CS294-158 Deep Unsupervised Learning

Lecture 9 Video Generation

Pieter Abbeel, Wilson Yan, Kevin Frans, Philipp Wu

Outline

- **Basics**
- Improving Video Generation
- Applications
  - Video Generation Models as Physical Simulators
  - Video Editing

Video as a Modality

Videos encoded using standard codecs: H264, H265, AV1, etc.

1 minute of encoded HD (1080p) 24FPS video is ~20MB

pixels fps sec rgb

Raw data size (RGB uint8): 1920 x 1080 x 60 x 60 x 3 = **~8GB**

After normalization (uint8 {0, …, 255} -> float32 [-1, 1]) = **~32GB**

Video Codecs

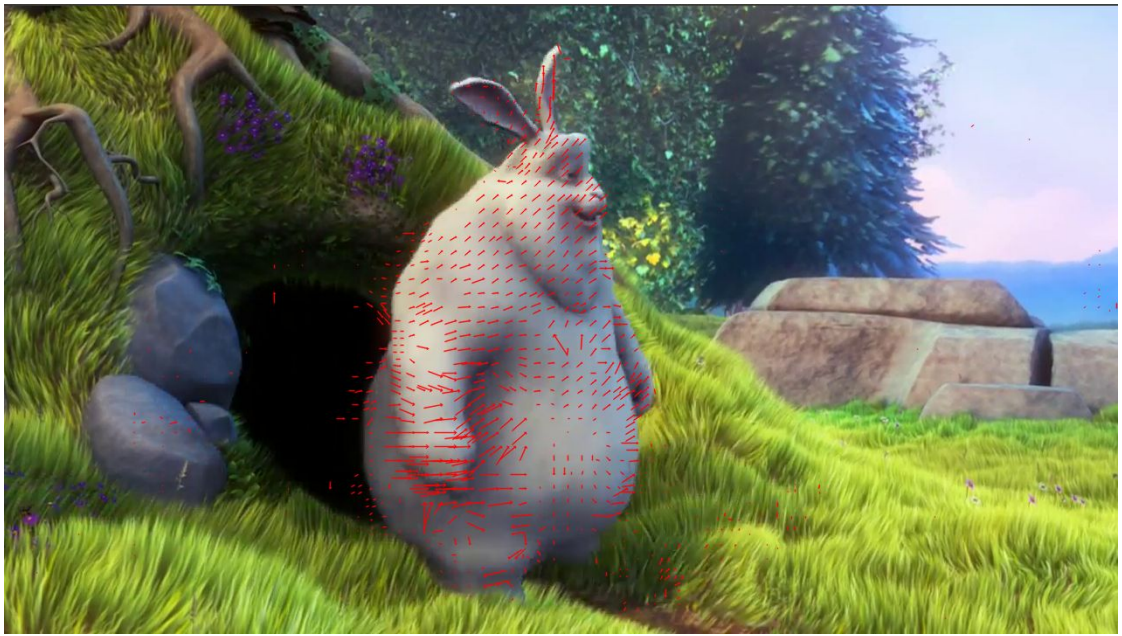
How do some of the codecs compress?

Each frame of the video is categorized as:

- *I-Frame*: encode the full frame (key frame)
- *P(redicted)-Frame*: define the content as relative to another prior P-Frame or I-Frame
- B(idirectional)-Frame: define the context

Video Codecs

Define relative encoding as *motion vectors* that describe translational relationships for 16 x 16 macroblocks between frames



Video Codecs

I-Frames take advantage of *spatial* redundancy to compress

P/B-Frames take advantage of *temporal* redundancy to compress

Can we take inspiration from this to build more efficient video models?

Factorization as a Cascade

Break up the problem into smaller, independent pieces consisting of spatial and temporal superresolution models

Papers:

- ImagenVideo (Sep 2022)
- Make-a-Video (Sep 2022)
- PYoCo (May 2023)
- Lumiere (Jan 2024)

ImagenVideo

14M text-video pairs, 500M text-image pairs

Ho, Jonathan, et al. "Imagen video: High definition video generation with diffusion models." *arXiv preprint arXiv:2210.02303* (2022).

ImagenVideo

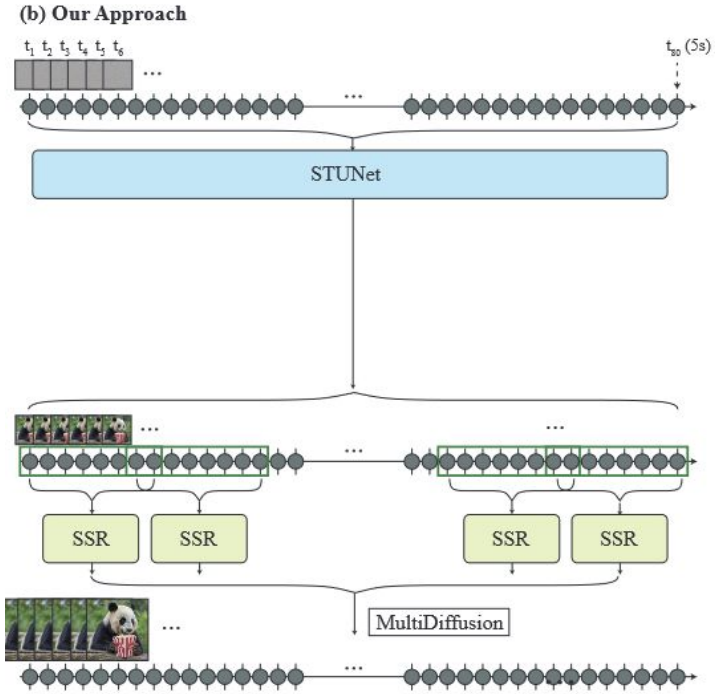
Lumiere

Base model: 80 x 128 x 128 (16fps)

SSR: 128 x 128 -> 1024 x 1024

30M text-video pairs

Initialized from text-image model



Bar-Tal, Omer, et al. "Lumiere: A space-time diffusion model for video generation." *arXiv preprint arXiv:2401.12945* (2024).

Lumiere

Per-Frame Latent Space Models

We can take a space-time factorized approach

- Learn a per-frame autoencoder (*spatial compression*)
  - VQGAN (for AR), VAE (for diffusion)
- Learn a base video model on **key frames** of the video (*temporal compression*)
- Learn frame interpolation model(s) to upsample FPS

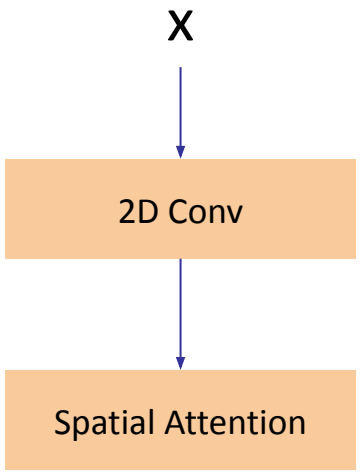
Per-Frame Latent Space Models

Papers

- Align Your Latents (Apr 2023)
- Emu Video (Nov 2023)
- Stable Video Diffusion (Nov 2023)

