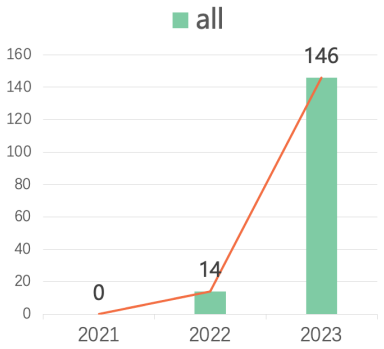
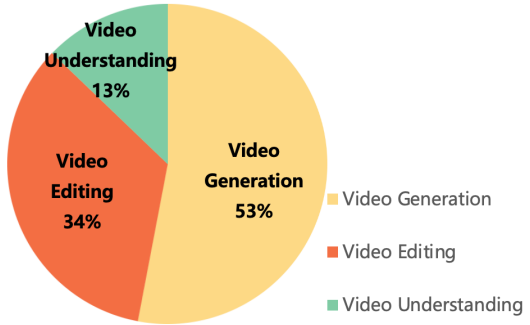
A Survey on Video Diffusion Models

Zhen Xing[1], Qijun Feng[1], Haoran Chen[1], Qi Dai[2], Han Hu[2],
Hang Xu[3], Zuxuan Wu[1], Yu–Gang Jiang[1]

Fudan University[1], Microsoft Research Asia [2] , Huawei Noah's Ark Lab [3]



(a) Number of research work



(b) Ratio of Different Directions

Presentation: Zhen Xing

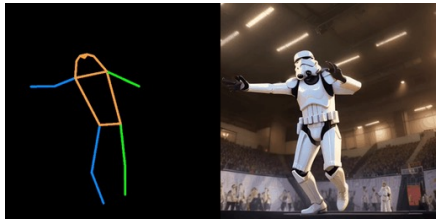
2023/12/06

An angry Godzilla is roaring and raising hands, big fire is burning

fireworks being displayed for a crowd of people.

A cute golden hamster throwing punches wearing pair of boxing gloves in a boxing ring

1. General Text-to-video Generation



(a) Pose Guided

(b) Depth Guided

(c) Image Guided

2. Video Generation with other conditions

3. Video Editing

(a) Text guided Video Editing

A man is skiing

Spider Man is skiing on the beach, cartoon style

(b) Instruction guided Video Editing

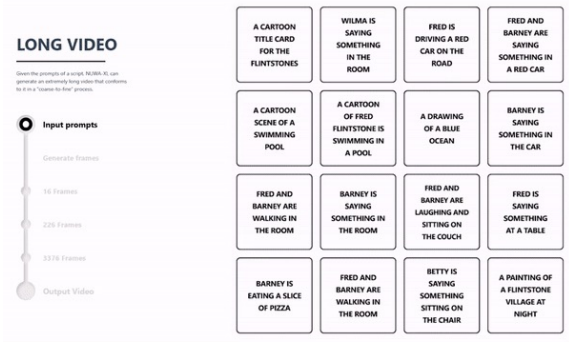
Instruction: Turn the video to Oil Painting Style

4. Future Work

(a) Open-sourced Video Dataset

Dataset	Year	Text	Domain	#Clips	Resolution
MSR-VTT [84]	2016	Manual	Open	10K	240P
DideMo [85]	2017	Manual	Flickr	27K	-
LSMDC [86]	2017	Manual	Movie	118K	1080P
ActivityNet [87]	2017	Manual	Action	100K	-
YouCook2 [88]	2018	Manual	Cooking	14K	-
How2 [89]	2018	Manual	Instruct	80K	-
VATEX [90]	2019	Manual	Action	41K	240P
HowTo100M [91]	2019	ASR	Instruct	136M	240P
WTS70M [92]	2020	Metadata	Action	70M	-
YT-Temporal [93]	2021	ASR	Open	180M	-
WebVid10M [94]	2021	Alt-text	Open	10.7M	360P
Echo-Dynamic [95]	2021	Manual	Echocardiogram	10K	-
Tiktok [96]	2021	Mannual	Action	0.3K	-
HD-VILA [97]	2022	ASR	Open	103M	720P
VideoCC3M [98]	2022	Transfer	Open	10.3M	-
HD-VG-130M [30]	2023	Generated	Open	130M	720P
InternVid [99]	2023	Generated	Open	234M	720P
CelebV-Text [100]	2023	Generated	Face	70K	480P

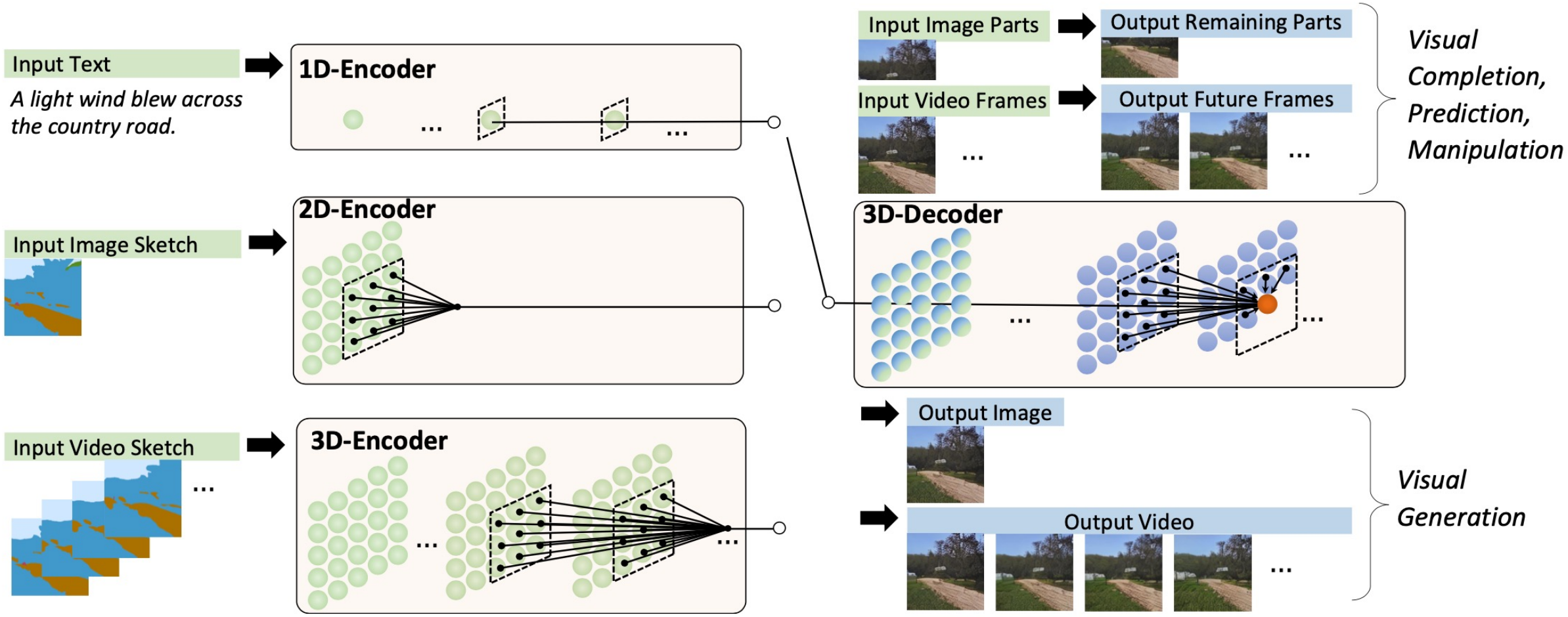
(b) Long Video Generation



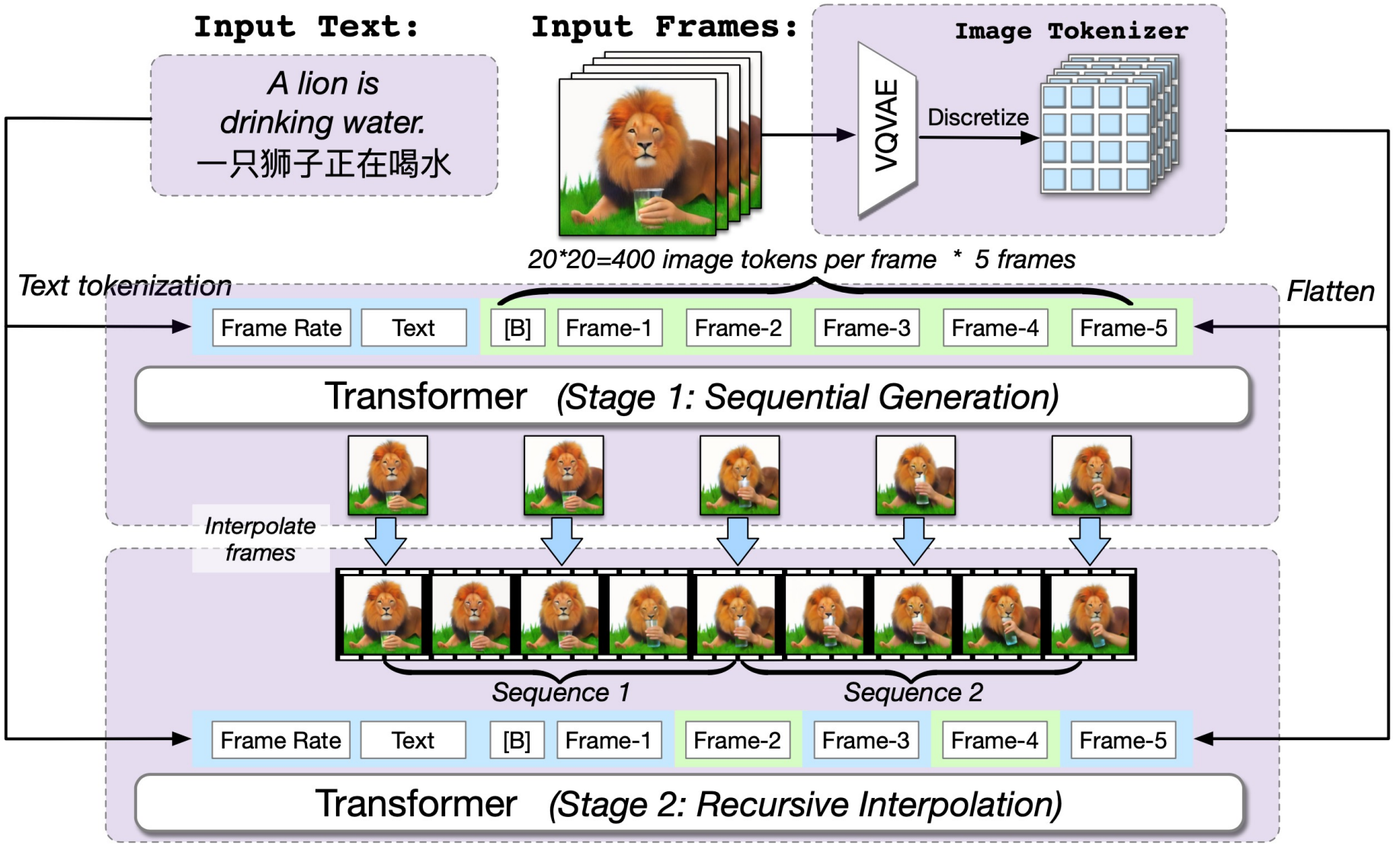
Content

- General Text-to-video Generation
  - Non-diffusion Methods
  - Diffusion-based Methods
- Conditional Video Generation
- Video Editing
- Future Work

Non-diffusion T2V Methods(VQ-GAN)

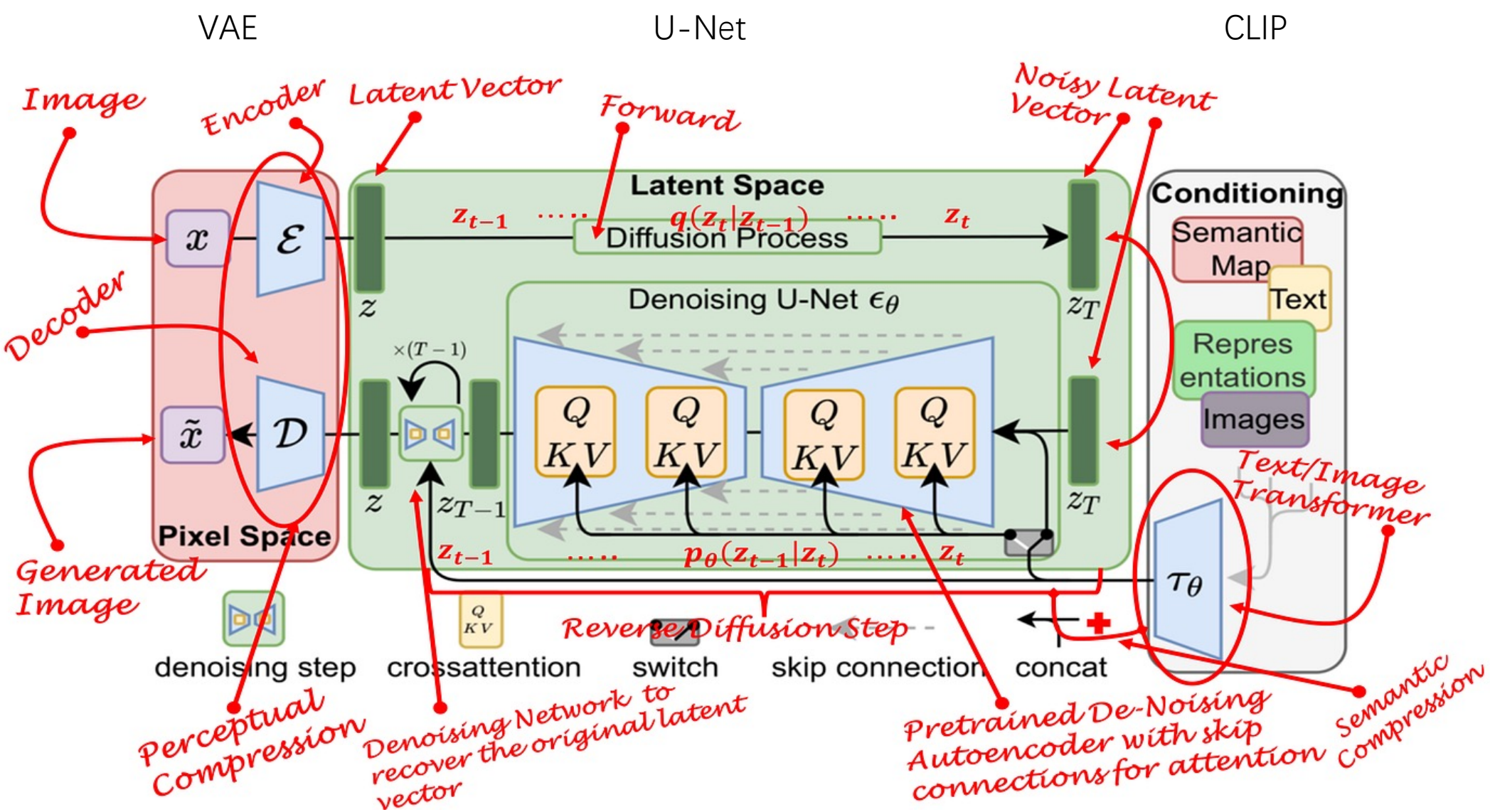


Non-diffusion T2V Methods (Autoregressive)

