A Survey on Video Diffusion Models

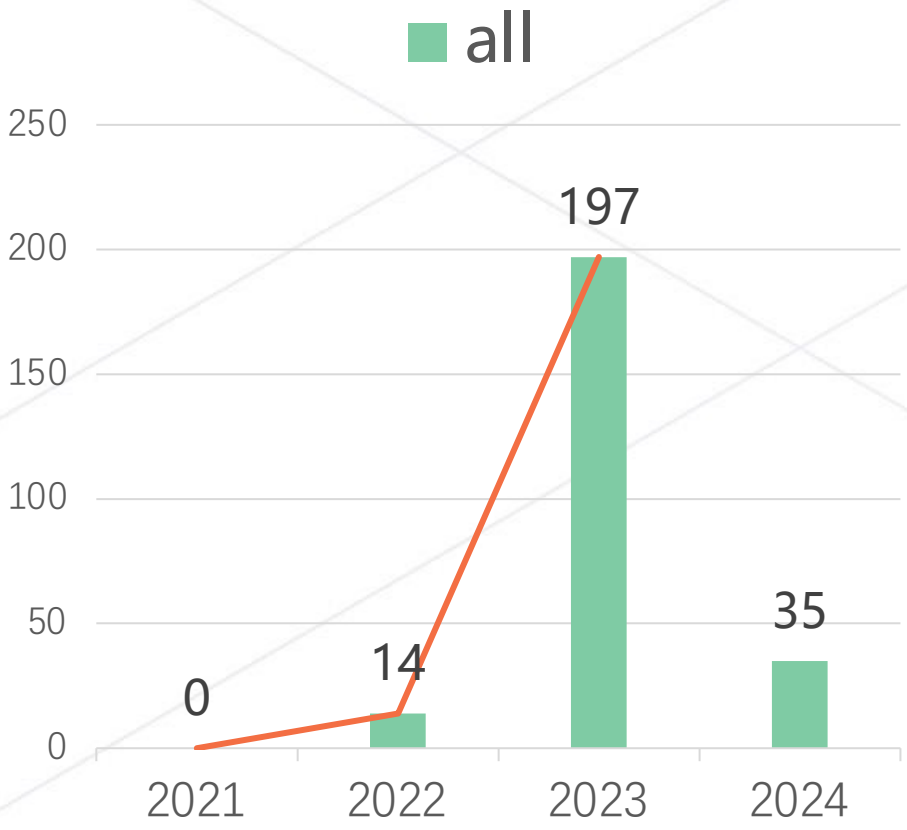
邢桢

2024/02/28

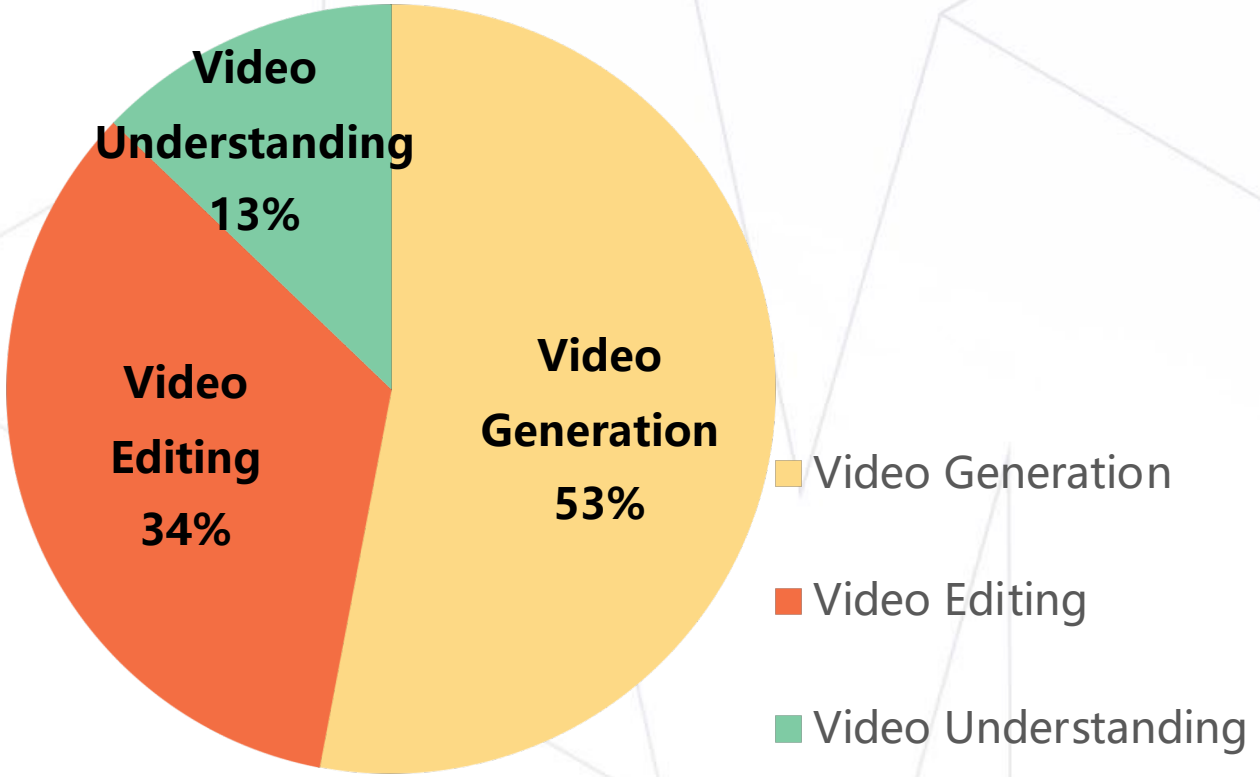
【社区开放麦】面向所有社区成员的技术分享活动，每周四晚八点准点播出。旨在搭建一个知识分享的舞台，在这里，社区里的每个人都能拿起话筒分享你的知识和见解。我们一直认为，分享与交流能更好地促进知识的传播；平等与共建能更好地维持社区的氛围。

加入社群，与讲师 1V1 沟通

Overview

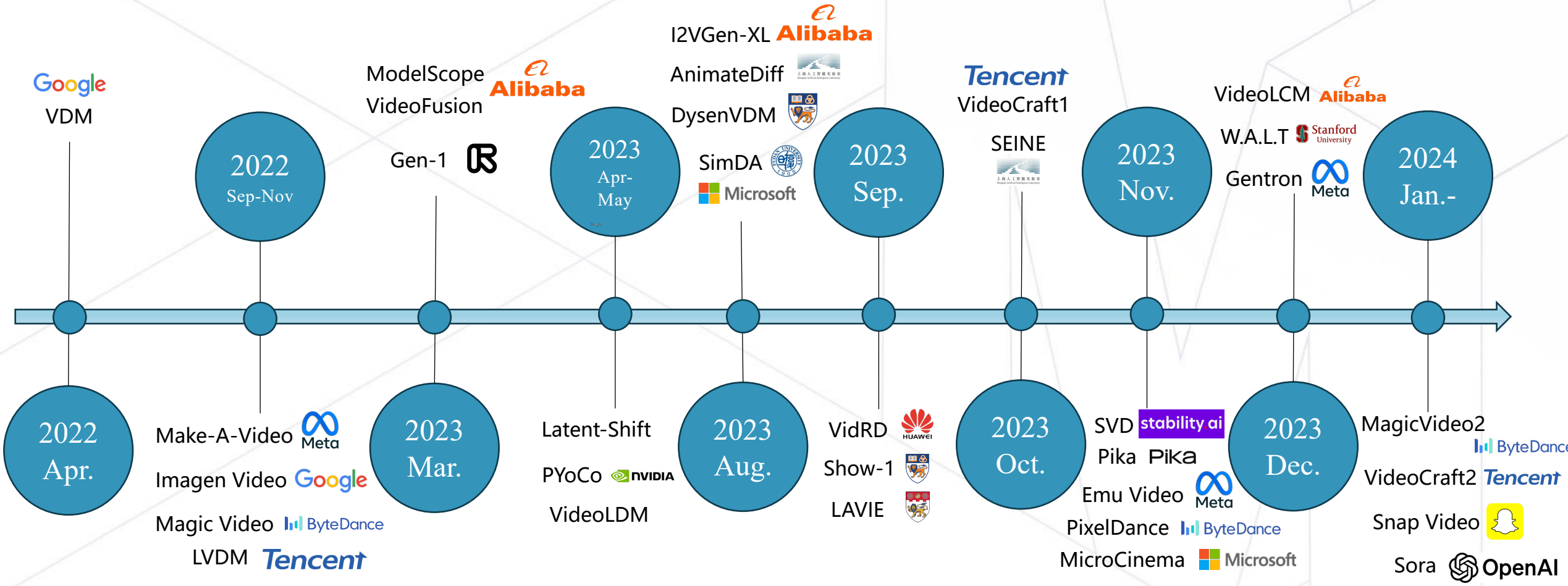


(a) Number of research work



(b) Ratio of Different Directions

Video Generation Models



Video Generation Models

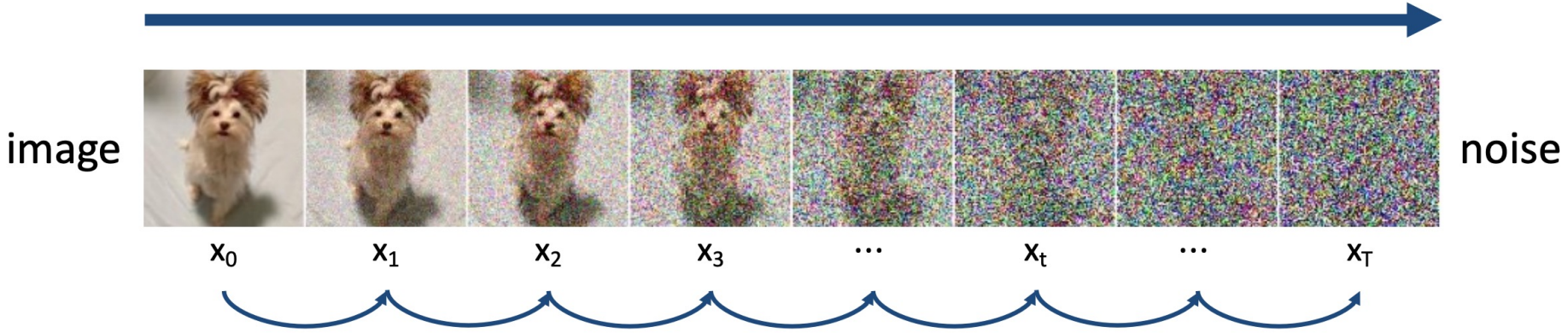
Video Diffusion Models（Google）
2022-04

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

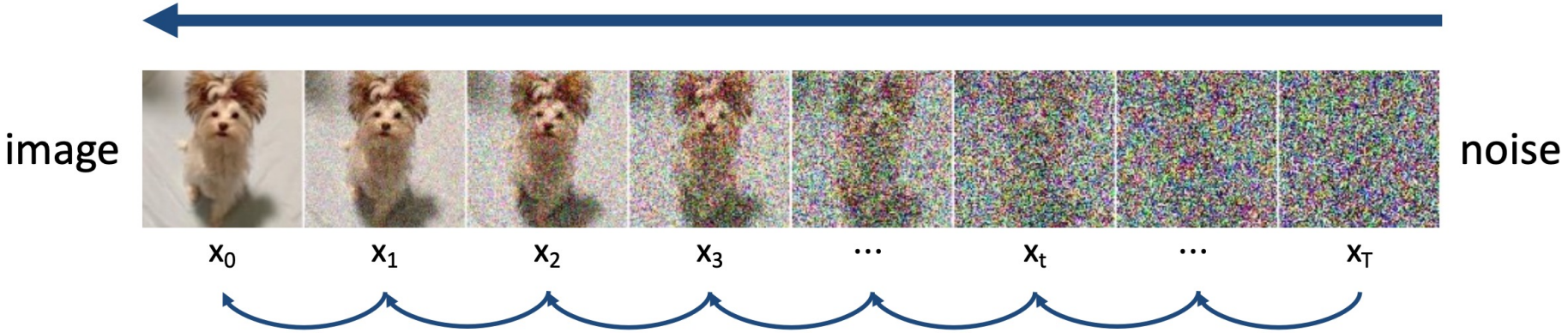
DDPM

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

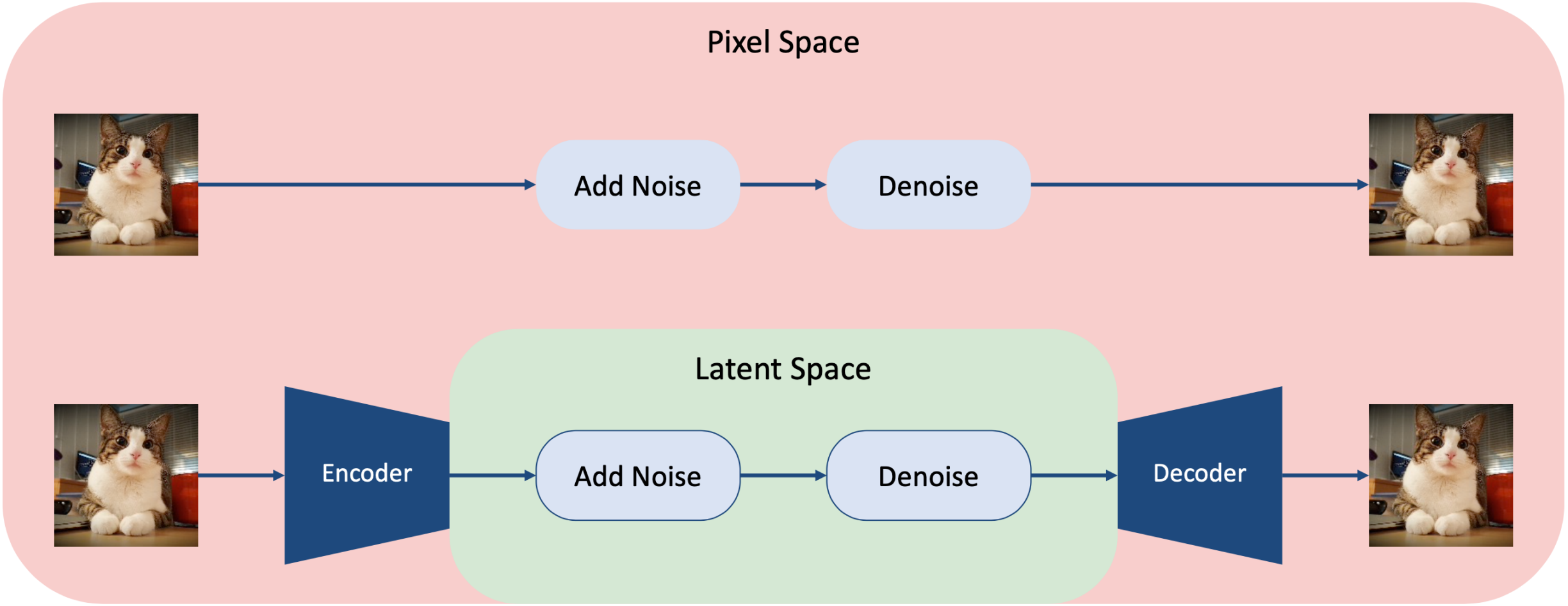


$$p_\theta(x_{t-1}|x_t)$$

Train a neural network model to remove noise gradually

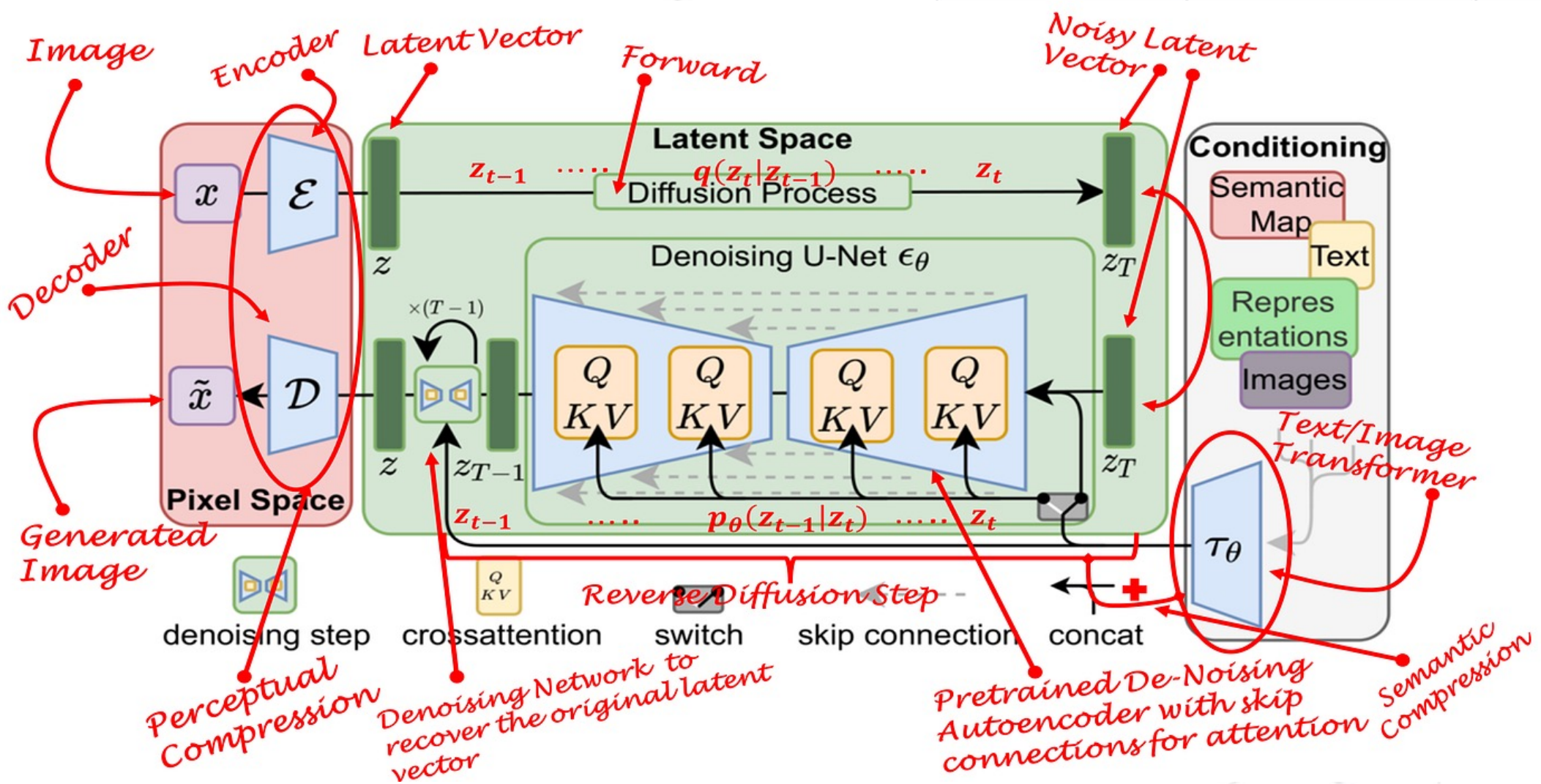
Ho et al., "Denoising Diffusion Probabilistic Models," NeurIPS 2020.
Sohl-Dickstein et al., "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," ICML 2015.
Song et al., "Score-Based Generative Modeling through Stochastic Differential Equations," ICLR 2021. Vahdat et al., "Denoising Diffusion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

Latent Diffusion

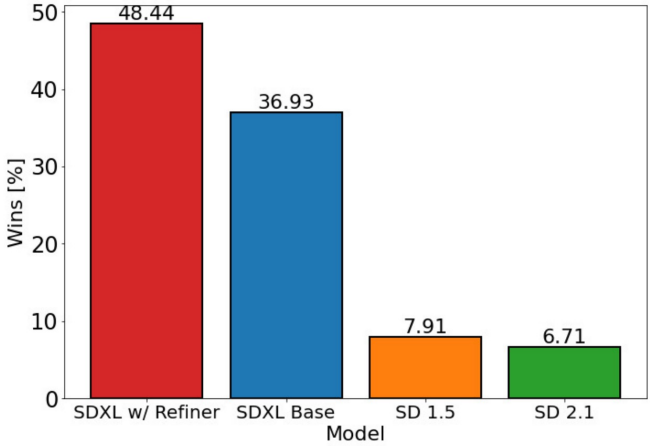


https://sites.google.com/view/showlab/tutorial

Latent Diffusion Model(Stable Diffusion)



Stable Diffusion XL



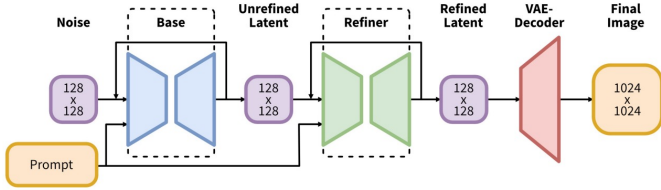


Table 1: Comparison of *SDXL* and older *Stable Diffusion* models.

