Make-A-Video)

Multi-stage Generation



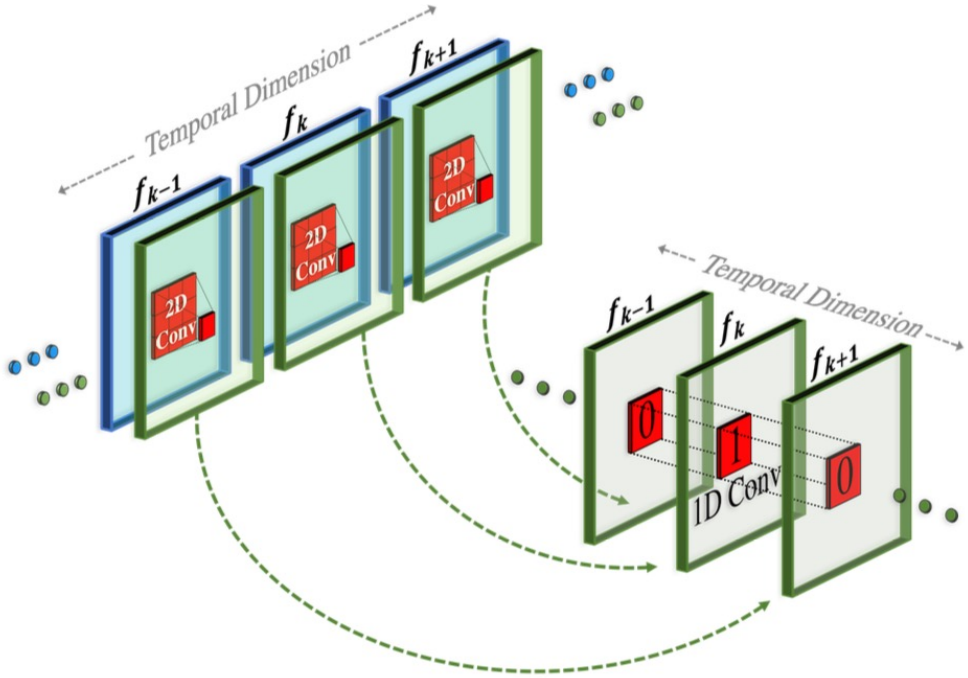


MAKE-A-VIDEO: TEXT-TO-VIDEO GENERATION WITHOUT TEXT-VIDEO DATA （ICLR2023）

Background (Make-A-Video)

Pseudo-3D Convolution Layers

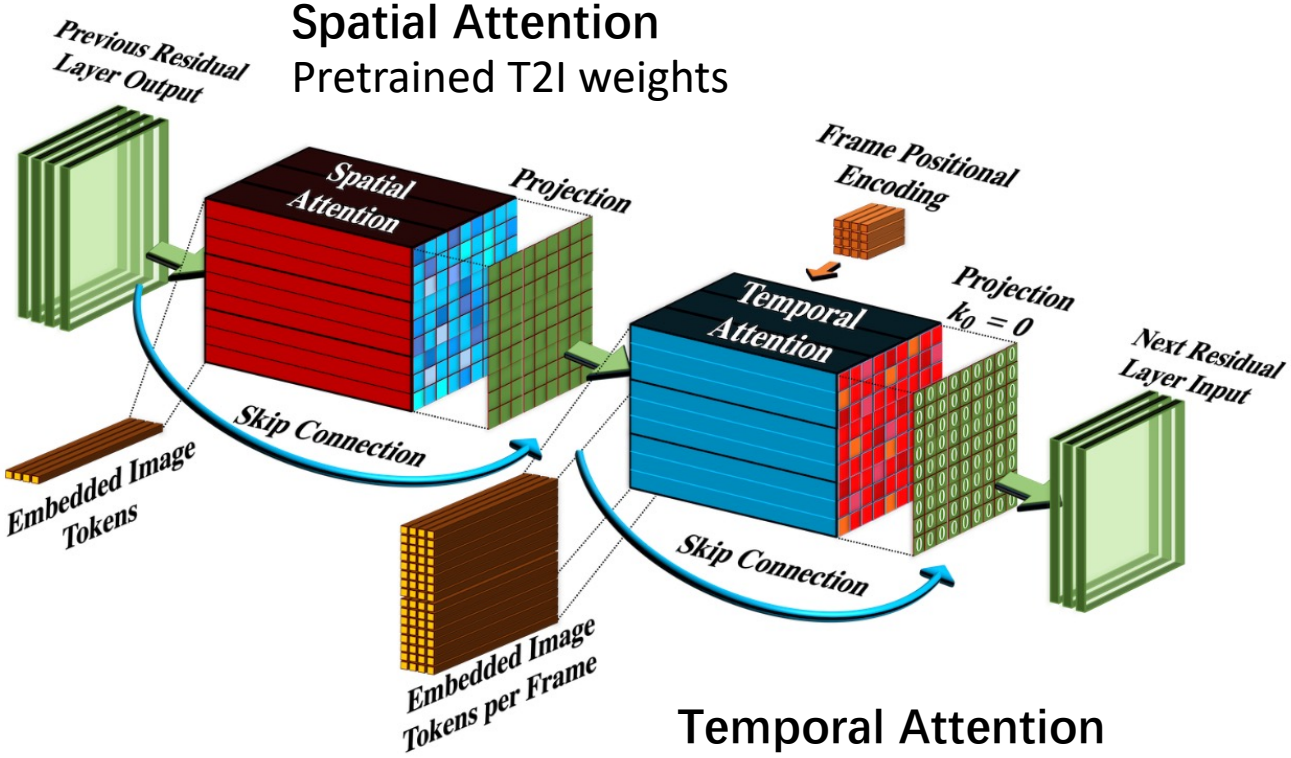
Spatial Convolution
Pretrained T2I weights