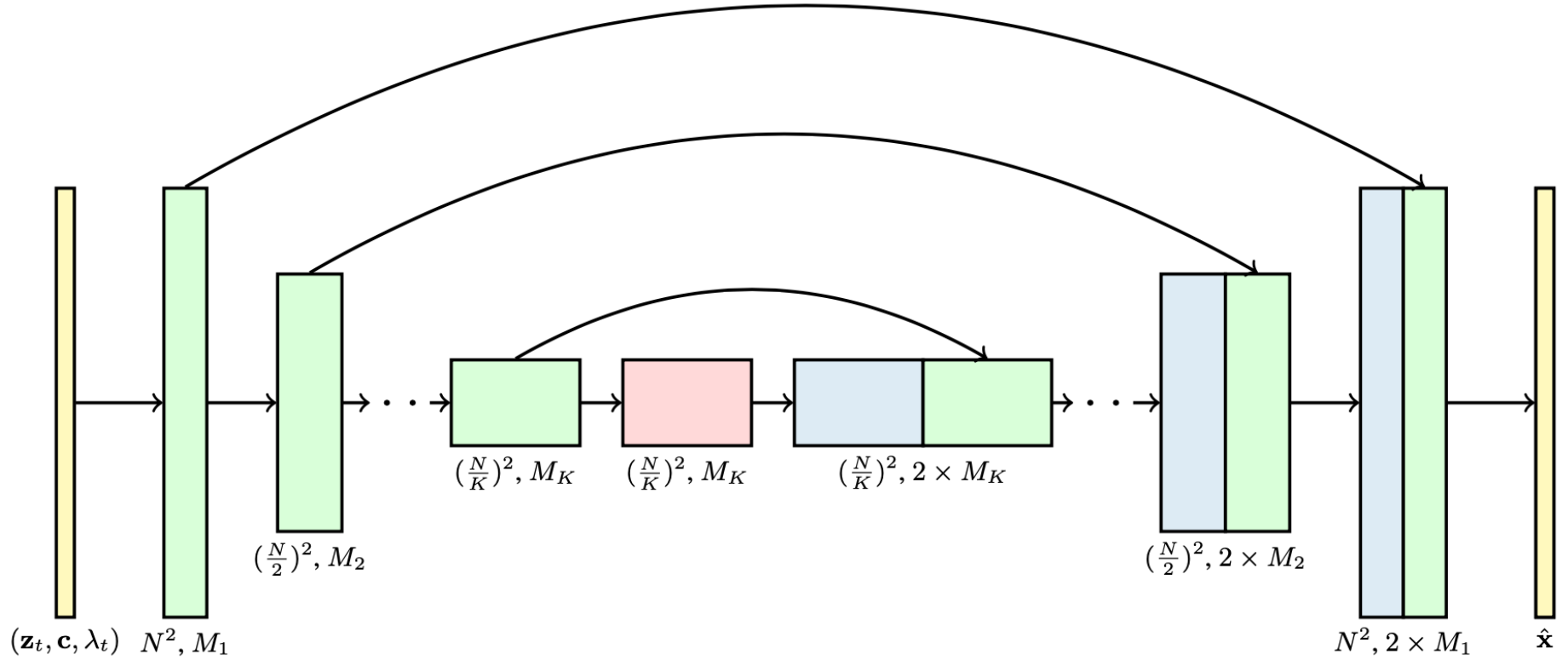


CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers（ICLR2023）

Latent Diffusion Model (Stable Diffusion)



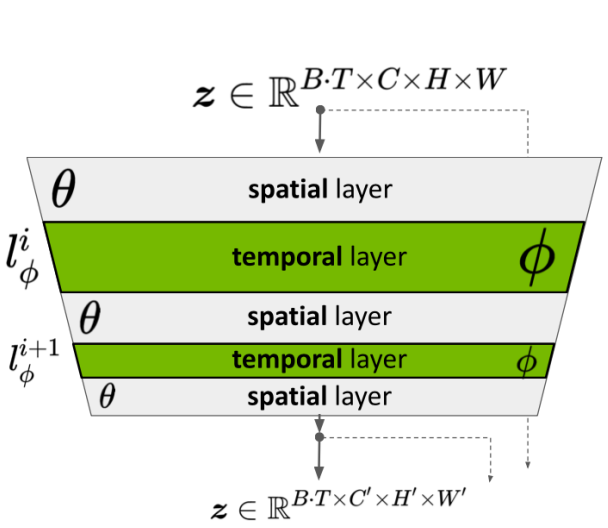
Diffusion-based T2V Methods (First VDM)

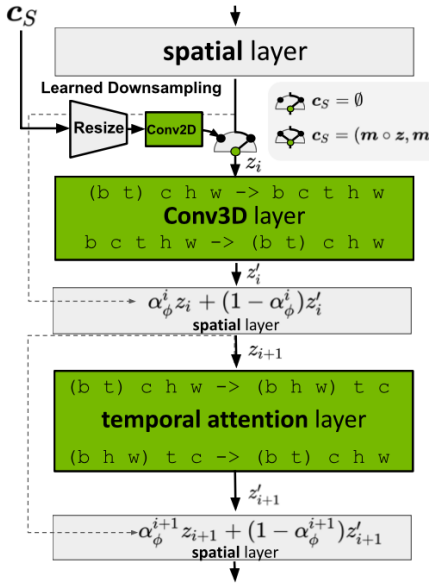


- Conv2D → 3D （3x3 -> 1x3x3）
- Space Attention →Divided Space-Temporal Attention
- Joint training on video and image modeling

Video Diffusion Models （NeurIPS 2022）

Diffusion-based T2V Methods （LDM-based）

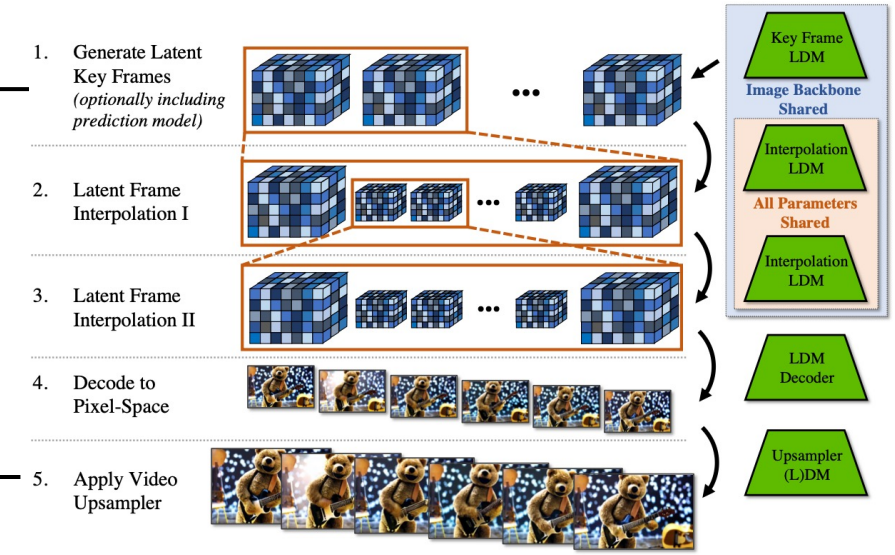




- Trained on **256** GPUs
- Batch Size 768, 402k steps
- **1.71B** Parameters

- 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）

Diffusion-based T2V Methods （Datasets Contribution）

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	✓

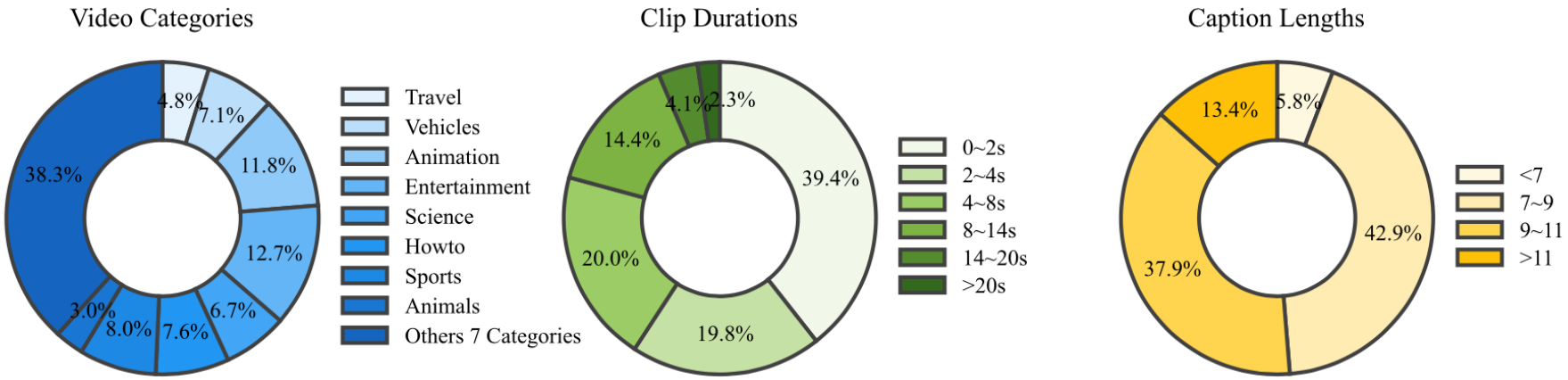
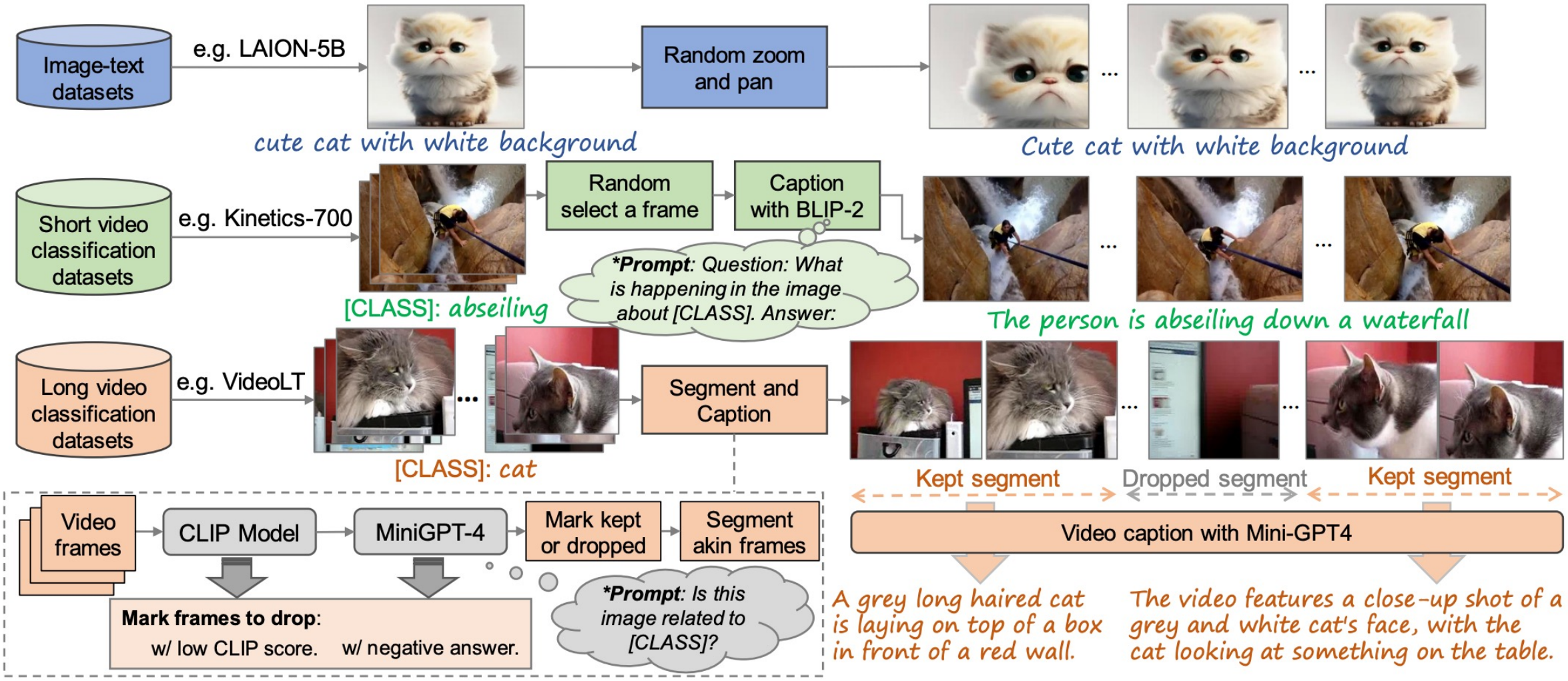