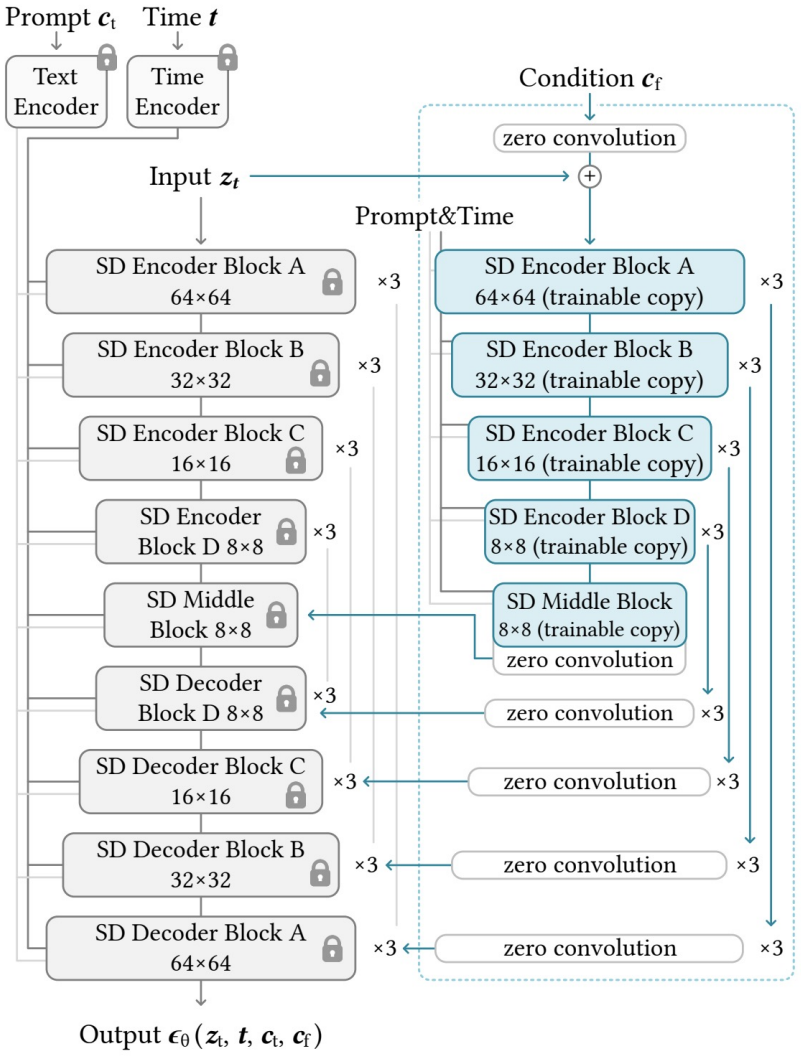
Model	*SDXL*	SD 1.4/1.5	SD 2.0/2.1
# of UNet params	2.6B	860M	865M
Transformer blocks	[0, 2, 10]	[1, 1, 1, 1]	[1, 1, 1, 1]
Channel mult.	[1, 2, 4]	[1, 2, 4, 4]	[1, 2, 4, 4]
Text encoder	CLIP ViT-L & OpenCLIP ViT-bigG	CLIP ViT-L	OpenCLIP ViT-H
Context dim.	2048	768	1024
Pooled text emb.	OpenCLIP ViT-bigG	N/A	N/A

Stable Diffusion 3

The Stable Diffusion 3 suite of models currently ranges from 800M to 8B parameters. This approach aims to align with our core values and democratize access, providing users with a variety of options for scalability and quality to best meet their creative needs. Stable Diffusion 3 combines a diffusion transformer architecture and flow matching. We will publish a detailed technical report soon.

We believe in safe, responsible AI practices. This means we have taken and continue to take reasonable steps to prevent the misuse of Stable Diffusion 3 by bad actors. Safety starts when we begin training our model and continues throughout the testing, evaluation, and deployment. In preparation for this early preview, we've introduced numerous safeguards. By continually collaborating with researchers, experts, and our community, we expect to innovate further with integrity as we approach the model's public release.

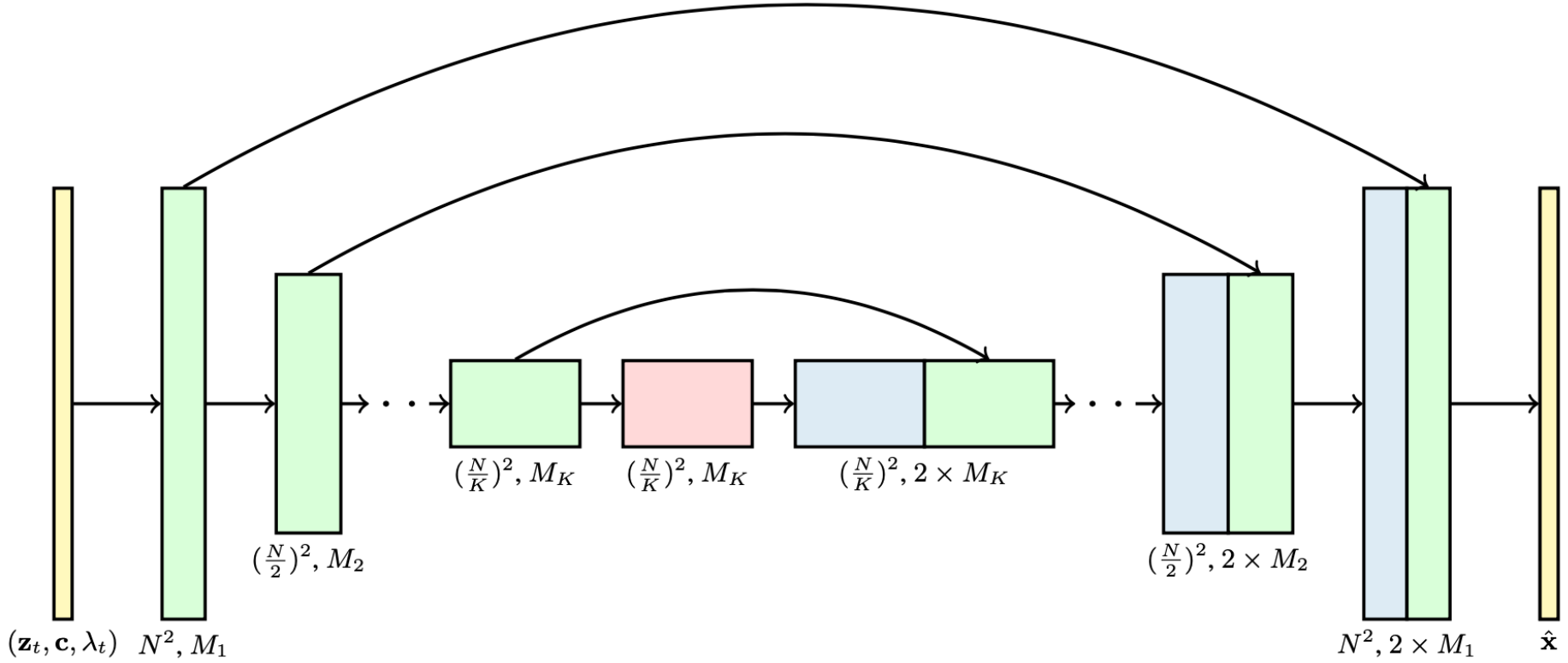
ControlNet



(a) Stable Diffusion (b) ControlNet

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

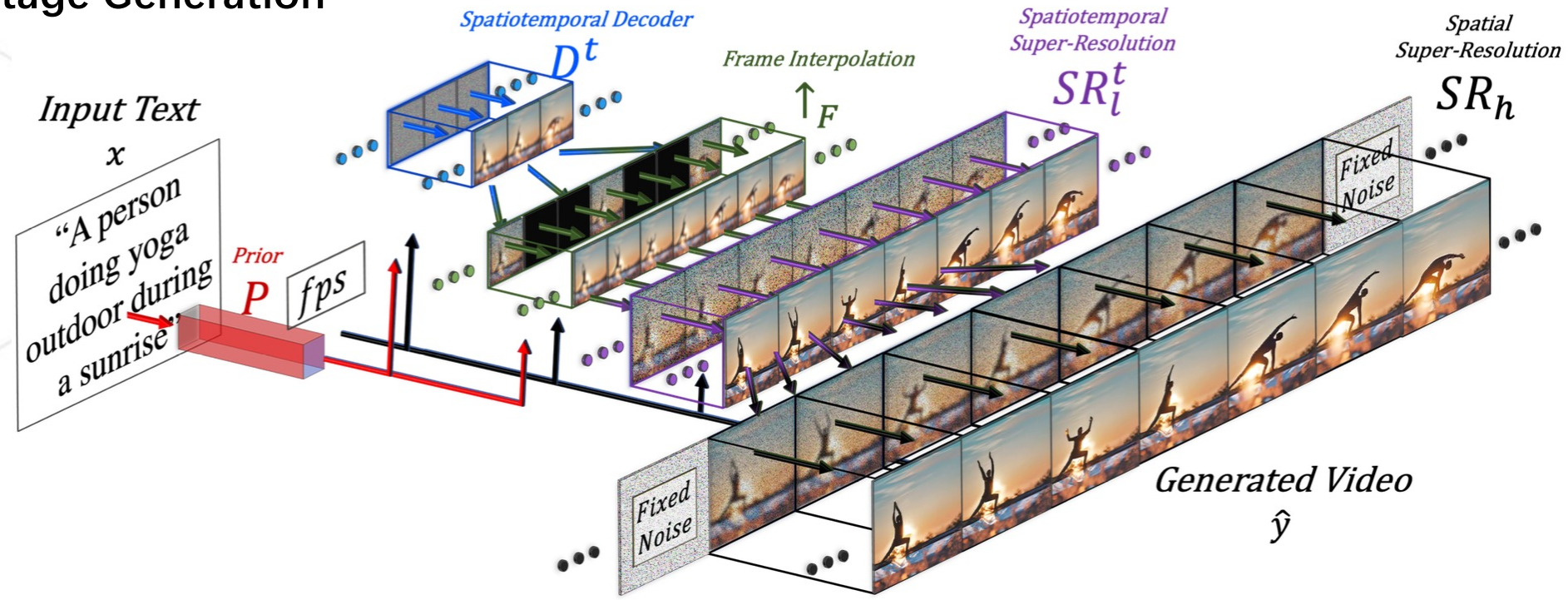
First Video Diffusion Model

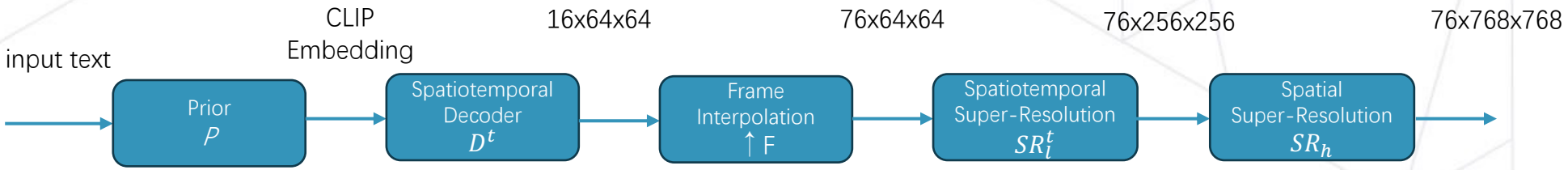


- Conv2D → 3D （3x3 -> 1x3x3）
- Space Attention →Divided Space-Temporal Attention
- Joint training on video and image modeling

Make-A-Video

Multi-stage Generation

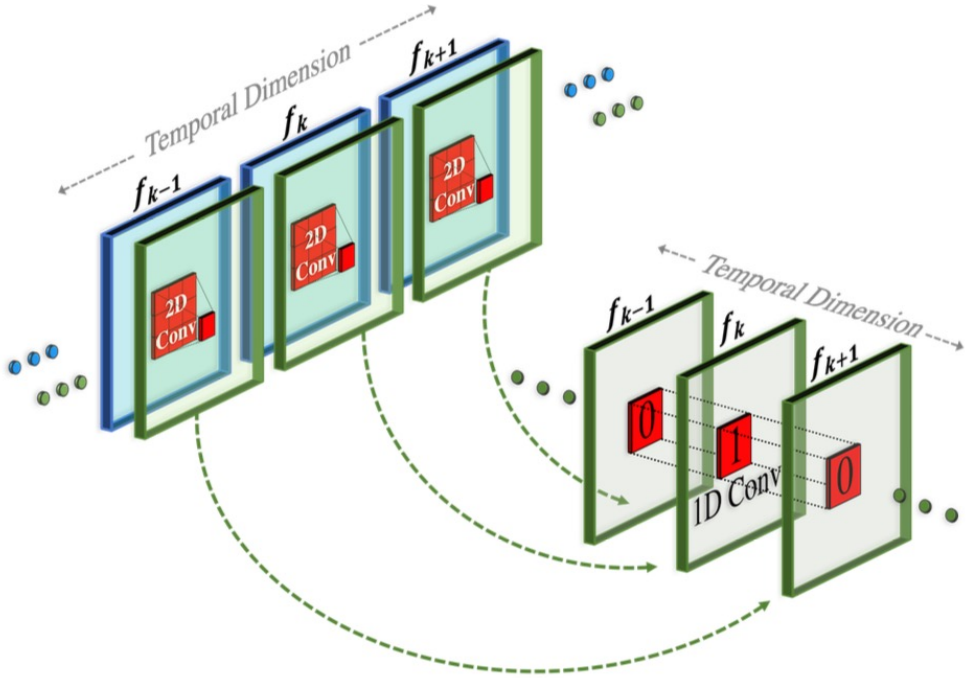




Make-A-Video

Pseudo-3D Convolution Layers

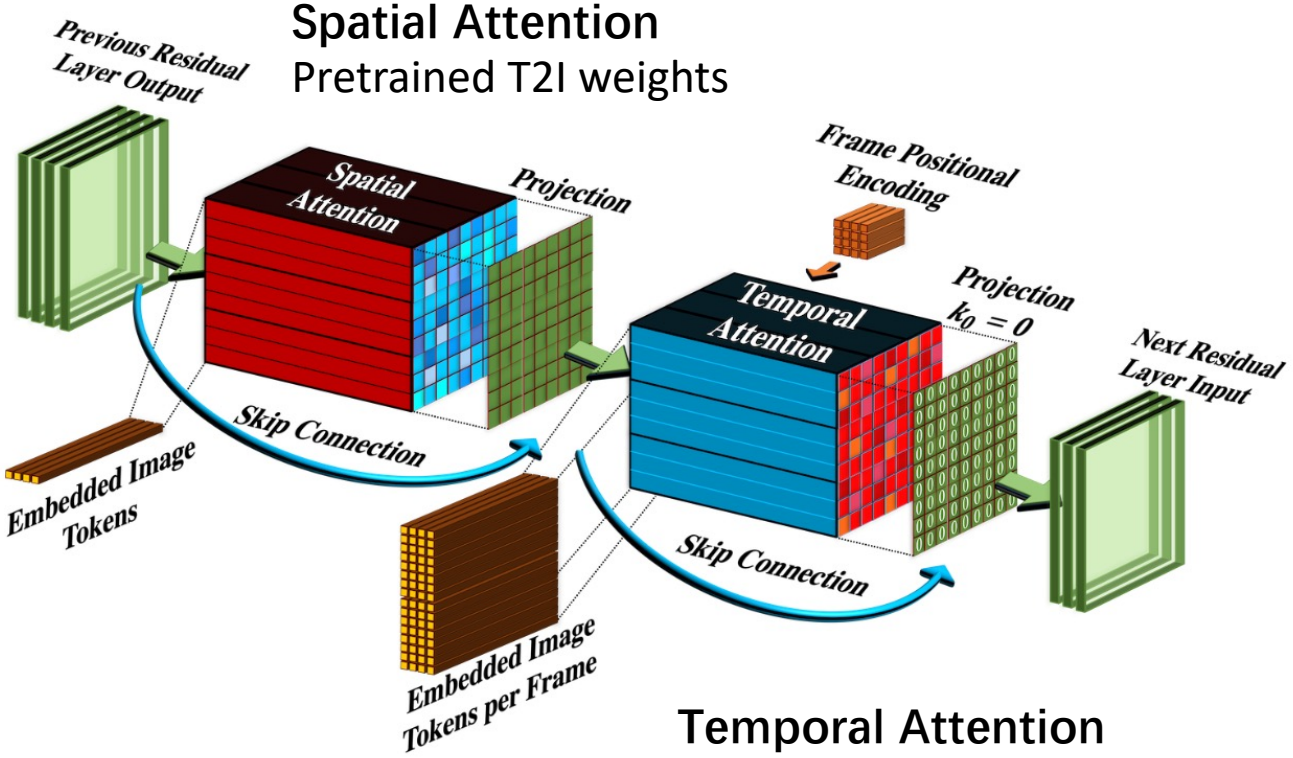
Spatial Convolution
Pretrained T2I weights