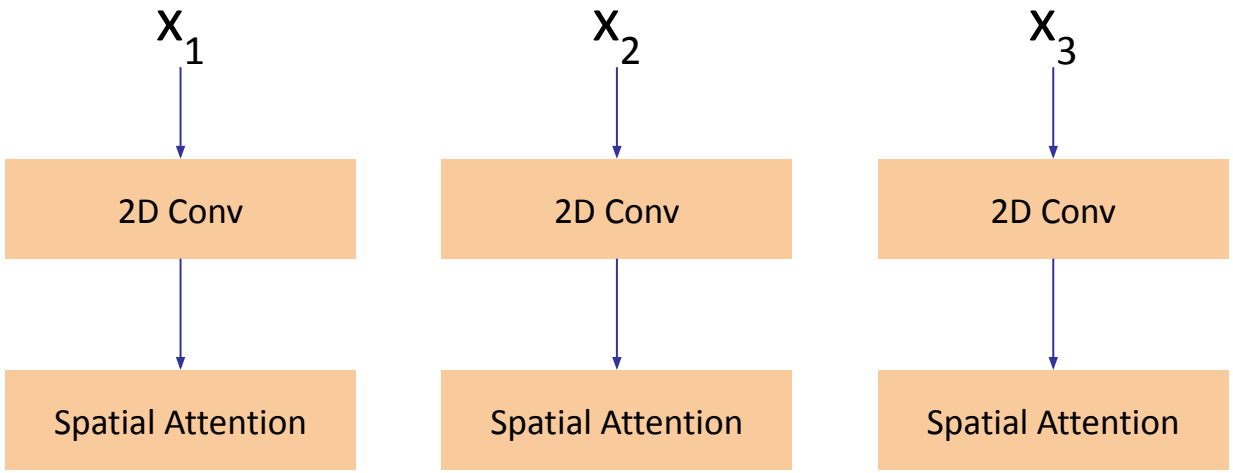
Leveraging Text-Image Pretraining

Key idea: Initialize from a pretrained text-image model



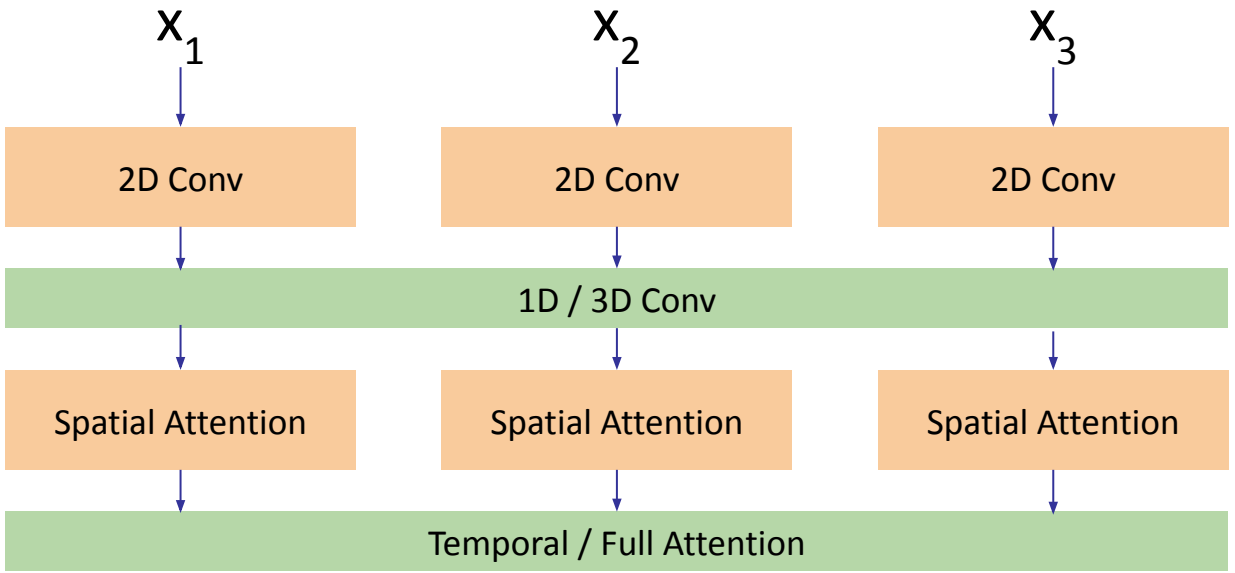
Leveraging Text-Image Pretraining

Key idea: Initialize from a pretrained text-image model



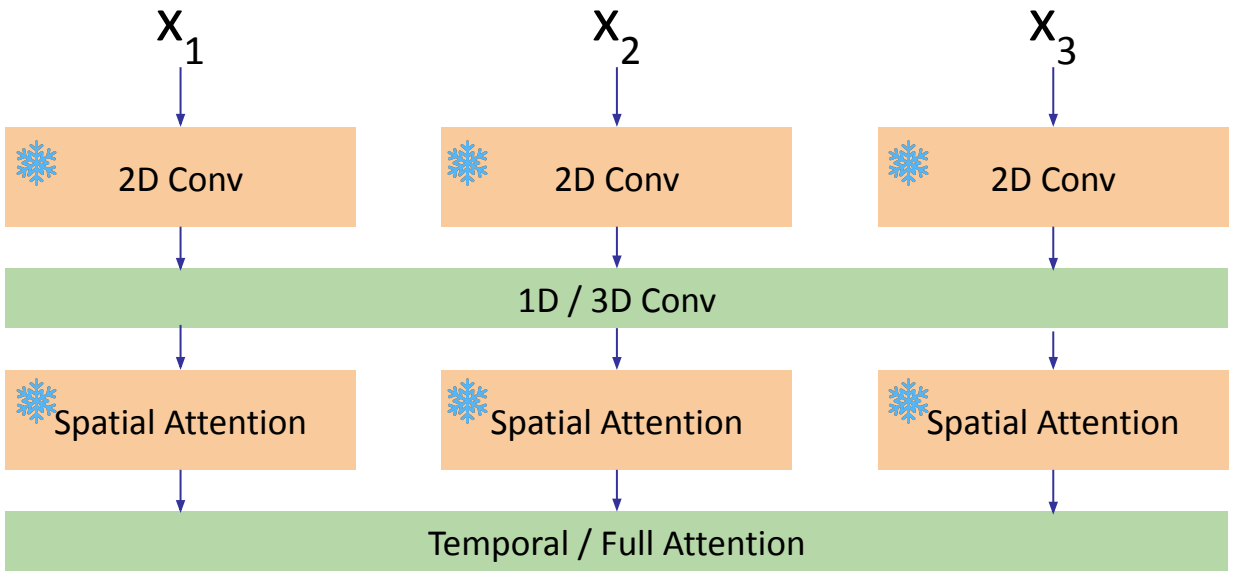
Leveraging Text-Image Pretraining

- **Key idea: Insert temporal operations**



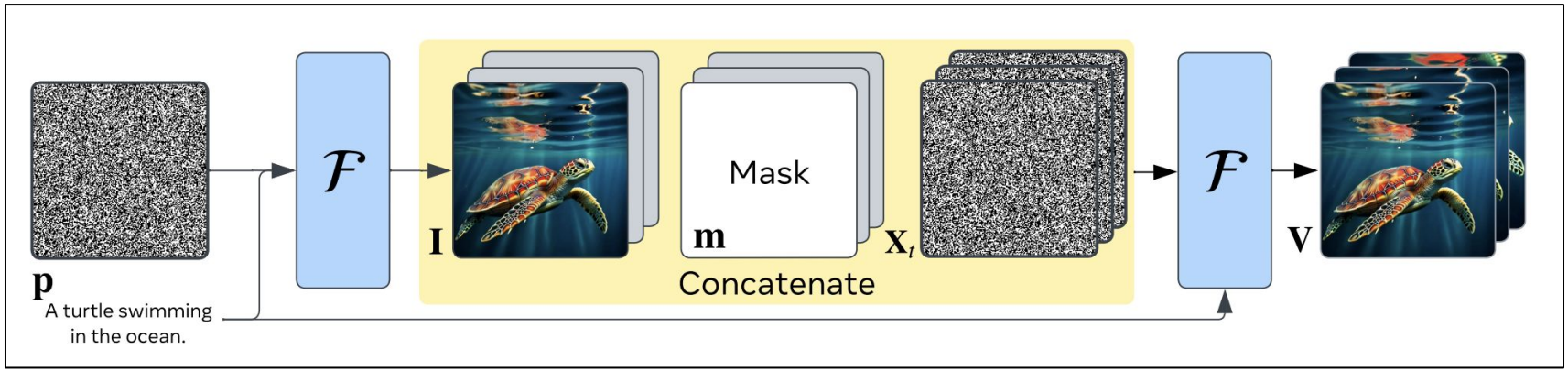
Leveraging Text-Image Pretraining

- **Key idea: Optionally freeze spatial parameters**



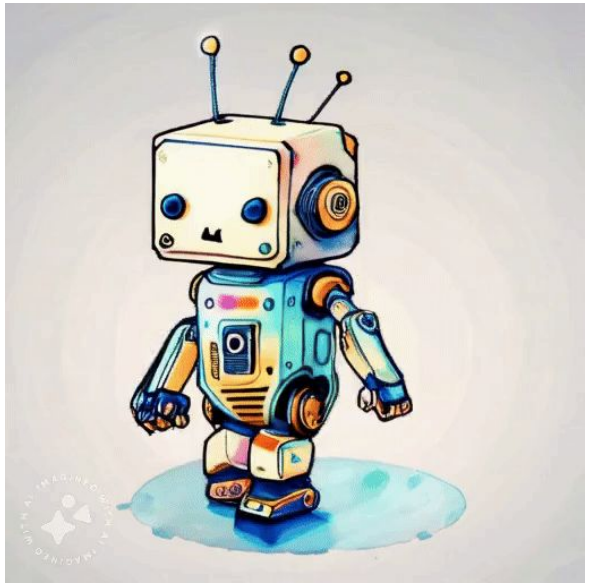
Emu Video

Generate image -> Generate rest of video (4 FPS) -> temporal upsampling to 16 FPS



Blattmann, Andreas, et al. "Align your latents: High-resolution video synthesis with latent diffusion models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

Emu Video



Spatio-Temporal Latent Spaces

Can we better temporally downsample our data?