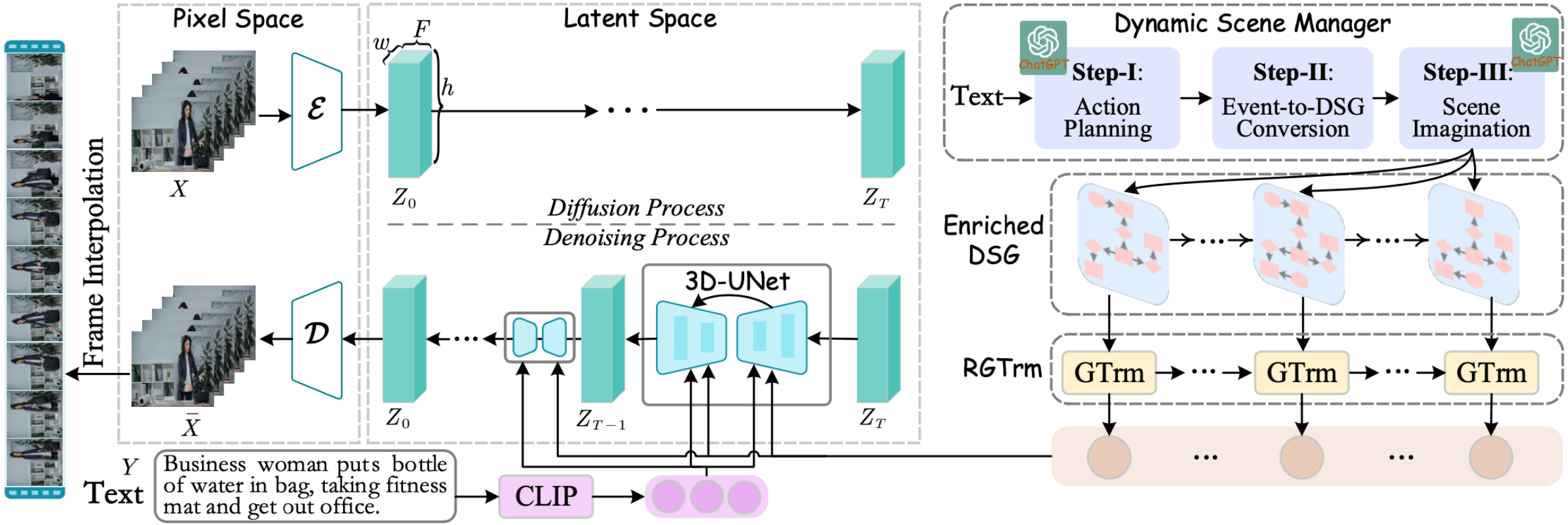
Figure 2: Statistics of video categories, clip durations, and caption word lengths in HD-VG-130M. HD-VG-130M covers a wide range of video categories.

Diffusion-based T2V Methods （Dataset Contribution）



REUSE AND DIFFUSE: ITERATIVE DENOISING FOR TEXT-TO-VIDEO GENERATION（Arxiv 2023）

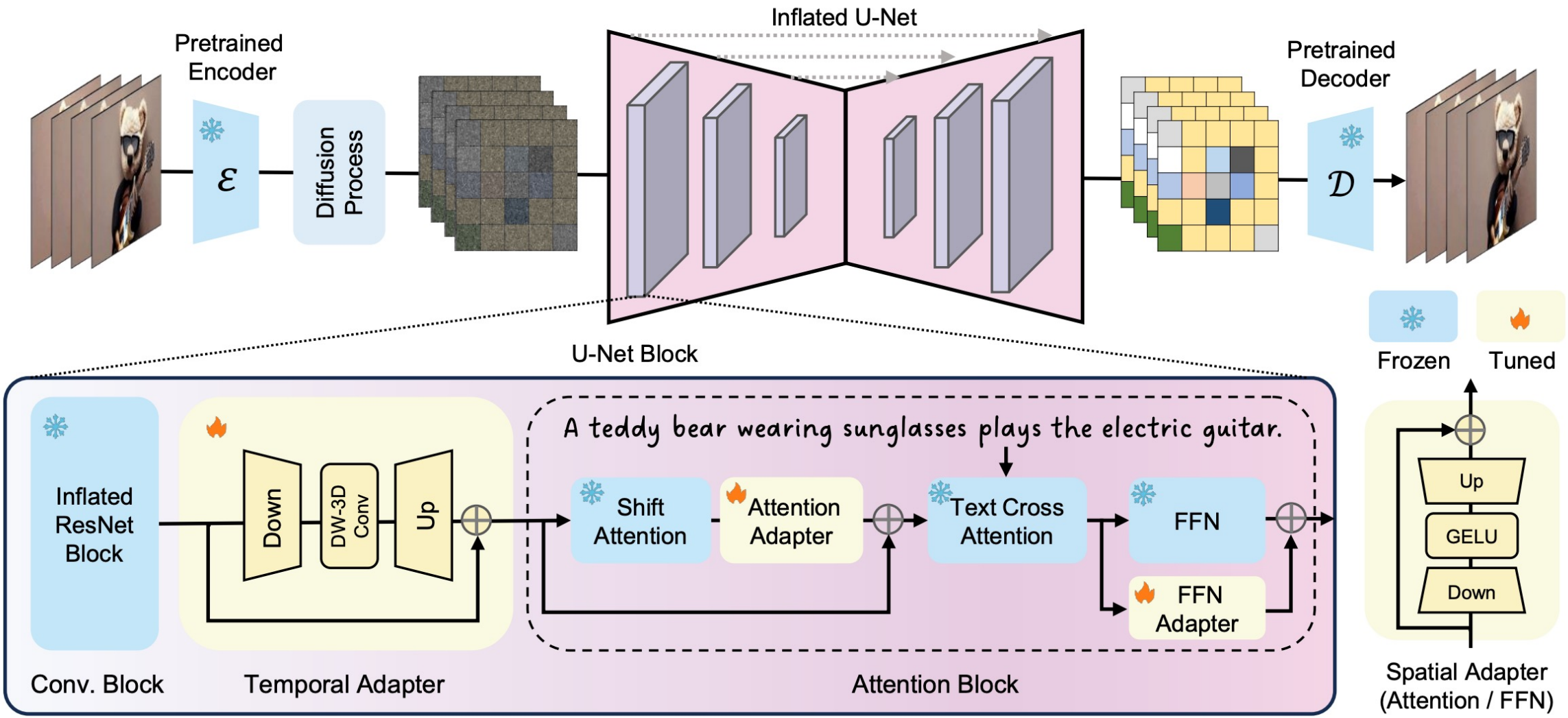
Diffusion-based T2V Methods （LLM guided）



Method	UCF-101		MSR-VTT	
	IS (↑)	FVD (↓)	FID (↓)	CLIPSIM (↑)
CogVideo (Hong et al., 2022)	25.27	701.59	23.59	0.2631
MagicVideo (Zhou et al., 2022)	/	699.00	/	/
MakeVideo (Singer et al., 2022)	33.00	367.23	13.17	0.3049
AlignLatent (Blattmann et al., 2023)	33.45	550.61	/	0.2929
Latent-VDM (Rombach et al., 2022a)	/	/	14.25	0.2756
Latent-Shift (An et al., 2023)	/	/	15.23	0.2773
Dysen-VDM	**35.57**	**325.42**	**12.64**	**0.3204**

EMPOWERING DYNAMICS-AWARE TEXT-TO-VIDEO DIFFUSION WITH LARGE LANGUAGE MODELS （Arxiv 2023）

Diffusion-based T2V Methods （Efficient Training）



SimDA: Simple Diffusion Adapter for Efficient Video Generation（Arxiv 2023）

Diffusion-based T2V Methods （Efficient Training）

An astronaut flying in space,4k,high resolution

Xmas Christmas tree holiday celebration winter snow animation gold background

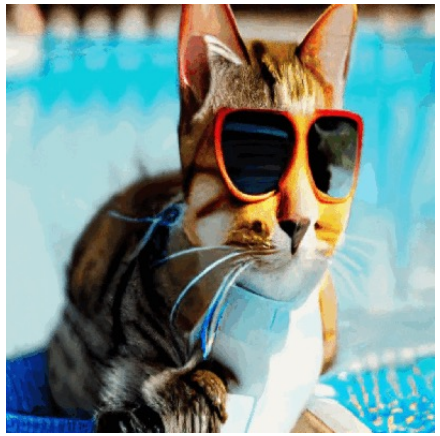
Standing on top a mountainside watching the sunset with the vivid pinks red orange showing from the fire colored sky.

Sea waves with foam on white tropical sandy beach.

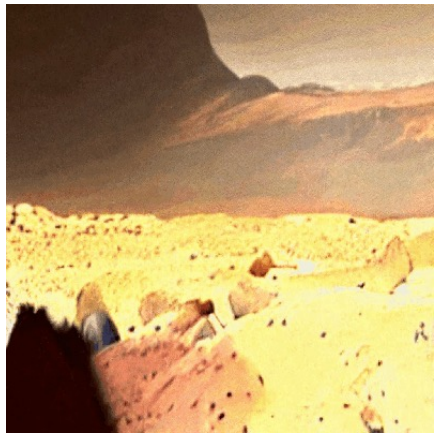
Beer pouring into glass, low angle video shot

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.

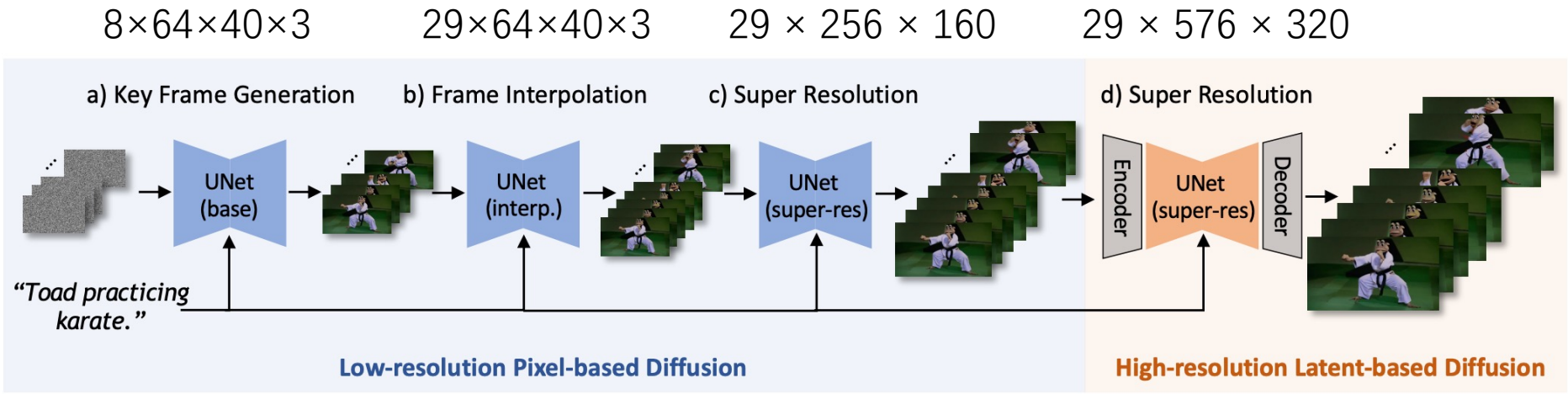


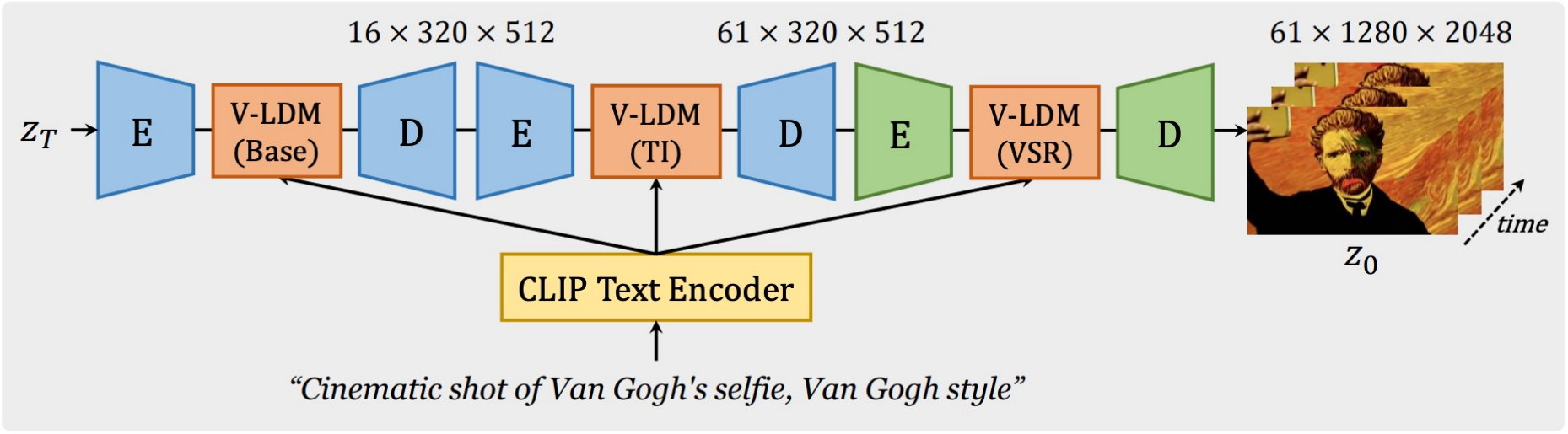
A beautiful sunrise on mars, Curiosity rover.

Coffee pouring into a cup.