Temporal Convolution

Divided Spatial Temporal Attention Layers

Spatial Attention
Pretrained T2I weights



Temporal Attention
Initialized with zero temporal projection

Imagen & Imagen Video

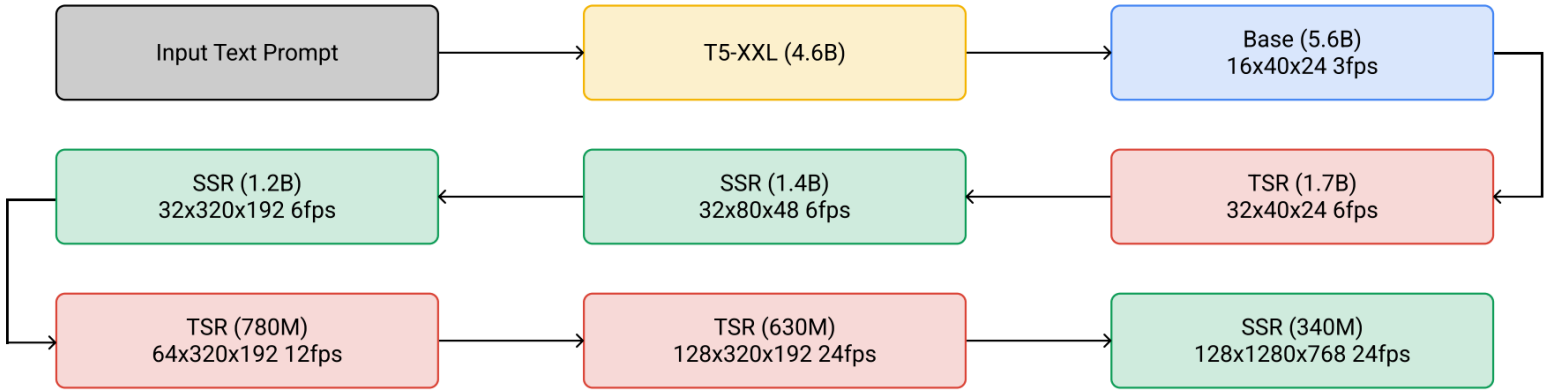
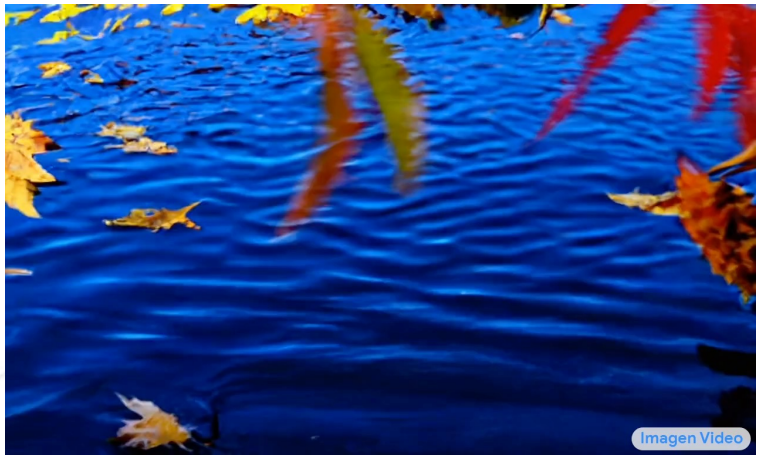
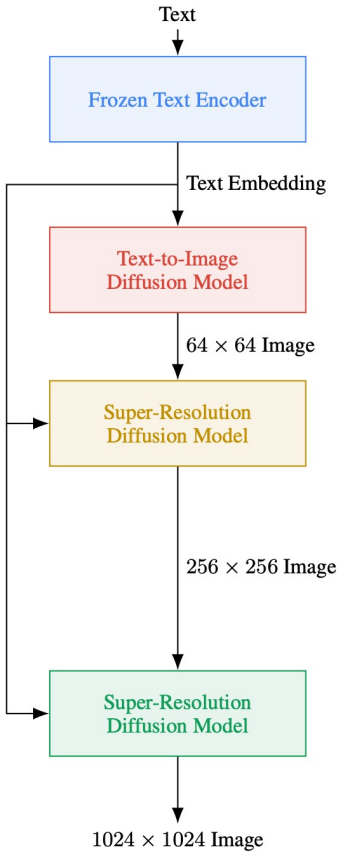


Imagen Video



Imagen



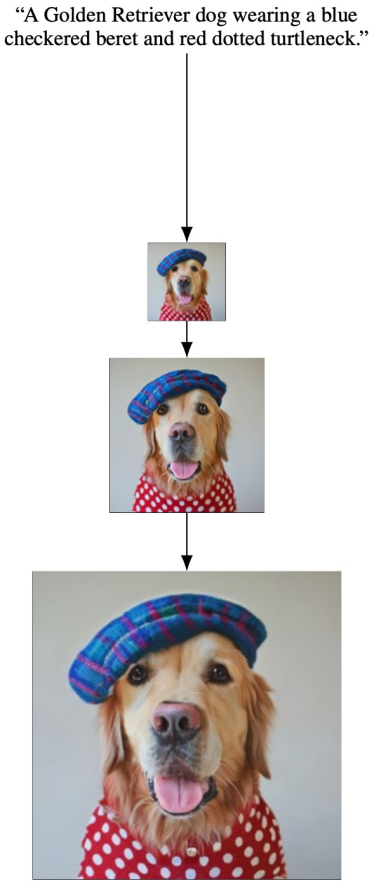
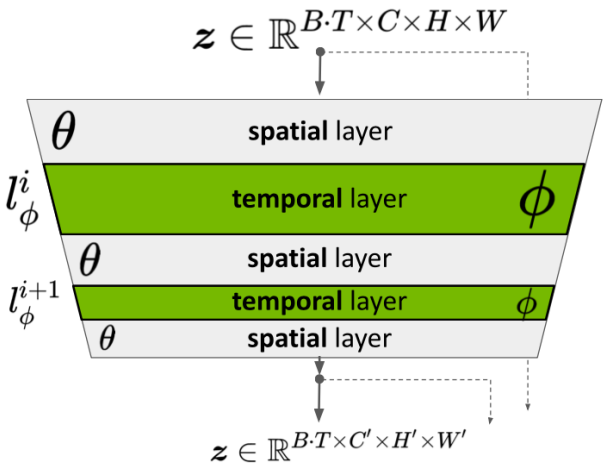
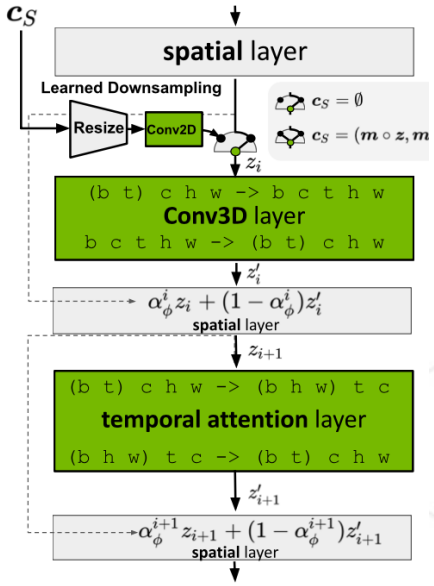


Imagen Video: High Definition Video Generation with Diffusion Models （arXiv 2023）

Latent Video Diffusion Model

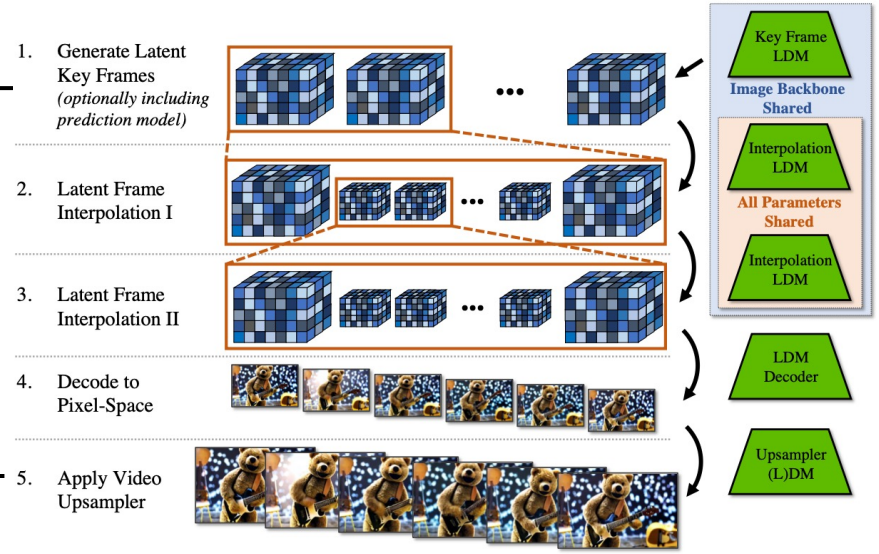




- Trained on **256** GPUs
- Batch Size 768, 402k steps
- **1.71B** Parameters

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

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



Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models （CVPR2023）

Training Dataset

WebVid-10M

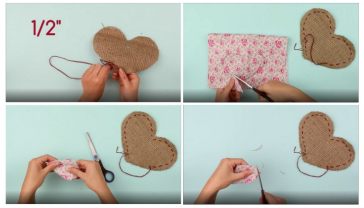
"Runners feet in a sneakers close up. realistic three dimensional animation."

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

"Billiards, concentrated young woman playing in club"

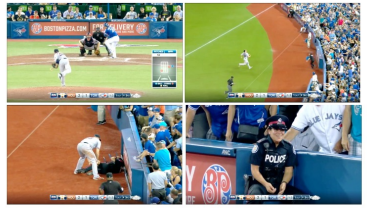
Top2 retrieval result: A dog is running on the road.

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.

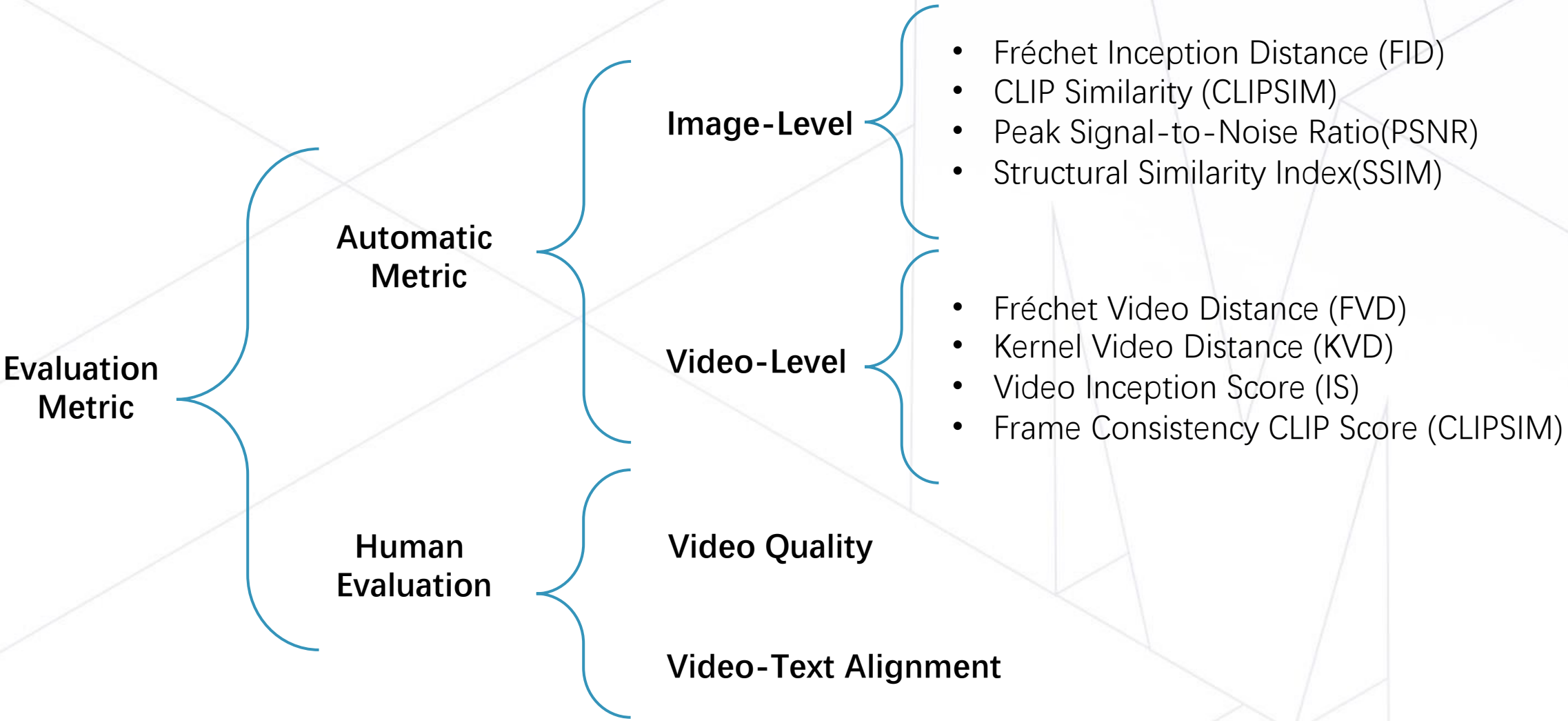
Dataset	Domain	#Video clips	#Sentence	Avg len(sec)	Sent len	Duration(h)	Resolution
MSR-VTT [59]	open	10K	200K	15.0	9.3	40	240p
DideMo [2]	Flickr	27K	41K	6.9	8.0	87	-
LSMDC [45]	movie	118K	118K	4.8	7.0	158	1080p
YouCook II [69]	cooking	14K	14K	19.6	8.8	176	-
How2 [47]	instructional	80K	80K	90.0	20.0	2K	-
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 ）

Evaluation Metric



Evaluation Benchmark

Method	Year	Training Data	Extra Dependency	Resolution	Params(B)	MSRVTT [235]			UCF-101 [188]		
						FID(↓)	FVD(↓)	CLIPSIM(↑)	FID(↓)	FVD(↓)	IS(↑)
Non-diffusion based method											
CogVideo [78]	2022	[5](5.4M)	-	256 × 256	15.5	23.59	1294	0.2631	179.00	701.59	25.27
MMVG [52]	2023	[5](2.5M)	-	256 × 256	-	-	-	0.2644	-	-	-
Diffusion based method											
LVDM [69]	2022	[5](2M)	-	256 × 256	1.16	-	742	0.2381	-	641.8	-
MagicVideo [266]	2022	[5](10M)	-	256 × 256	-	-	998	-	145.00	699.00	-
Make-A-Video [181]	2022	[5, 236]	-	256 × 256	9.72	13.17	-	0.3049	-	367.23	33.00
ED-T2V [129]	2023	[5](10M)	-	256 × 256	1.30	-	-	0.2763	-	-	-
InternVid [220]	2023	[5](10M) + 18M*	-	256 × 256	-	-	-	0.2951	60.25	616.51	21.04
Video-LDM [10]	2023	[5](10M)	-	256 × 256	4.20	-	-	0.2929	-	550.61	33.45
VideoComposer [217]	2023	[5](10M)	-	256 × 256	1.85	-	580	0.2932	-	-	-
Latent-shift [2]	2023	[5](10M)	-	256 × 256	1.53	15.23	-	0.2773	-	-	-
VideoFusion [135]	2023	[5](10M)	-	256 × 256	1.83	-	581	0.2795	75.77	639.90	17.49
Make-Your-Video [230]	2023	[5](10M)	Depth Input	256 × 256	-	-	-	-	-	330.49	
PYoCo [55]	2023	[5] (22.5M)	-	256 × 256	-	**9.73**	-	-	-	355.19	47.76
CoDi [194]	2023	[5, 236]	-	512 × 512	-	-	-	0.2890	-	-	-
NExT-GPT [229]	2023	[5, 236]	-	320 × 576	1.83	13.04	-	0.3085	-	-	-
SimDA [232]	2023	[5](10M)	-	256 × 256	**1.08**	-	456	0.2945	-	-	-
Dysen-VDM [48]	2023	[5](10M)	ChatGPT	256 × 256	-	12.64		**0.3204**	-	325.42	35.57
VideoFactory [215]	2023	[5, 236]	-	256 × 256	2.04	-	-	0.3005	-	410.00	-
ModelScope [211]	2023	[5](10M)	-	256 × 256	1.70	11.09	550	0.2930	-	410.00	-
VideoGen [118]	2023	[5](10M)	Reference Image	256 × 256	-	-	-	0.3127	-	554.00	**71.61**
Animate-A-Story [68]	2023	[5](10M)	Depth Input	256 × 256	-	-	-	-	-	515.15	
VidRD [62]	2023	[5, 17, 259](5.3M*)	-	256 × 256	-	-	-	-	-	363.19	39.37
LAVIE [219]	2023	[5](10M)+25M*	-	320 × 512	3.00	-	-	0.2949	-	526.30	-
VideoDirGPT [123]	2023	[5](10M)	GPT-4	256 × 256	1.92	12.22	550	0.2860	-	-	-
Show-1 [257]	2023	[5](10M)	-	320 × 576	-	13.08	538	0.3072	-	394.46	35.42
Dynamicrafter [231]	2023	[5](10M)	Reference Image	256 × 256	-	-	**234**	-	-	429.23	-
EMU-Video [231]	2023	34M*	Reference Image	256 × 256	-	-	-	-	-	606.20	42.70
PixelDance [253]	2023	[5](10M)+50W*	Reference Image	256 × 256	1.50	-	381	0.3125	**49.36**	242.82	42.10
MicroCinema [218]	2023	[5](10M)	Reference Image	256 × 256	2.42	-	377	0.2967	-	342.86	37.46
ART·V [218]	2023	[5](5M)	Reference Image	256 × 256	-	-	291	0.2859	-	315.69	50.34
SVD [218]	2023	577M*	Reference Image	256 × 384	-	-	-	-	-	**242.02**	-

Table 1: Comparison of different video datasets. Existing text-video datasets are always limited in either scale or quality, while our HD-VG-130M includes 130M text-video pairs from open-domain in high-definition, widescreen and watermark-free formats.

Dataset	Video clips	Resolution	Domain	Text	Watermark-free
MSR-VTT [55]	10K	240p	open	caption	✓
UCF101 [42]	13K	240p	human action	class label	✓
HowTo100M [28]	136M	240p	instructional	subtitle	✓
HD-VILA-100M [57]	103M	720p	open	subtitle	✓
WebVid-10M [2]	10M	360p	open	caption	✗
HD-VG-130M (Ours)	130M	720p	open	caption	✓

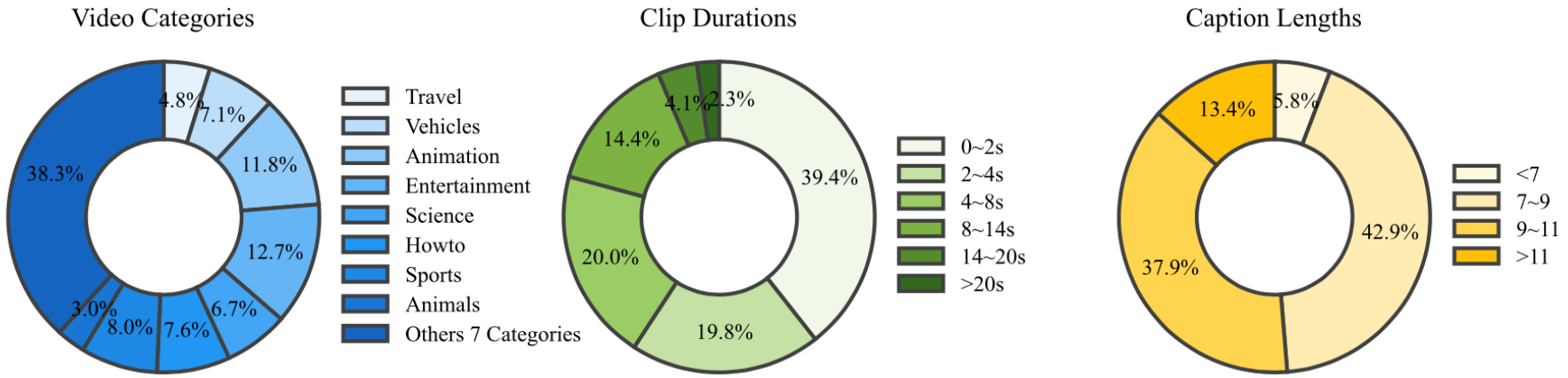


Figure 2: Statistics of video categories, clip durations, and caption word lengths in HD-VG-130M. HD-VG-130M covers a wide range of video categories.