Temporal Convolution

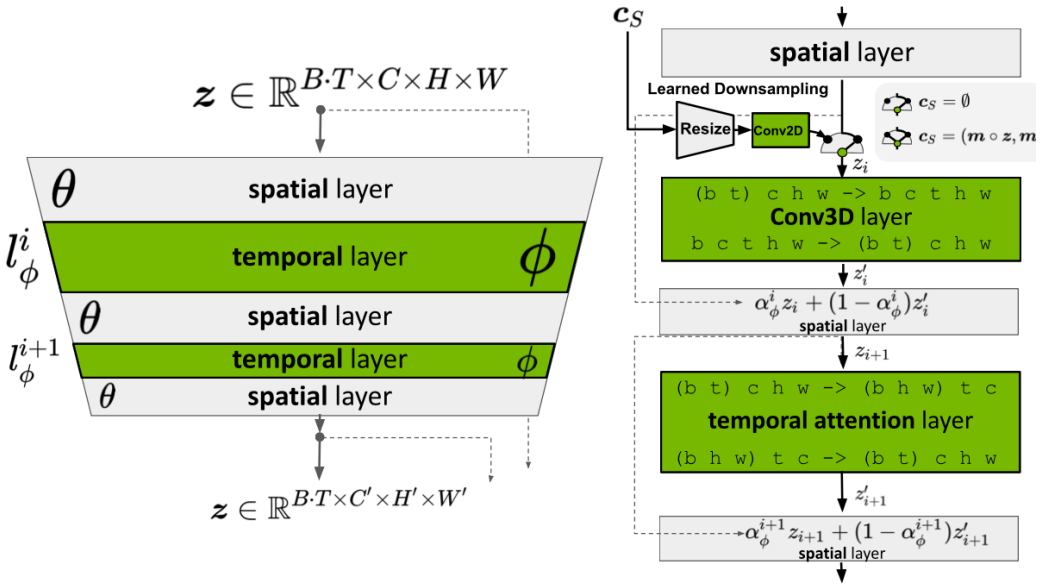
Divided Spatial Temporal Attention Layers

Spatial Attention
Pretrained T2I weights



Temporal Attention
Initialized with zero temporal projection

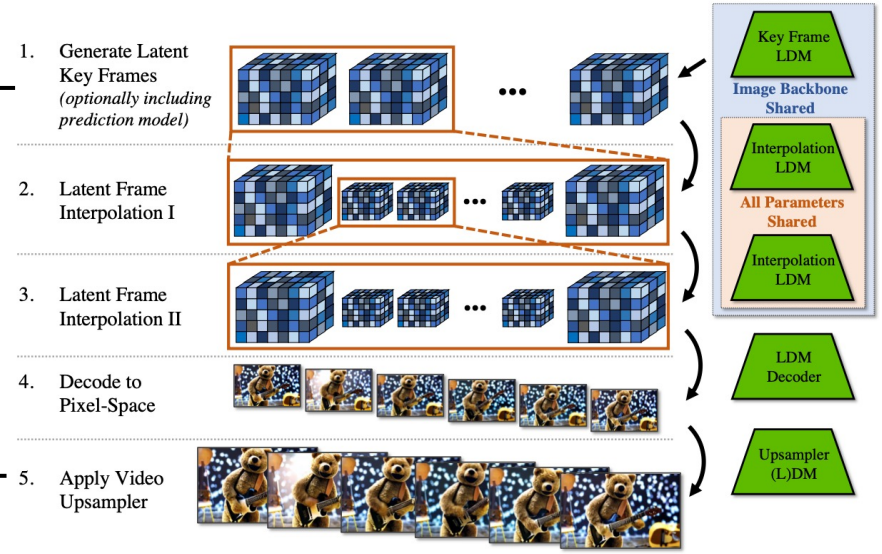
Background (Video Latent Diffusion Model)



- Trained on **256** GPUs
- Batch Size 768, 402k steps
- **1.71B** Parameters

- Trained on **128** GPUs
- Batch Size 1028, 95k steps
- **1.51B** Parameters

- Trained on **32** GPUs
- Batch Size 256, 10k steps
- **0.98B** Parameters

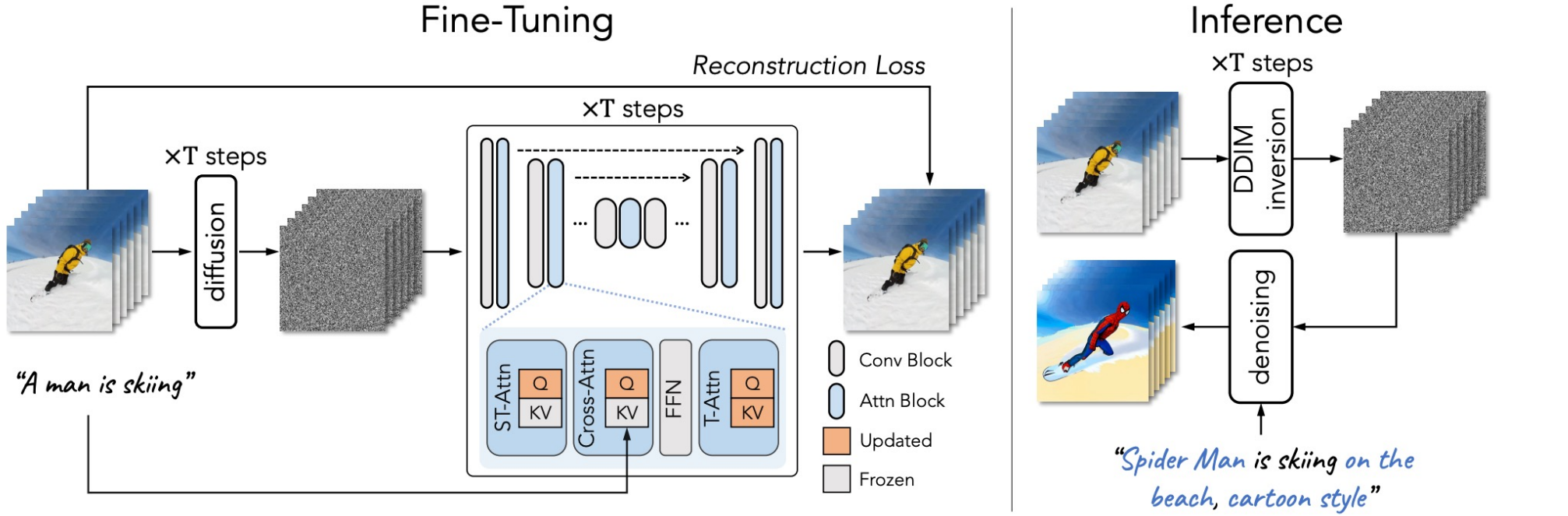


Temporal Layer: Conv3D+Temporal Attention

5-stage Process

Blattmann, Andreas, et al. “Align your latents: High-resolution video synthesis with latent diffusion models.” (CVPR 2023)

Background & Motivation (Tune-A-Video)



A man is skiing

A man, wearing pink clothes, is skiing at sunset

Wonder Woman, wearing a cowboy hat, is skiing

Spider Man is skiing on the beach, cartoon style

Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation (ICCV 2023)

Background (Training Datasets)

WebVid-10M

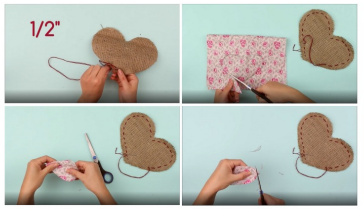
"Runners feet in a sneakers close up. realistic three dimensional animation."