Just learn a 3D autoencoder!

3D Autoencoders

Downsample over **time and space**

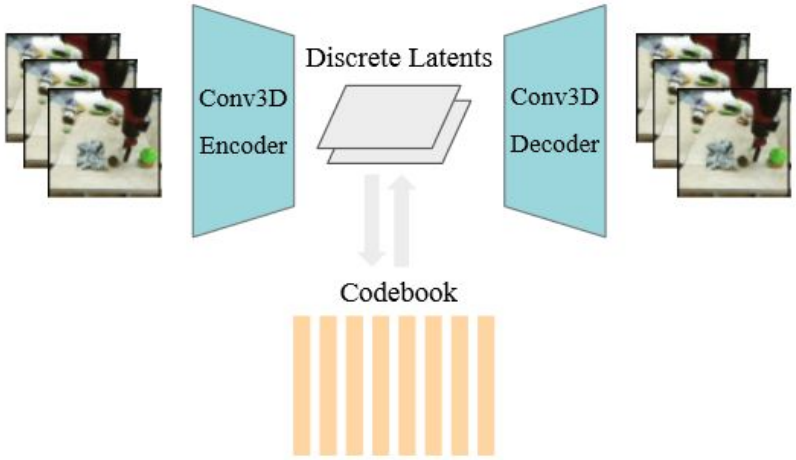
- E.g. 16 x 256 x 256 -> 4 x 16 x 16

VideoGPT (2021) / TATS (2022)

- 3D CNN VQ-VAE/VQ-GAN, learn an AR prior

LVDM (2022)

- 3D CNN VAE, learn a diffusion prior



3D Autoencoders

Prior video generation works have found large benefits from jointly training on images + video when starting from scratch.

- More text-image data, have better coverage of text-vision concepts that can be carried on to text-video

If we downsample temporally, how can we encode images?

3D Autoencoders

Treat the first frame differently

- E.g. 17 x 256 x 256 -> (1 + 4) x 16 x 16

Phenaki (Sept 2022)

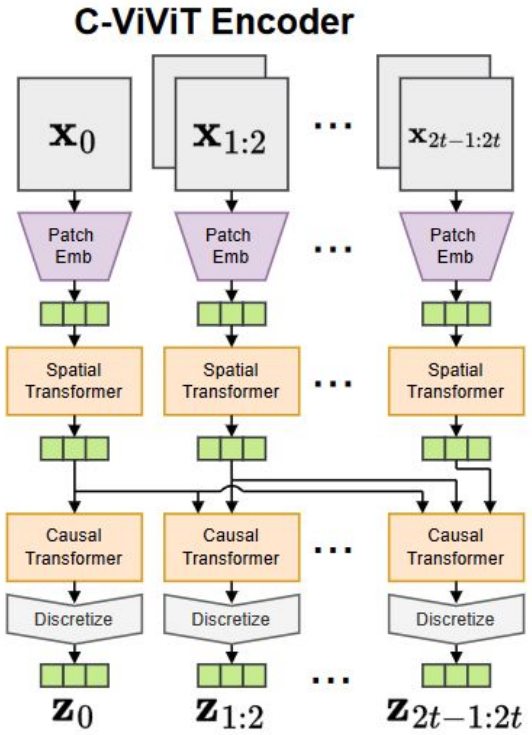
- 3D ViT-VQ, learn a MaskGit prior

MAGVIT-v2 (Oct 2023) / VideoPoet (Dec 2023)

- Causal 3D CNN LFQ, learn a MaskGit prior / AR prior

WALT (Dec 2023)

