भारतीय प्रौद्योगिकी संस्थान दिल्ली
Indian Institute of Technology Delhi

COV877
Special Module on Visual Computing

Generative AI for Visual Content Creation: Image, Video, and 3D

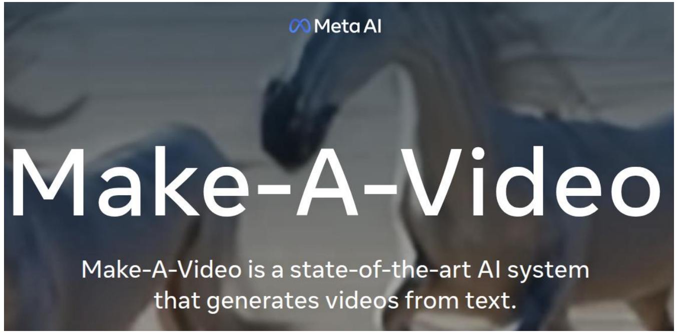
Text-to-Video Generation

Instructor:

Dr. Lokender Tiwari

Research Scientist

Text to Video Generation


Meta AI

Make-A-Video

Make-A-Video is a state-of-the-art AI system that generates videos from text.

NVIDIA Toronto AI Lab

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

Andreas Blattmann[1 *,†] Robin Rombach[1 *,†] Huan Ling[2,3,4 *] Tim Dockhorn[2,3,5 *,†] Seung Wook Kim[2,3,4] Sanja Fidler[2,3,4] Karsten Kreis[2]

[1] LMU Munich, [2] NVIDIA, [3] Vector Institute, [4] University of Toronto, [5] University of Waterloo

[*] Equal contribution.

[†] Andreas, Robin and Tim did the work during internships at NVIDIA.

IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2023

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

ICCV 2023

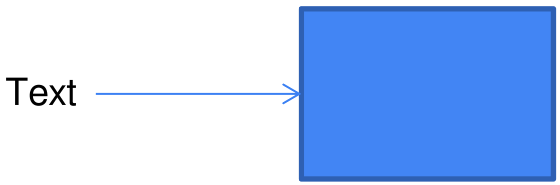
Jay Zhangjie Wu[1] Yixiao Ge[3] Xintao Wang[3] Stan Weixian Lei[1] Yuchao Gu[1] Yufei Shi[1]
Wynne Hsu[2] Ying Shan[3] Xiaohu Qie[4] Mike Zheng Shou[1]

[1]Show Lab,[2]National University of Singapore [3]ARC Lab,[4]Tencent PCG

... many more

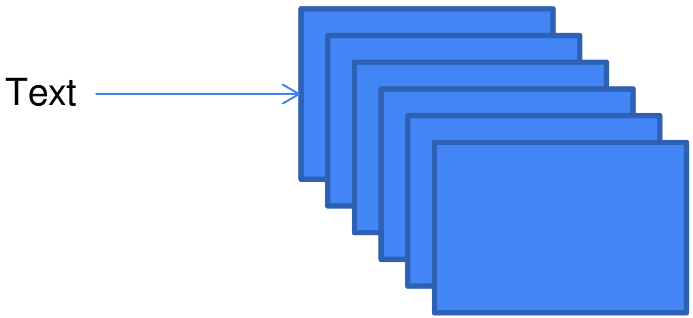
Fundamentals of Video Generation

Text to Image



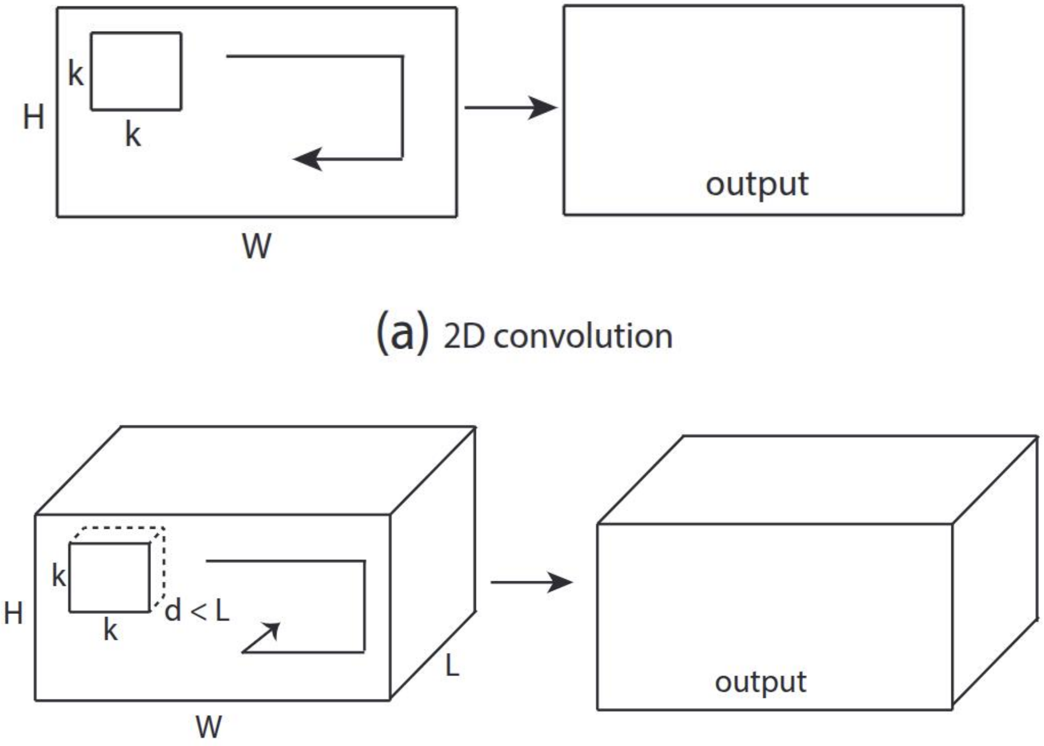
Text

Text to Video



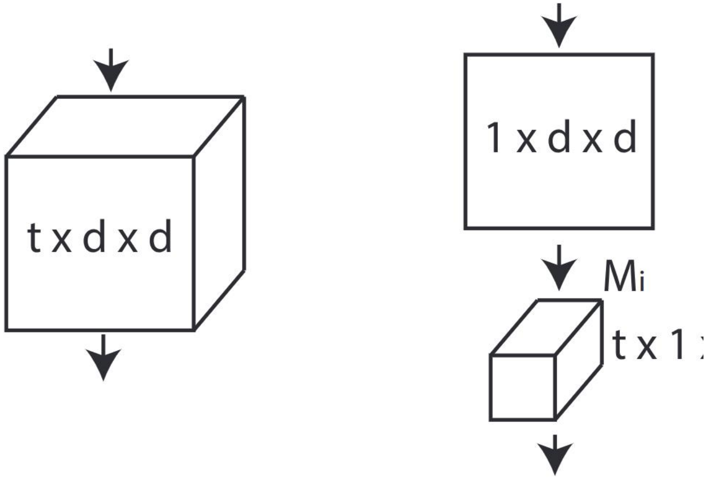
Text

Fundamentals of Video Generation



(a) 2D convolution

(c) 3D convolution



a) 3D Conv

b) (2+1) D Conv

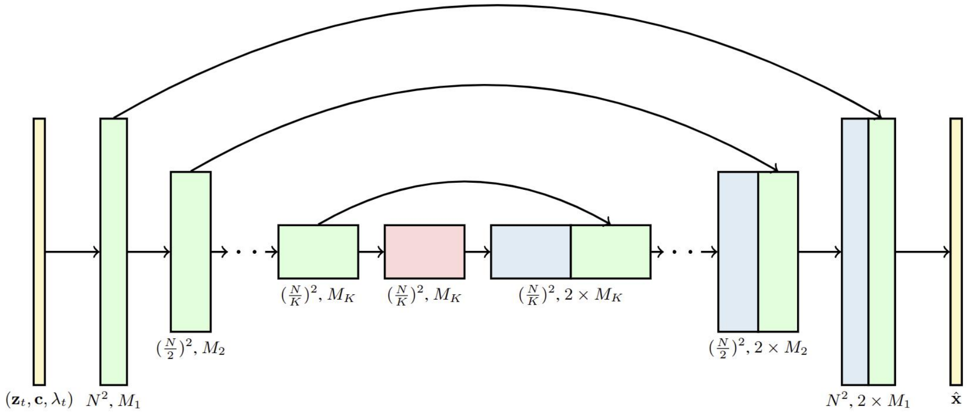
[1] Du et al., “Learning Spatiotemporal Features with 3D Convolutional Networks,” ICCV 2015
[2] Du et al., “A Closer Look at Spatiotemporal Convolutions for Action Recognition,” CVPR 2018

Video Diffusion Model

- Extends 2D U-net , for 3D data
- Each feature map is a 4D tensor (frames x height x width x channels)
- 3D U-net is factorized over space and time - Each layer operates either in the space or time dimension