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

"Billiards, concentrated young woman playing in club"

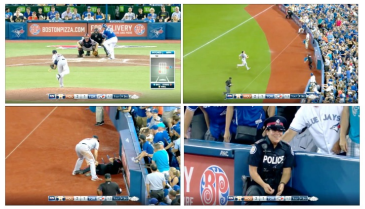
Top2 retrieval result: A dog is running on the road.

HD-VILA-100M



Keep half an inch allowance with filler draw a smaller heart on the pattern fabric. Cut it out to make the heart sides identical. Fold it in half and trim.

Applying the powder with a stippling motion instead of a sweeping motion, because I do not want to disturb my foundation brushes.



A little slapstick comedy watch. Josh Donaldson hits a foul ball to the first base side and AJ Read knocks over a police officer.

Mexican food is all around us. In Los Angeles, there are Taco stands on every corner.

Some of you guys have seen the things I have in here are my husky collection and my stuffed animals.

The gauntlet allows the wearer to wield all of the stones powers at once with one snap of his fingers.

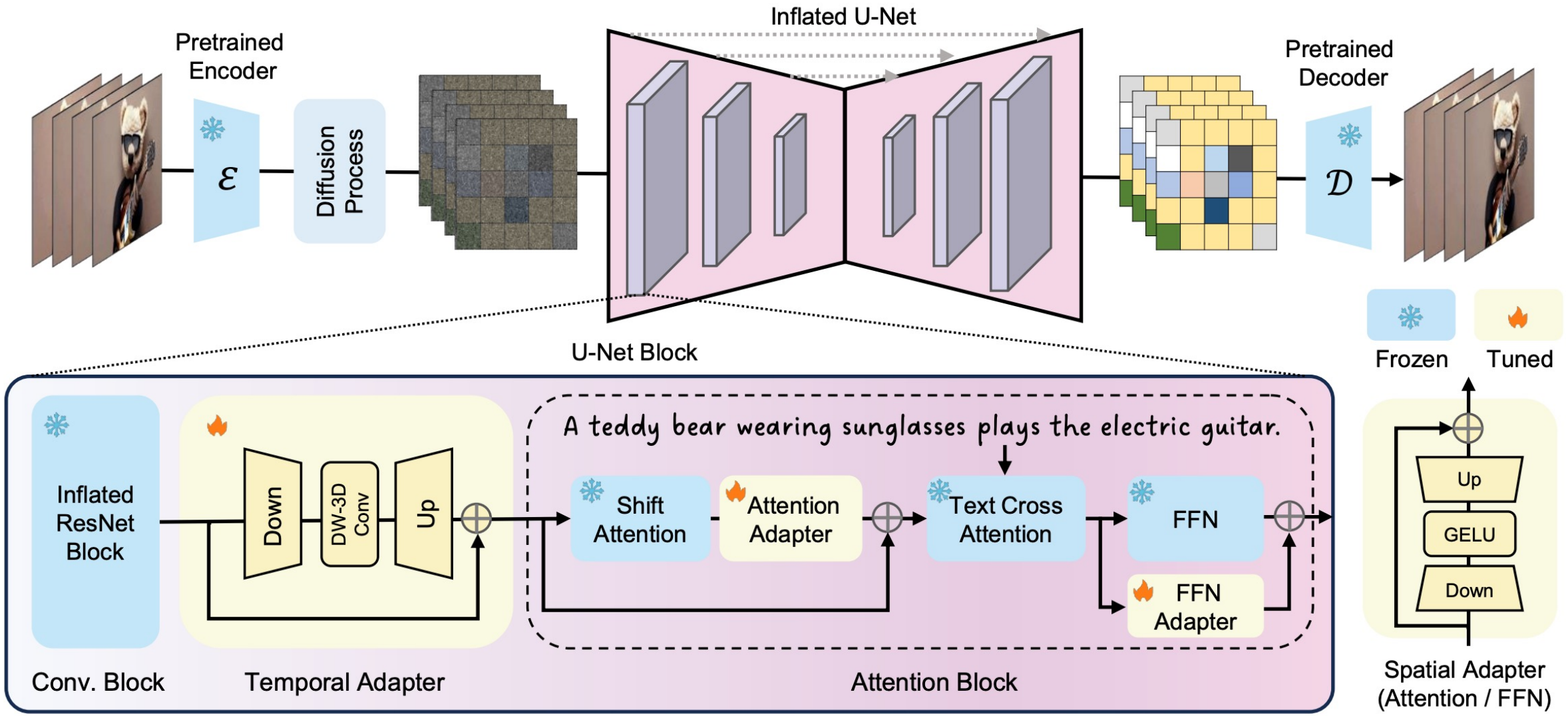
Dataset	Domain	#Video clips	#Sentence	Avg len(sec)	Sent len	Duration(h)	Resolution
MSR-VTT [59]	open	10K	200K	15.0	9.3	40	240p
DideMo [2]	Flickr	27K	41K	6.9	8.0	87	-
LSMDC [45]	movie	118K	118K	4.8	7.0	158	1080p
YouCook II [69]	cooking	14K	14K	19.6	8.8	176	-
How2 [47]	instructional	80K	80K	90.0	20.0	2K	-
ActivityNet Caption [28]	action	100K	100K	36.0	13.5	849	-
WebVid-2M [3]	open	2.5M	2.5M	18.0	12.0	13K	360p
HowTo100M [41]	instructional	136M	136M	3.6	4.0	134.5K	240p
HD-VILA-100M (Ours)	open	103M	103M	13.4	32.5	371.5K	720p

Table 1. Statistics of HD-VILA-100M and its comparison with existing video-language datasets.

Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval（ICCV 2021 ）

Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions（CVPR 2022 ）

Pipeline of SimDA



Spatial Adapter

$$\texttt{S-Adapter}(\mathbf{X}) = \mathbf{X} + \mathbf{W}_{\text{up}}(\texttt{GELU}(\mathbf{W}_{\text{down}}(\mathbf{X}))),$$

Temporal Adapter

$$\texttt{T-Adapter}(\mathbf{X}) = \mathbf{X} + \mathbf{W}_{\text{up}}(\texttt{3D-Conv}(\mathbf{W}_{\text{down}}(\mathbf{X}))).$$