- Causal 3D CNN VAE, learn a diffusion prior



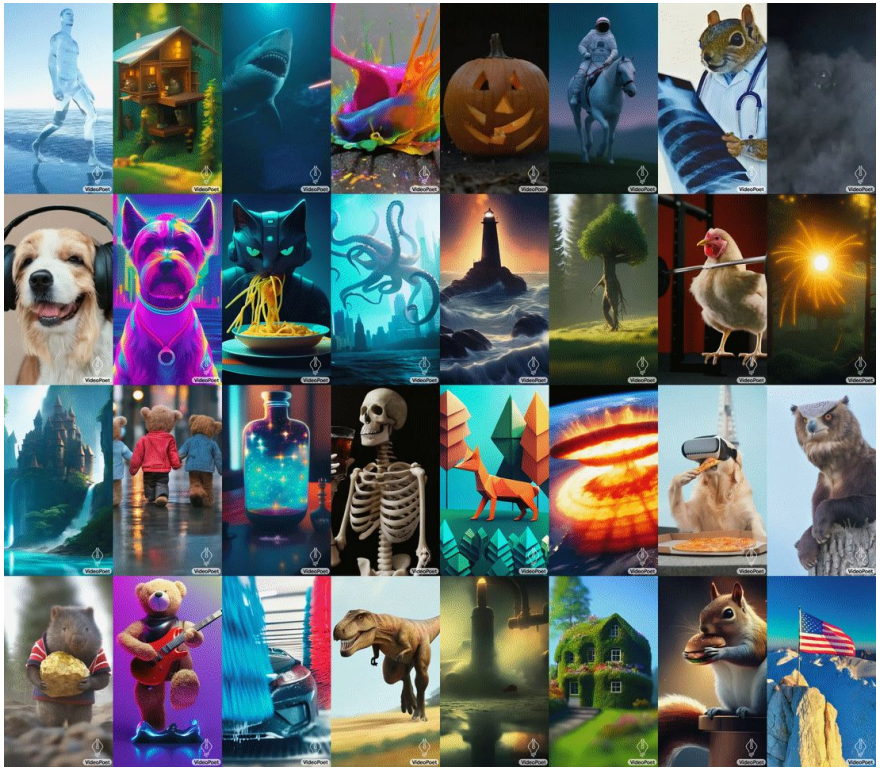
Phenaki

Villegas, Ruben, et al. "Phenaki: Variable length video generation from open domain textual descriptions." *International Conference on Learning Representations*. 2022.

WALT

Gupta, Agrim, et al. "Photorealistic video generation with diffusion models." *arXiv preprint arXiv:2312.06662* (2023).

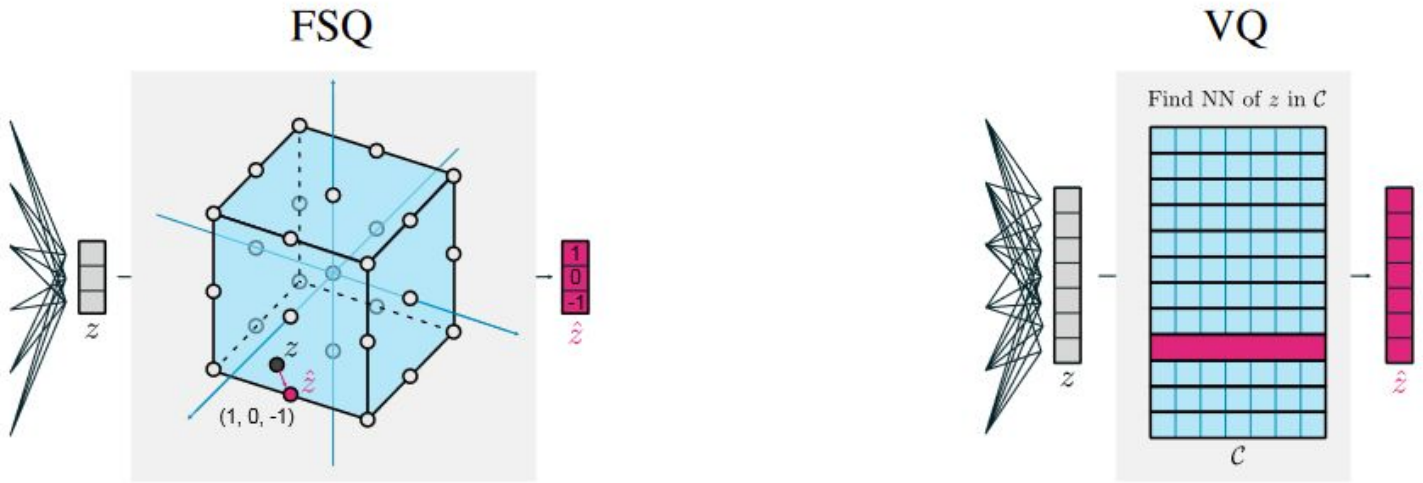
VideoPoet



Kondratyuk, Dan, et al. "Videopoet: A large language model for zero-shot video generation." *arXiv preprint arXiv:2312.14125* (2023).

LFQ / FSQ Tokenizer

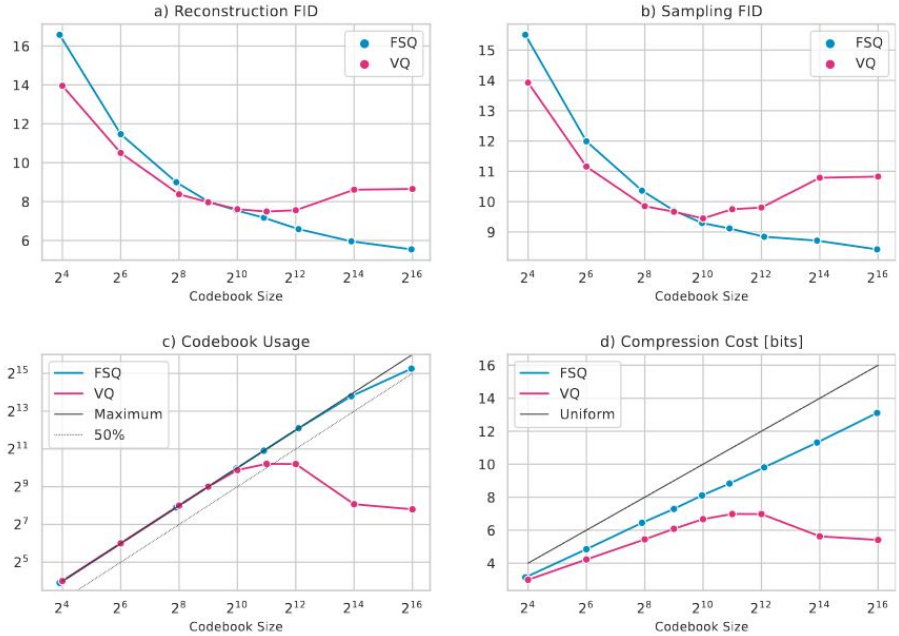
No more VQ - just round your representations for quantization!



Yu, Lijun, et al. "Language Model Beats Diffusion--Tokenizer is Key to Visual Generation." *arXiv preprint arXiv:2310.05737* (2023).
Mentzer, Fabian, et al. "Finite scalar quantization: Vq-vae made simple." *arXiv preprint arXiv:2309.15505* (2023).

LFQ / FSQ Tokenizer

LFQ / FSQ methods allow much larger codebook sizes



Outline

- Basics
- **Improving Video Generation**
- Applications
  - Video Generation Models as Physical Simulators
  - Video Editing

How to Improve Video Models?