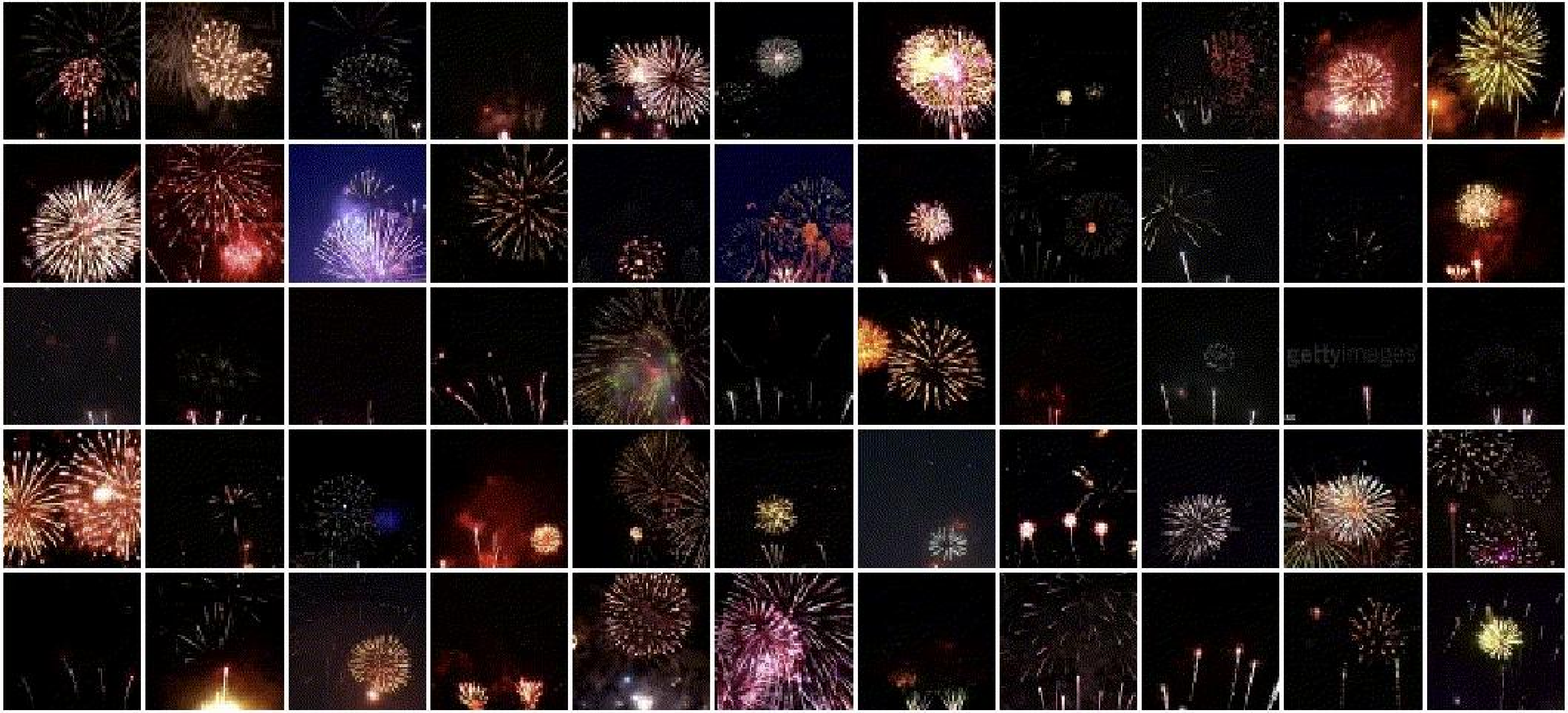
- 2D conv in the 2D U-net is extended to be space-only 3D conv - 3x3 conv become 1x3x3 conv
- Each spatial attention block remains as attention over space - here first axis (frames) is treated as batch dim

- A temporal attention block is added after each spatial attention block.
- Perform attention over the first axis (frames) and treats spatial axes as the batch dim
- The relative position embedding is used for track the order of frames.
- The temporal attention block is important for the model to capture good temporal coherence.



[1] Ho et al., “Video Diffusion Models,” NeurIPS 2022

Video Diffusion Model - Sample Results



Samples from a text-conditioned video diffusion model, conditioned on the string *fireworks*.

Make-a-Video (by Meta)

Make-A-Video consists of three main components:

1. A base T2I model trained on text-image pairs

2. Spatiotemporal convolution and attention layers that extend the networks’ building blocks to the temporal dimension

3. A frame interpolation network for high frame rate generation

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

Make-a-Video (by Meta)

$$\hat{y}_t = \mathrm{SR}_h \circ \mathrm{SR}_l^t \circ \uparrow_F \circ \mathrm{D}^t \circ \mathrm{P} \circ (\hat{x}, \mathrm{C}_x(x)),$$

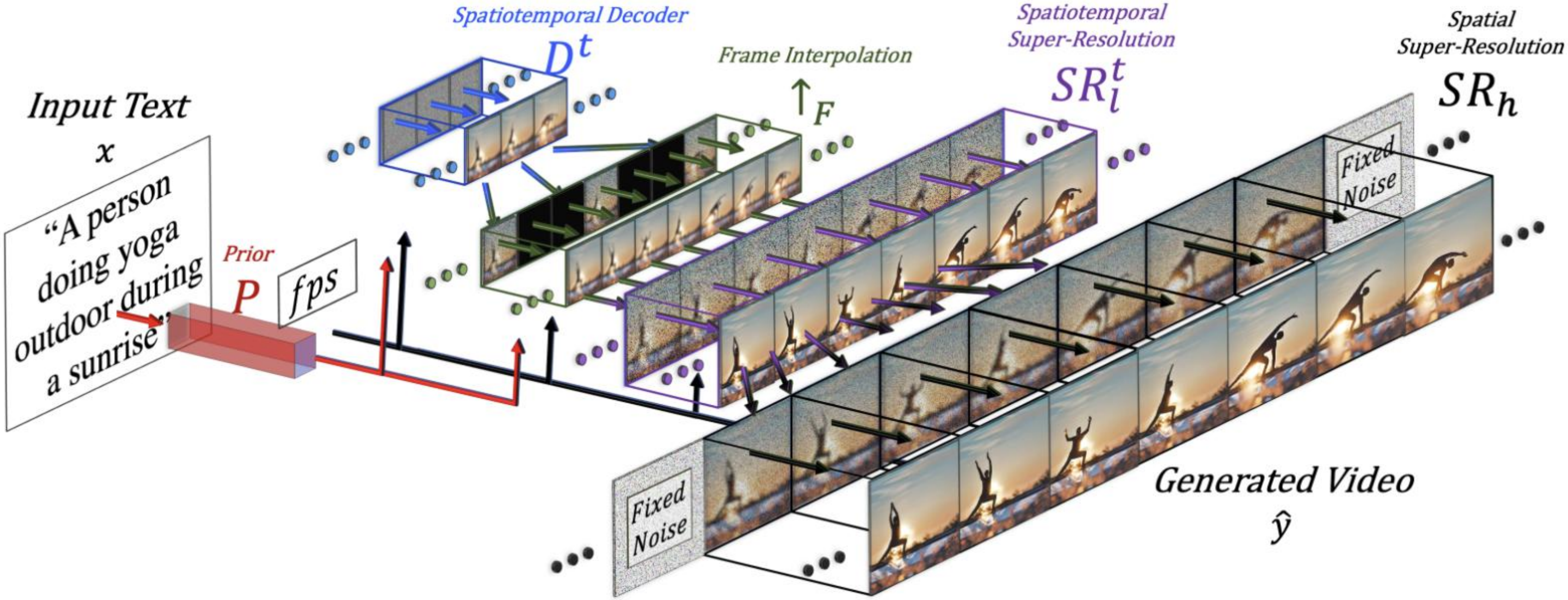
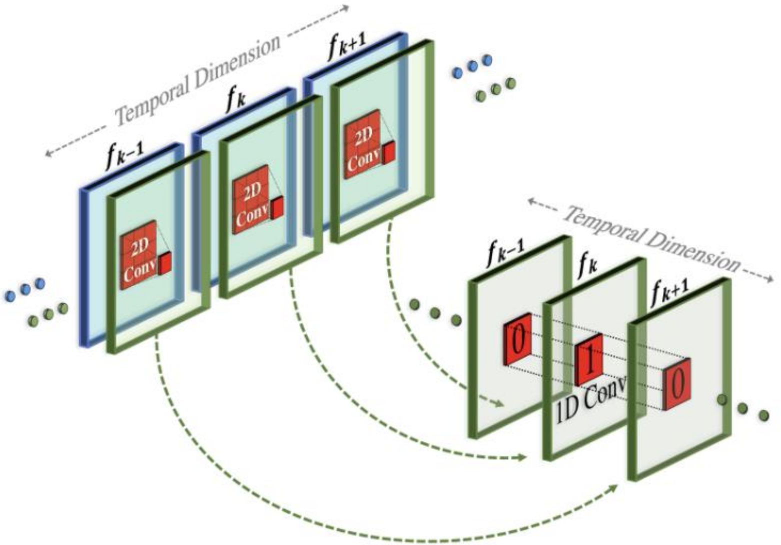
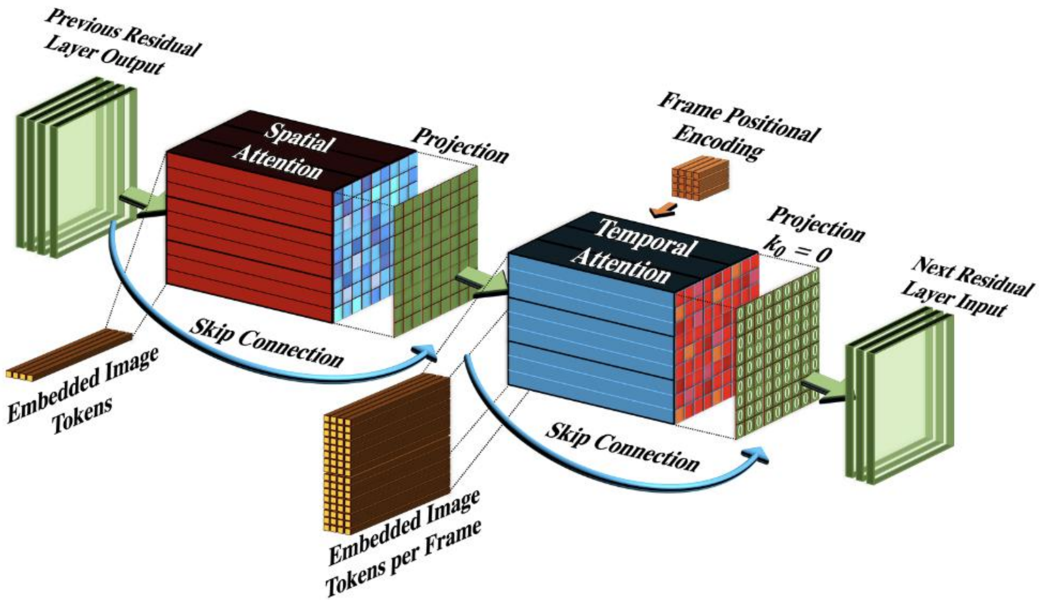