SimDA: Simple Diffusion Adapter for Efficient Video Generation （Arxiv 2023）

Diffusion-based T2V Methods (Multi-stage)



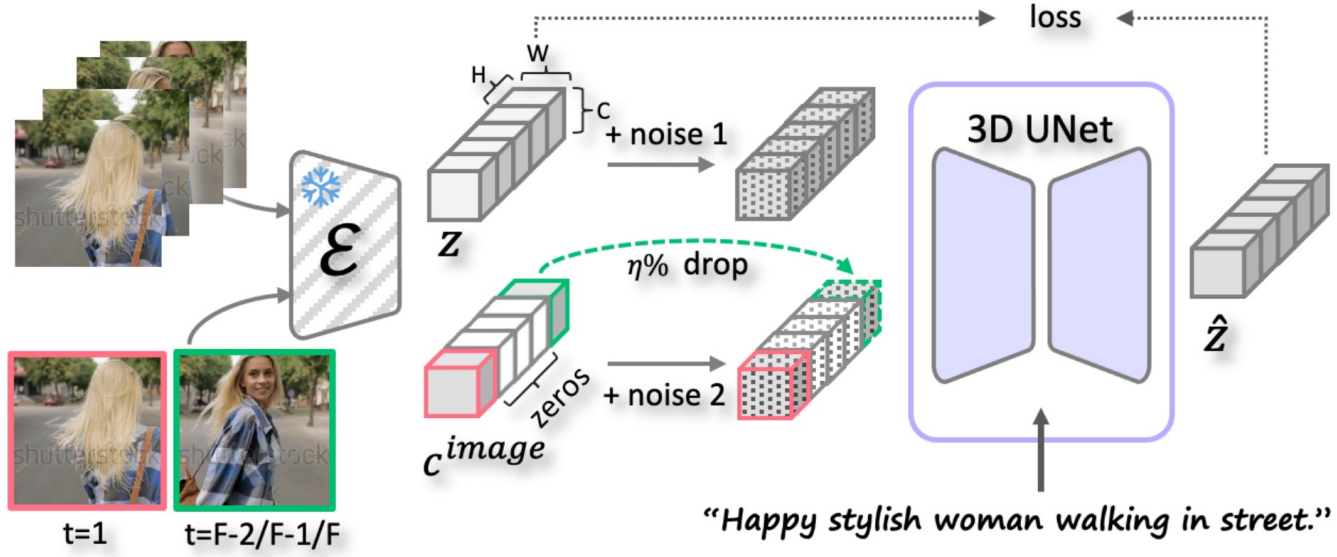


SHOW-1: MARRYING PIXEL AND LATENT DIFFUSION MODELS FOR TEXT-TO-VIDEO GENERATION
LAVIE: HIGH-QUALITY VIDEO GENERATION WITH CASCADED LATENT DIFFUSION MODELS (Arxiv 2023)

Diffusion-based T2V Methods （Image Conditioned）

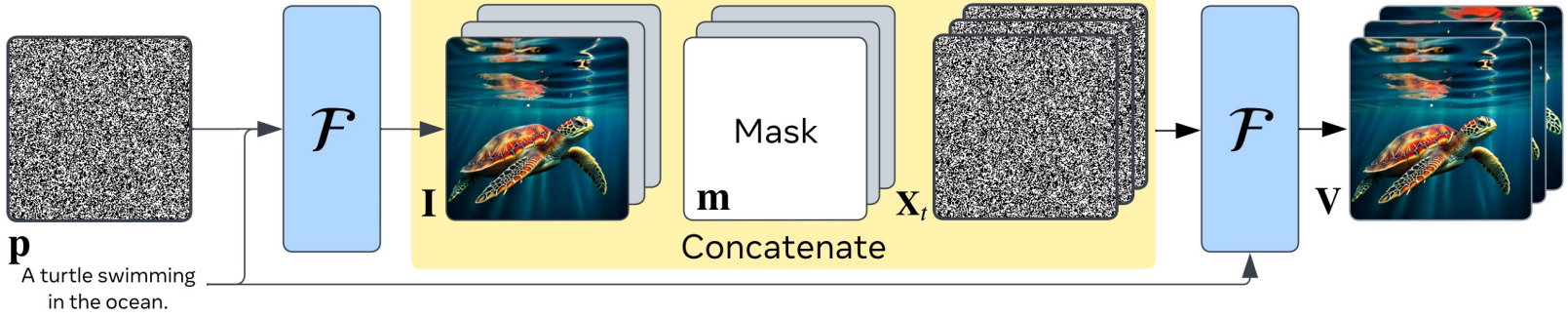
A beautiful woman with long golden hair is driving a red convertible sports car.

The skull is burining, red fires, the skull exploding



fireworks

A panda bear driving a car



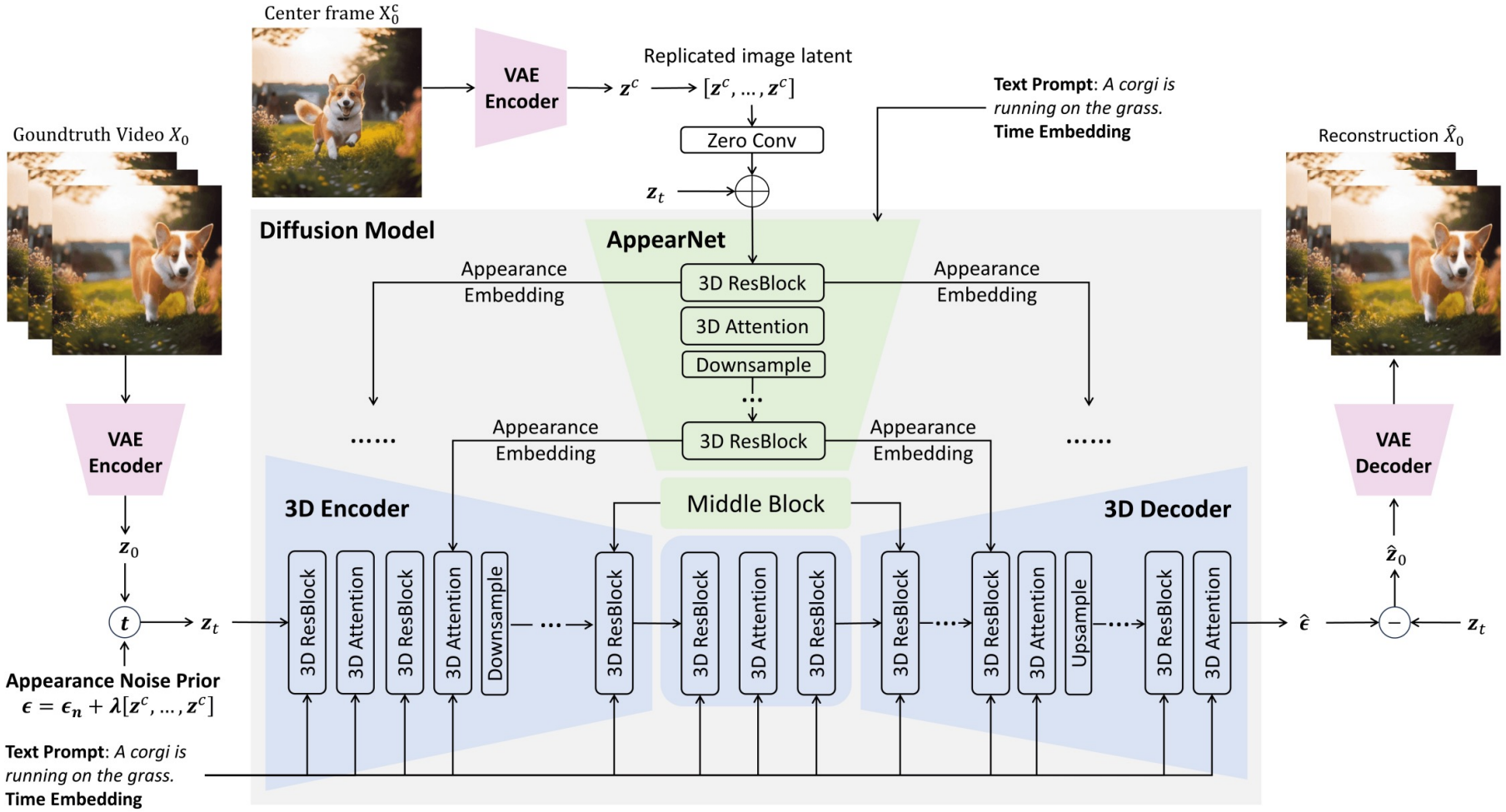
Make Pixels Dance: High-Dynamic Video Generation (Arxiv 2023)

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

Diffusion-based T2V Methods (Image Conditioned)

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



MicroCinema: A Divide-and-Conquer Approach for Text-to-Video Generation (Arxiv 2023)

Diffusion-based T2V Methods (Stable Video Diffusion)

- 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.

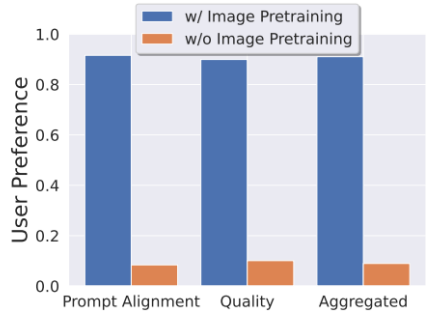
Video Caption Method

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

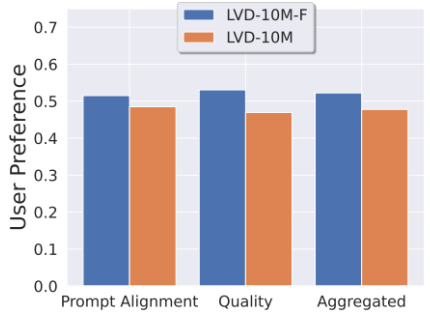
Source Video	Caption: CoCa	Caption: VBLIP	Caption: LLM
	there is a piece of wood on the floor next to a tape measure .	a person is using a ruler to measure a piece of wood	A person is using a ruler to measure a piece of wood on the floor next to a tape measure.
	two men sitting on a rock near a river . 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?	✓	✓	✓	✓	✗	✗



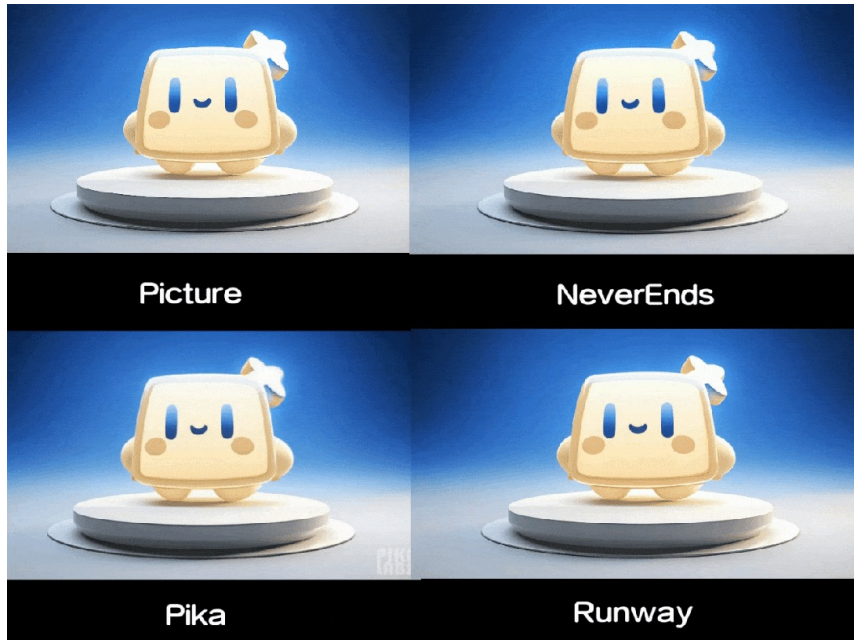
(a) Initializing spatial layers from pretrained images models greatly improves performance.

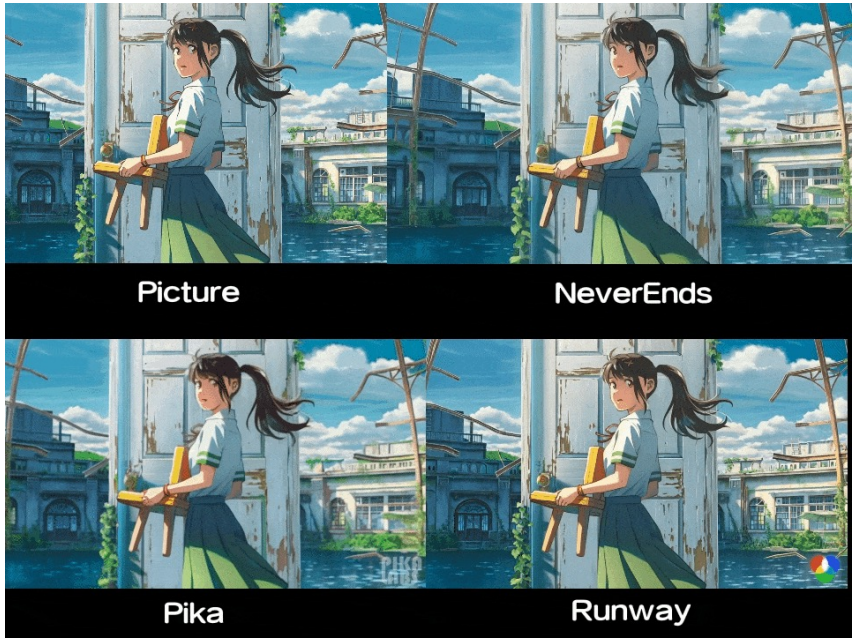


(b) Video data curation boosts performance after video pretraining.