Three Key Axes

- Scale - larger models
- Representation - more compressed latent spaces
- Data - better data

How to Improve Video Models?

Three Key Axes

- **Scale - larger models**
- **Representation - more compressed latent spaces**
- Data - better data

How to Improve Video Models?

Three Key Axes

- Scale - larger models
- Representation - more compressed latent spaces
- **Data - better data**

Data

Extremely impactful on resulting video generation quality

- The data is essentially the model

Data Filtering

What kind of videos / data do we want?

- **Good motion** - filter out static videos (e.g. optical flow score)
- **Good text-video alignment** - filter based on CLIP score
- **Good quality** - filter based on aesthetic score, resolution, other metadata (likes, views, etc.)

Synthetic Data

Good text-video data is comparatively harder to find on the web

Solution: *Synthetically annotate the data using a VLM*

- Using off-the-shelf labelers (CoCa, LLaVA, ShareCaptioner, Video-LLaVA, etc.)
- Collect high-quality video captions, and finetune a VLM to caption data

Synthetic Data

Image			
Alt Text	now at victorian plumbing.co.uk	is he finished...just about!	23 (19 of 30) 1200
SSC	a white modern bathtub sits on a wooden floor.	a quilt with an iron on it.	a jar of rhubarb liqueur sitting on a pebble background.
DSC	this luxurious bathroom features a modern freestanding bathtub in a crisp white finish. the tub sits against a wooden accent wall with glass-like panels, creating a serene and relaxing ambiance. three pendant light fixtures hang above the tub, adding a touch of sophistication. a large window with a wooden panel provides natural light, while a potted plant adds a touch of greenery. the freestanding bathtub stands out as a statement piece in this contemporary bathroom.	a quilt is laid out on a ironing board with an iron resting on top. the quilt has a patchwork design with pastel-colored strips of fabric and floral patterns. the iron is turned on and the tip is resting on top of one of the strips. the quilt appears to be in the process of being pressed, as the steam from the iron is visible on the surface. the quilt has a vintage feel and the colors are yellow, blue, and white, giving it an antique look.	rhubarb pieces in a glass jar, waiting to be pickled. the colors of the rhubarb range from bright red to pale green, creating a beautiful contrast. the jar is sitting on a gravel background, giving a rustic feel to the image.

Synthetic Data

Stable Video Diffusion used a purely synthetically labelled dataset of ~150M text-video pairs

Blattmann, Andreas, et al. "Stable video diffusion: Scaling latent video diffusion models to large datasets." *arXiv preprint arXiv:2311.15127* (2023).

Finetuning

Further finetuning the model can also dramatically produce results

- Use a small set of *extremely* high quality video data

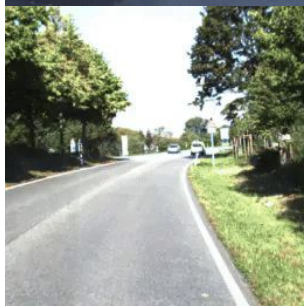
Outline

- Basics
- Improving Video Generation
- **Applications**
  - **Video Generation Models as Physical Simulators**
  - Video Editing

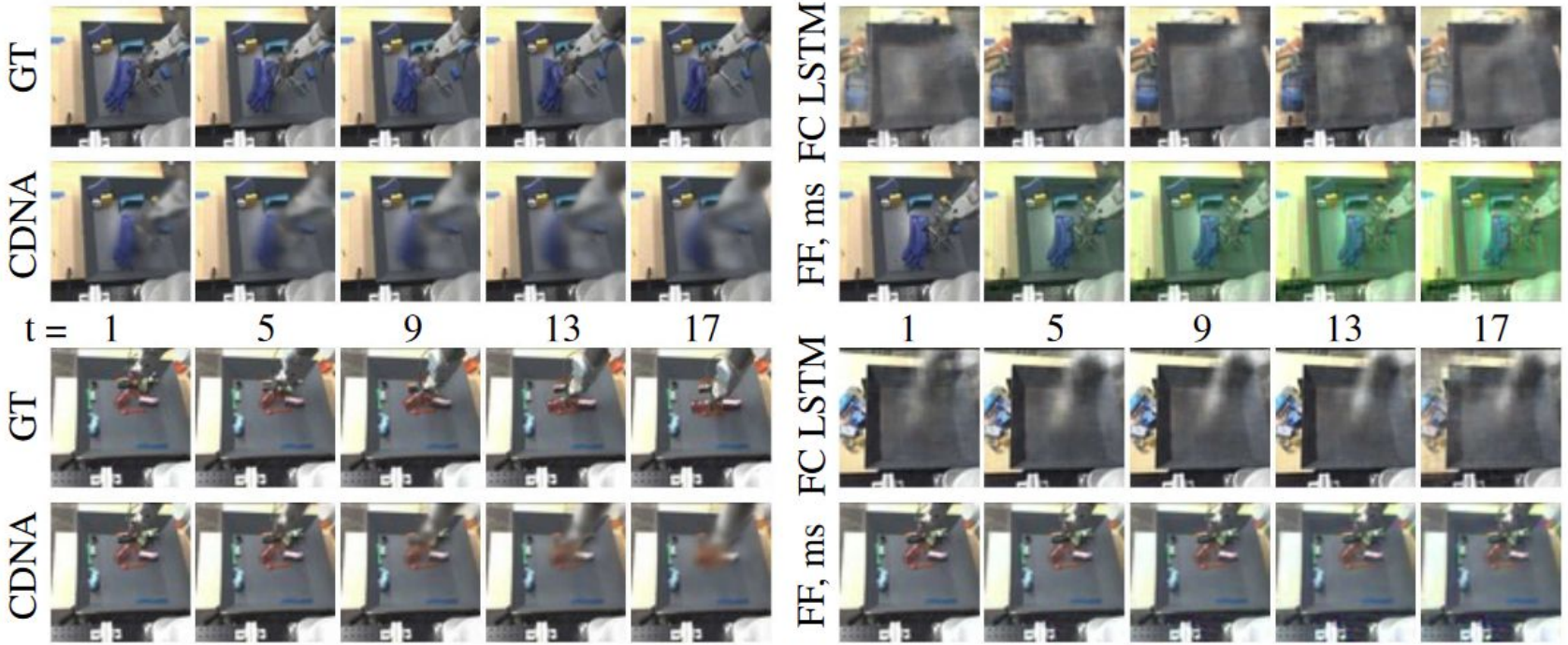
Video Generation as Physical Simulators

A large amount of initial work in video prediction was motivated to simulate the physical (or digital) world