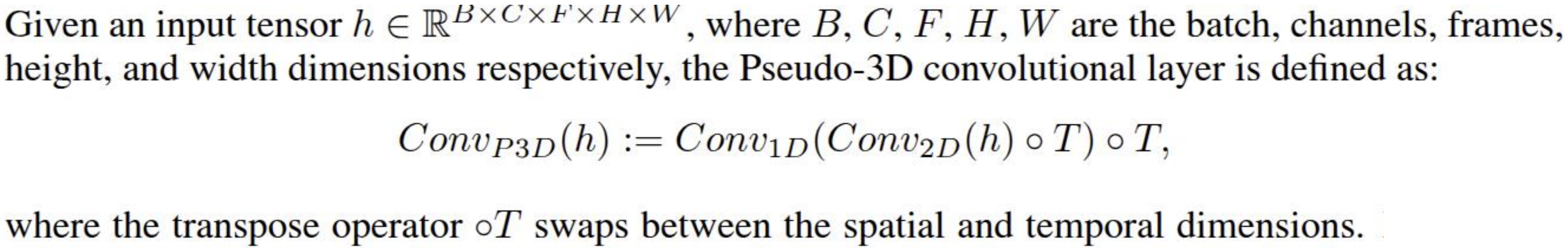
Figure 2: **Make-A-Video high-level architecture.** Given input text x translated by the prior P into an image embedding, and a desired frame rate fps, the decoder D^t generates 16 64 × 64 frames, which are then interpolated to a higher frame rate by $\uparrow_F$, and increased in resolution to 256 × 256 by SR_l^t and 768 × 768 by SR_h, resulting in a high-spatiotemporal-resolution generated video $\hat{y}$.

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

Make-a-Video (by Meta)







Given an input tensor $h \in \mathbb{R}^{B \times C \times F \times H \times W}$, where B, C, F, H, W are the batch, channels, frames, height, and width dimensions respectively, the Pseudo-3D convolutional layer is defined as:

$$Conv_{P3D}(h) := Conv_{1D}(Conv_{2D}(h) \circ T) \circ T,$$

where the transpose operator $\circ T$ swaps between the spatial and temporal dimensions.

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

Make-a-Video (by Meta)

$$ATTN_{P3D}(h) = unflatten(ATTN_{1D}(ATTN_{2D}(flatten(h)) \circ T) \circ T).$$

$$h' \in R^{B \times C \times F \times HW}$$

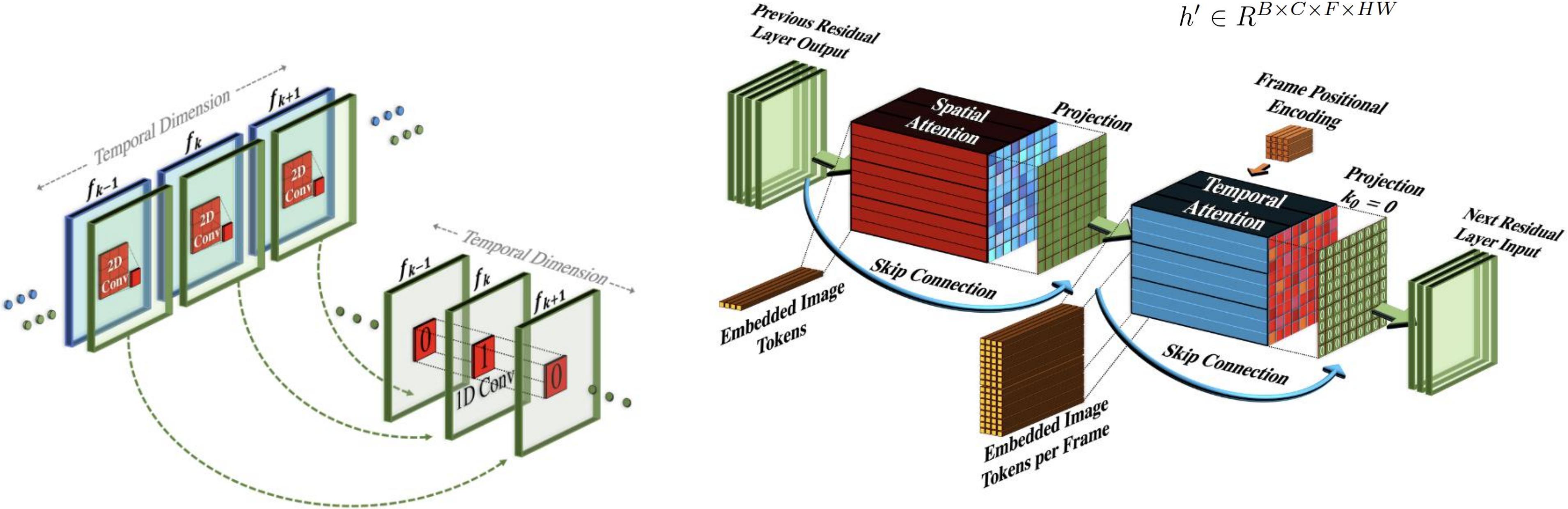
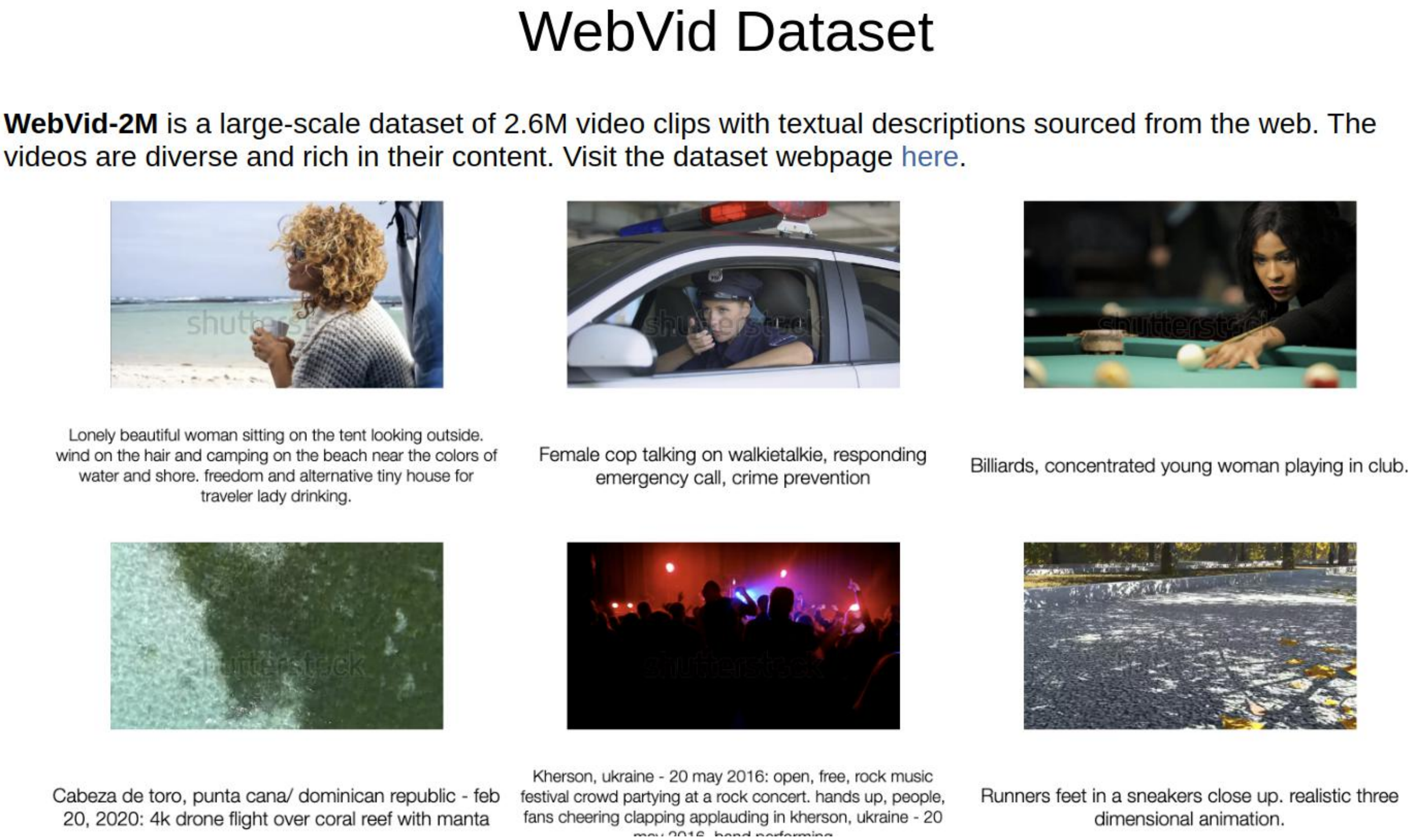