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

Diffusion-based T2V Methods （Image Conditioned）





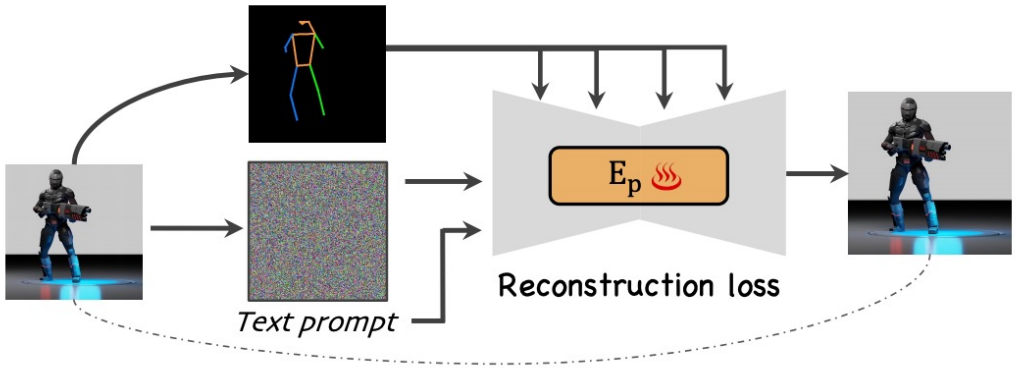
Content

- General Text-to-video Generation
- Conditional Video Generation
  - Pose Guided Video Generation
  - Depth Guided Video Generation
  - Multi-modal Video Generation
  - Uni Audio-Video Generation
- Video Editing
- Future Work

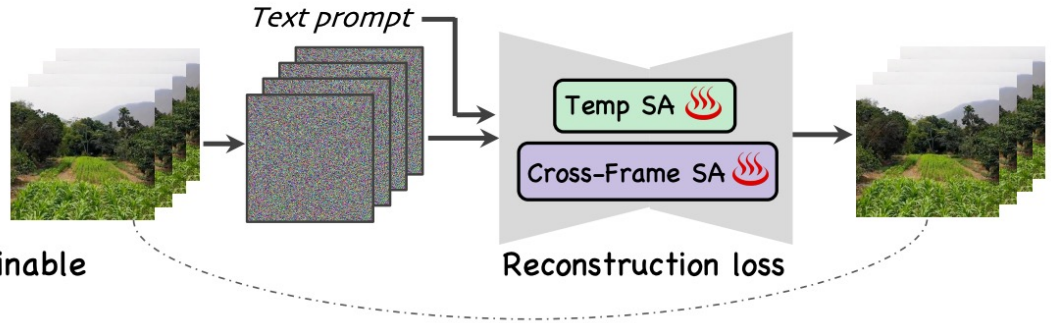
Pose Guided Video Generation

2-Stage Training

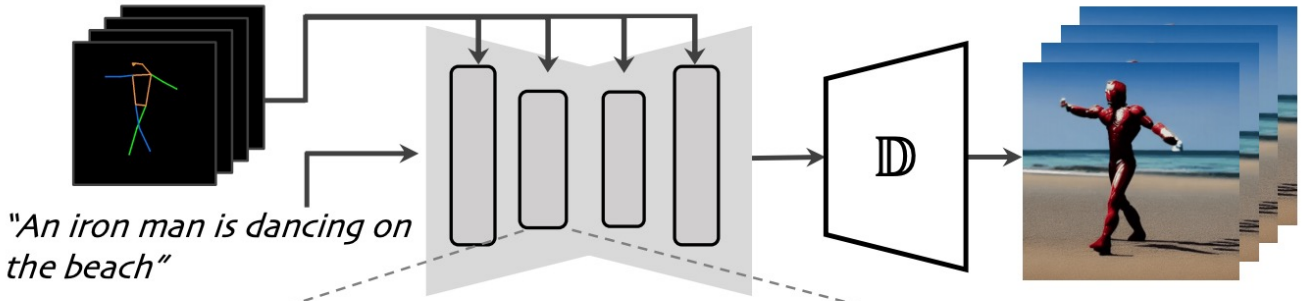
Stage-1: Optimization with keypoints-image pairs



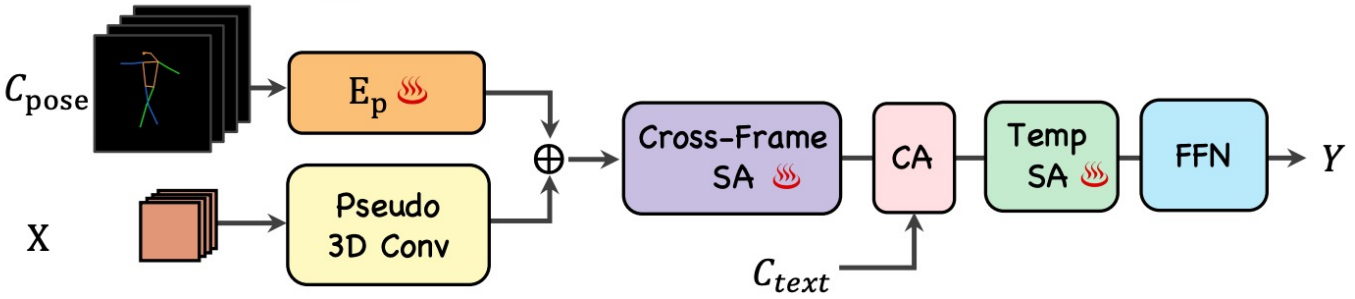
Stage-2: Optimization with Pose-free videos

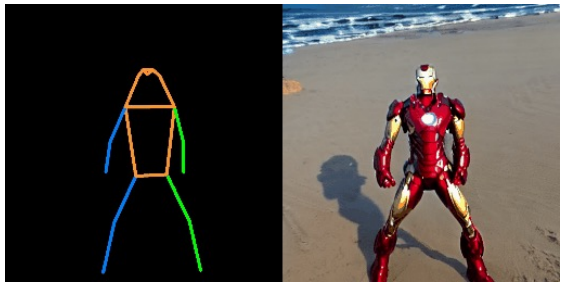


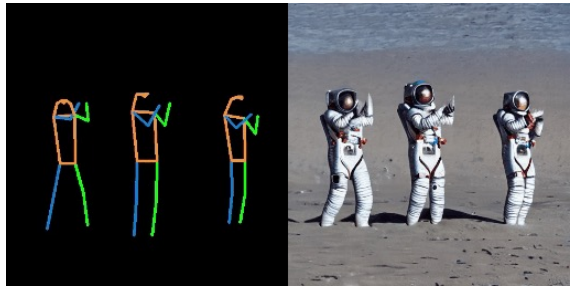
Inference

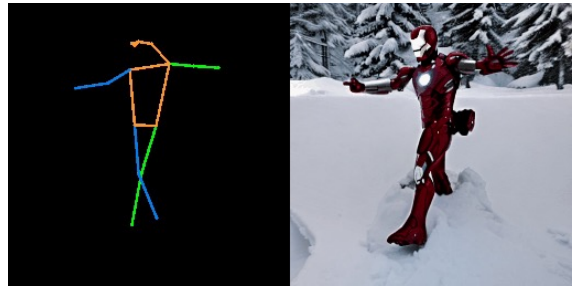


Pose-guided Video Diffusion Block



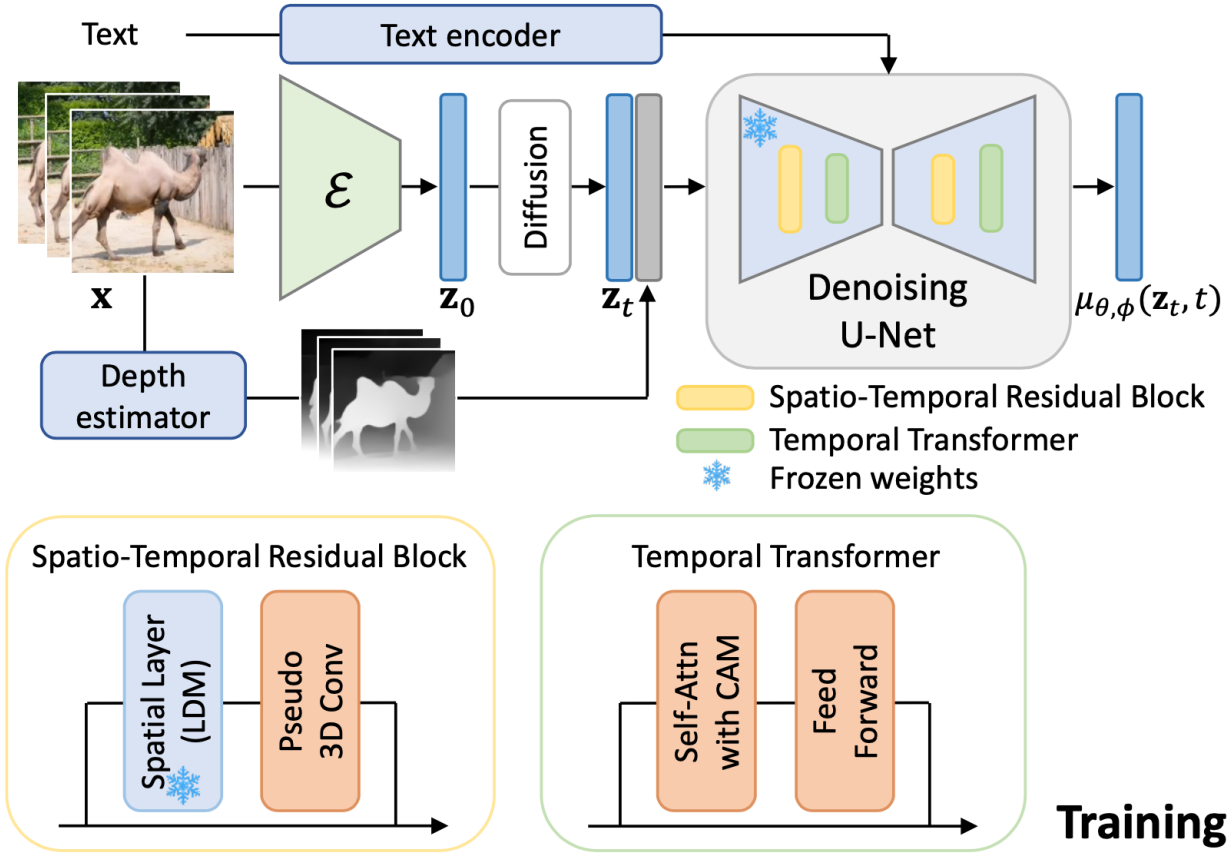


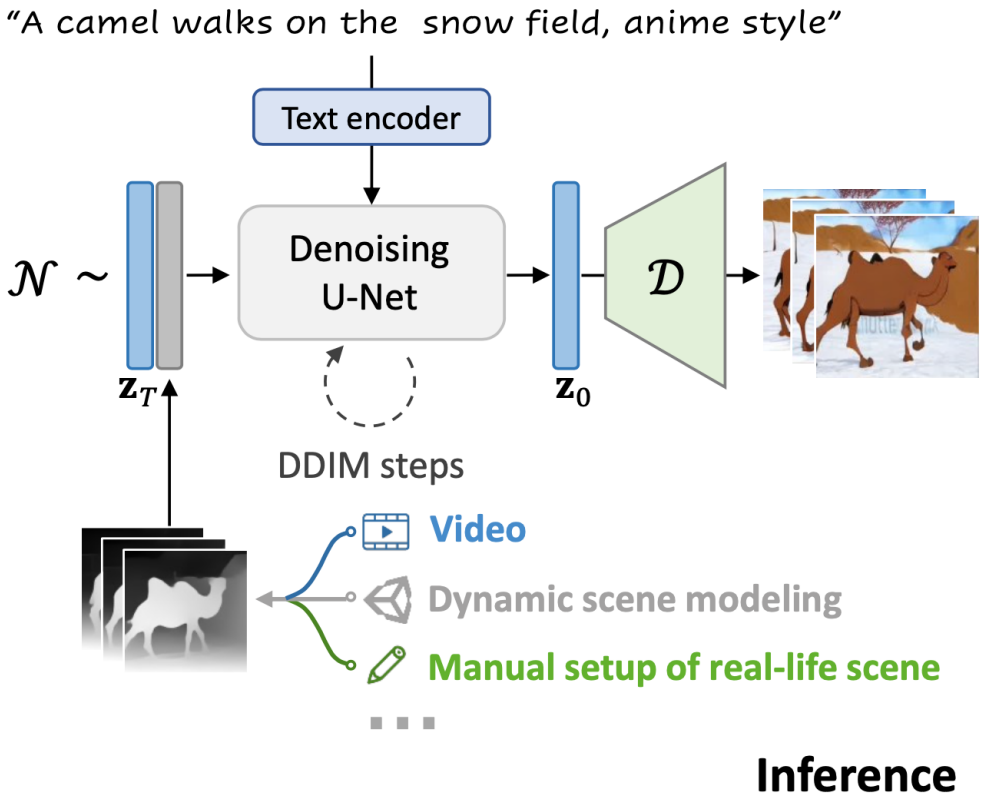




Follow Your Pose: Pose-Guided Text-to-Video Generation using Pose-Free Videos（Arxiv 2023）