- Robotics
- Self-driving
- Games

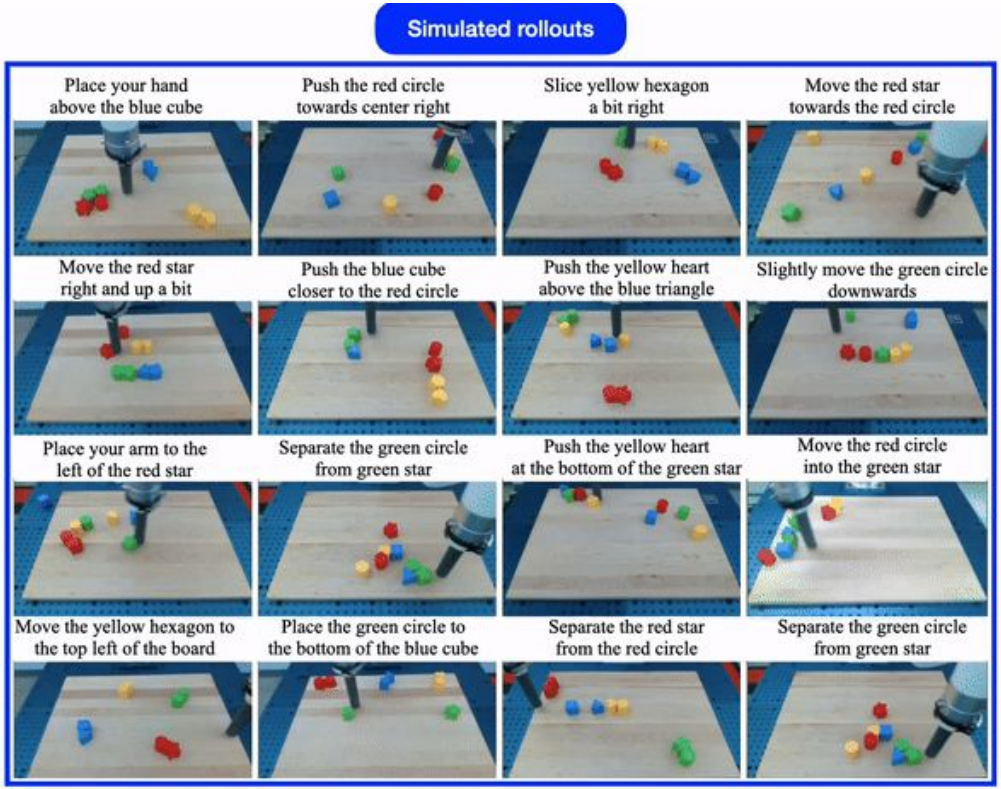


Video Generation as Physical Simulators



Finn, Chelsea, Ian Goodfellow, and Sergey Levine. "Unsupervised learning for physical interaction through video prediction." *Advances in neural information processing systems* 29 (2016).

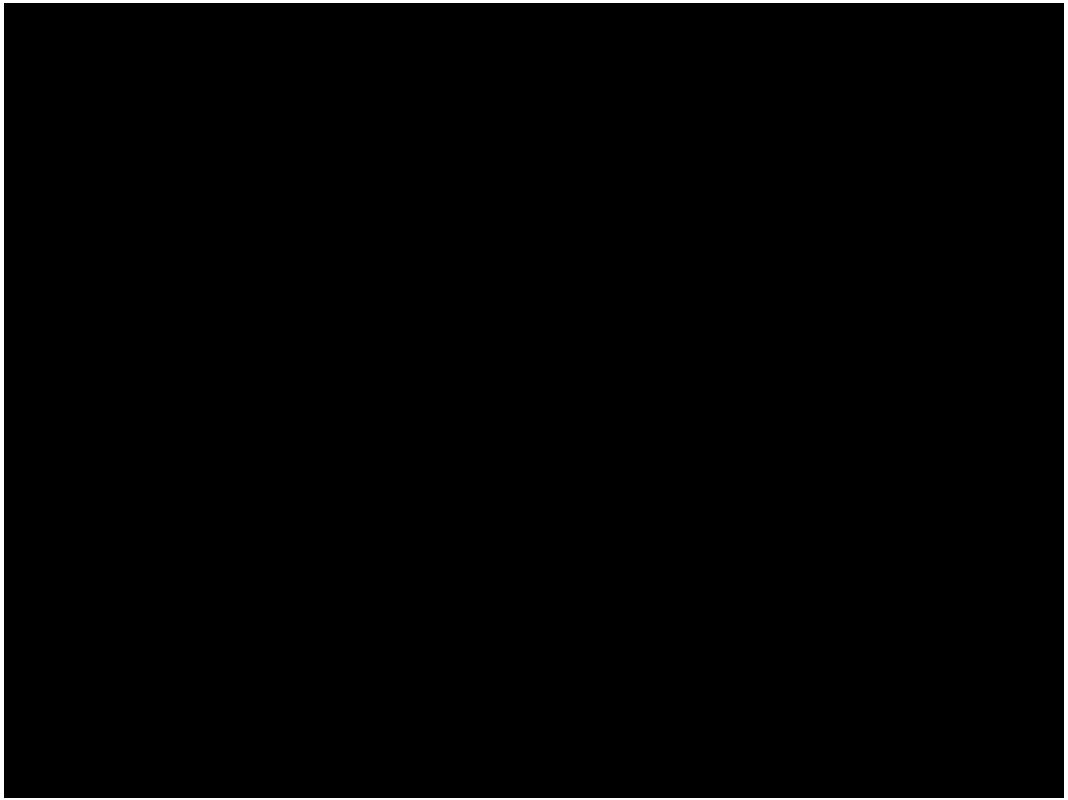
UniSim: Learning Interactive Real-World Simulators



Yang, Mengjiao, et al. "Learning interactive real-world simulators." *arXiv preprint arXiv:2310.06114* (2023).

UniSim: Learning Interactive Real-World Simulators

GAIA-1 (Wayve)

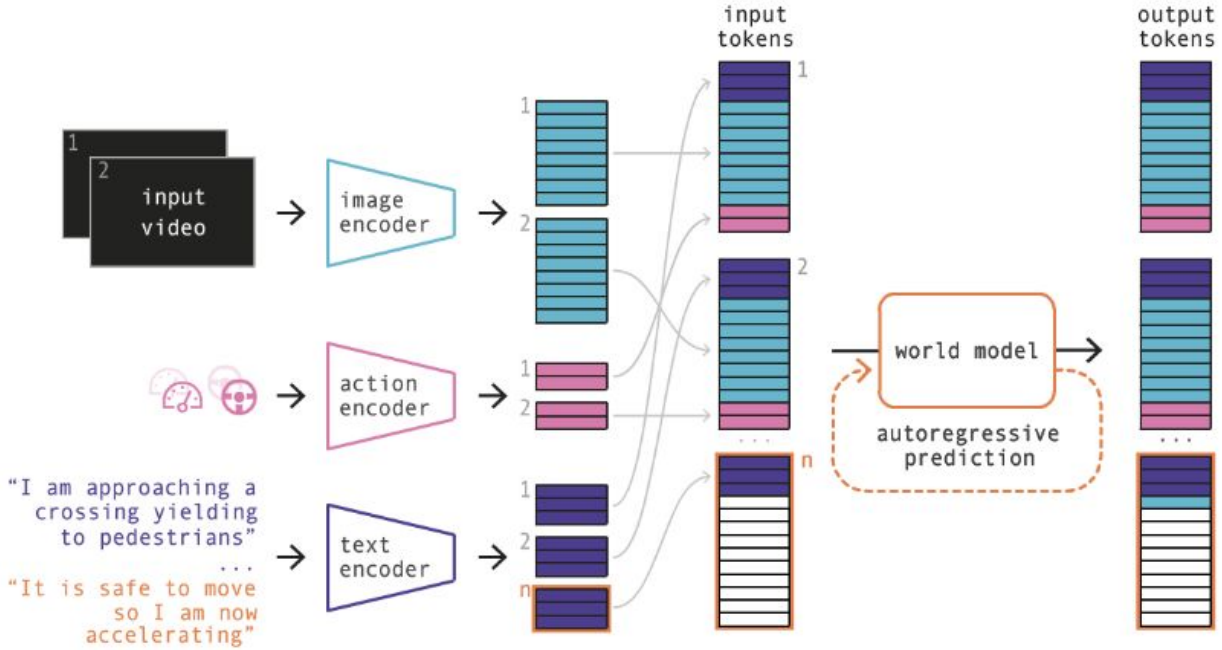


Hu, Anthony, et al. "Gaia-1: A generative world model for autonomous driving." *arXiv preprint arXiv:2309.17080* (2023).