Figure 3: **The architecture and initialization scheme of the Pseudo-3D convolutional and attention layers, enabling the seamless transition of a pre-trained Text-to-Image model to the temporal dimension.** **(left)** Each spatial 2D conv layer is followed by a temporal 1D conv layer. The temporal conv layer is initialized with an identity function. **(right)** Temporal attention layers are applied following the spatial attention layers by initializing the temporal projection to zero, resulting in an identity function of the temporal attention blocks.

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

Make-a-Video (by Meta) - Training



WebVid Dataset

WebVid-2M is a large-scale dataset of 2.6M video clips with textual descriptions sourced from the web. The videos are diverse and rich in their content. Visit the dataset webpage here.

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.

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

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

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

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

Make-a-Video (by Meta) - Results

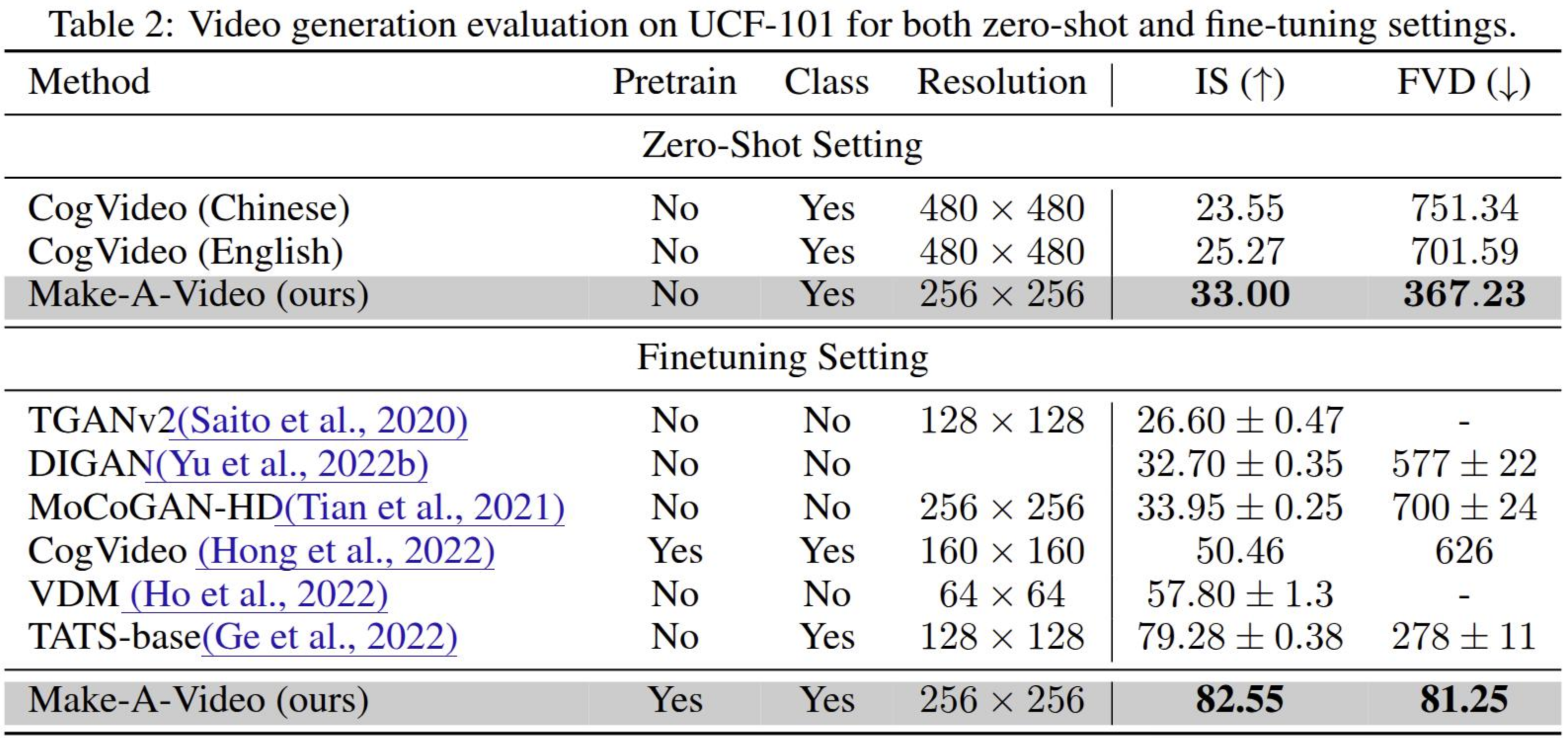


Table 2: Video generation evaluation on UCF-101 for both zero-shot and fine-tuning settings.

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**

https://makeavideo.studio/

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

Tune-a-Video

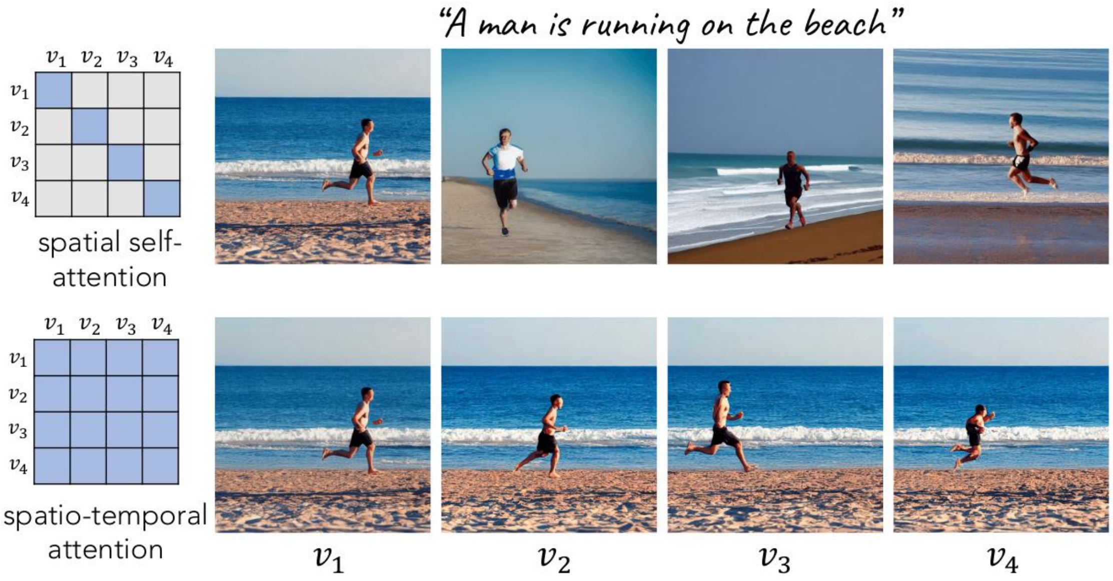
One-Shot Video Tuning,

- Only one text-video pair is presented

- Model is built on state-of-the-art T2I diffusion models pre-trained on massive image data.