VidRD

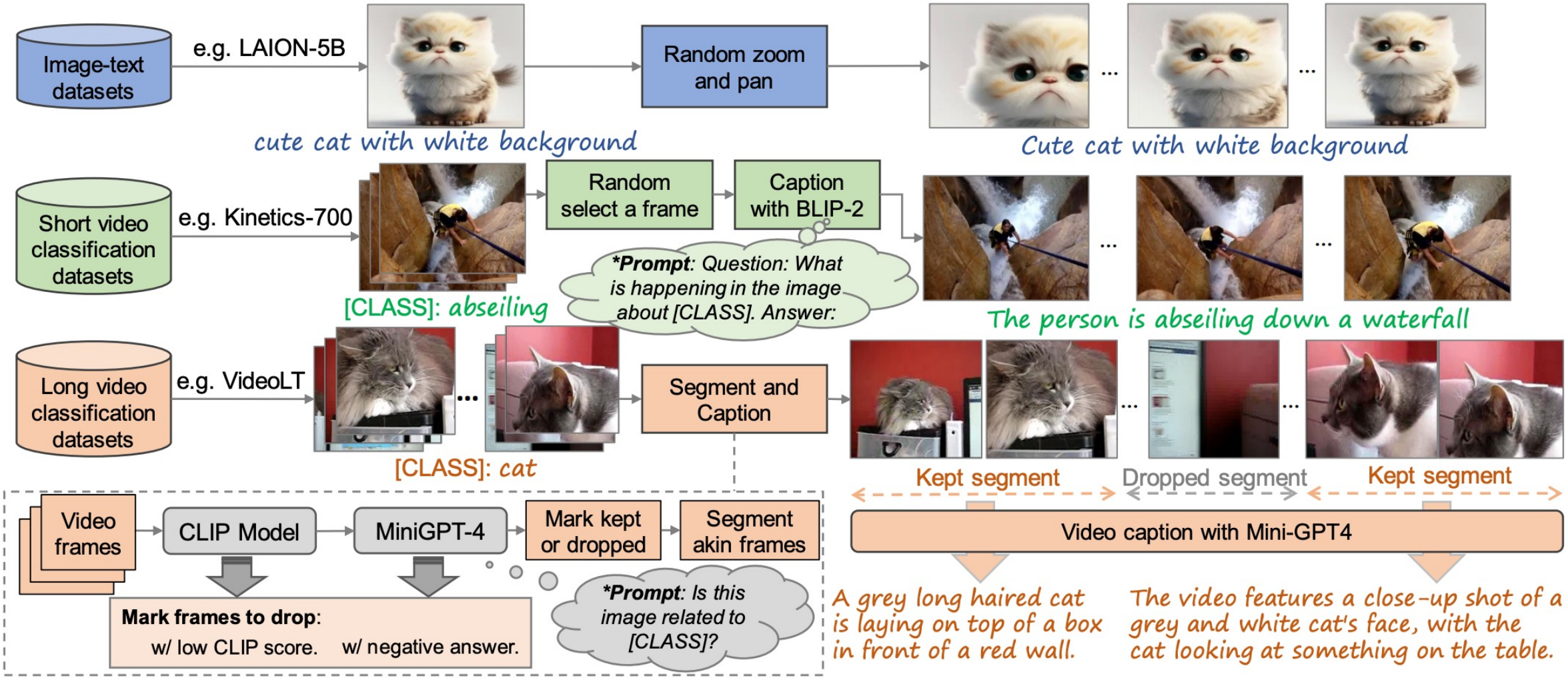


Image-text datasets
e.g. LAION-5B
Random zoom and pan
cute cat with white background
Cute cat with white background
Short video classification datasets
e.g. Kinetics-700
Random select a frame
Caption with BLIP-2
*Prompt: Question: What is happening in the image about [CLASS]. Answer:
[CLASS]: abseiling
The person is abseiling down a waterfall
Long video classification datasets
e.g. VideoLT
Segment and Caption
[CLASS]: cat
Kept segment
Dropped segment
Kept segment
Video frames
CLIP Model
MiniGPT-4
Mark kept or dropped
Segment akin frames
Video caption with Mini-GPT4
Mark frames to drop:
w/ low CLIP score.
w/ negative answer.
*Prompt: Is this image related to [CLASS]?
A grey long haired cat is laying on top of a box in front of a red wall.
The video features a close-up shot of a grey and white cat's face, with the cat looking at something on the table.

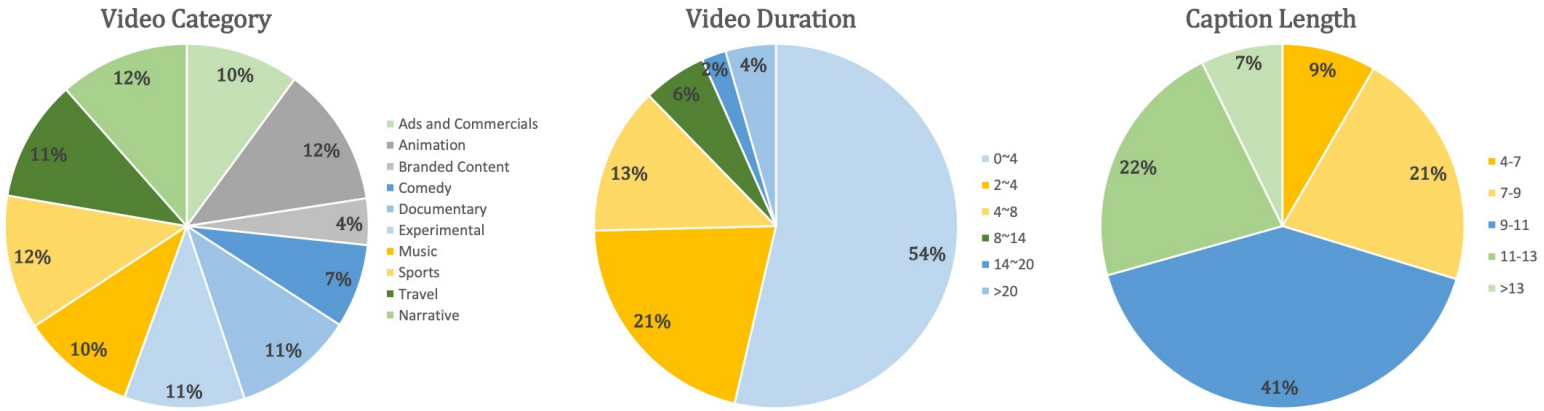
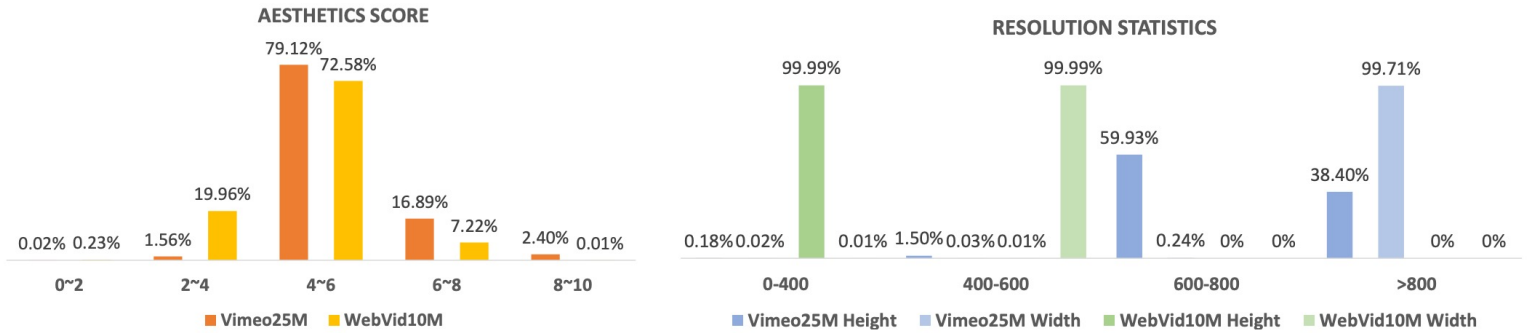


Figure 6: **Vimeo25M general information statistics.** We show statistics of video categories, clip durations, and caption word lengths in Vimeo25M.

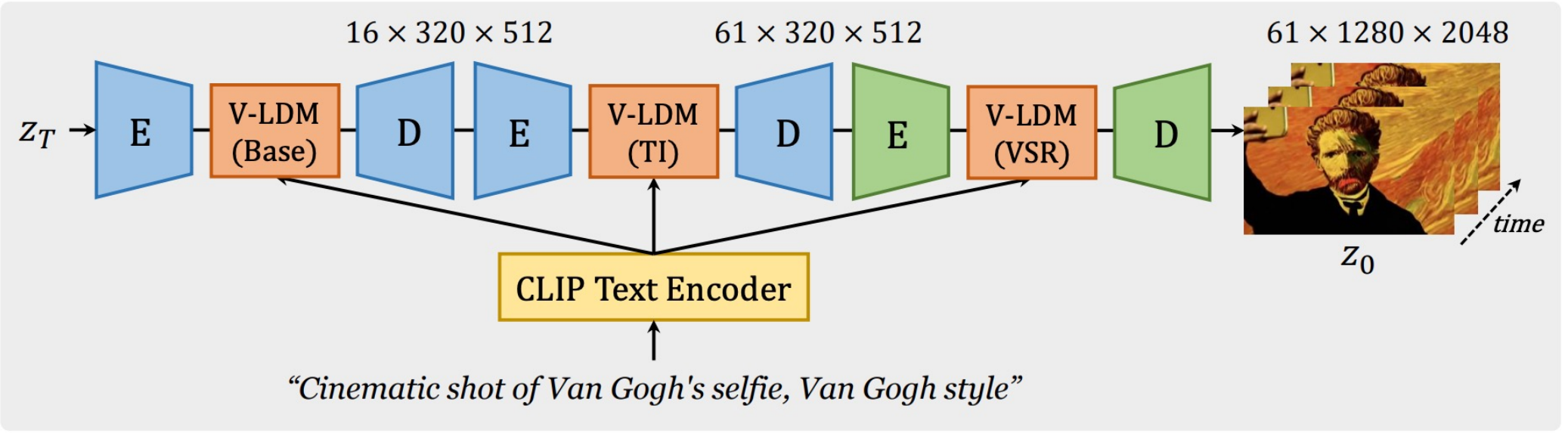


(a) Aesthetics score statistics

(b) Resolution statistics

Figure 7: **Aesthetics score, video resolution statistics.** We compare Vimeo25M with WebVid10M in terms of (a) aesthetics score and (b) video spatial resolution.

LAVIE (open-sourced)



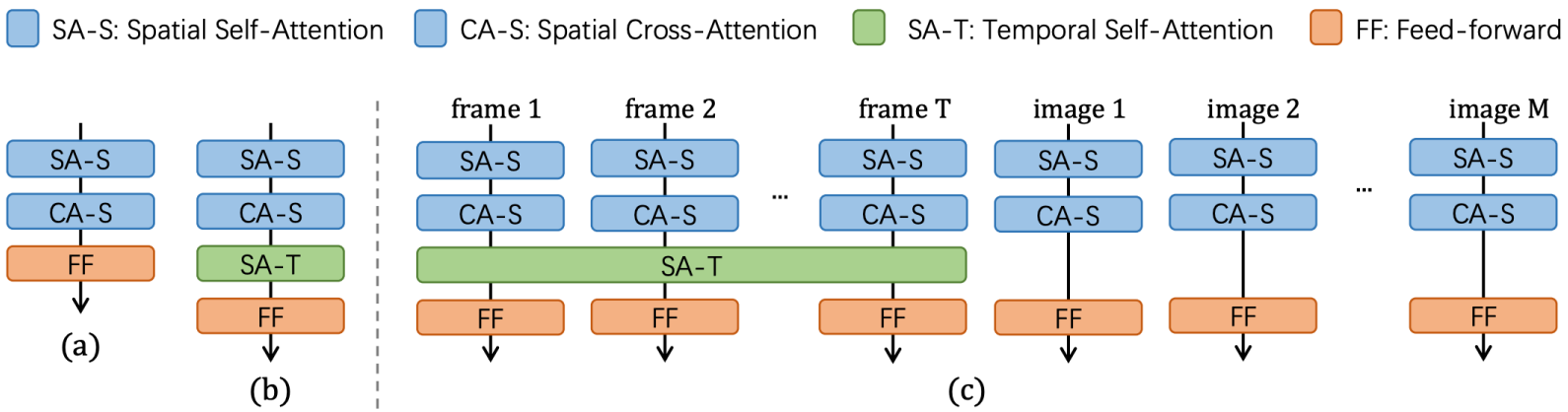
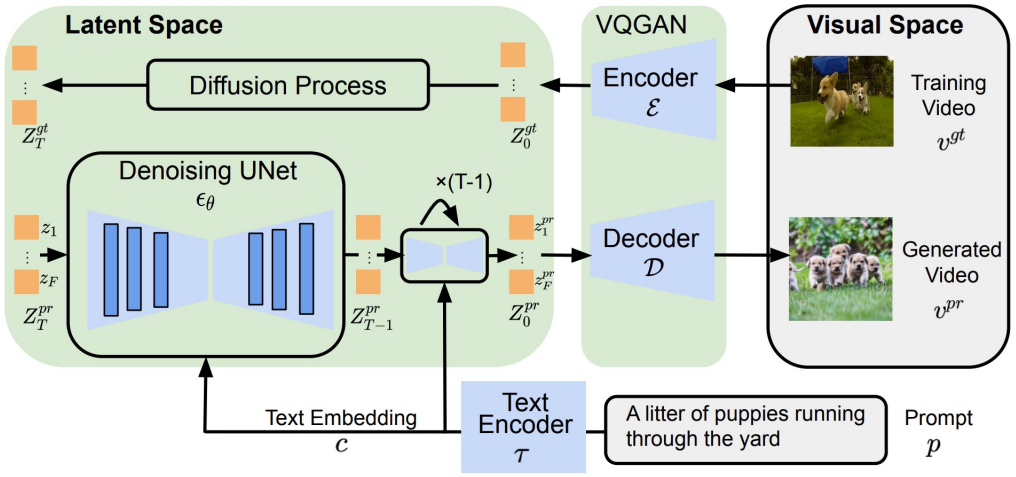


Figure 4: **Spatio-temporal module.** We show the Transformer block in Stable Diffusion in (a), our proposed ST-Transformer block in (b), and our joint image-video training scheme in (c).

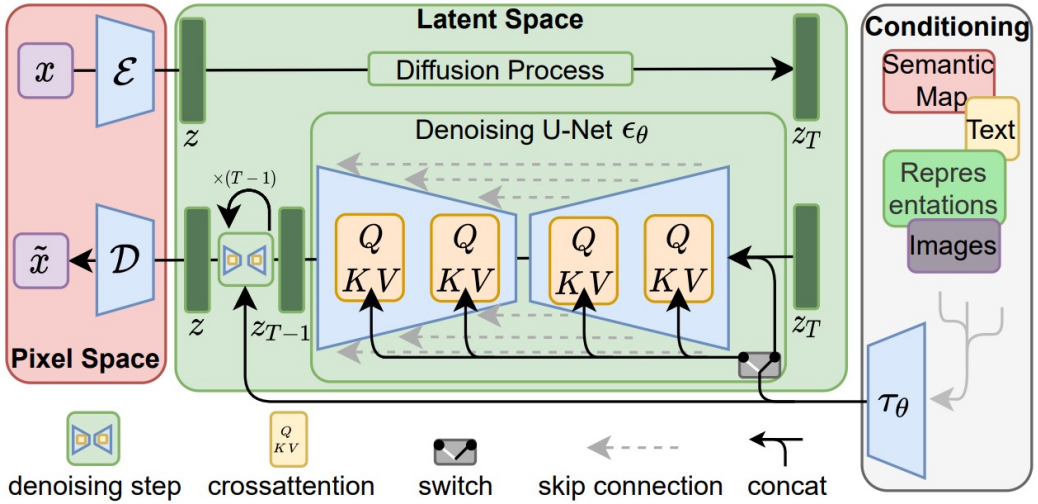
ModelScope(open-sourced)