Latent Shift Attention

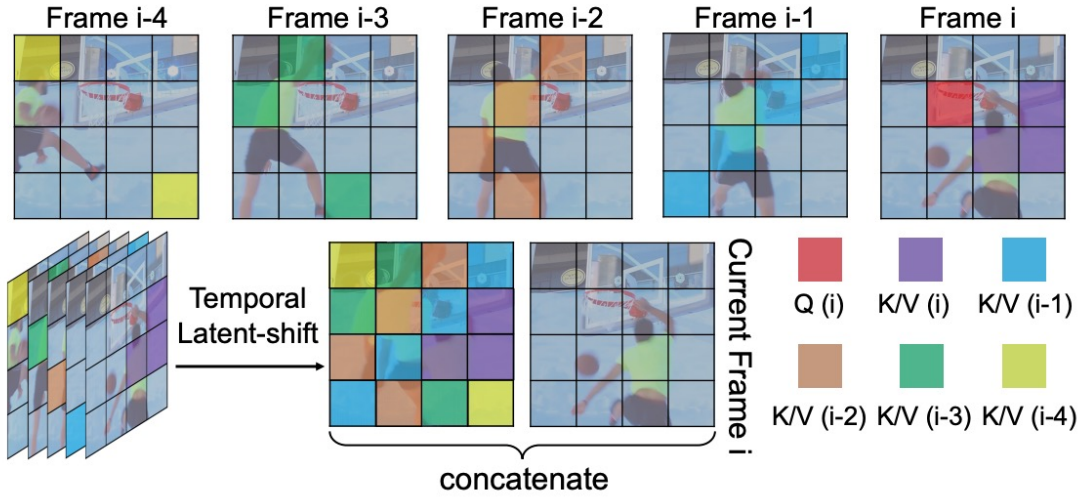


Figure 3. The overview of Temporal Latent-shift Attention module. It is noted that the Latent-shift attention is performed on latent space, but the visualization overview is shown on image-level for clear presentation.

$$\mathbf{Q} = \mathbf{W}_{\mathrm{q}}(\boldsymbol{x}_{z_i}),$$

$$\mathbf{K} = \mathbf{W}_{\mathrm{k}}[\boldsymbol{x}_{z_i}, \boldsymbol{x}_{z_{shift}}],$$

$$\mathbf{V} = \mathbf{W}_{\mathrm{v}}[\boldsymbol{x}_{z_i}, \boldsymbol{x}_{z_{shift}}],$$

$$O(L^2N^2) \text{ --> } O(2LN^2)$$

Video Editing Task

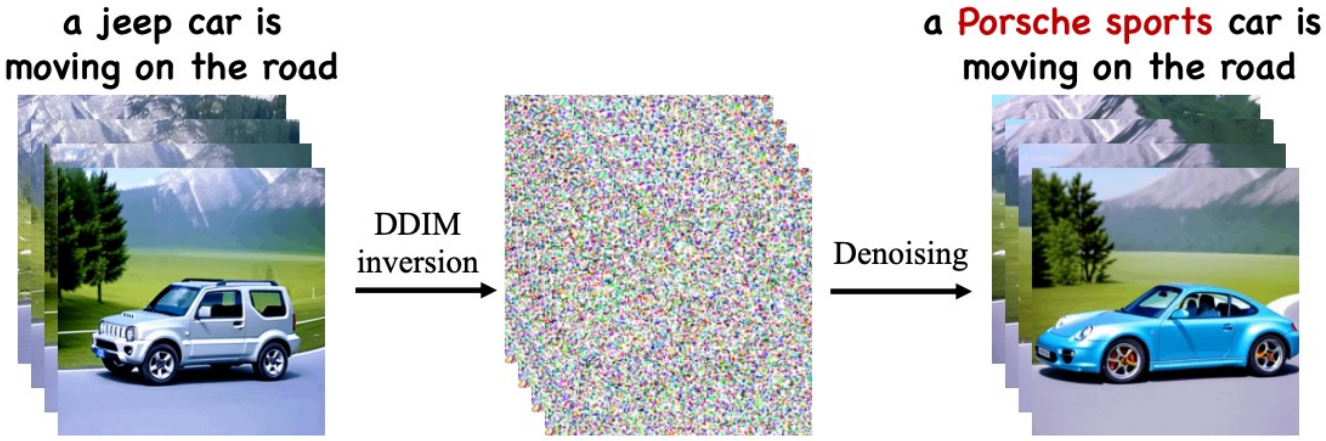


Figure 4. During inference, we sample a novel video from the latent noise inverted from the input video, guided by an edited prompt (*e.g.*, "a Porsche sports car is moving on the road").

Video Super Resolution Task

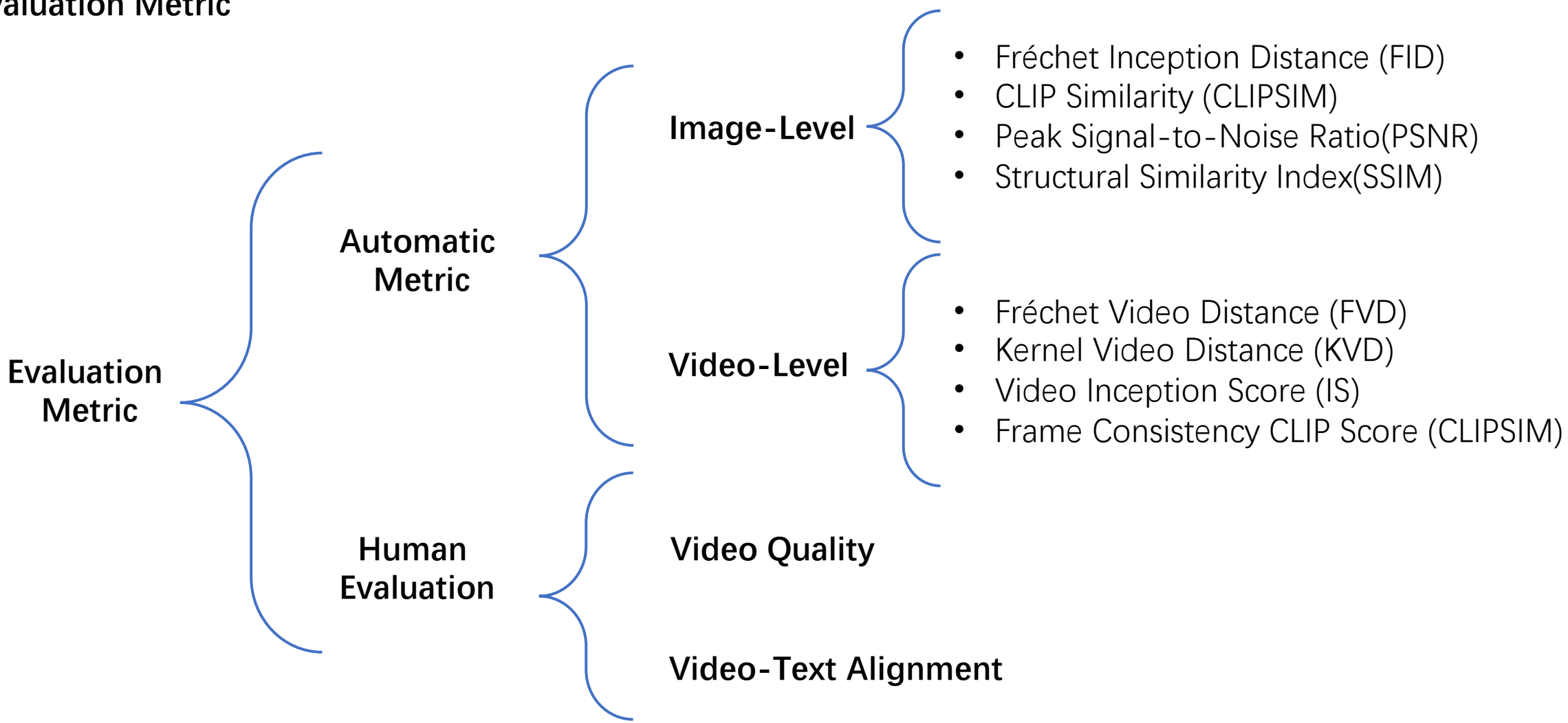
$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta([\boldsymbol{x}_t, \boldsymbol{x}_{low}], \boldsymbol{c}, t)||_2^2],$$