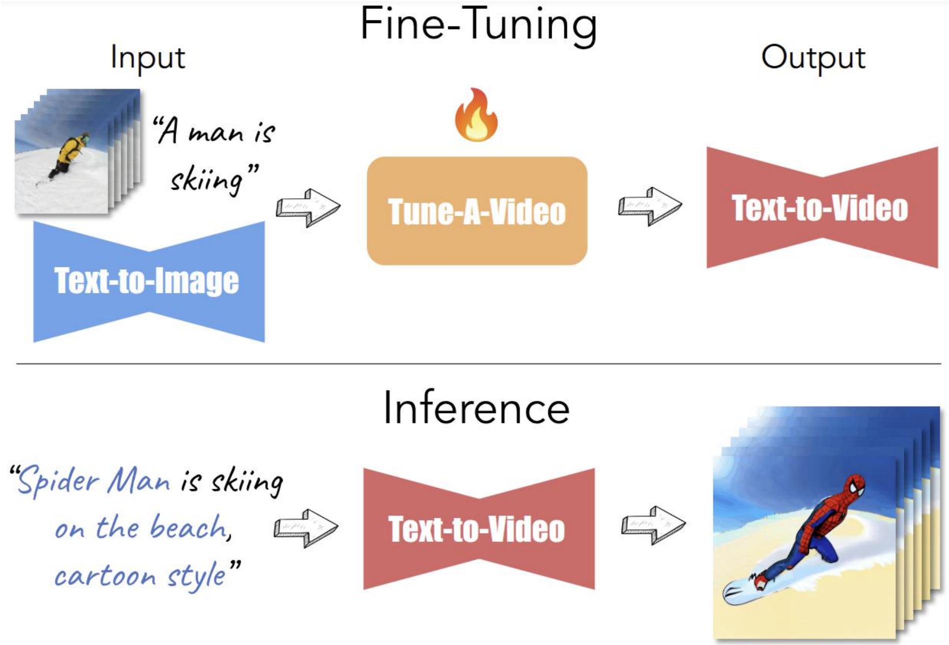
- Two key observations:
  - 1) T2I models can generate still images that represent verb terms
  - 2) extending T2I models to generate multiple images concurrently exhibits surprisingly good content consistency.

- To learn continuous motion, Tune-A-Video is introduced, which involves a tailored spatio-temporal attention mechanism and an efficient one-shot tuning strategy.



Wu, Jay Zhangjie, et al. "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation." in ICCV 2023.

Tune-a-Video



Given a captioned video,

- finetune a pre-trained T2I model (e.g.,Stable Diffusion) for T2V modeling.
- During inference, generate novel videos that represent the edits in text prompt while preserving the temporal consistency of input video.

Wu, Jay Zhangjie, et al. "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation." in ICCV 2023.

Tune-a-Video

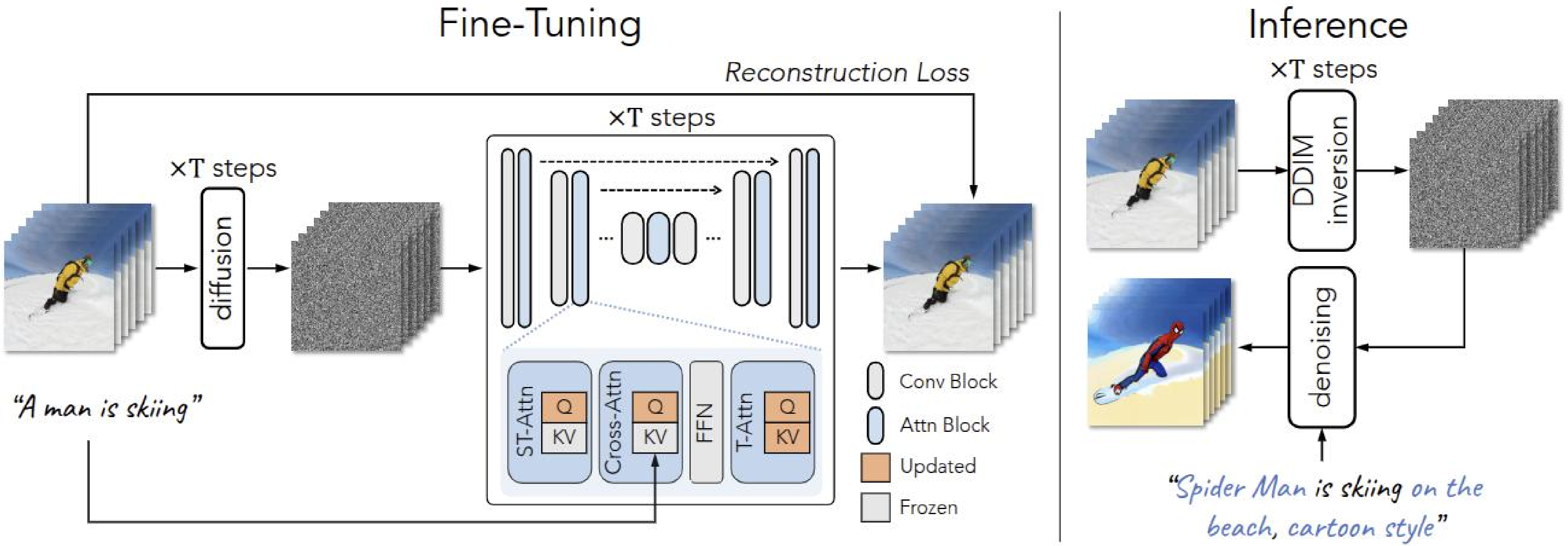


Figure 4: ***Pipeline of Tune-A-Video:*** Given a text-video pair (*e.g.*, "a man is skiing") as input, our method leverages the pretrained T2I diffusion models for T2V generation. During fine-tuning, we update the projection matrices in attention blocks using the standard diffusion training loss. During inference, we sample a novel video from the latent noise inverted from the input video, guided by an edited prompt (*e.g.*, "Spider Man is surfing on the beach, cartoon style").

Wu, Jay Zhangjie, et al. "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation." in ICCV 2023.

Tune-a-Video

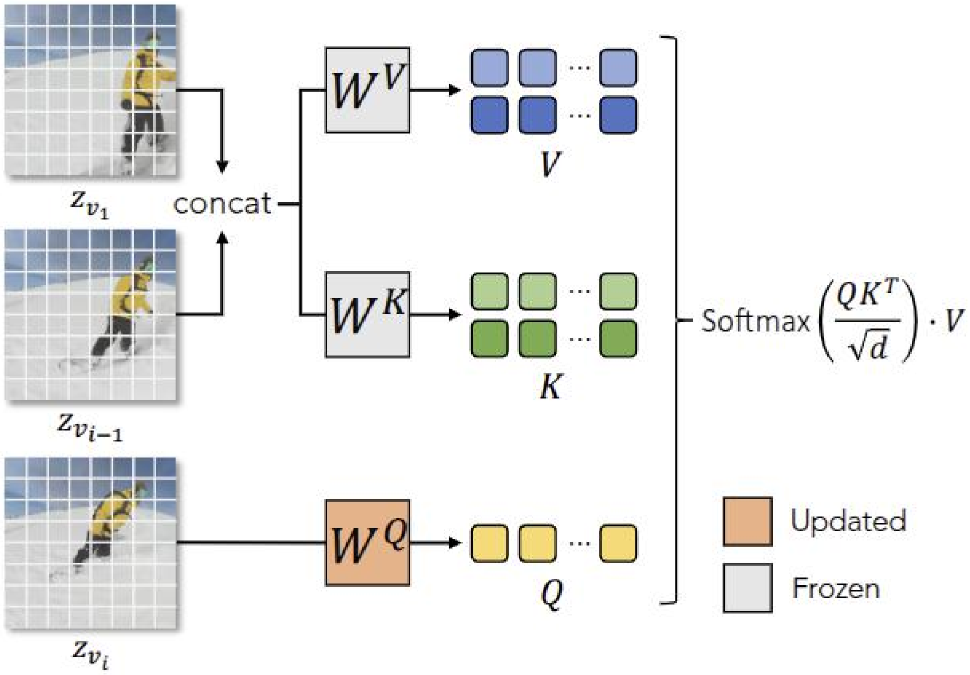
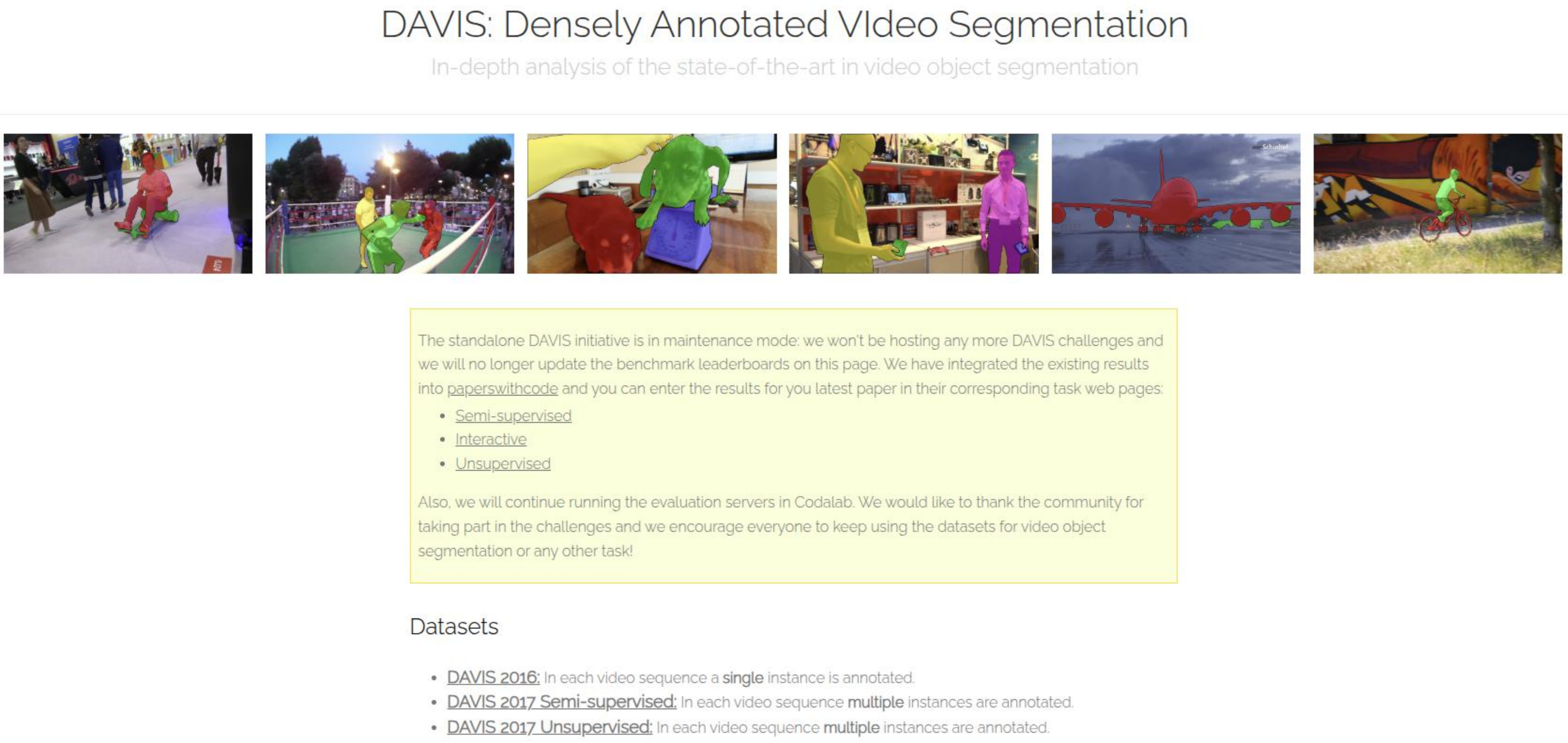