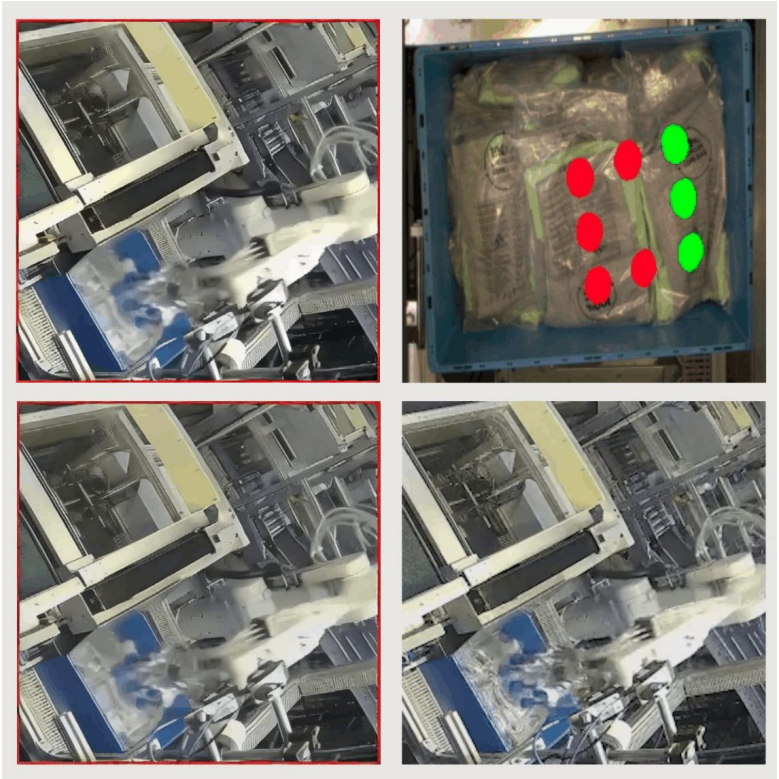
GAIA-1 (Wayve)


WAYVE
Go Left
Go Right
Go Left (strong)
Go Right (strong)
GIFRUN.COM

GAIA-1 (Wayve)

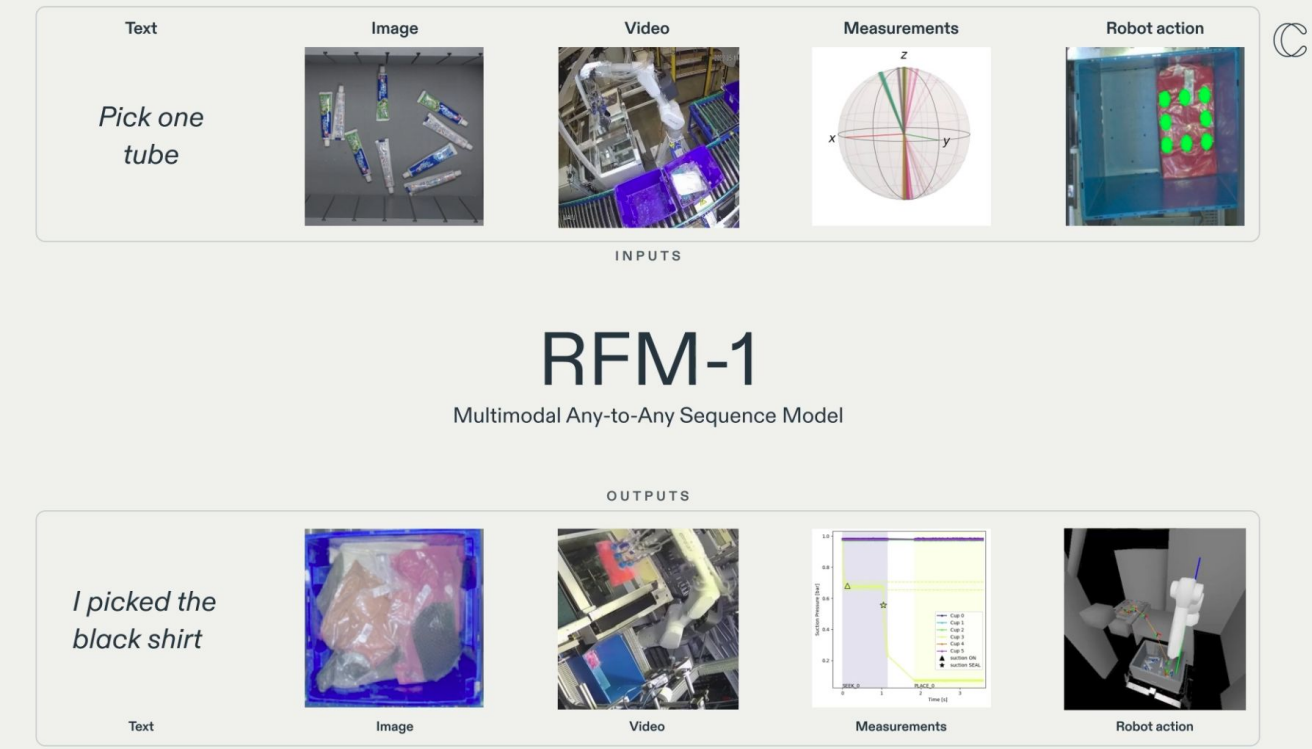

input video
image encoder
action encoder
"I am approaching a crossing yielding to pedestrians"
...
"It is safe to move so I am now accelerating"
text encoder
input tokens
world model
autoregressive prediction
output tokens

RFM-1 (Covariant)



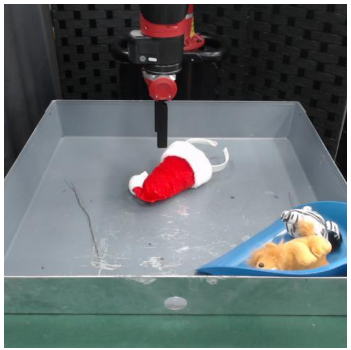
https://covariant.ai/insights/introducing-rfm-1-giving-robots-human-like-reasoning-capabilities/