Model Size and Inference Speed Comparisons

Table 1. Model size and inference speed comparisons. The speed is measured in seconds on one A100 (80GB) GPU. The majority of results are sourced from [1].

Method	**Parameters (Billion)**								**Speed (s)**
	T2V Core	Auto Encoder	Text Encoder	Prior Model	Super Resolution	Frame Interpolation	Overall	Tuned	
CogVideo [39]	7.7	0.10	–	–	–	7.7	15.5	15.5	434.53
Make-A-Video [82]	3.1	–	0.12	1.3	$1.4 + 0.7$	3.1	9.72	9.72	–
Imagen Video [35]	5.6	–	4.6	–	$1.2 + 1.4 + 0.34$	$1.7 + 0.78 + 0.63$	16.25	16.25	–
Video LDM [8]	1.51	0.08	0.12	–	0.98	1.51	4.20	2.65	–
Latent-VDM [1]	0.92	0.08	0.58	–	–	–	1.58	0.92	28.62
Latent-Shift [1]	0.88	0.08	0.58	–	–	–	1.53	0.88	23.40
LVDM [32]	0.96	0.08	0.12	–	–	–	1.16	1.04	21.23
SimDA (Ours)	0.88	0.08	0.12	–	–	–	**1.08**	**0.025**	11.20

Evaluation Metric



Text2Video Result Comparisons

Table 2. Text-to-Video generation comparison on MSR-VTT [106] dataset. We report the Fréchet Video Distance (FVD) scores and CLIPSIM scores.

Method	Training Data	Resolution	Zero-shot	Params(B)	FVD(↓)	CLIPSIM(↑)
GODIVA [98]	MSR-VTT	128x128	No	-	-	0.2402
NÜWA [99]	MSR-VTT	128x128	No	0.87	-	0.2439
Make-A-Video [82]	WebVid-10M + HD-VILA-10M	256x256	Yes	9.72	-	0.3049
VideoFactory [95]	WebVid-10M + HD-VG-130M	256x256	Yes	2.04	-	0.3005
LVDM [32]	WebVid-2M	256x256	Yes	1.16	742	0.2381
MMVG [22]	WebVid-2.5M	256x256	Yes	-	-	0.2644
CogVideo [39]	WebVid-5.4M	256x256	Yes	15.5	1294	0.2631
ED-T2V [54]	WebVid-10M	256x256	Yes	1.30	-	0.2763
MagicVideo [121]	WebVid-10M	256x256	Yes	-	998	-
Video-LDM [8]	WebVid-10M	256x256	Yes	4.20	-	0.2929
VideoComposer [96]	WebVid-10M	256x256	Yes	1.85	580	0.2932
Latent-Shift [1]	WebVid-10M	256x256	Yes	1.53	-	0.2773
VideoFusion [58]	WebVid-10M	256x256	Yes	1.83	581	0.2795
SimDA (Ours)	WebVid-10M	256x256	Yes	**1.08**	**456**	**0.2945**

Table 3. Text-to-video generation on the validation set of WebVid [5]. We report the FVD and CLIPSIM scores.

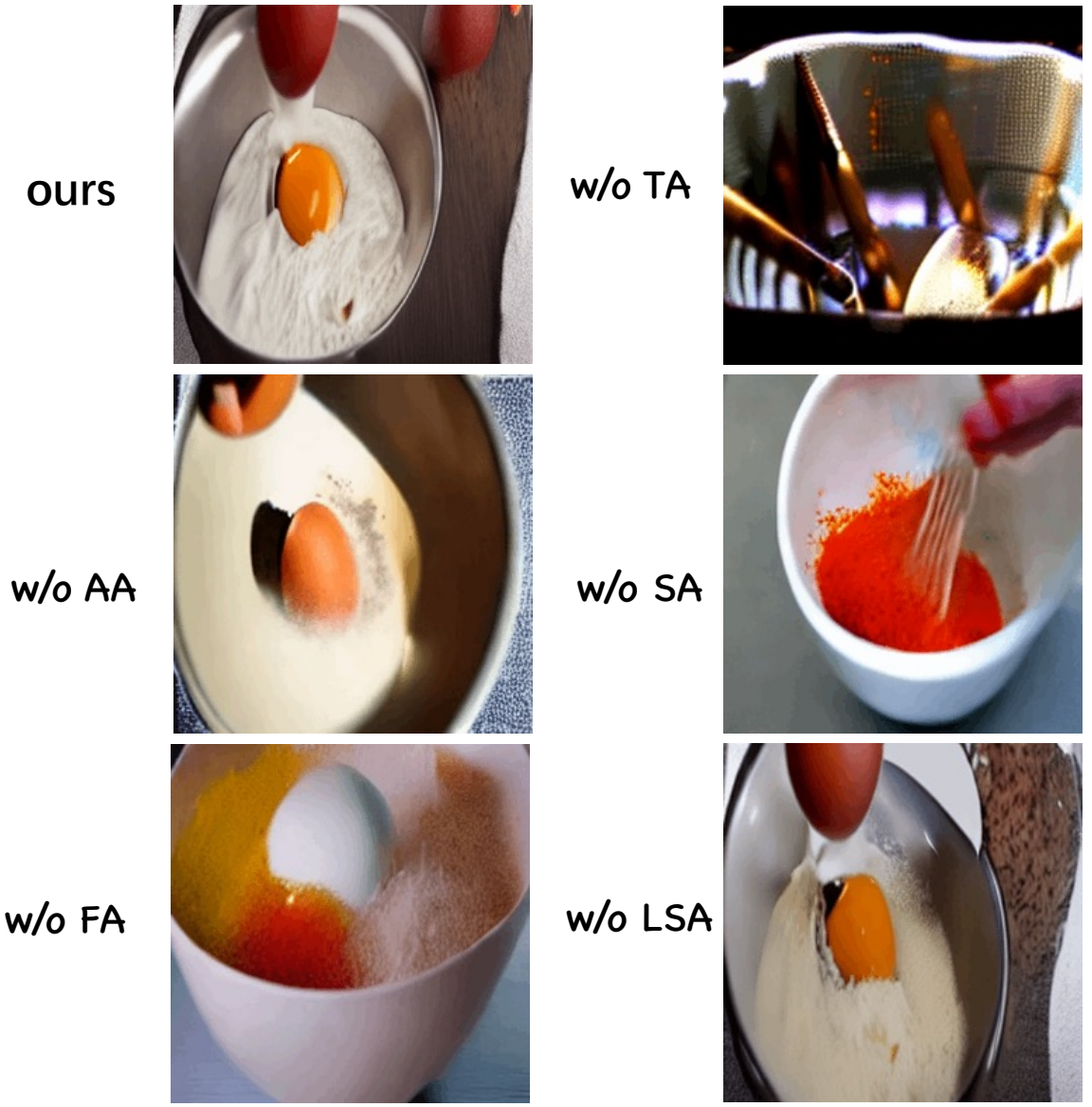
Method	Params(B)	FVD(↓)	CLIPSIM(↑)
LVDM [32]	1.16	455.53	0.2751
VideoFusion [58]	1.83	414.11	0.3000
VideoFactory [95]	2.04	292.35	0.3070
SimDA (Ours)	**1.08**	**363.98**	**0.3054**