Figure 5: ***Illustration of our ST-Attn:*** Latent features of frame v_i, previous frames v_{i-1} and v_1 are projected to query Q, key K and value V. Output is a weighted sum of the values, weighted by the similarity between the query and key features. We highlight the updated parameter W^Q.

Wu, Jay Zhangjie, et al. "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation." in ICCV 2023.

Tune-a-Video - Training



DAVIS: Densely Annotated VIdeo Segmentation

In-depth analysis of the state-of-the-art in video object segmentation

The standalone DAVIS initiative is in maintenance mode: we won't be hosting any more DAVIS challenges and we will no longer update the benchmark leaderboards on this page. We have integrated the existing results into paperswithcode and you can enter the results for you latest paper in their corresponding task web pages:

- Semi-supervised
- Interactive
- Unsupervised

Also, we will continue running the evaluation servers in Codalab. We would like to thank the community for taking part in the challenges and we encourage everyone to keep using the datasets for video object segmentation or any other task!

Datasets

- DAVIS 2016: In each video sequence a **single** instance is annotated.
- DAVIS 2017 Semi-supervised: In each video sequence **multiple** instances are annotated.
- DAVIS 2017 Unsupervised: In each video sequence **multiple** instances are annotated.

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

Tune-a-Video - Results

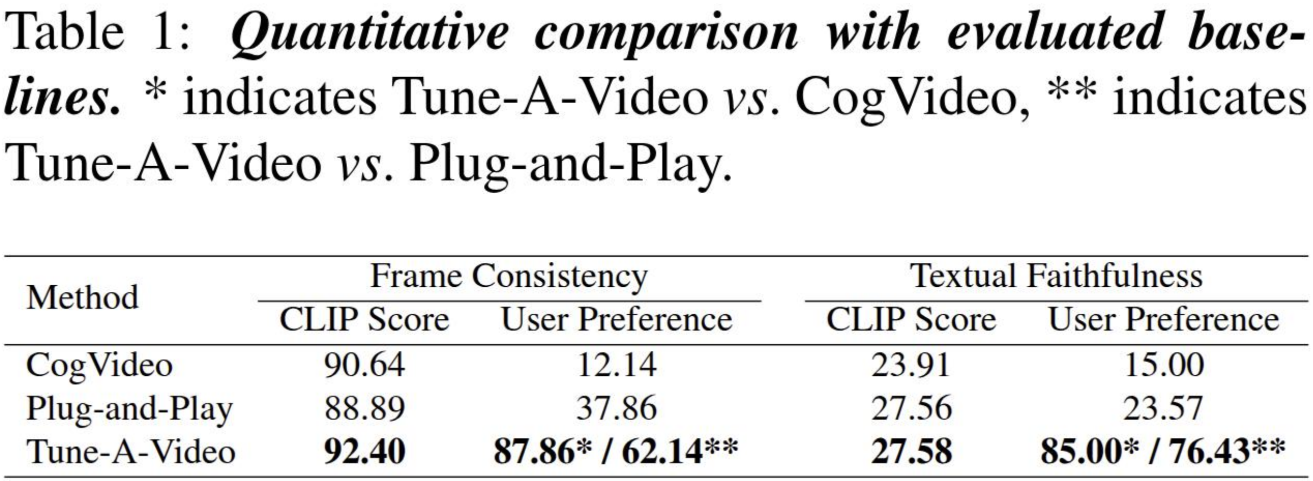


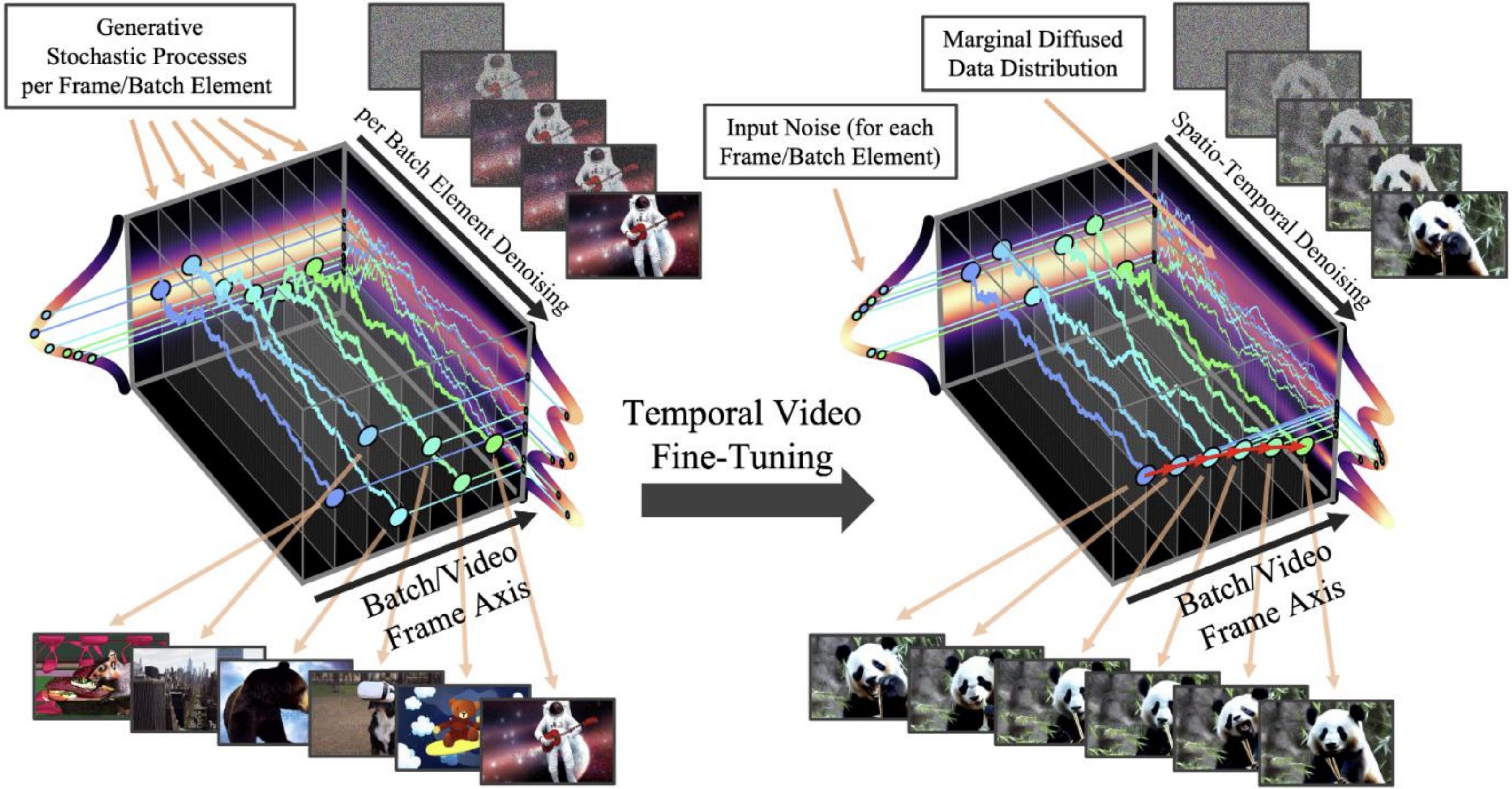
Table 1: ***Quantitative comparison with evaluated baselines.*** * indicates Tune-A-Video *vs*. CogVideo, ** indicates Tune-A-Video *vs*. Plug-and-Play.