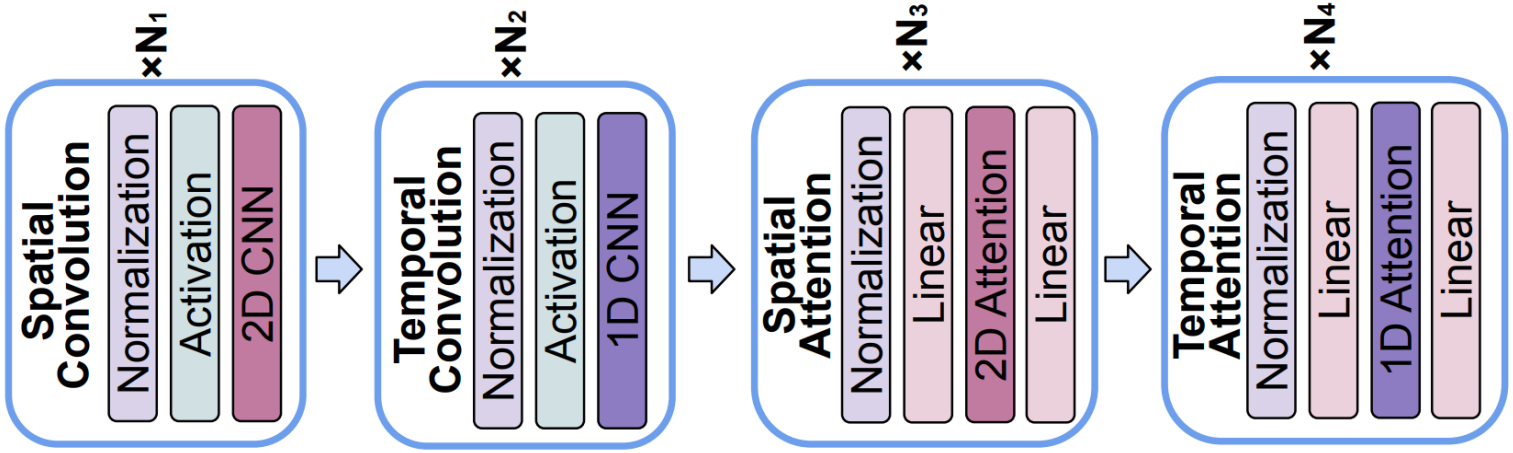
OpenMMLab

ModelScope T2V



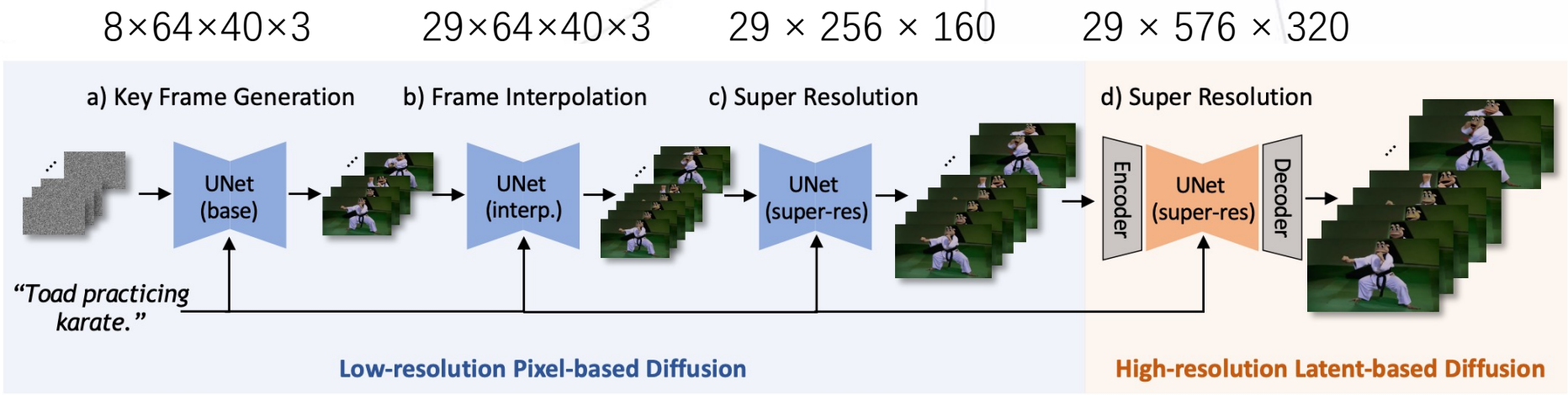
StableDiffusion T2I

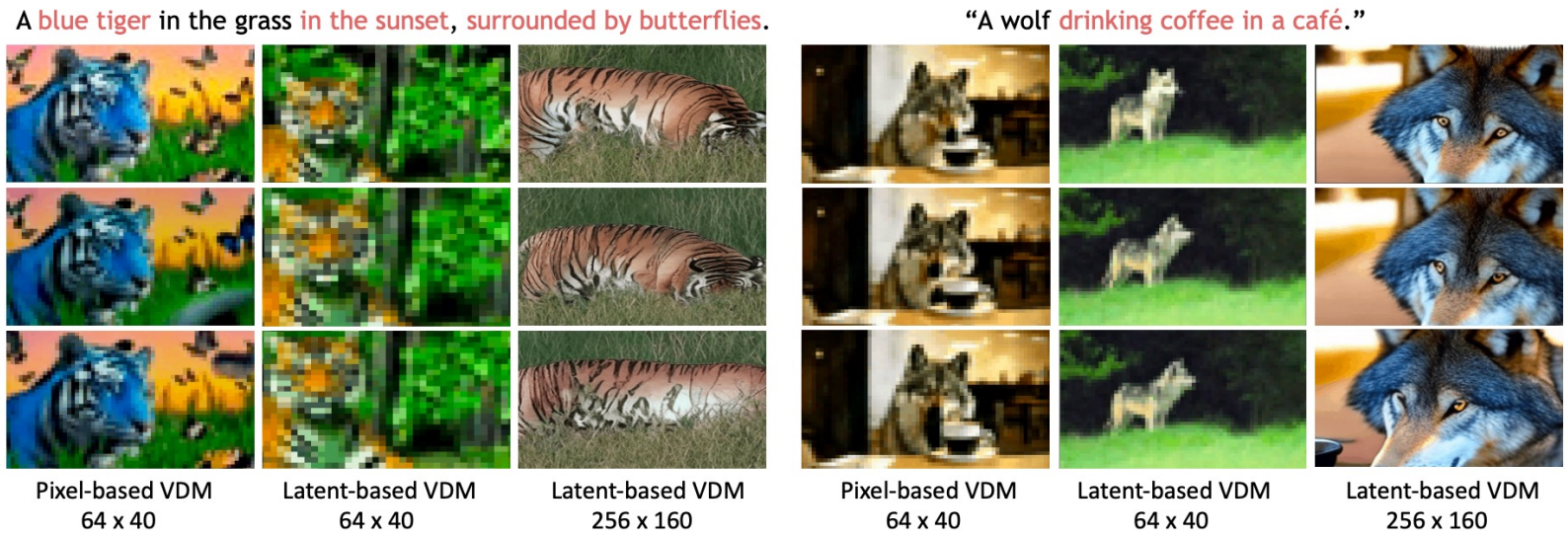




ModelScope Text-to-Video Technical Report (Arxiv 2023)

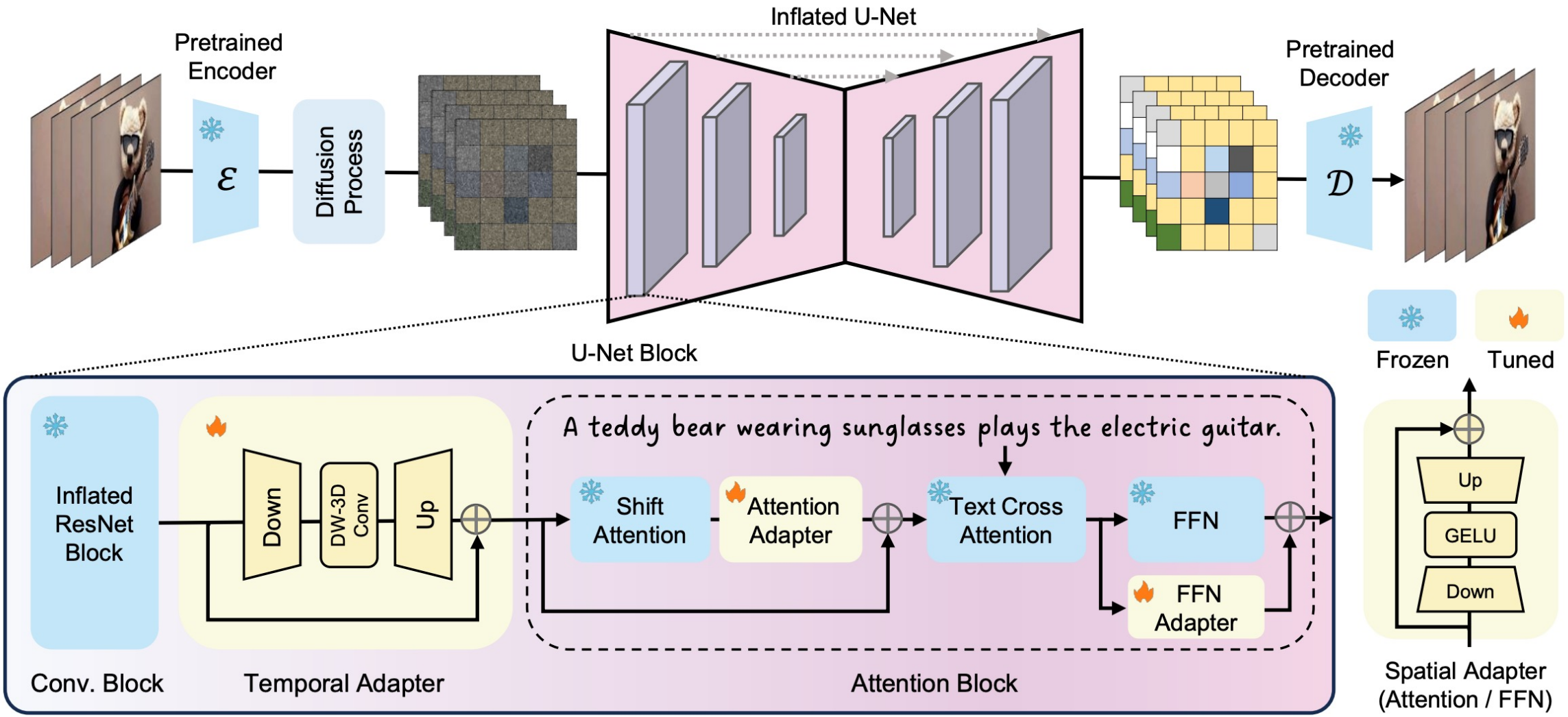
Show-1(open-sourced)





SHOW-1: MARRYING PIXEL AND LATENT DIFFUSION MODELS FOR T2V GENERATION (Arxiv 2023)

SimDA


Pretrained Encoder
Diffusion Process
Inflated U-Net
Pretrained Decoder
U-Net Block
Frozen
Tuned
Inflated ResNet Block
Down
DW-3D Conv
Up
A teddy bear wearing sunglasses plays the electric guitar.
Shift Attention
Attention Adapter
Text Cross Attention
FFN
FFN Adapter
Up
GELU
Down
Conv. Block
Temporal Adapter
Attention Block
Spatial Adapter (Attention / FFN)

SimDA

An astronaut flying in space,4k,high resolution

Xmas Christmas tree holiday celebration winter snow animation gold background

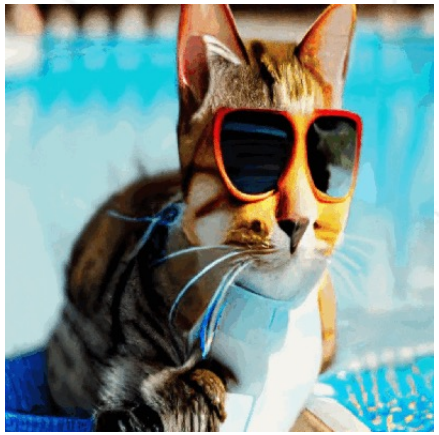
Standing on top a mountainside watching the sunset with the vivid pinks red orange showing from the fire colored sky.

Sea waves with foam on white tropical sandy beach.

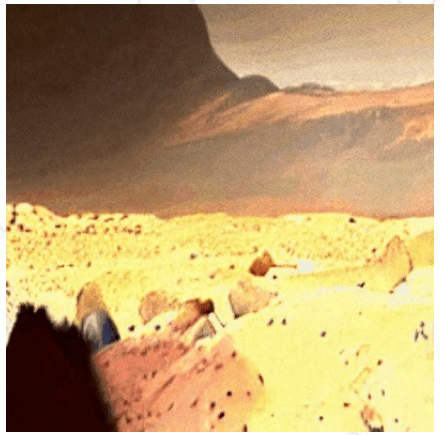
Beer pouring into glass, low angle video shot

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

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



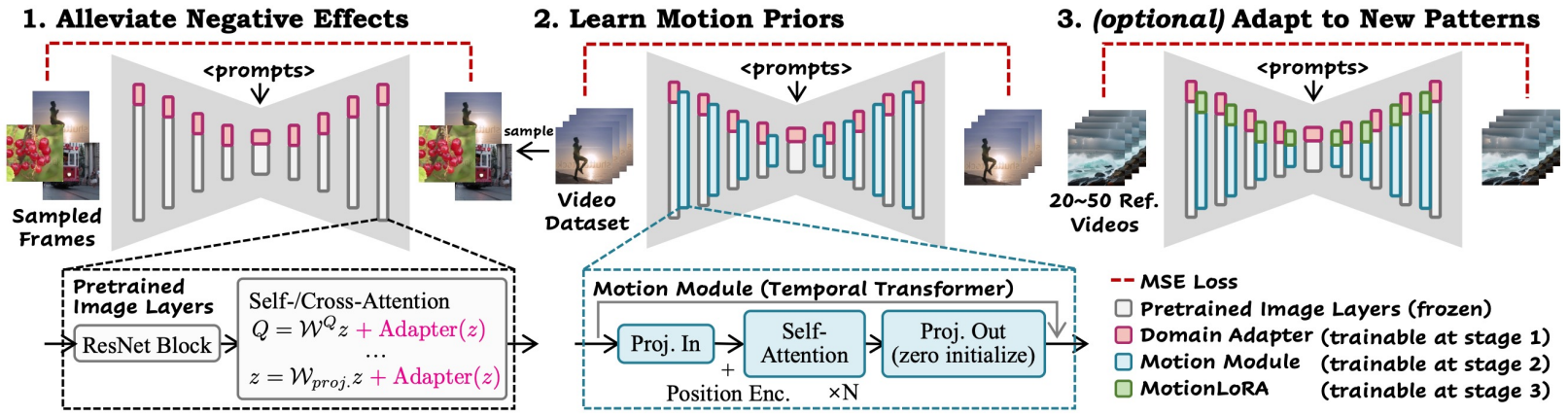
A cat wearing sunglasses and working as a lifeguard at a pool.

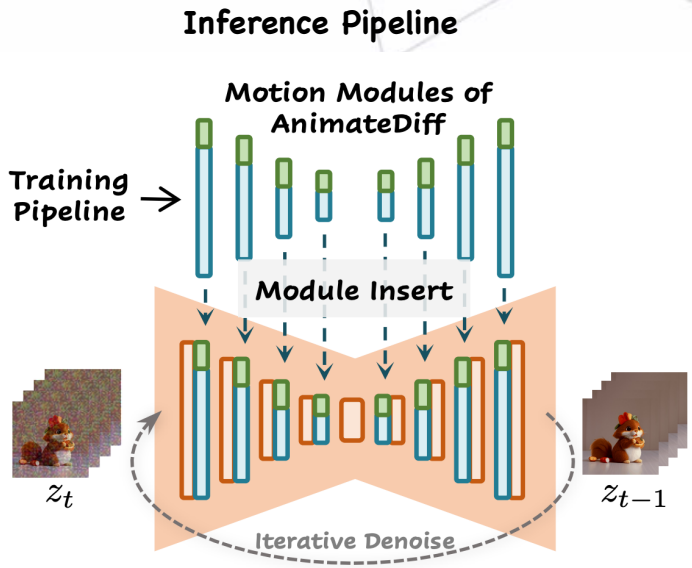


A beautiful sunrise on mars, Curiosity rover.

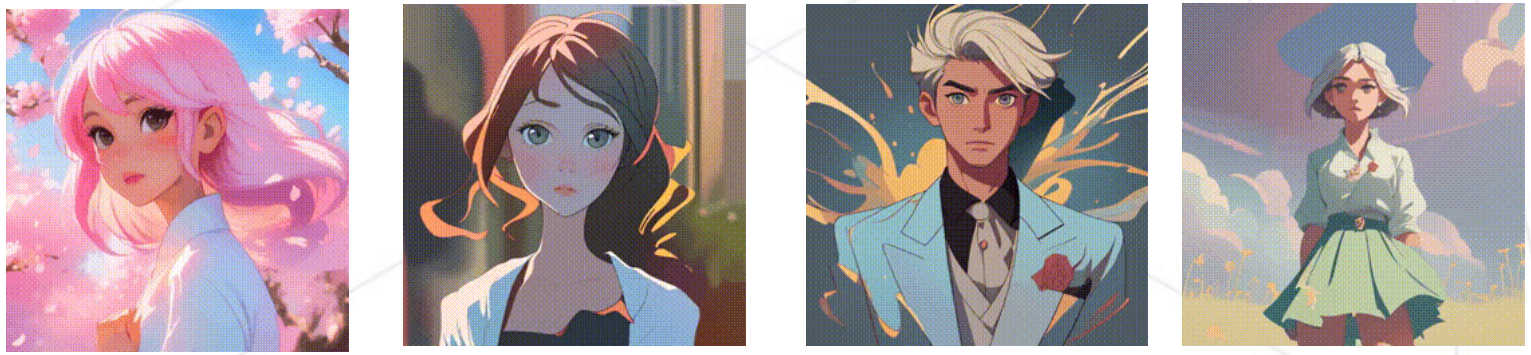
Coffee pouring into a cup.