Sample	Method	Param Ratio	Quality	Faithfulness
Open Website	VDM [38]	0.83×	85.2%	81.4%
	Latent-Shift [1]	1.41×	81.5%	79.3%
Pretrained Model	VideoFusion [58]	1.69×	78.3%	79.5%
	LVDM [32]	1.07×	83.4%	84.7%

Ablation Study

Table 6. Ablation study on different modules. We report the FVD [90] and CLIPSIM [70] on $1K$ samples from the validation set of WebVid-10M [5]. TA, SA, AA, FA and LSA represent Temporal Adapter, Spatial Adapter, Attention Adapter, FFN Adapter and Latent-shift Attention, respectively.

	TA	AA	FA	LSA	FVD(↓)	CLIPSIM(↑)
w/o TA		✓	✓	✓	1470.1	0.2629
w/o SA	✓			✓	811.3	0.2822
w/o AA	✓		✓	✓	764.8	0.2851
w/o FA	✓	✓		✓	623.7	0.2962
w/o LSA	✓	✓	✓		618.2	0.3011
Ours	✓	✓	✓	✓	**530.2**	**0.3034**



Prompt: Mixer in a bowl to beat the milk and egg on a black table. slow motion.

T2V Examples

An astronaut flying in space,4k,high resolution

Xmas Christmas tree holiday celebration winter snow animation gold background

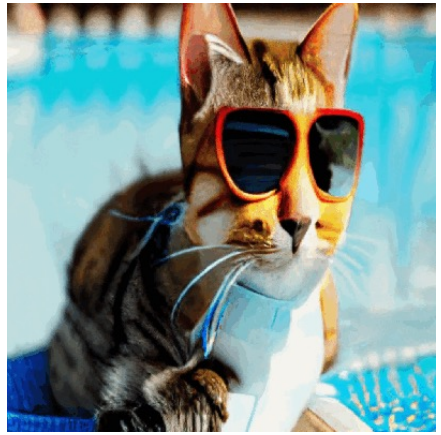
Standing on top a mountainside watching the sunset with the vivid pinks red orange showing from the fire colored sky.

Sea waves with foam on white tropical sandy beach.

Beer pouring into glass, low angle video shot

Time lapse at a fantasy landscape, 4k, high resolution.

A red Cardinal on a tree branch stands out when the snow is falling.



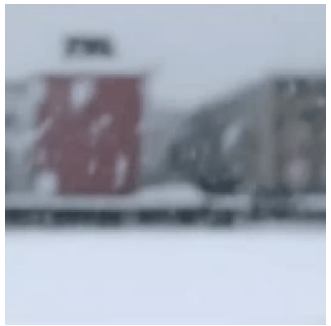
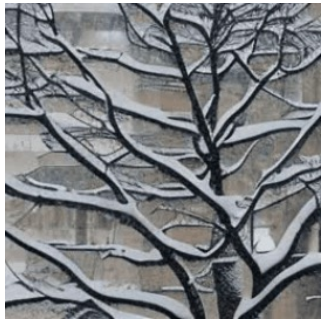
A cat wearing sunglasses and working as a lifeguard at a pool.

A beautiful sunrise on mars, Curiosity rover.

Coffee pouring into a cup.

Homepage: https://ChenHsing.github.io/SimDA

T2V Comparison Examples

	CogVideo	VDM	LVDM	ModelScope	SimDA(ours)
4K illuminated Christmas Tree at Night During Snowstorm					
Irrigation Canal in Western USA Water sourced from the Colorado River 4K Aerial Video					
Snowfall in city					

Homepage: https://ChenHsing.github.io/SimDA

Video Editing Examples

A jeep car is moving on the road.

A jeep car is moving on the road, Cartoon Style.

A Porsche Sports car is moving on the road.

An AE86 car is moving on the road.

A jeep car is moving on the road, in snowy day.

A jeep car is moving on the road, Van Gogh Style.

A jeep car is moving in a forest, in Autumn.

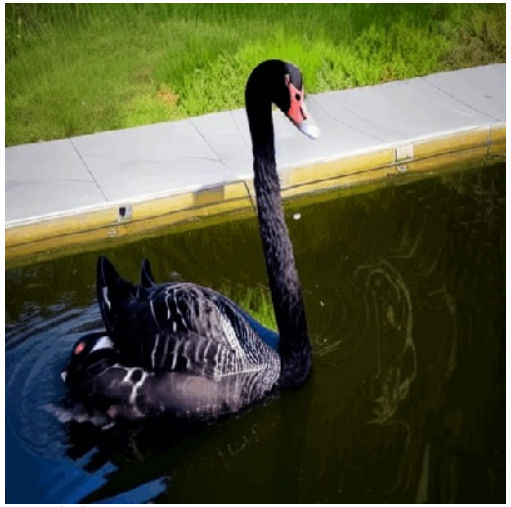
A jeep car is moving on the desert, at dark.

Video Editing Comparison

Table 5. Quantitative comparison with evaluated baseline [100]. The "Training" refers to the process of optimization, and "Memory" refers to the GPU memory.