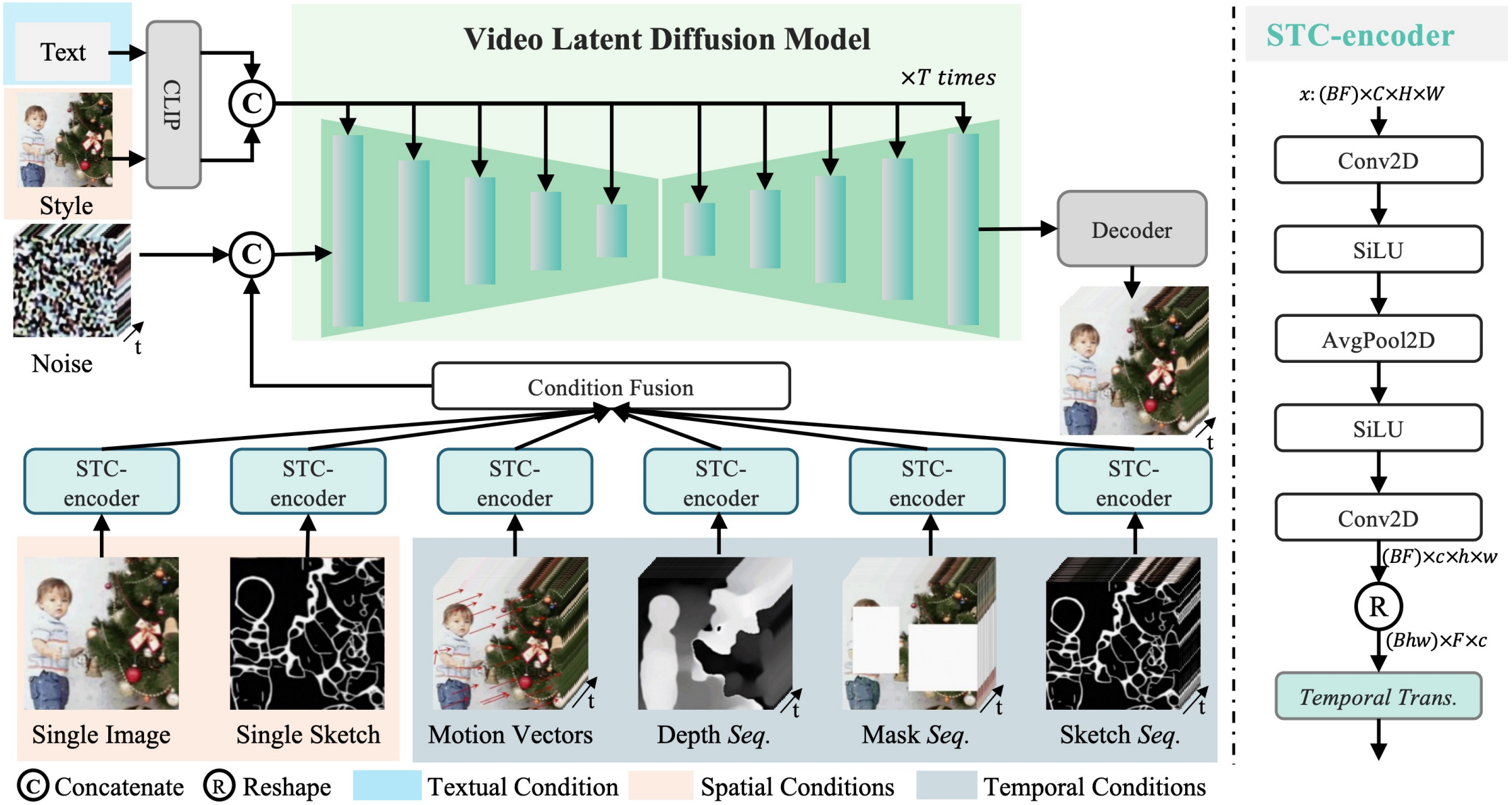
Depth Guided Video Generation





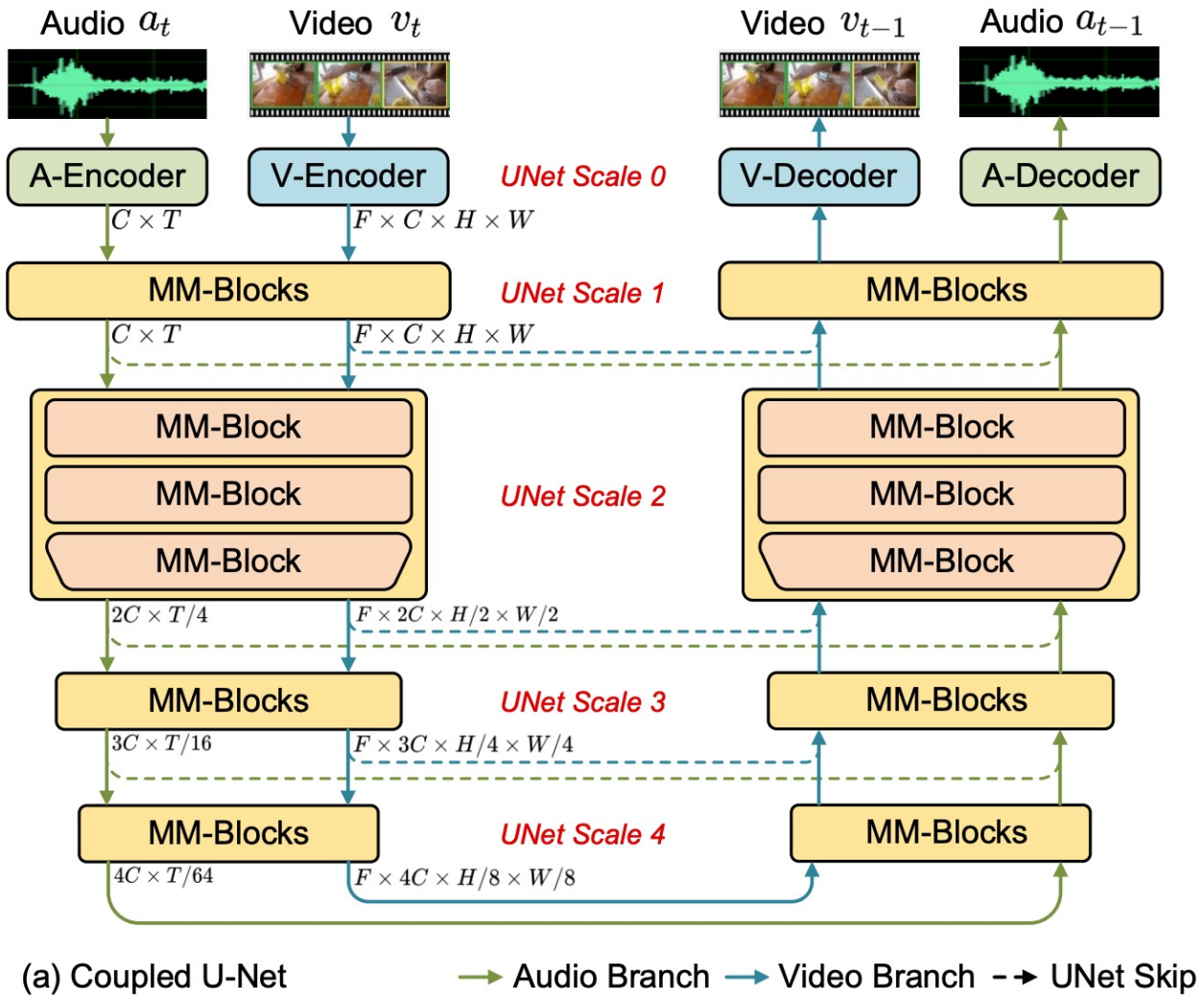
Make-Your-Video: Customized Video Generation Using Textual and Structural Guidance （Arxiv 2023 ）

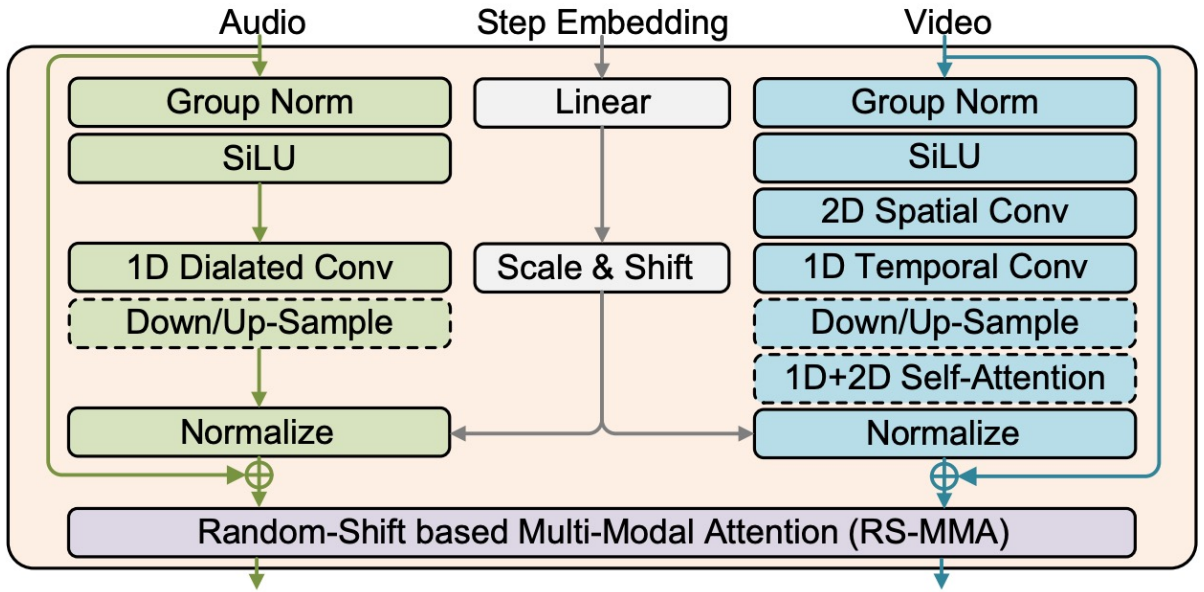
Multi-modal Guided Video Generation

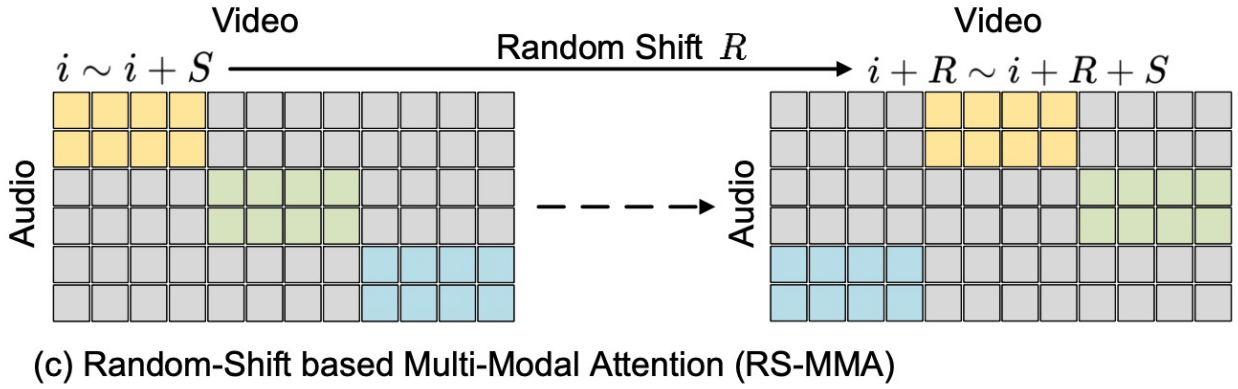


VideoComposer: Compositional Video Synthesis with Motion Controllability（Arxiv 2023）

Audio-Video Generation





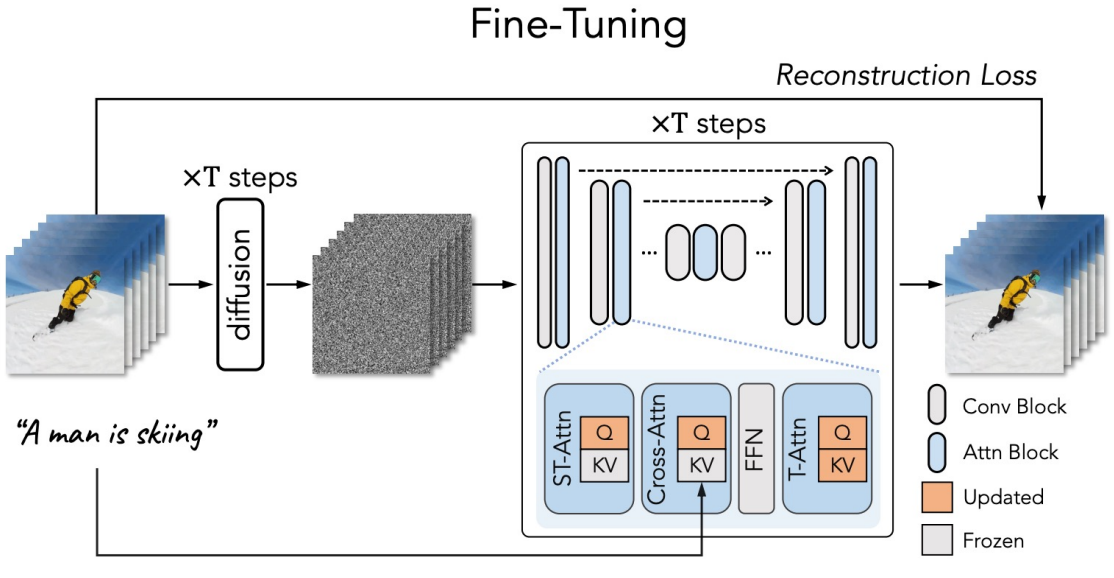


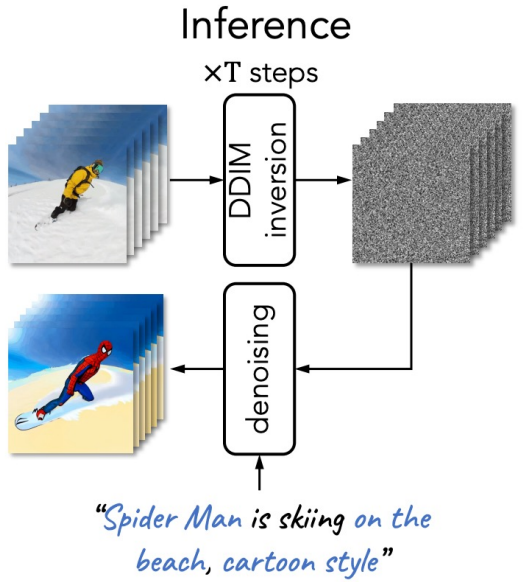
MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation（CVPR 2023）

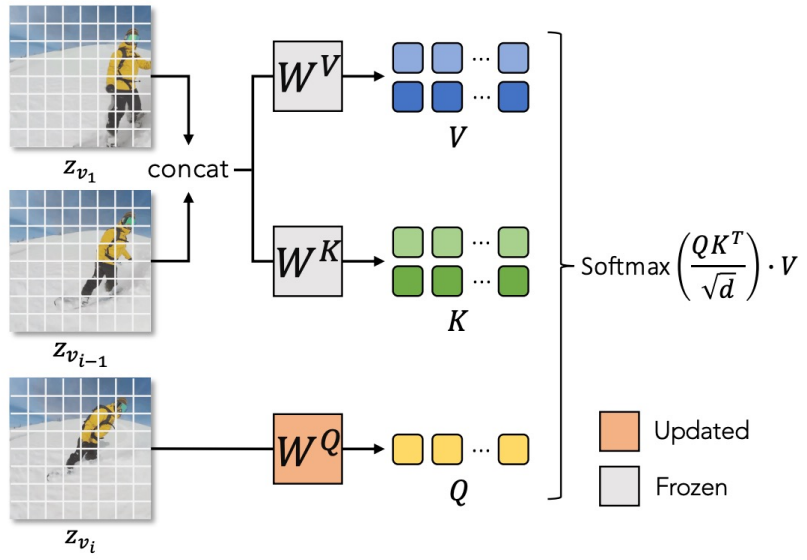
Content

- General Text-to-video Generation
- Conditional Video Generation
- Video Editing
  - One-shot Video Editing
  - General Video Editing
  - Instruction-guided Video Editing
- Future Work

One-Shot Video Editing







A rabbit is eating a watermelon.