AnimateDiff

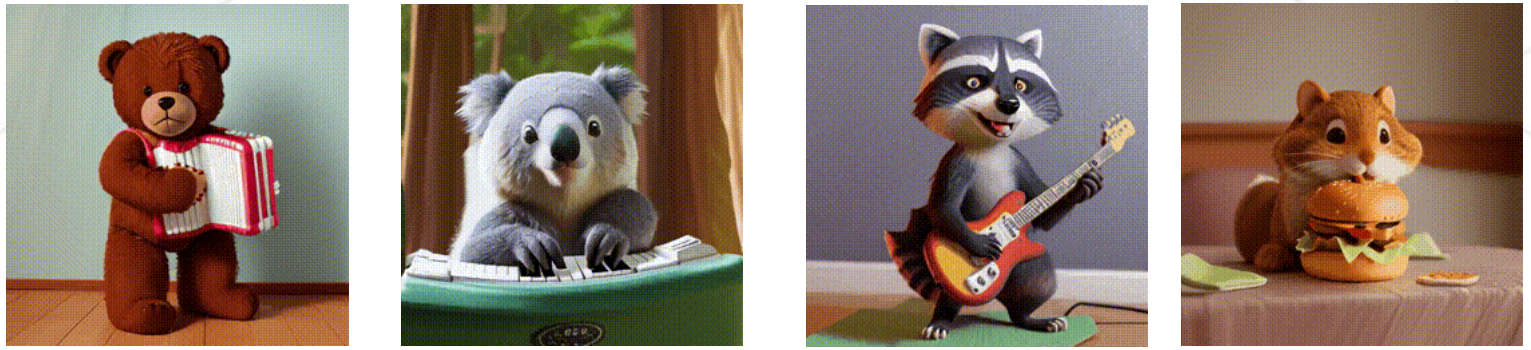




ToonYou

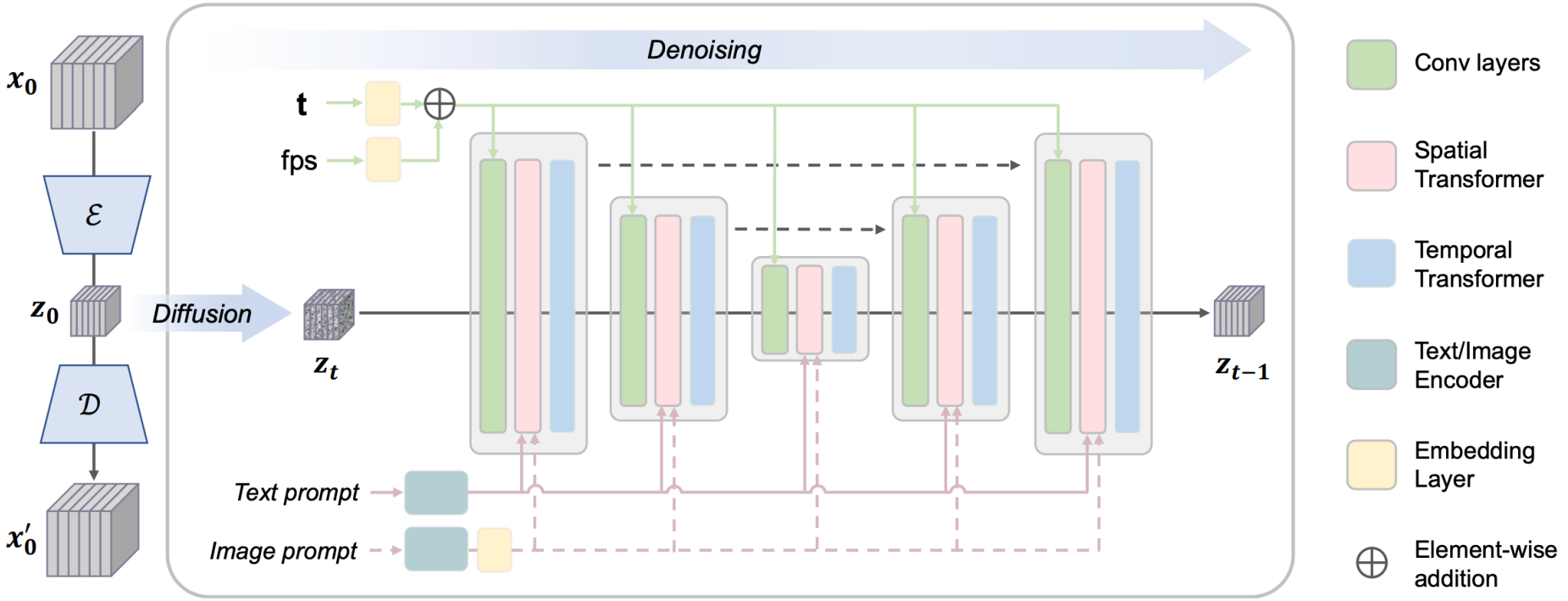


RCNZ Cartoon

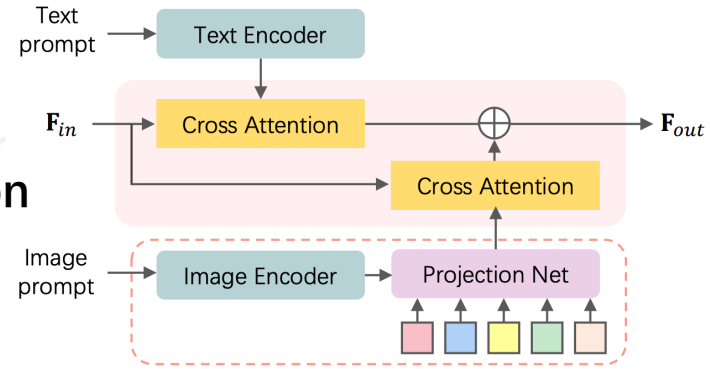


AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning (ICLR 2024)

VideoCraft-1 (Image Condition)



Condition

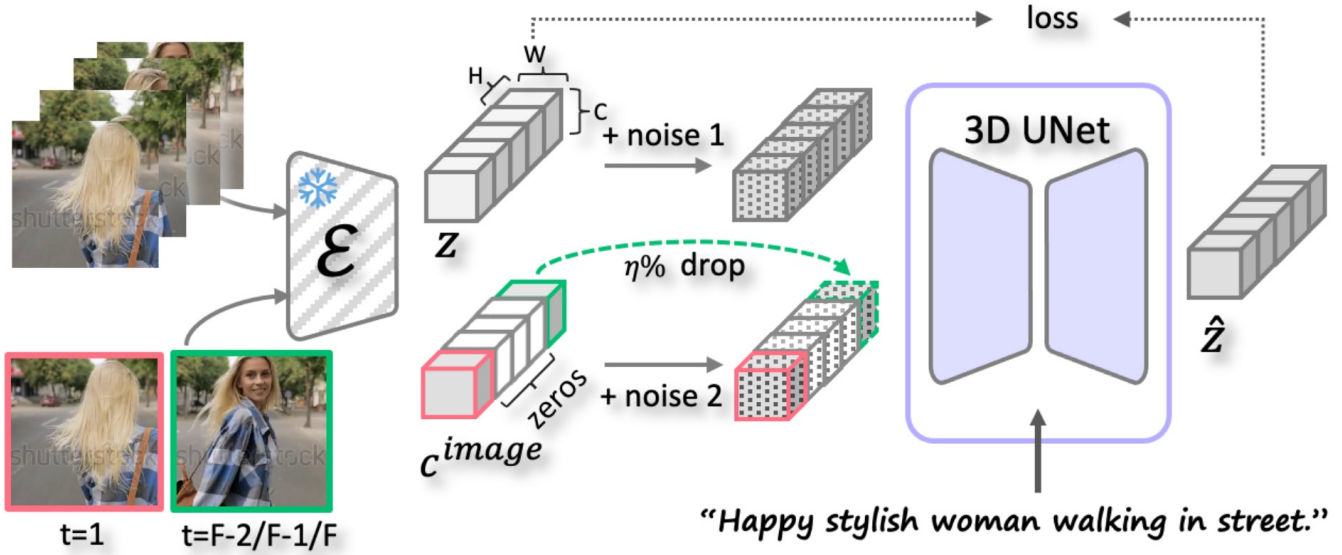


VideoCrafter1: Open Diffusion Models for High-Quality Video Generation (Arxiv 2023)

PixelDance & Emu Video (Image Condition)

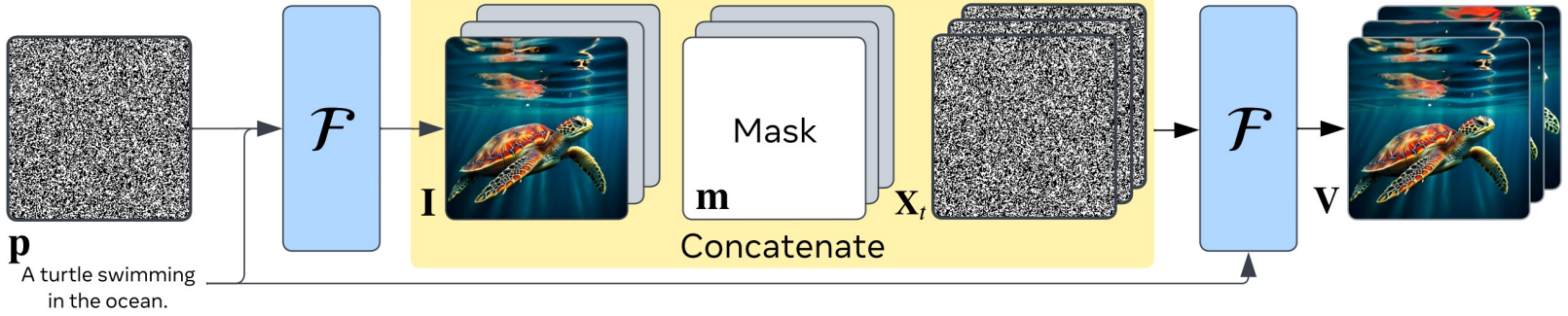
A beautiful woman with long golden hair is driving a red convertible sports car.

The skull is burining, red fires, the skull exploding



fireworks

A panda bear driving a car



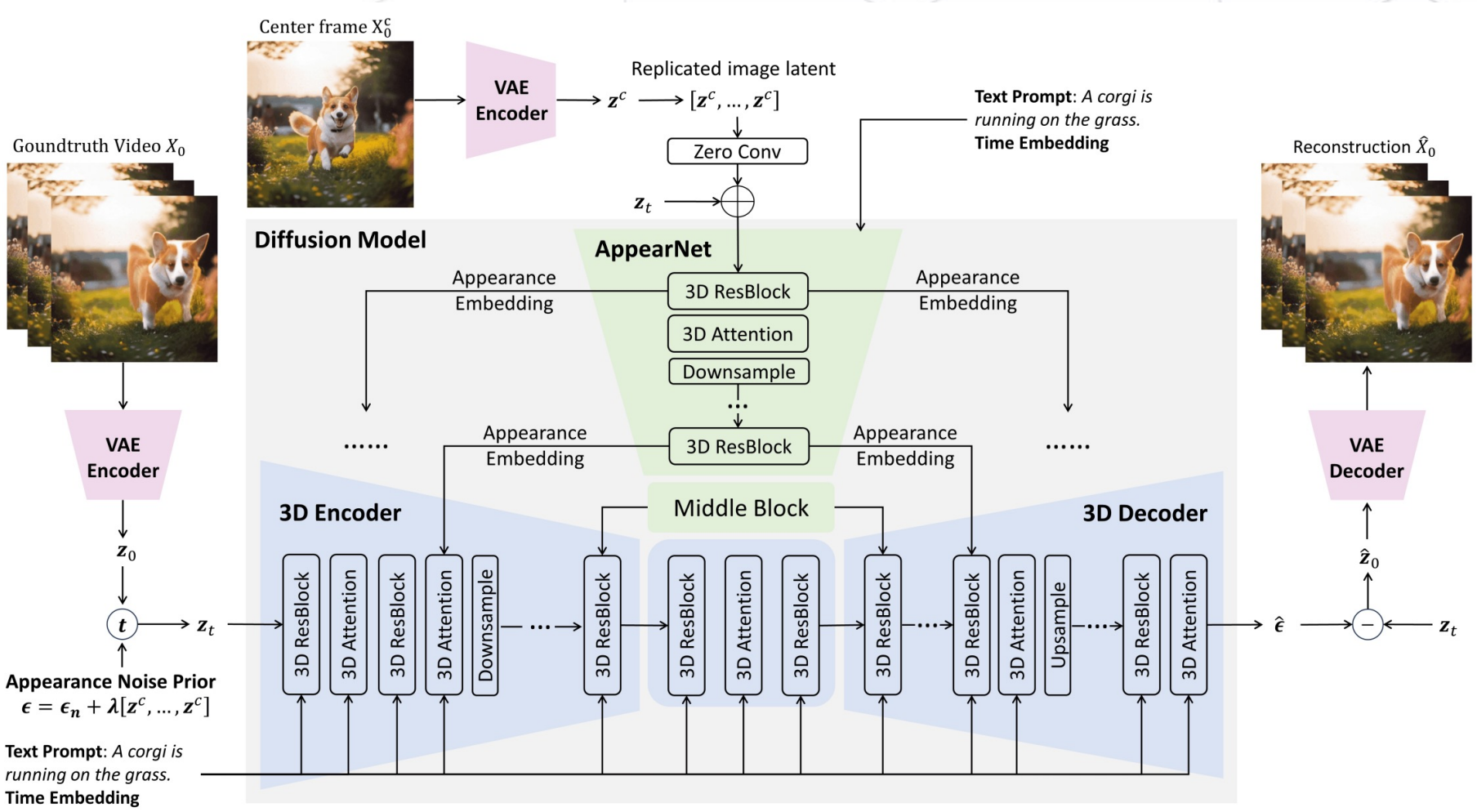
Make Pixels Dance: High-Dynamic Video Generation (CVPR 2024)

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

MicroCinema(Image Condition)

Disney animation style, One frosty day, when snow blanketed everything like a white quilt, a little girl named Zosia was coming home from school. With gloves keeping her hands warm and a cozy jacket, she walked along the path.

Teddy bear walking down 5th Avenue, front view, beautiful sunset, close up, high definition, 4K



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

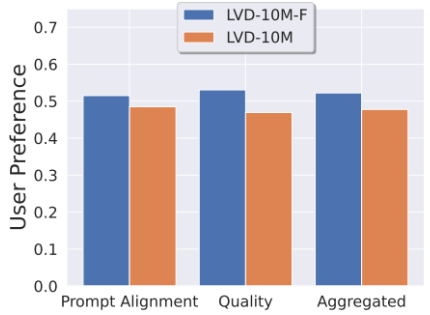
- Mid-frame caption, CoCa
- Video-based caption, V-BLIP
- LLM summarization

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.

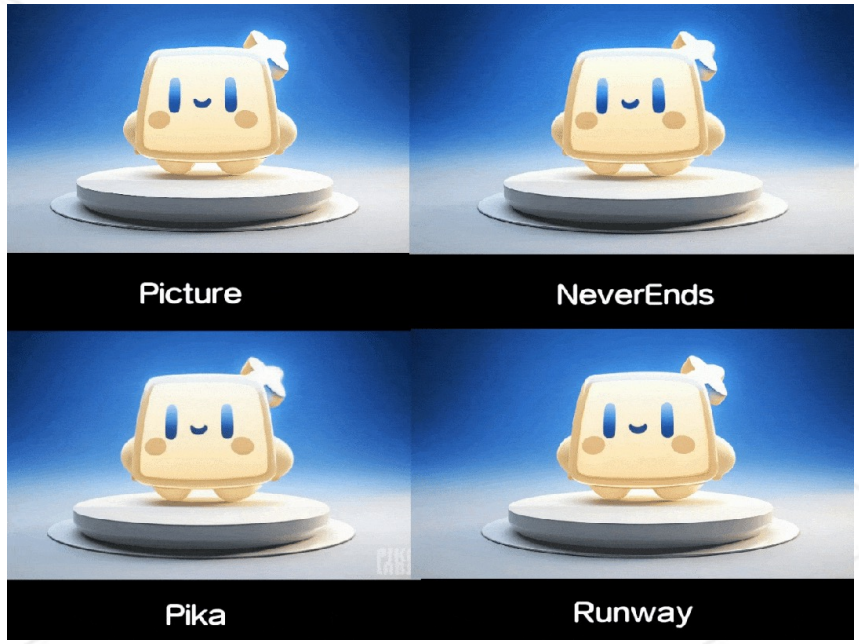
	LVD	*LVD-F*	*LVD-10M*	*LVD-10M-F*	*WebVid*	*InternVid*
#Clips	577M	152M	9.8M	2.3M	10.7M	234M
Clip Duration (s)	11.58	10.53	12.11	10.99	18.0	11.7
Total Duration (y)	212.09	50.64	3.76	0.78	5.94	86.80
Mean #Frames	325	301	335	320	-	-
Mean Clips/Video	11.09	4.76	1.2	1.1	1.0	32.96
Motion Annotations?	✓	✓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

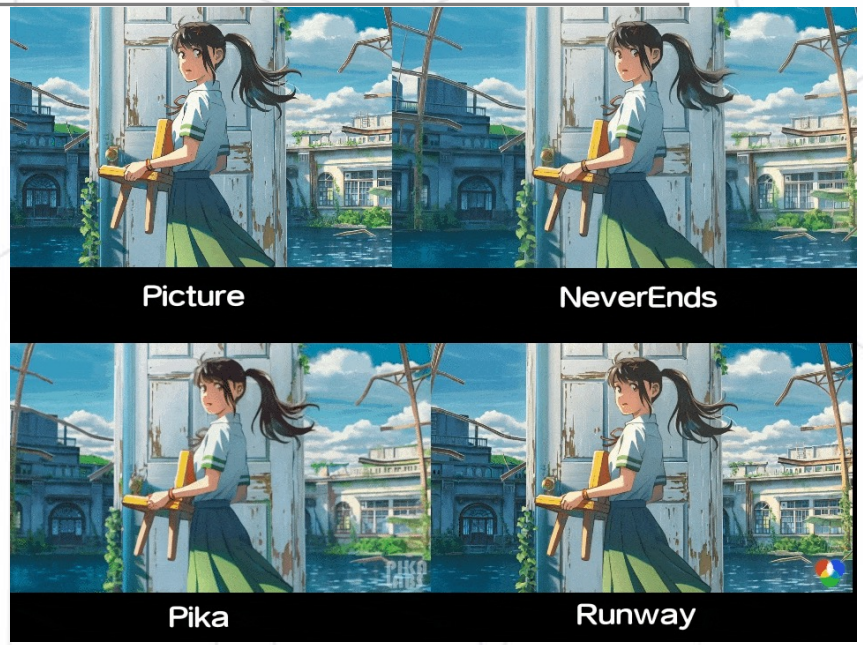
(a) Initializing spatial layers from pretrained images models greatly improves performance.



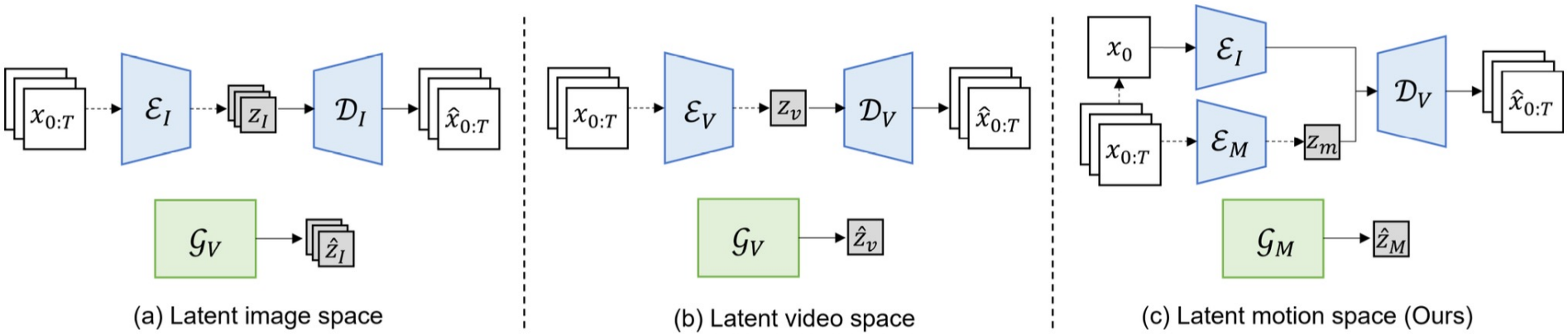
(b) Video data curation boosts performance after video pretraining.