Method	Frame consistency		Textual alignment		Runtime [min]		Memory [Gib]		Params [Mb]
	CLIP Score↑	User Vote↑	CLIP Score↑	User Vote↑	Training↓	Inference↓	Training↓	Inference↓	Tuned↓
Tune-A-Video [100]	94.1	31.2%	31.8	39.5%	9.3	0.8	31.3	11.4	74.4
SimDA(Ours)	**94.9**	**68.8%**	**31.9**	**60.5%**	**2.5**	**0.4**	**28.6**	**8.8**	**24.9**

a black swan is swimming in a pool.



a rabbit is eating a watermelon

Input

a swarovski blue crystal swan is swimming in a pool.

a cat is eating an apple

Tune-A-Video

a swarovski blue crystal swan is swimming in a pool.

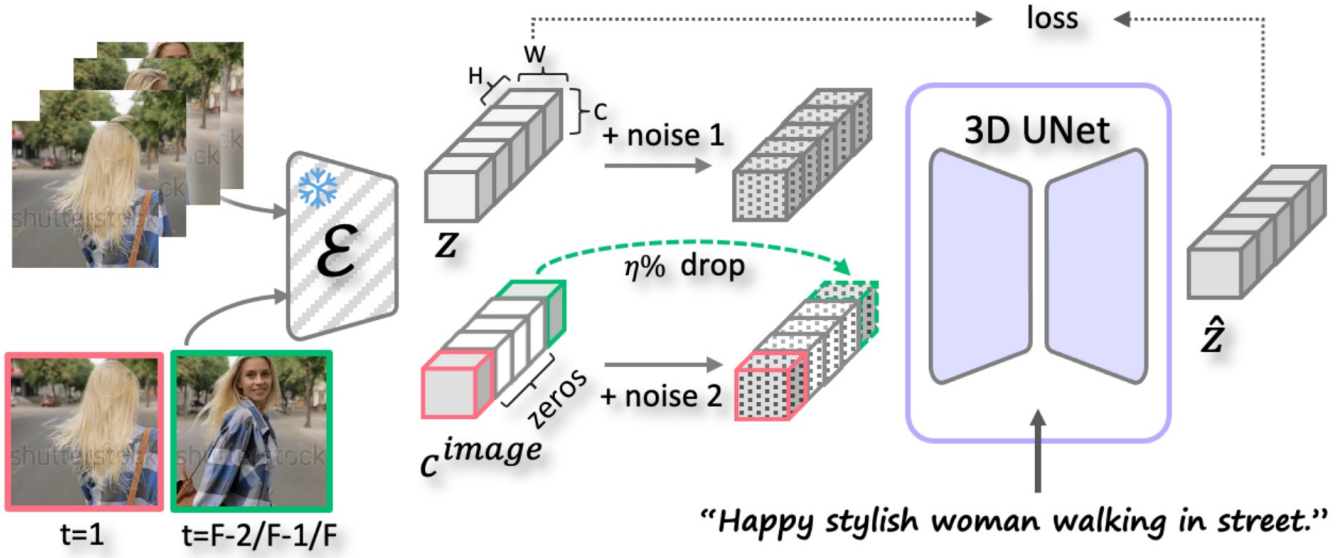
a cat is eating an apple

SimDA(ours)

PixelDance & Emu Video (Image Condition)

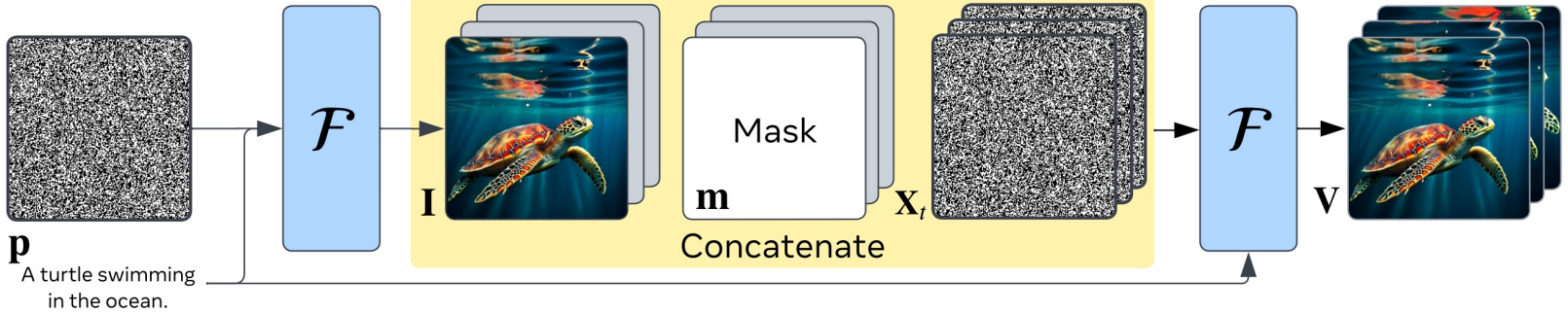
A beautiful woman with long golden hair is driving a red convertible sports car.

The skull is burining, red fires, the skull exploding



fireworks

A panda bear driving a car



Make Pixels Dance: High-Dynamic Video Generation (CVPR 2024)

EMU VIDEO: Factorizing Text-to-Video Generation by Explicit Image Conditioning（Arxiv 2023）

Stable Video Diffusion(Image Condition)

- Stage I: image pretraining, i.e. a 2D text-to-image diffusion model.
- Stage II: video pretraining, which trains on large amounts of videos.
- Stage III: video finetuning, which refines the model on a small subset of high-quality videos at higher resolution.

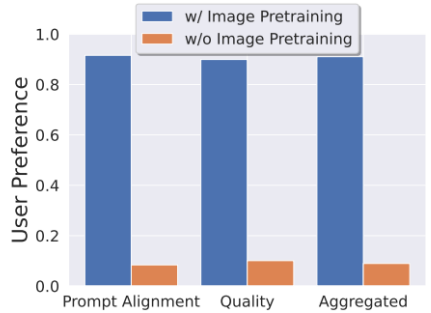
Video Caption Method

- Mid-frame caption, CoCa
- Video-based caption, V-BLIP
- LLM summarization

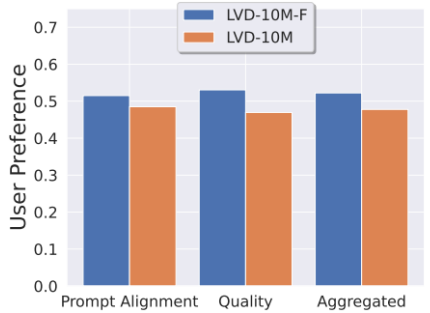
Source Video	Caption		
	CoCa	VBLIP	LLM
	there is a piece of wood on the floor next to a tape measure .	a person is using a ruler to measure a piece of wood	A person is using a ruler to measure a piece of wood on the floor next to a tape measure.
	two men sitting on a rock near a river . one is holding a stick and the other is holding a pole .	two people are fishing in a river	Two men are fishing in a river. One is holding a stick and the other is holding a pole.