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

A lion is roaring

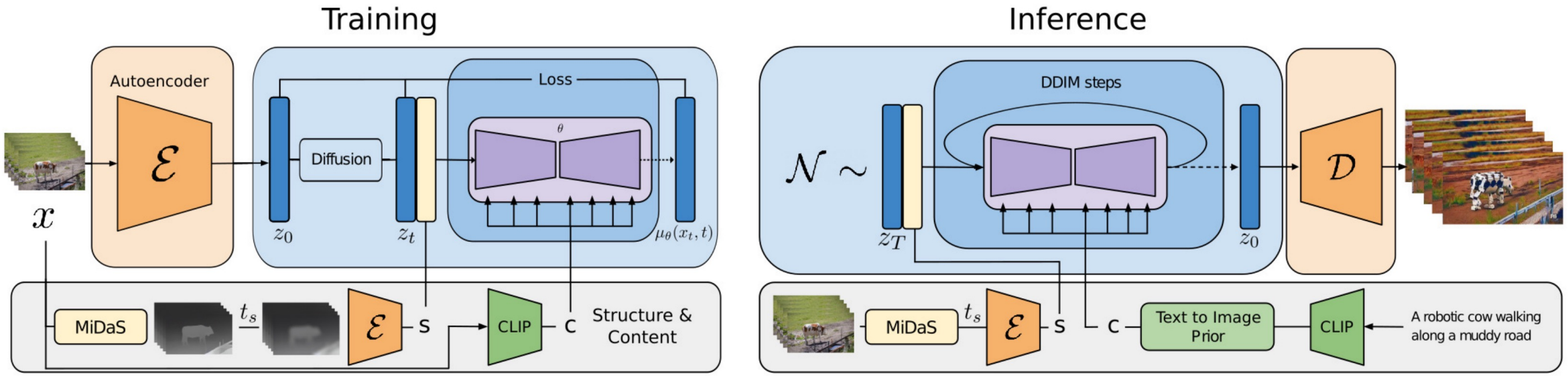
A lion is roaring, Van Gogh Style.

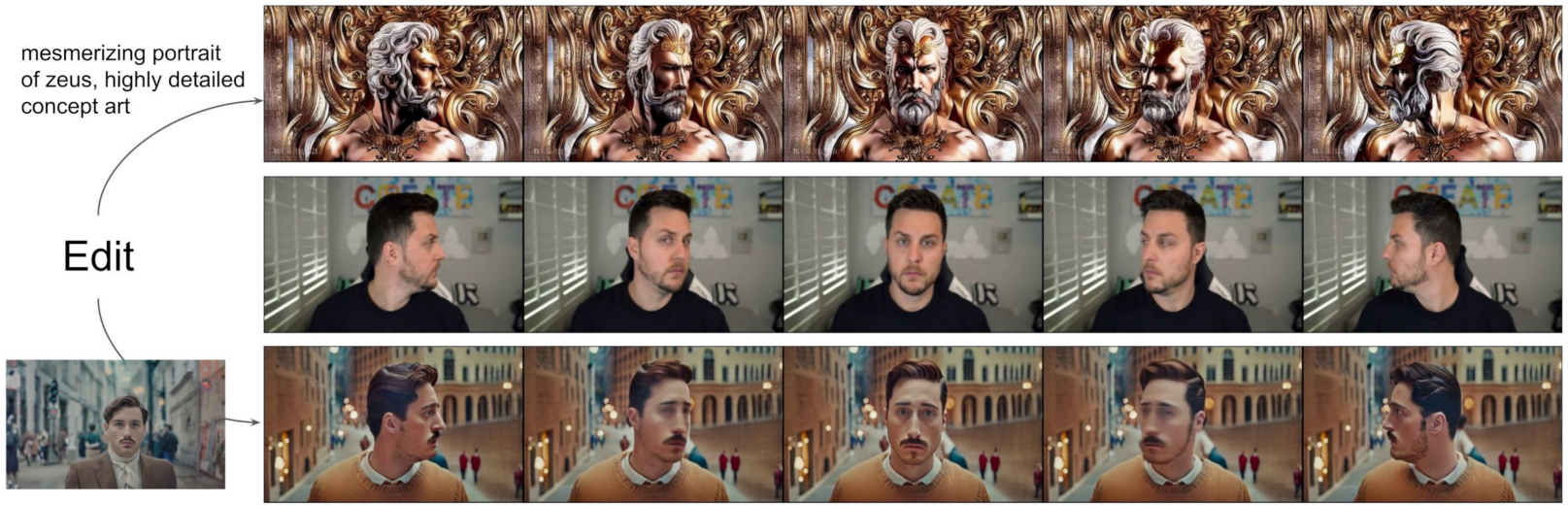
A jeep car is moving on the road.

A jeep car is moving on the road, cartoon style.

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

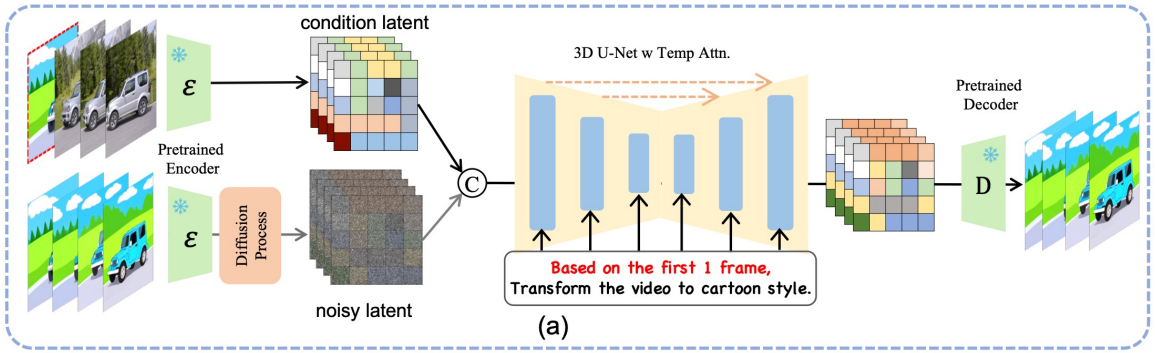
General Video Editing

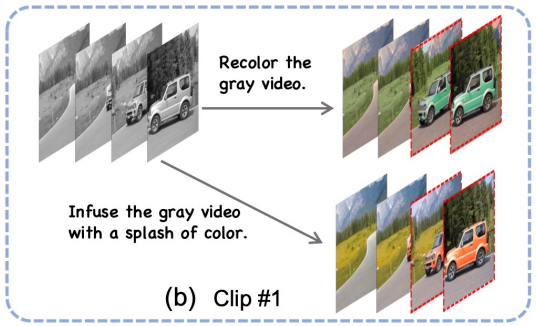


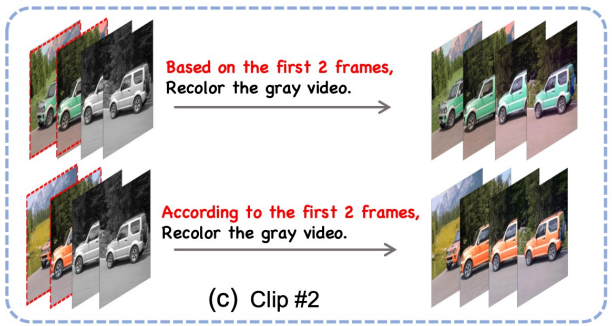


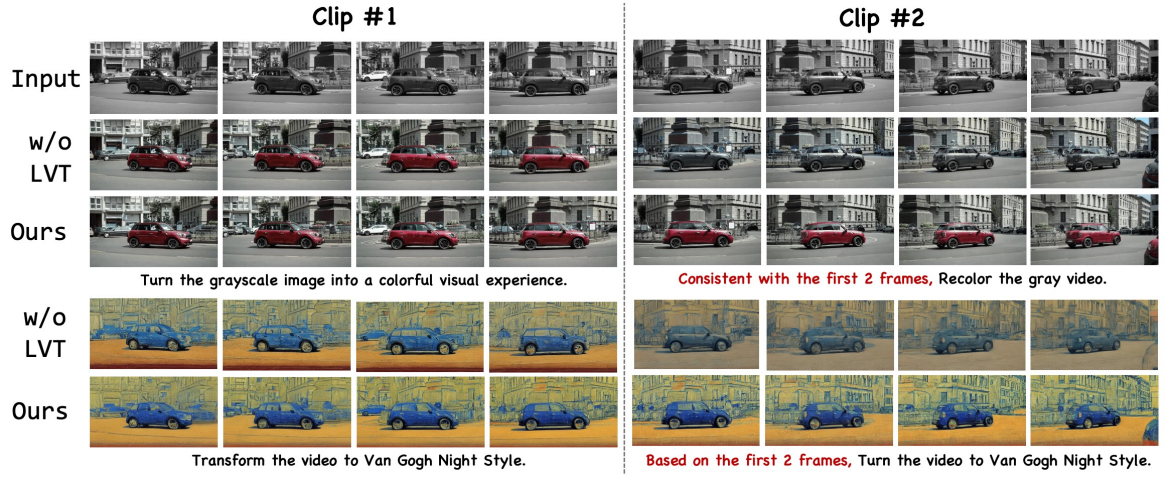
Structure and Content-Guided Video Synthesis with Diffusion Models（ICCV 2023）

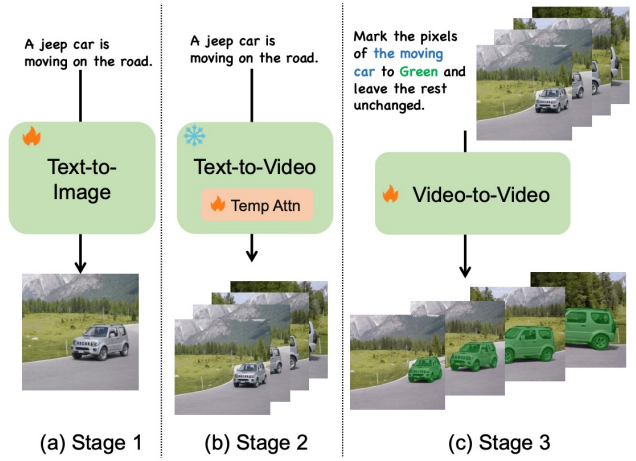
Instruction guided Video Editing











VIDiff: Translating Videos via Multi-Modal Instructions with Diffusion Models（Arxiv 2023）

Instruction guided Video Editing

VIDiff: Translating Videos via Multi-Modal Instructions with Diffusion Models（Arxiv 2023）

Content

- General Text-to-video Generation
- Control guided Video Generation
- Video Completion
- Future Work

Datasets & Metrics
Efficient Training & Inference
Long Video Generation
More Application

Dataset

"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"

"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"

"Kherson, ukraine - 20 may 2016: open, free, rock music festival crowd partying at a rock concert. hands up, people, fans cheering clapping applauding in kherson, ukraine - 20 may 2016. band performing"

"Cabeza de toro, punta cana/ dominican republic - feb 20, 2020: 4k drone flight over coral reef with manta"

Evaluation Metric

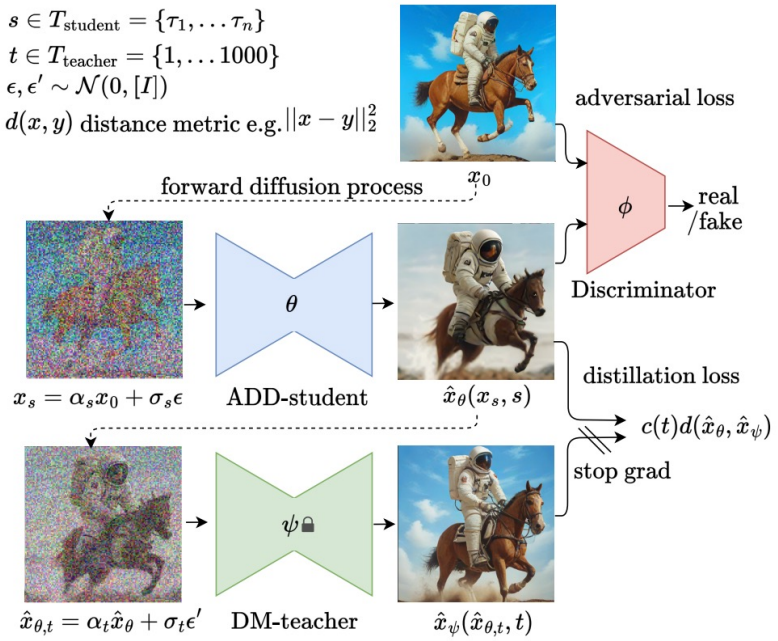
Method	Year	Training Data	Extra Dependency	Resolution	Params(B)	MSRVTT [235]			UCF-101 [188]		
						FID(↓)	FVD(↓)	CLIPSIM(↑)	FID(↓)	FVD(↓)	IS(↑)
Non-diffusion based method											
CogVideo [78]	2022	[5](5.4M)	-	256 × 256	15.5	23.59	1294	0.2631	179.00	701.59	25.27
MMVG [52]	2023	[5](2.5M)	-	256 × 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**	-