Method	Frame Consistency		Textual Faithfulness	
	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****

https://tuneavideo.github.io/

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

Align your Latents


Generative Stochastic Processes per Frame/Batch Element
per Batch Element Denoising
Batch/Video Frame Axis
Temporal Video Fine-Tuning
Input Noise (for each Frame/Batch Element)
Marginal Diffused Data Distribution
Spatio-Temporal Denoising
Batch/Video Frame Axis

Before temporal video fine-tuning, different batch samples are independent.

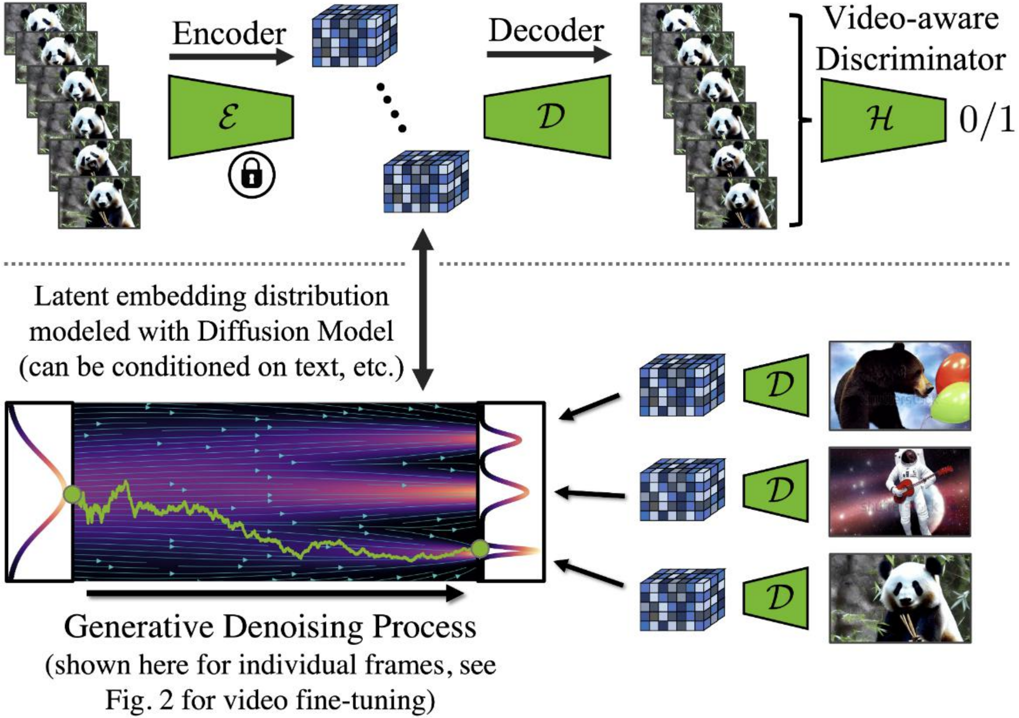
After temporal video fine-tuning, samples are aligned to form a video sequence (after applying the LDM decoder).

Blattmann, Andreas, et al. "Align your latents: High-resolution video synthesis with latent diffusion models." in CVPR 2023.

Align your Latents

Finetune temporal decoder :

- Process video sequences with a frozen per-frame encoder and enforce temporally coherent reconstructions across frames.
- Additionally a video-aware discriminator is also employed


Encoder
$\mathcal{E}$
Decoder
$\mathcal{D}$
Video-aware Discriminator
$\mathcal{H}$
0/1
Latent embedding distribution modeled with Diffusion Model (can be conditioned on text, etc.)
$\mathcal{D}$
$\mathcal{D}$
$\mathcal{D}$
Generative Denoising Process (shown here for individual frames, see Fig. 2 for video fine-tuning)

LDMs, a diffusion model is trained in latent space. It synthesizes latent features, which are then transformed through the decoder into images.

Blattmann, Andreas, et al. "Align your latents: High-resolution video synthesis with latent diffusion models." in CVPR 2023.

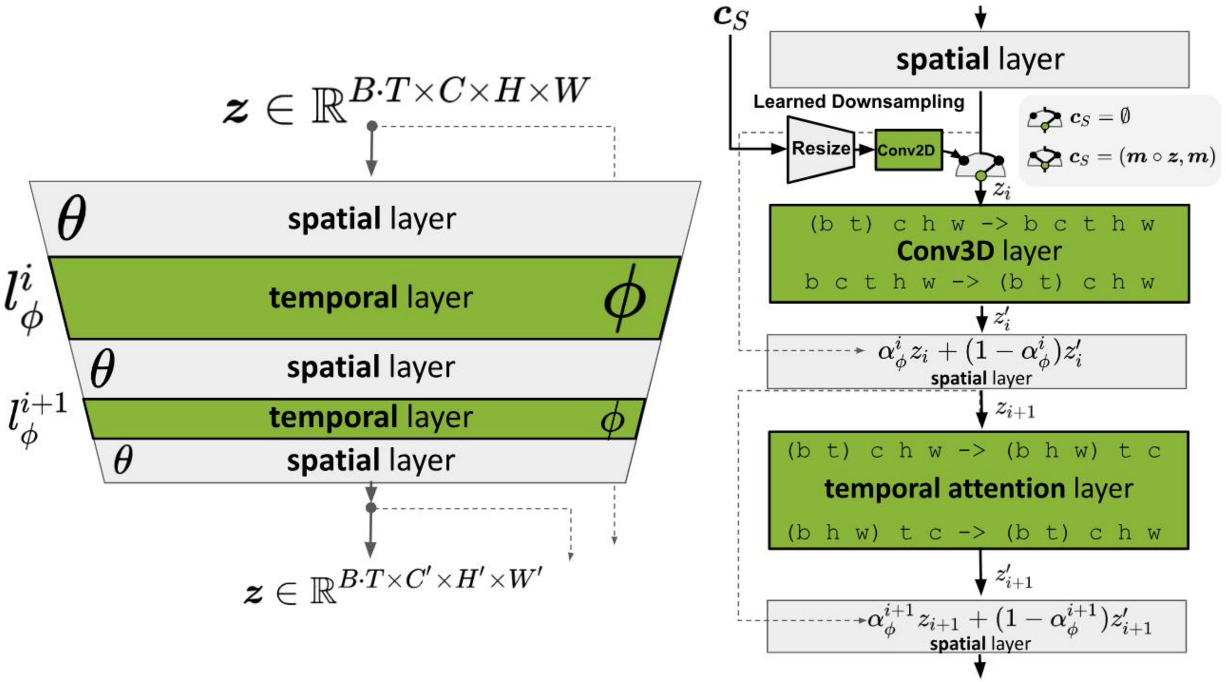
Align your Latents

Turn a pre-trained LDM into a video generator by inserting temporal layers that learn to align frames into temporally consistent sequences.

- During optimization, the image backbone θ remains fixed and only the parameters ϕ of the temporal layers ϕ are trained

- During training, the basevmodel θ interprets the input sequence of length T as a batch of images.

- For the temporal layers ϕ , these batches are reshaped into video format. Their output z′ is combined with the spatial output z, using a learned merge parameter α.



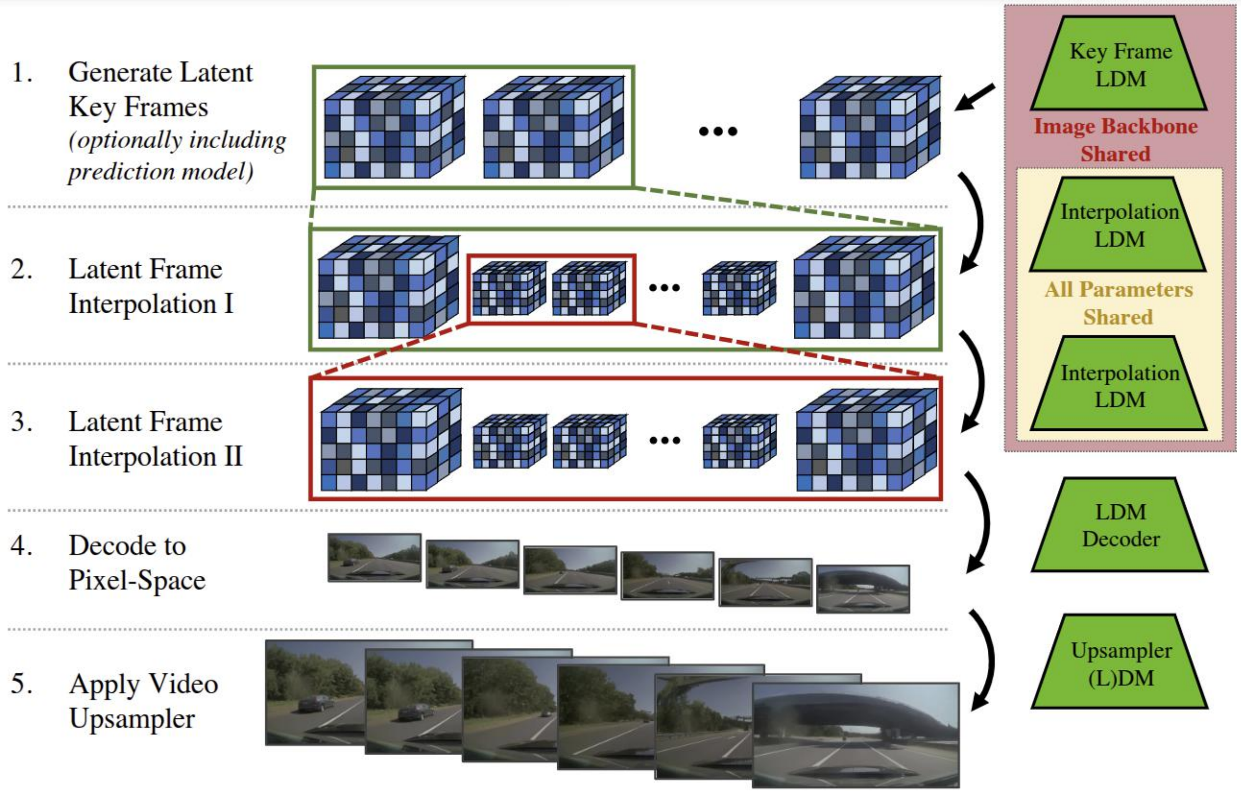
- During inference, skipping the temporal layers yields the original image model.