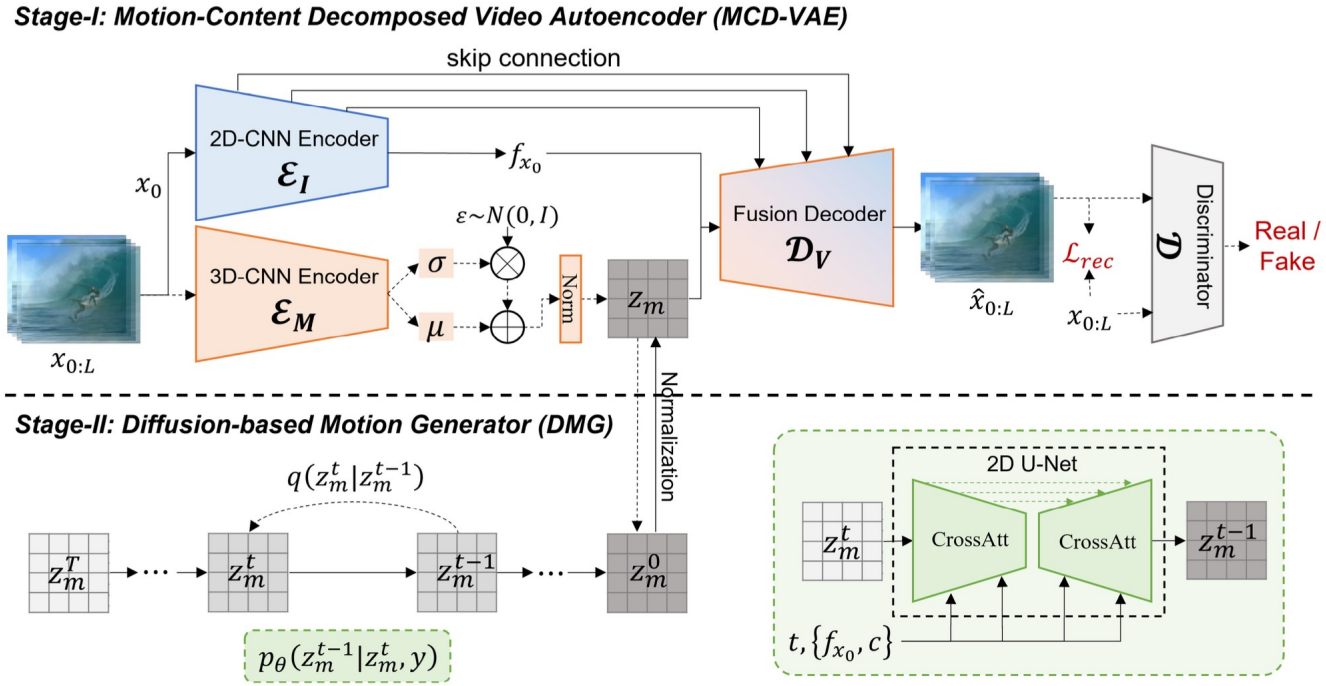
Stable Video Diffusion(Image Condition)

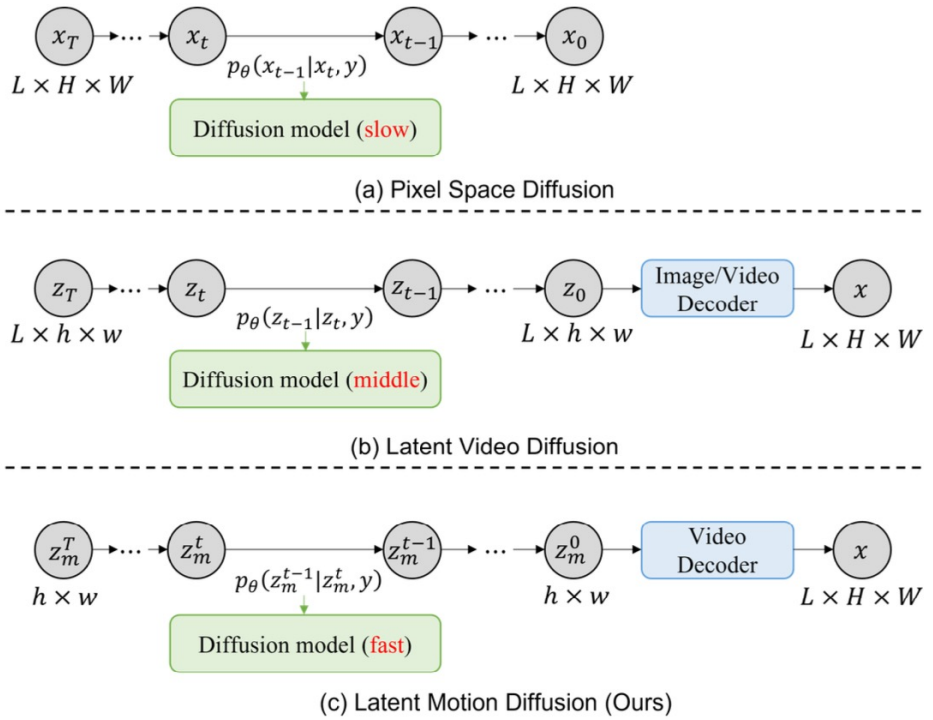




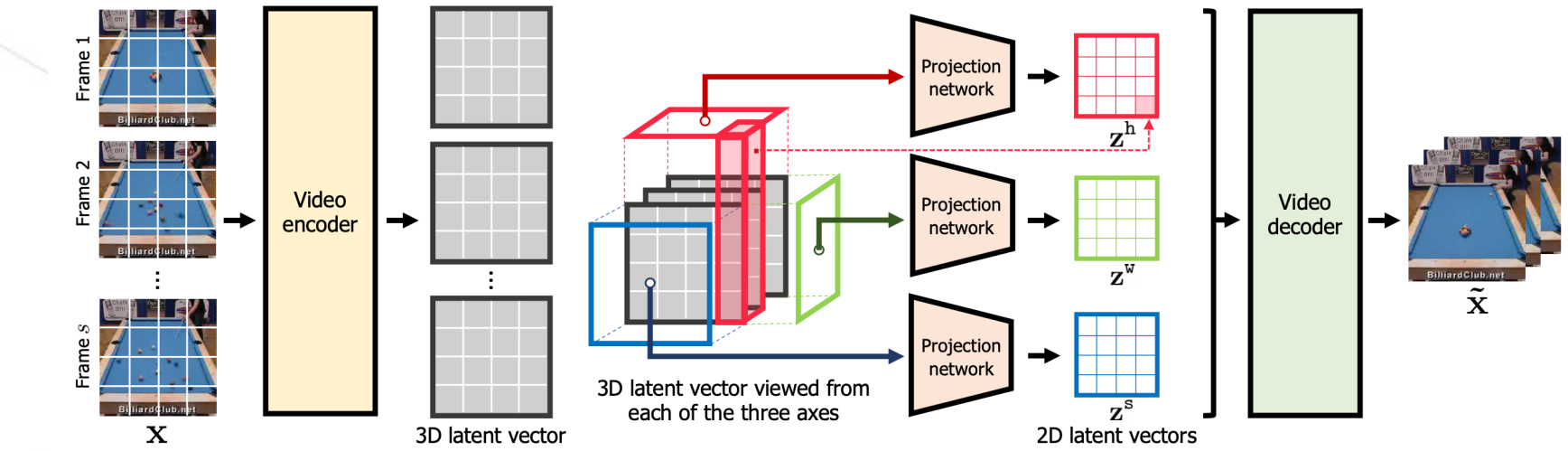
Architecture (AE)

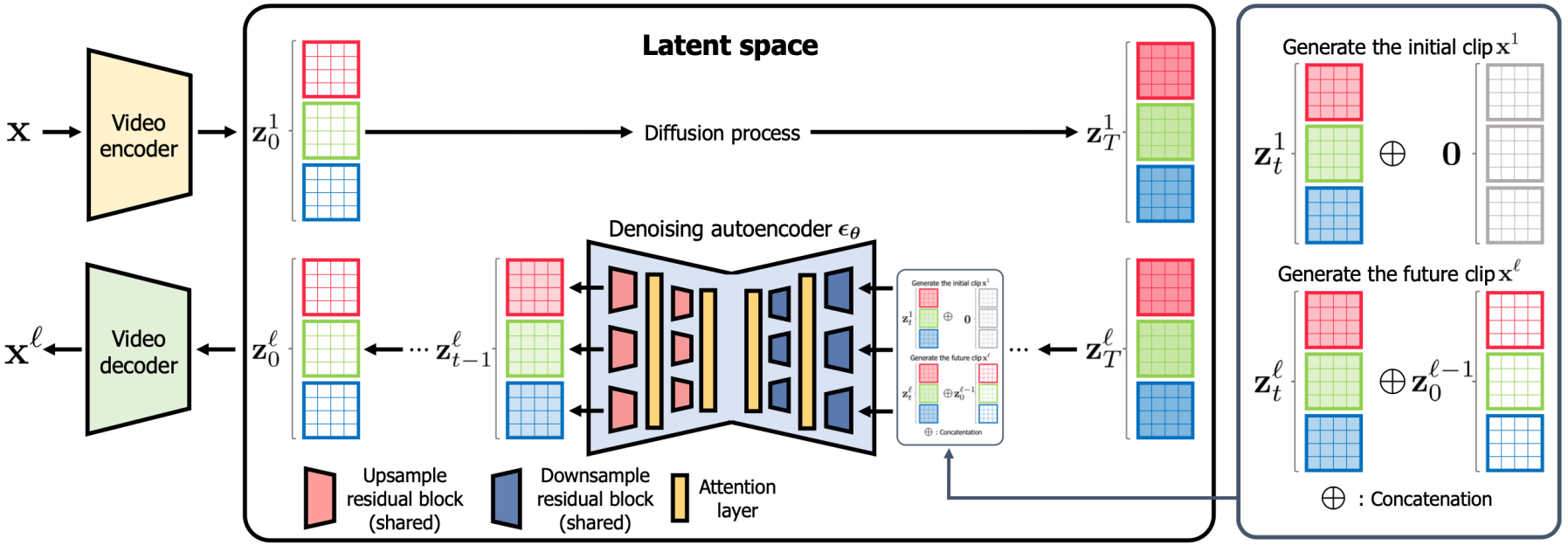






Architecture (AE)

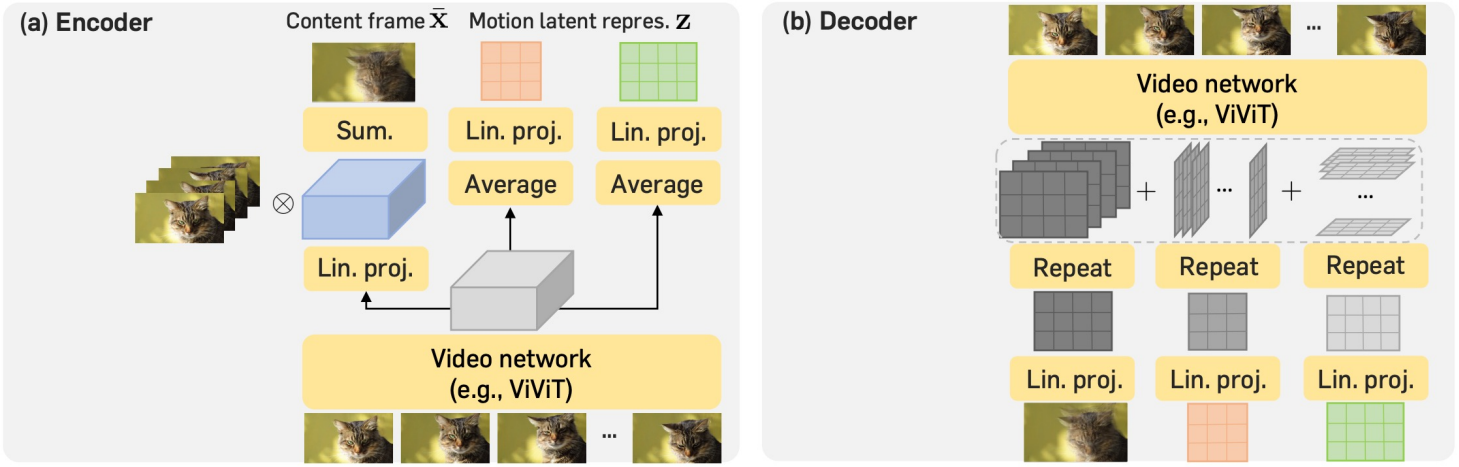


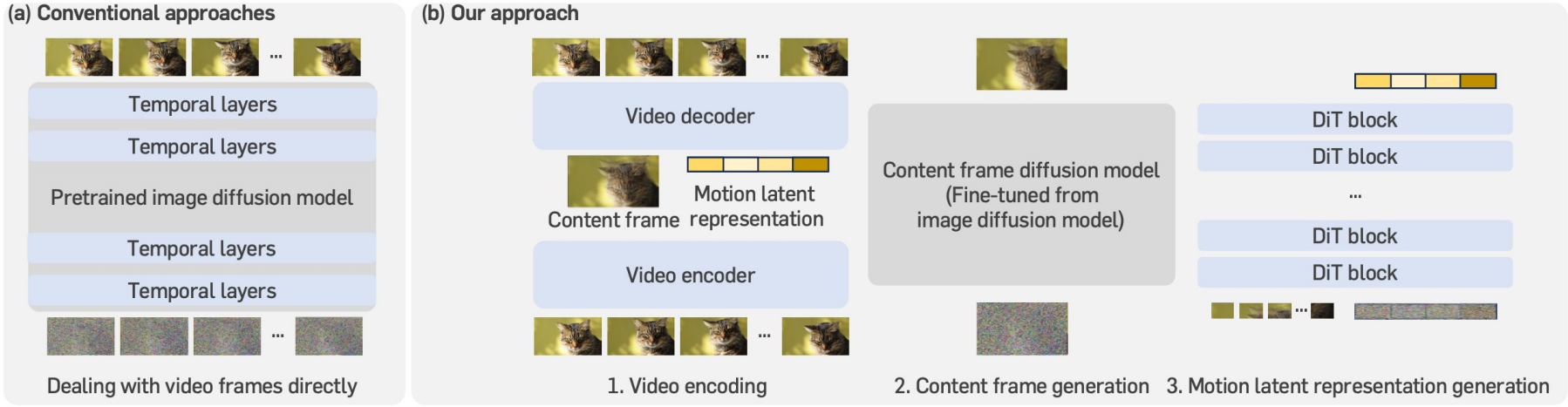


Architecture (AE)

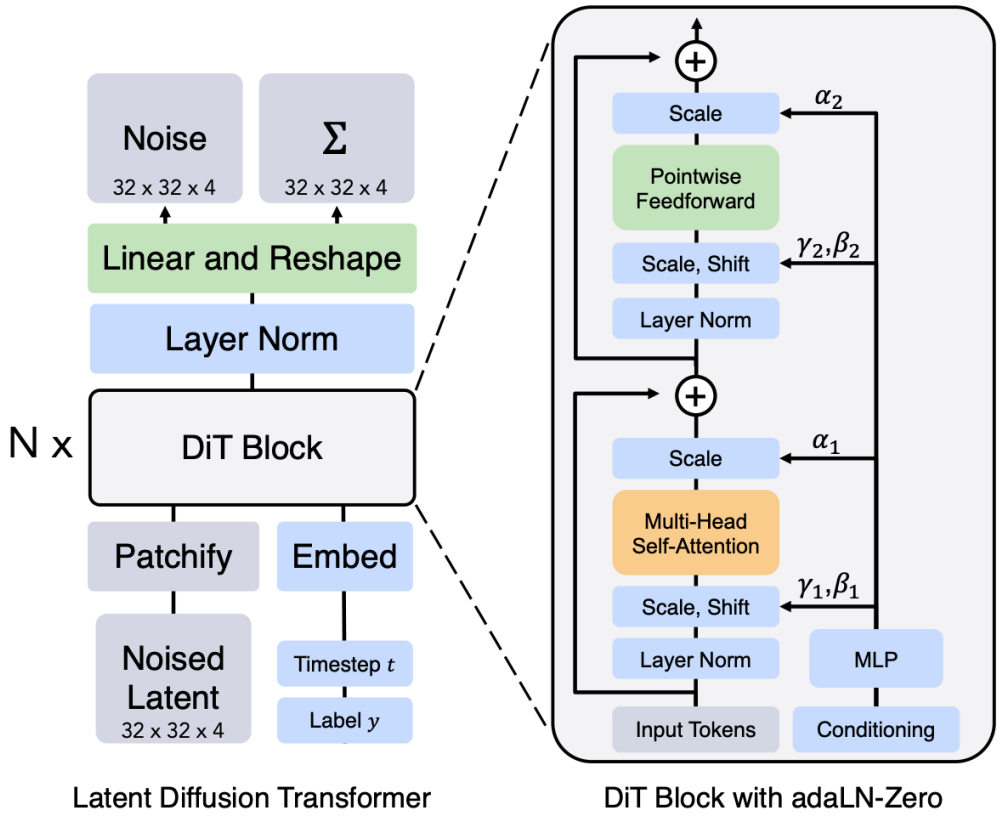
OpenMMLab

AutoEncoder Pretrain



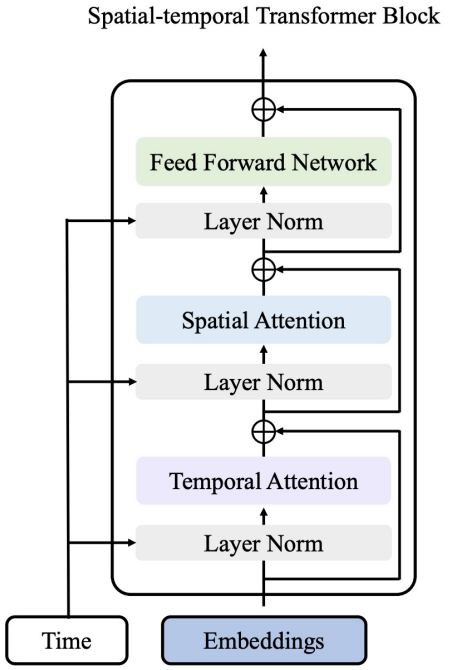


Architecture (Transformer)

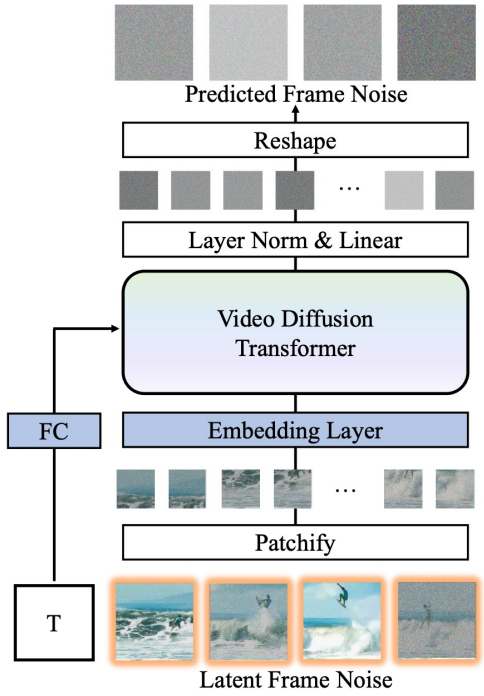


Latent Diffusion Transformer

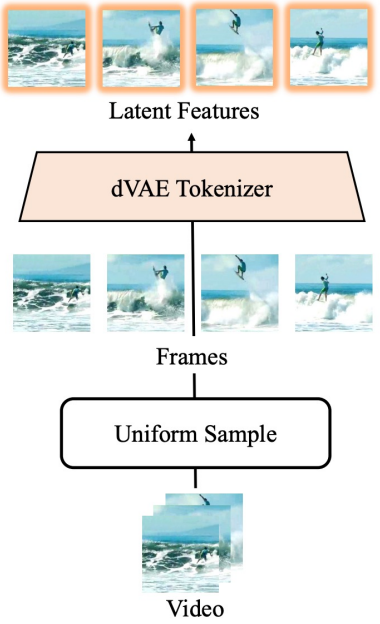
DiT Block with adaLN-Zero



(a) VDT Block



(b) Video Diffusion Transformer (VDT)



(c) Latent Feature Extraction

Model	Layers N	Hidden size d	Heads	Gflops (I=32, p=4)
DiT-S	12	384	6	1.4
DiT-B	12	768	12	5.6
DiT-L	24	1024	16	19.7
DiT-XL	28	1152	16	29.1

Table 1: Configurations of VDT. FVD results are reported on UCF101 unconditional generation.