Text-to-Video Zero-shot Evaluation on MSRVTT & UCF-101

Efficient Inference

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

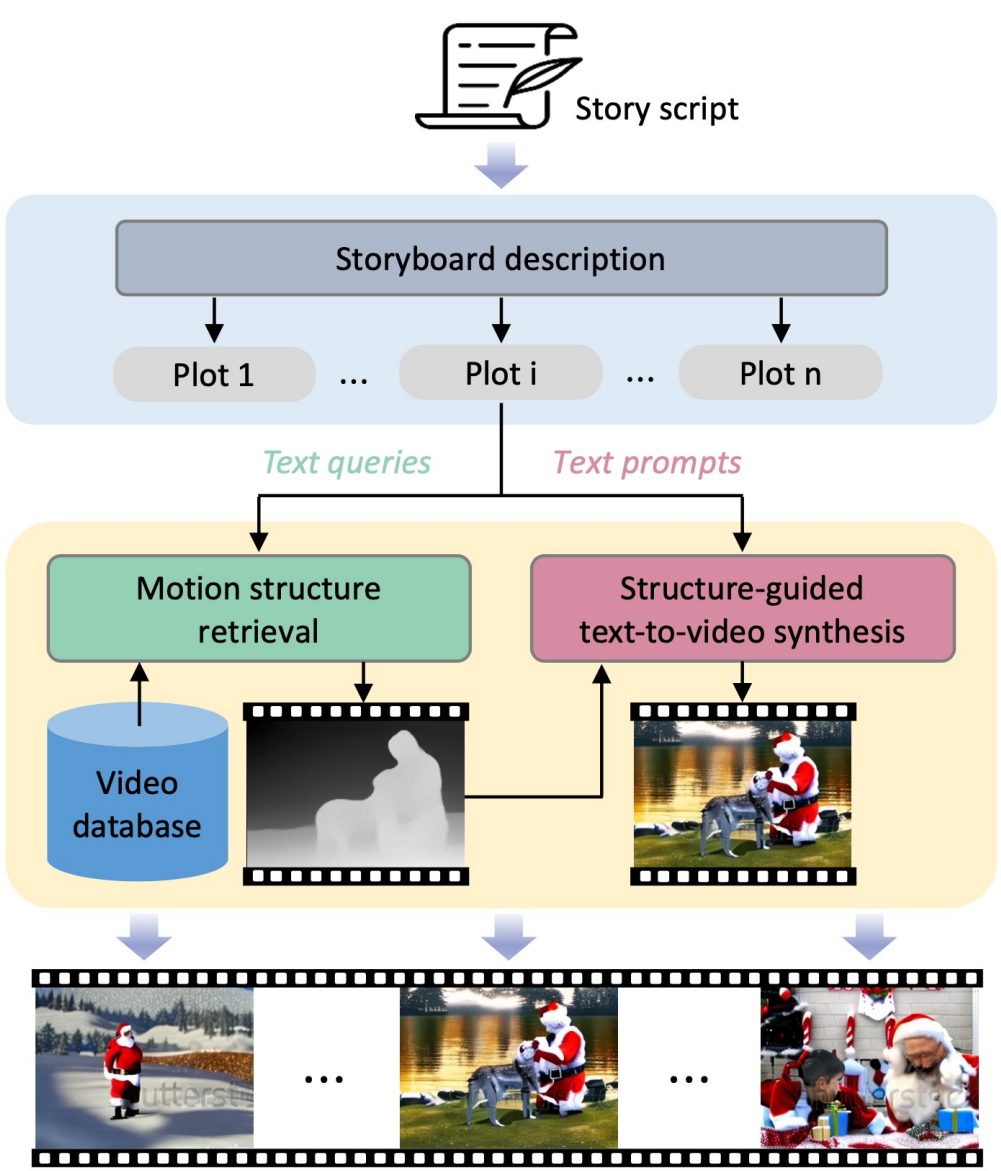


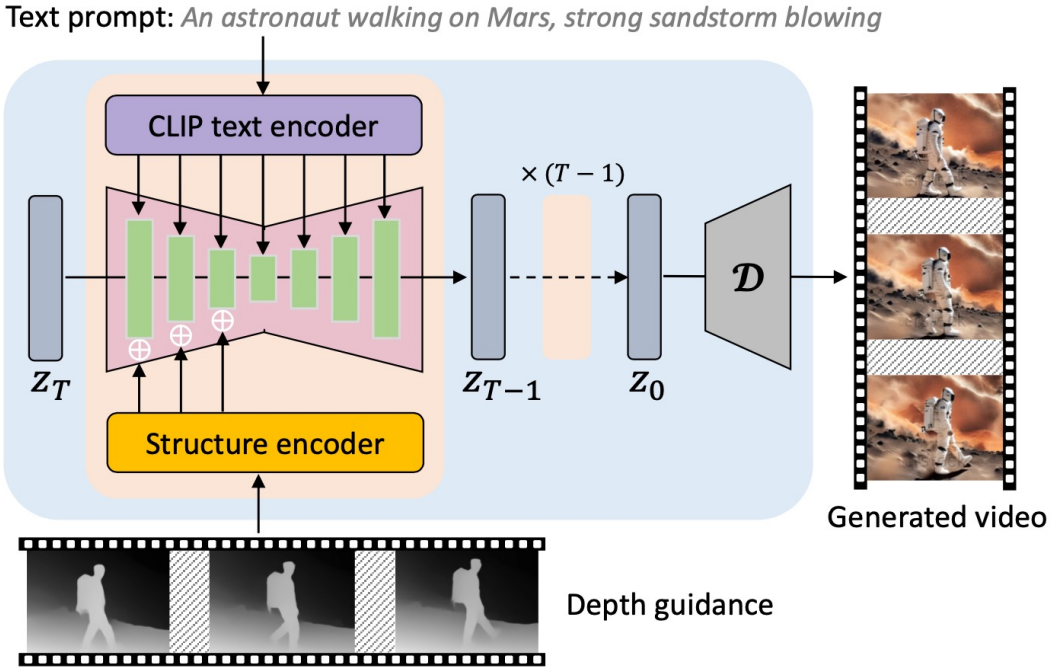
MSR-VTT	Speed		Image Quality				
	Step↓	Time↓	IQS↑	CLIP↑	IS↑	FID↓	NIQE↓
ModelScope	50	21.2	-0.518	0.293	**18.79**	44.85	6.57
Random	29.98	13.5	-0.723	0.293	18.22	47.41	6.75
AdaDiff	31.14	13.6	**-0.517**	**0.299**	18.74	**44.49**	**6.36**

Adversarial Diffusion Distillation （Arxiv 2023 ）

AdaDiff: Adaptive Step Selection for Fast Diffusion (Arxiv 2023)

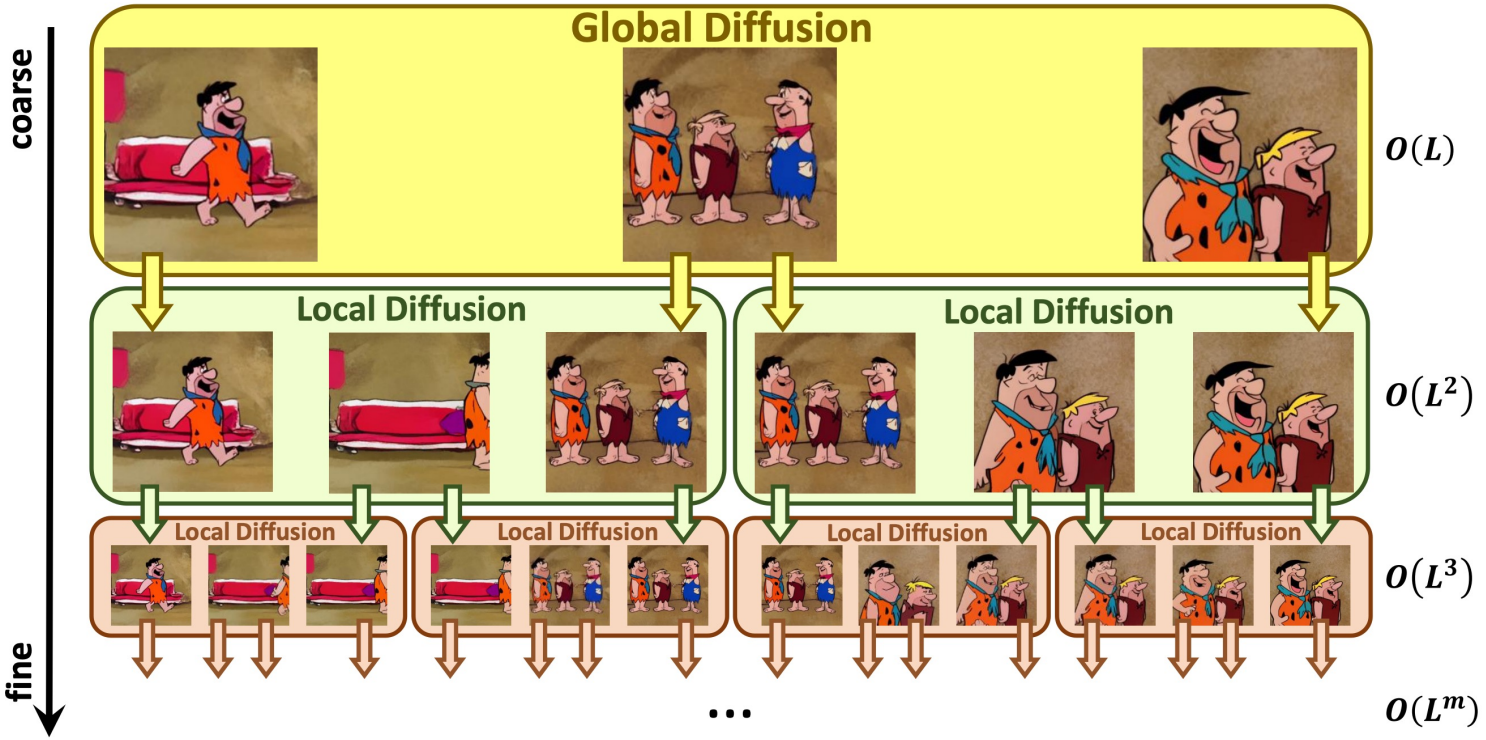
Long Video Generation (Retrieval based motion)

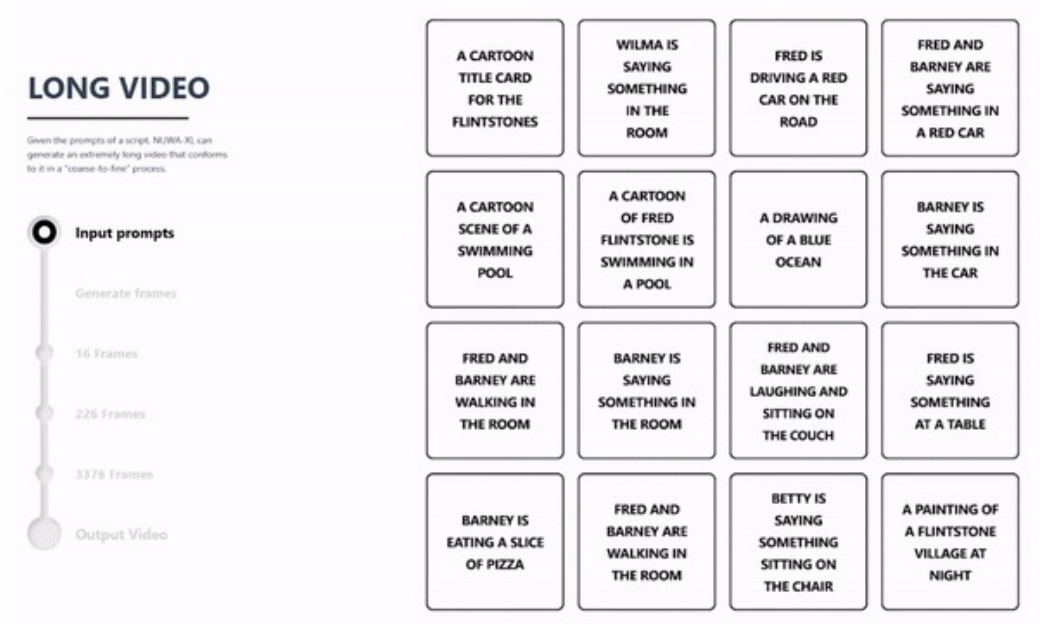




Animate-a-story: Storytelling with retrieval-augmented video generation (Arxiv 2023)

Long Video Generation (Coarse-to-fine)





NUWA-XL: Diffusion over Diffusion for eXtremely Long Video Generation （Arxiv 2023）

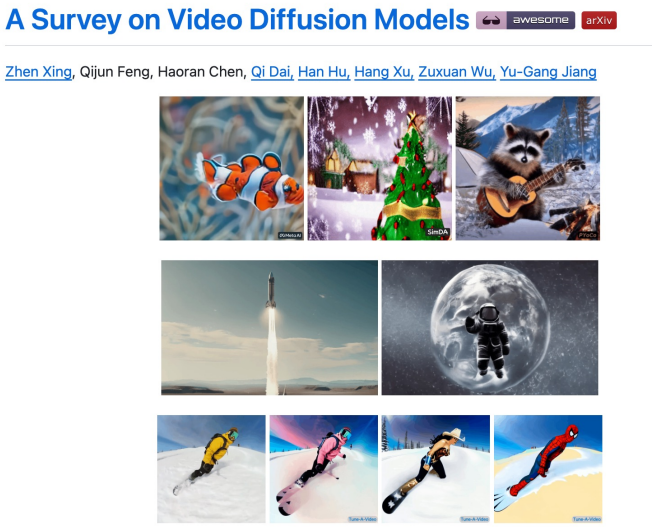
More Application (Animate Anyone)

Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation（Arxiv 2023）

Thanks

HomePage

https://github.com/ChenHsing/Awesome-Video-Diffusion-Models



Presentation: Zhen Xing

2023/12/06

Complex Text Understanding

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)