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?	✓	✓	✓	✓	✗	✗



(a) Initializing spatial layers from pretrained images models greatly improves performance.

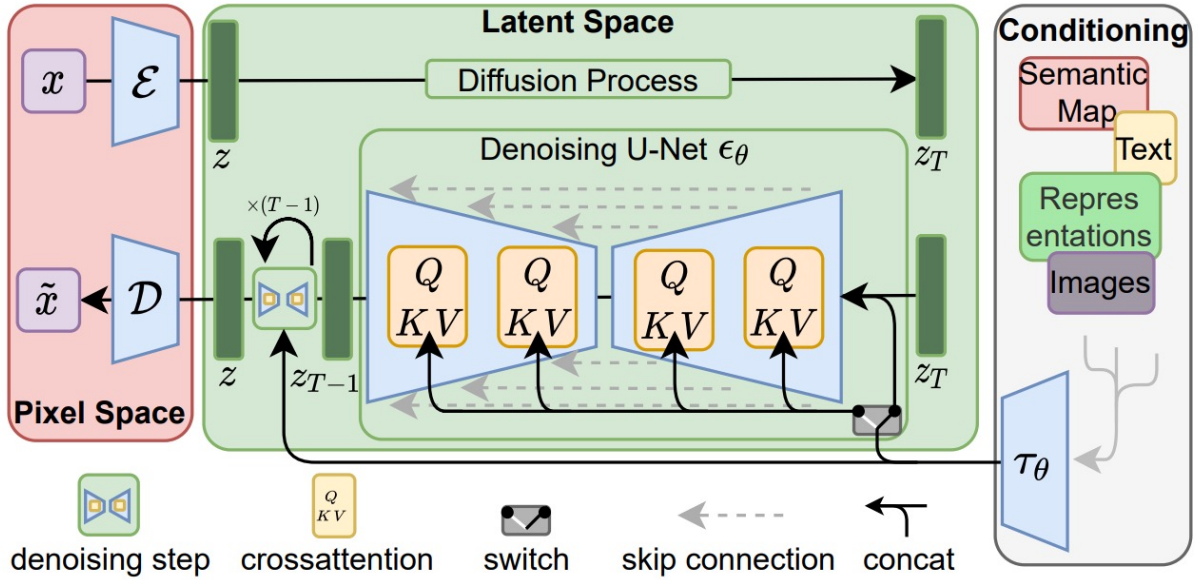


(b) Video data curation boosts performance after video pretraining.

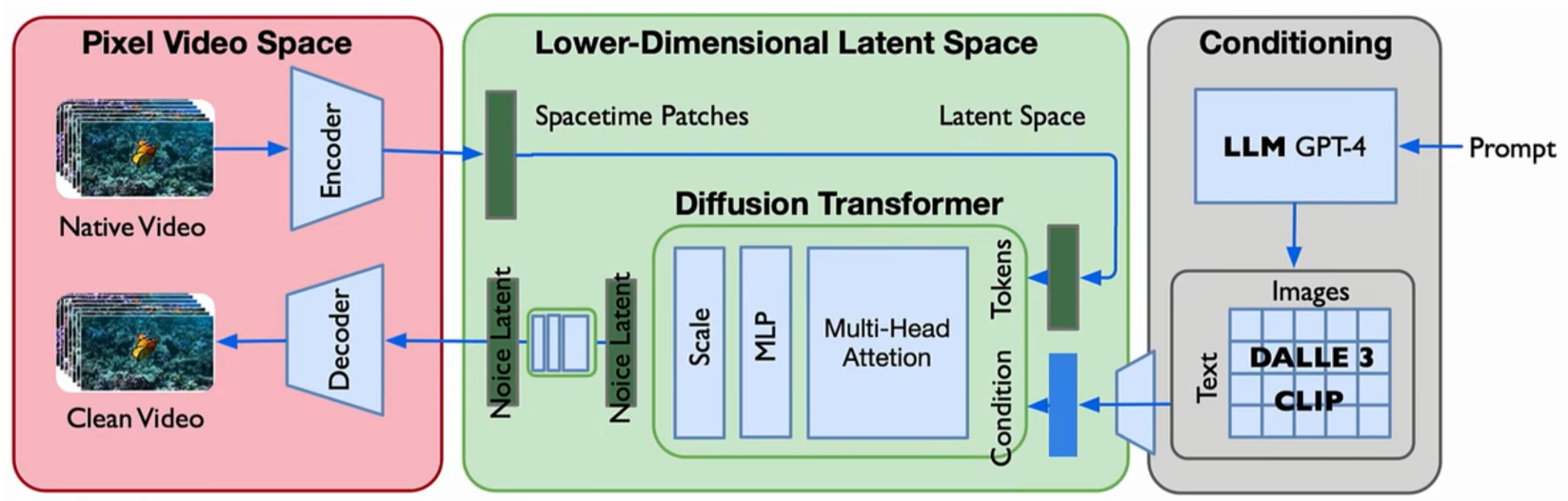
Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets (Arxiv 2023)

Sora (Architecture)

Stable Diffusion



Sora



1. https://openai.com/research/video-generation-models-as-world-simulators
2.https://www.bilibili.com/video/BV1Bx4y1k7BQ/?spm_id_from=333.337.search-card.all.click&vd_source=b94ff746cdc4d7a1b331e9f42ce92950

Sora (Datasets)

Language understanding

Training text-to-video generation systems requires a large amount of videos with corresponding text captions. We apply the re-captioning technique introduced in DALL·E 3[30] to videos. We first train a highly descriptive captioner model and then use it to produce text captions for all videos in our training set. We find that training on highly descriptive video captions improves text fidelity as well as the overall quality of videos.

Similar to DALL·E 3, we also leverage GPT to turn short user prompts into longer detailed captions that are sent to the video model. This enables Sora to generate high quality videos that accurately follow user prompts.

1. **Train a captioner model**
2. **GPT4 captioning**
3. **virtual engine**

Prompt: The camera follows behind a white vintage SUV with a black roof rack as it speeds up a steep dirt road surrounded by pine trees on a steep mountain slope, dust kicks up from it's tires, the sunlight shines on the SUV as it speeds along the dirt road, casting a warm glow over the scene. The dirt road curves gently into the distance, with no other cars or vehicles in sight. The trees on either side of the road are redwoods, with patches of greenery scattered throughout. The car is seen from the rear following the curve with ease, making it seem as if it is on a rugged drive through the rugged terrain. The dirt road itself is surrounded by steep hills and mountains, with a clear blue sky above with wispy clouds.

1. https://openai.com/research/video-generation-models-as-world-simulators

CVPR

FUDAN UNIVERSITY 1905 復旦大學 FUDAN UNIVERSITY

THANK YOU

A Survey on Video Diffusion Models

Zhen Xing, Fudan University, China

2024/04/24

SimDA Homepage