Model	Layer	Hidden State	Heads	MLP ratio	FVD ↓
VDT-S	12	384	6	4	425.6
VDT-L	28	1152	16	4	225.7

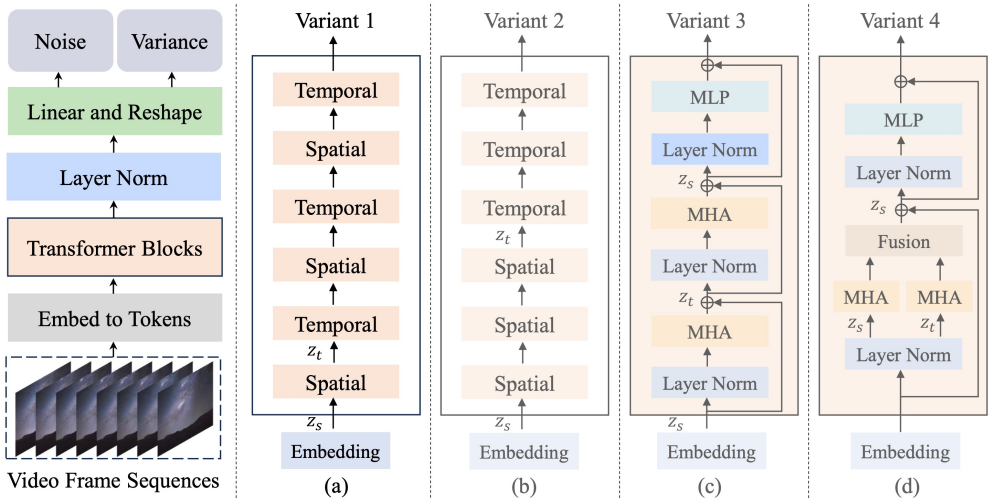
Scalable Diffusion Models with Transformers (ICCV 2023)

VDT: GENERAL-PURPOSE VIDEO DIFFUSION TRANSFORMERS VIA MASK MODELING (ICLR 2024)

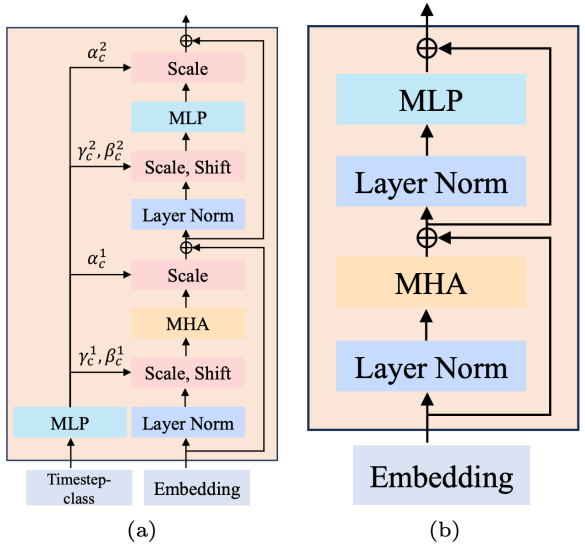
Architecture (Transformer)

OpenMMLab

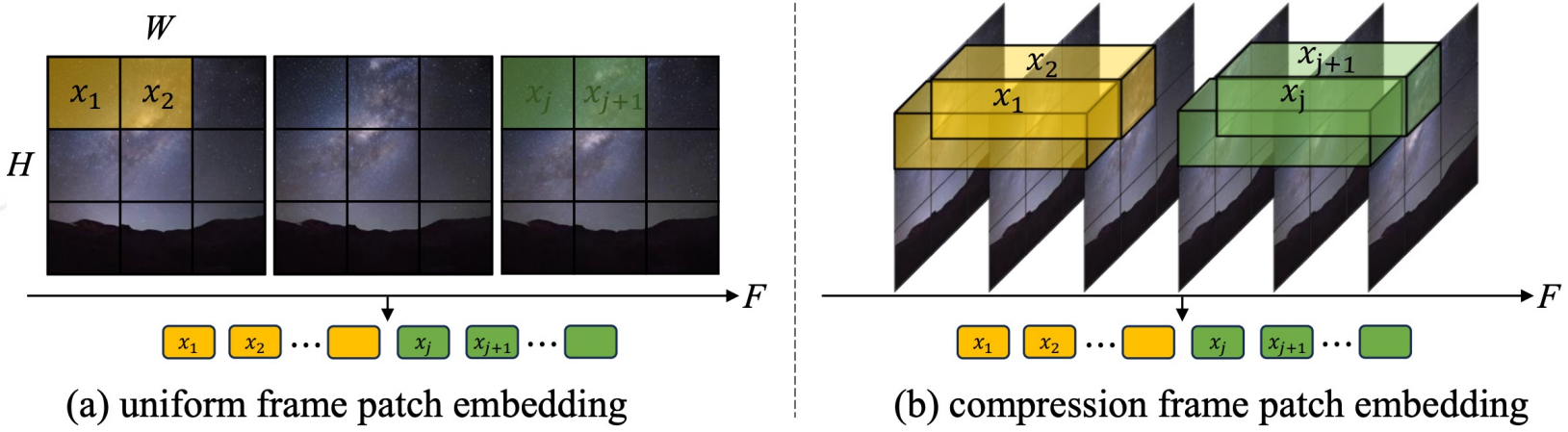
Architecture Variant



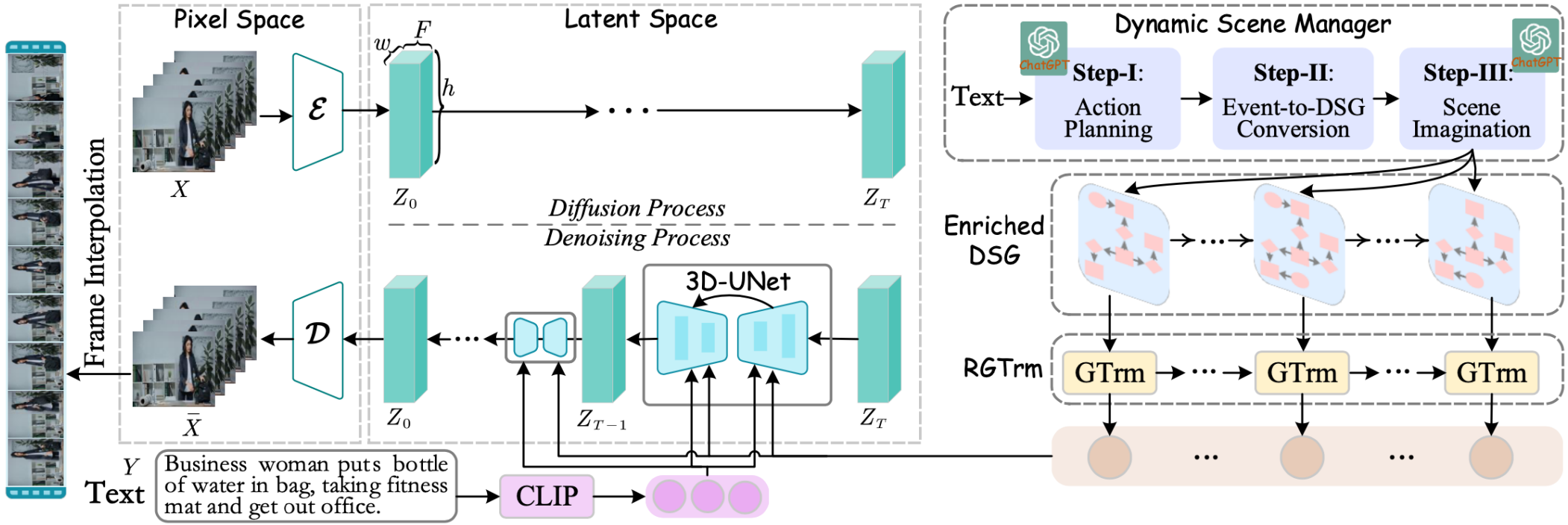
Condition Variant



Patch Embedding



Architecture (Text Condition)



Text Encoder

Stable Diffusion:	CLIP Text Encoder: ViT-H/14
DALL-E3:	T5-XXL
EMU-Video:	CLIP Text Encoder + T5-XL
PYoCo:	CLIP Text Encoder + T5-XL
Latent-Shift:	BERTEmbedder
Show-1:	CLIP Text Encoder + T5-XL
Dysen-VDM:	CLIP Text Encoder + DSG (dynamic scene graph)
Imagen Video:	T5-XXL (4.6B)

Sora （Connecting Video）

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

Sora (Video2Video Editing)

OpenMMLab

Sora

Input

Make it go underwater

Rewrite the video in a pixel art style

VIDiff

Input

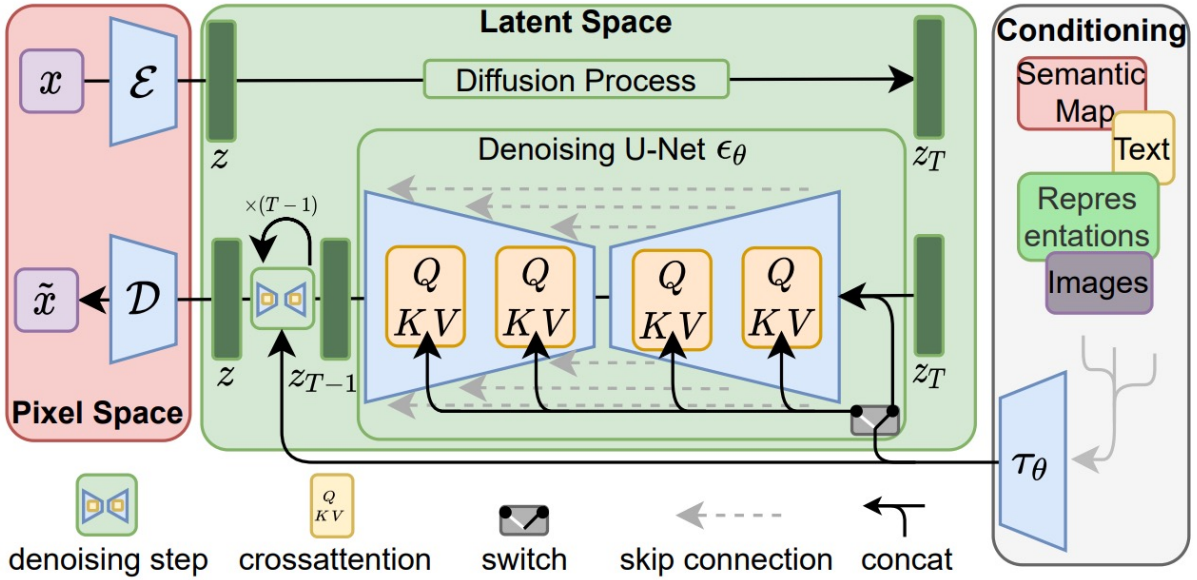
Turn the video to sketch style

Transform the video to animation style

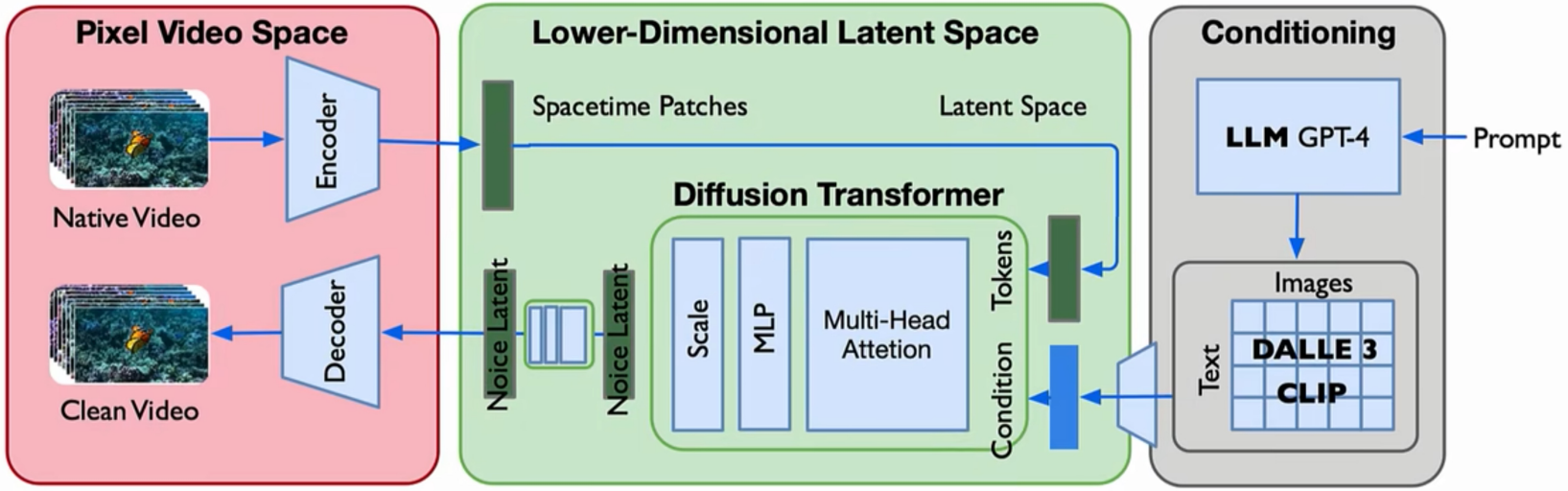
1. https://openai.com/research/video-generation-models-as-world-simulators
2. VIDiff: Translating Videos via Multi-Modal Instructions with Diffusion Models (Arxiv)

Sora (Architecture)

Stable Diffusion



Sora

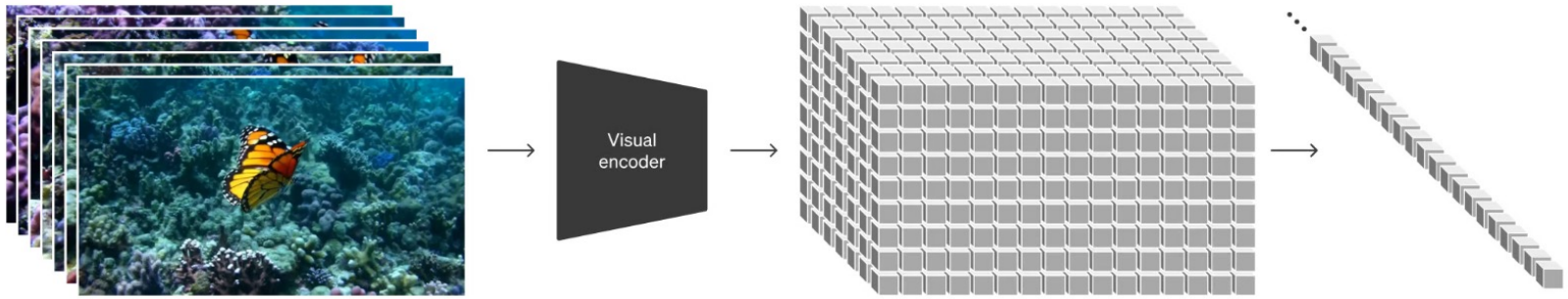


1. https://openai.com/research/video-generation-models-as-world-simulators
2.https://www.bilibili.com/video/BV1Bx4y1k7BQ/?spm_id_from=333.337.search-card.all.click&vd_source=b94ff746cdc4d7a1b331e9f42ce92950

Sora (Architecture)

OpenMMLab

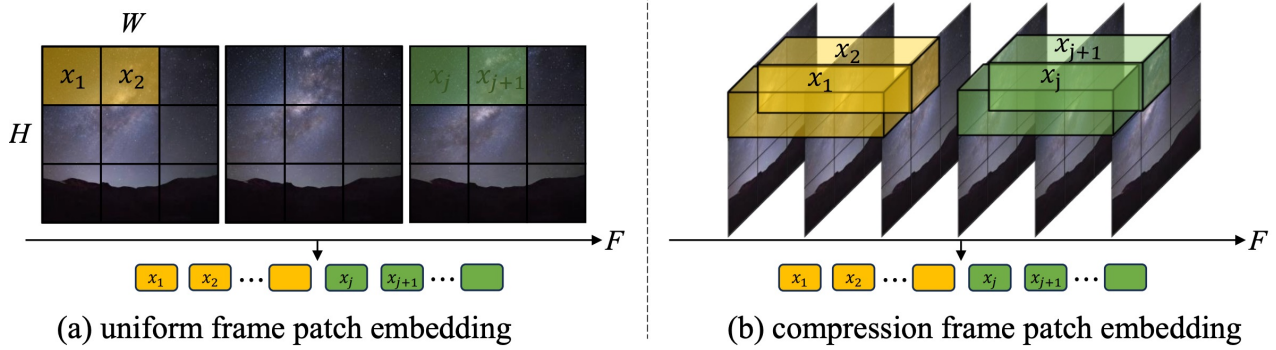
Video Compression



1min 60x30x1024x1024 —**Compress**→ 180 x 32 x 32 —**Tokenizer** (2 x 2 x 2)→ 90 x 16 x 16

Seconds x fps x H x W — Lengths x H_{low} x W_{low} — Token length

Tokenizer



(a) uniform frame patch embedding (b) compression frame patch embedding

1. https://openai.com/research/video-generation-models-as-world-simulators
2. Patch n'Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution (Arxiv)

Sora (Inference)

OpenMMLab

SimDA

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

VideoLCM

Table 1. **Inference latency** on text-to-video generation task. All experiments are performed on an NVIDIA A100 GPU. The inference overhead of generating eight videos at a time is reported.

Method	Step	Resolution	Latency
Baseline	DDIM 50-step	16 × 256 × 256	60s
VideoLCM	LCM 4-step	16 × 256 × 256	**10s**
Baseline	DDIM 50-step	16 × 448 × 256	104s
VideoLCM	LCM 4-step	16 × 448 × 256	**16s**

AdaDiff

MSR-VTT	Speed		Image Quality				
	Step↓	Time↓	IQS↑	CLIP↑	IS↑	FID↓	NIQE↓
ModelScope	50	21.2	-0.518	0.293	**18.79**	44.85	6.57
Random	29.98	13.5	-0.723	0.293	18.22	47.41	6.75
AdaDiff	31.14	13.6	**-0.517**	**0.299**	18.74	**44.49**	**6.36**

Table 2. Comparison of AdaDiff on video generation.

1. SimDA: Simple Diffusion Adapter for Efficient Video Generation
2. AdaDiff: Adaptive Step Selection for Fast Diffusion
3. VideoLCM: Video Latent Consistency Model

Sora (Dataset)

OpenMMLab

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

Sora (Scaling Law)

In this work, we find that diffusion transformers scale effectively as video models as well. Below, we show a comparison of video samples with fixed seeds and inputs as training progresses. Sample quality improves markedly as training compute increases.

Base compute

4x compute

32x compute

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

A Survey on Video Diffusion Models

Q & A

OpenMMLab 公众号
回复‘社区开放麦’领取PPT

视频生成交流群

原文&主页网址

欢迎体验，觉得好用就点亮小星星吧~