- For illustration purposes, only a single UNet Block is shown. B denotes batch size, T sequence length, C input channels and H and W the spatial dimensions of the input.

Blattmann, Andreas, et al. "Align your latents: High-resolution video synthesis with latent diffusion models." in CVPR 2023.

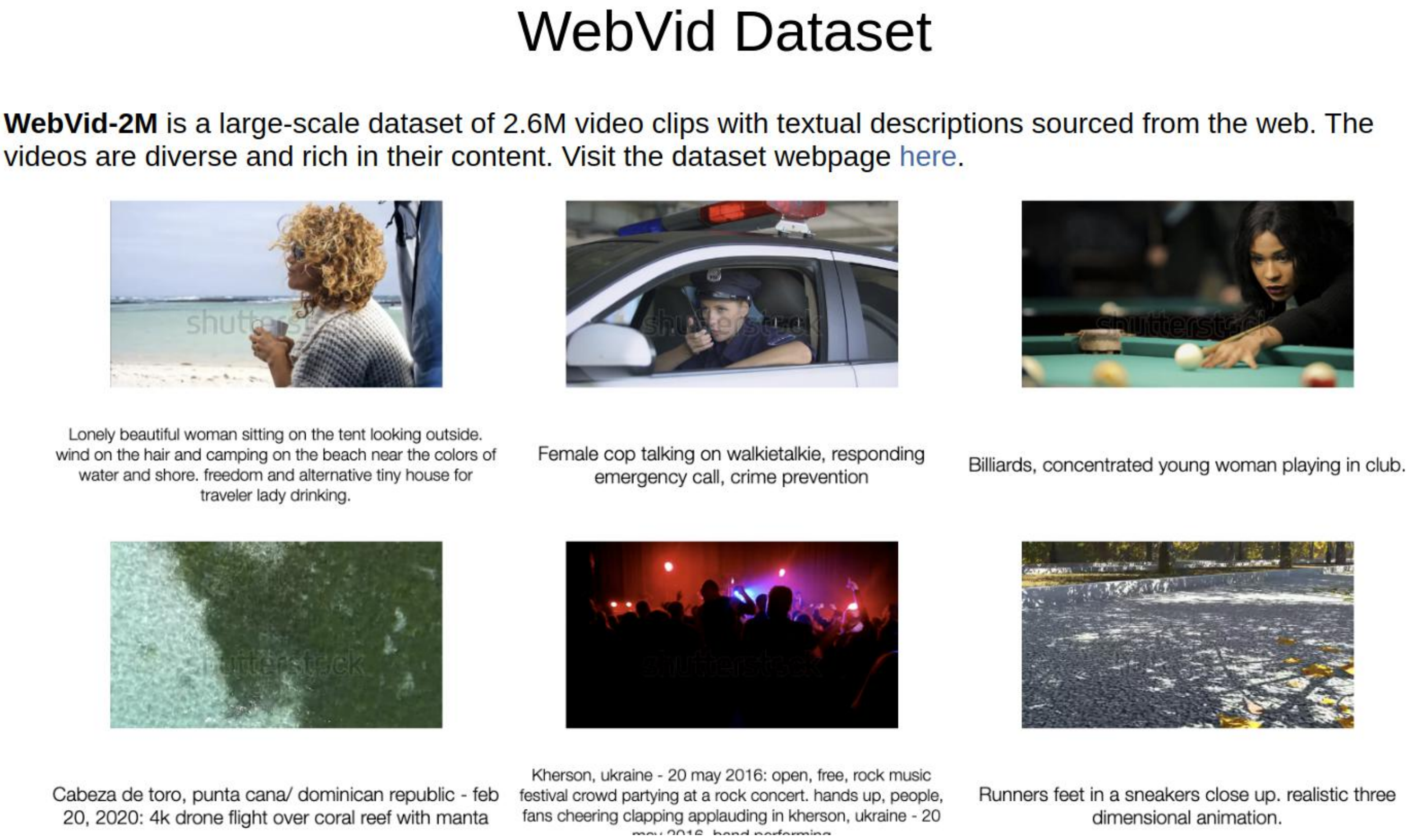
Align your Latents

- initially generates sparse keyframes at low frame rates,
- temporally upsample them twice by another interpolation latent diffusion model.



Blattmann, Andreas, et al. "Align your latents: High-resolution video synthesis with latent diffusion models." in CVPR 2023.

Align your Latents - Training



WebVid Dataset

WebVid-2M is a large-scale dataset of 2.6M video clips with textual descriptions sourced from the web. The videos are diverse and rich in their content. Visit the dataset webpage here.

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.

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

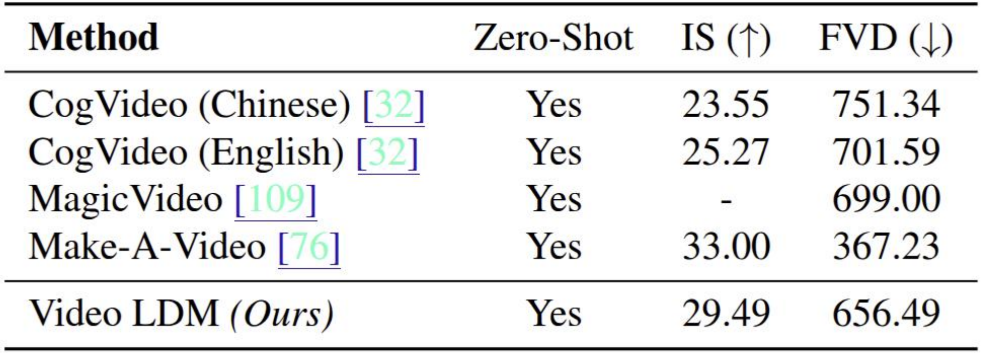
Kherson, ukraine - 20 may 2016: open, free, rock music festival crowd partying at a rock concert. hands up, people, fans cheering clapping applauding in kherson, ukraine - 20 may 2016. band performing

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

https://github.com/m-bain/webvid

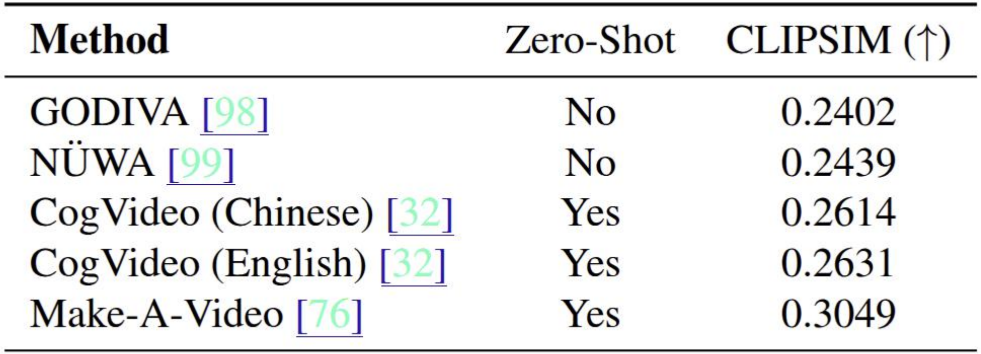
Align your Latents - Results

Table 4. UCF-101 text-to-video generation.



Method	Zero-Shot	IS (↑)	FVD (↓)
CogVideo (Chinese) [32]	Yes	23.55	751.34
CogVideo (English) [32]	Yes	25.27	701.59
MagicVideo [109]	Yes	-	699.00
Make-A-Video [76]	Yes	33.00	367.23
Video LDM *(Ours)*	Yes	29.49	656.49

Table 5. MSR-VTT text-to-video generation performance.



Method	Zero-Shot	CLIPSIM (↑)
GODIVA [98]	No	0.2402
NÜWA [99]	No	0.2439
CogVideo (Chinese) [32]	Yes	0.2614
CogVideo (English) [32]	Yes	0.2631
Make-A-Video [76]	Yes	0.3049

https://research.nvidia.com/labs/toronto-ai/VideoLDM/samples.html

Blattmann, Andreas, et al. "Align your latents: High-resolution video synthesis with latent diffusion models." in CVPR 2023.