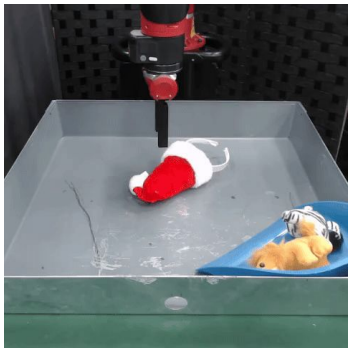
RFM-1 (Covariant)

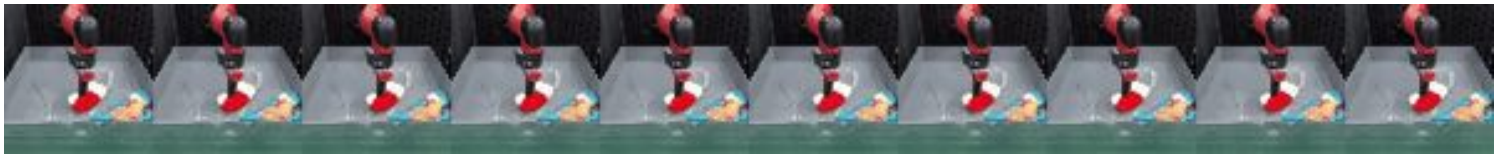


Video Generation as Physical Simulators

Video generation can be used in visual planning



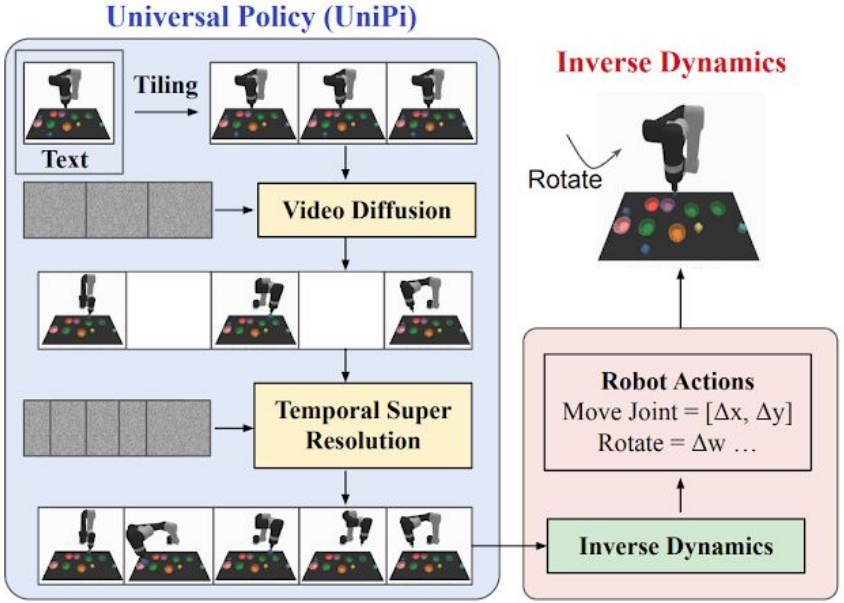




Gupta, Agrim, et al. "Maskvit: Masked visual pre-training for video prediction." *arXiv preprint arXiv:2206.11894* (2022).

Video Generation as Physical Simulators

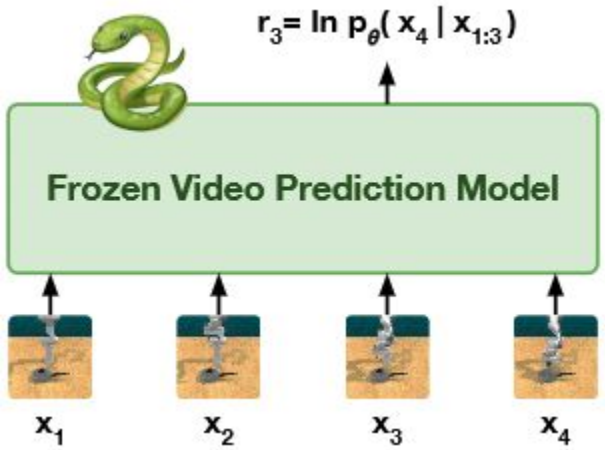
They can also be used to generate visual plans



Du, Yilun, et al. "Learning universal policies via text-guided video generation." *Advances in Neural Information Processing Systems* 36 (2024).

Video Generation as Physical Simulators

Likelihoods can also be used as a reward function



Escontrela, Alejandro, et al. "Video prediction models as rewards for reinforcement learning." *Advances in Neural Information Processing Systems* 36 (2024).

Outline

- Basics
- Improving Video Generation
- **Applications**
  - Video Generation Models as Physical Simulators
  - **Video Editing**

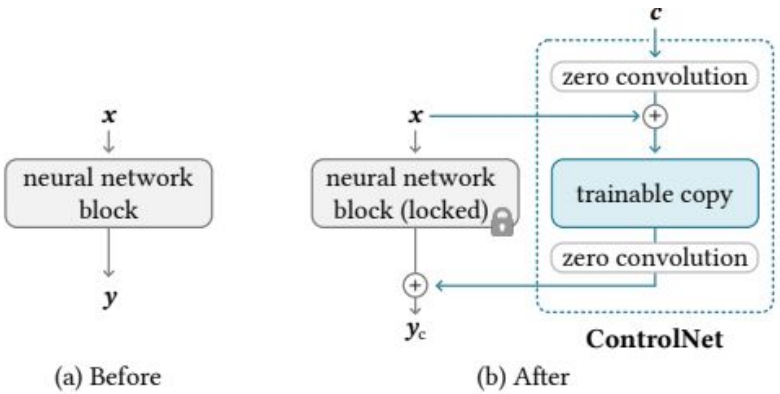
ControlNet

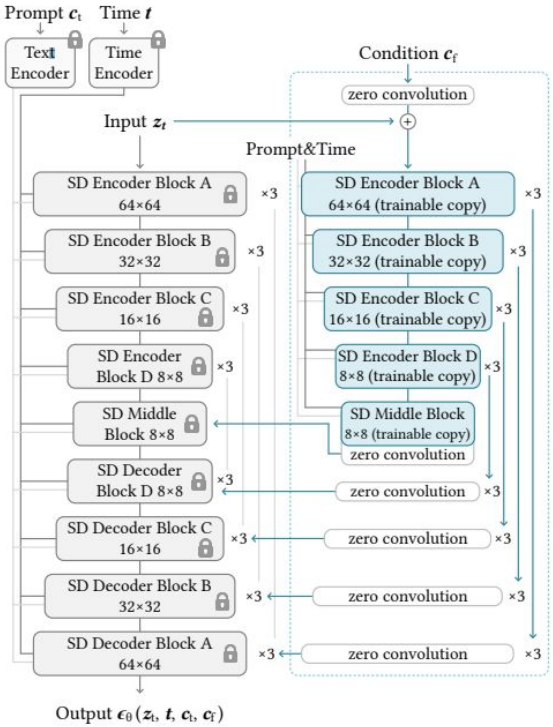
Goal: Finetune a generative model to include extra conditioning

- **Canny edge map**
- **Depth**
- **Human Pose**
- **Lower Resolution Image**
- **Segmentation**
- **Sketch**

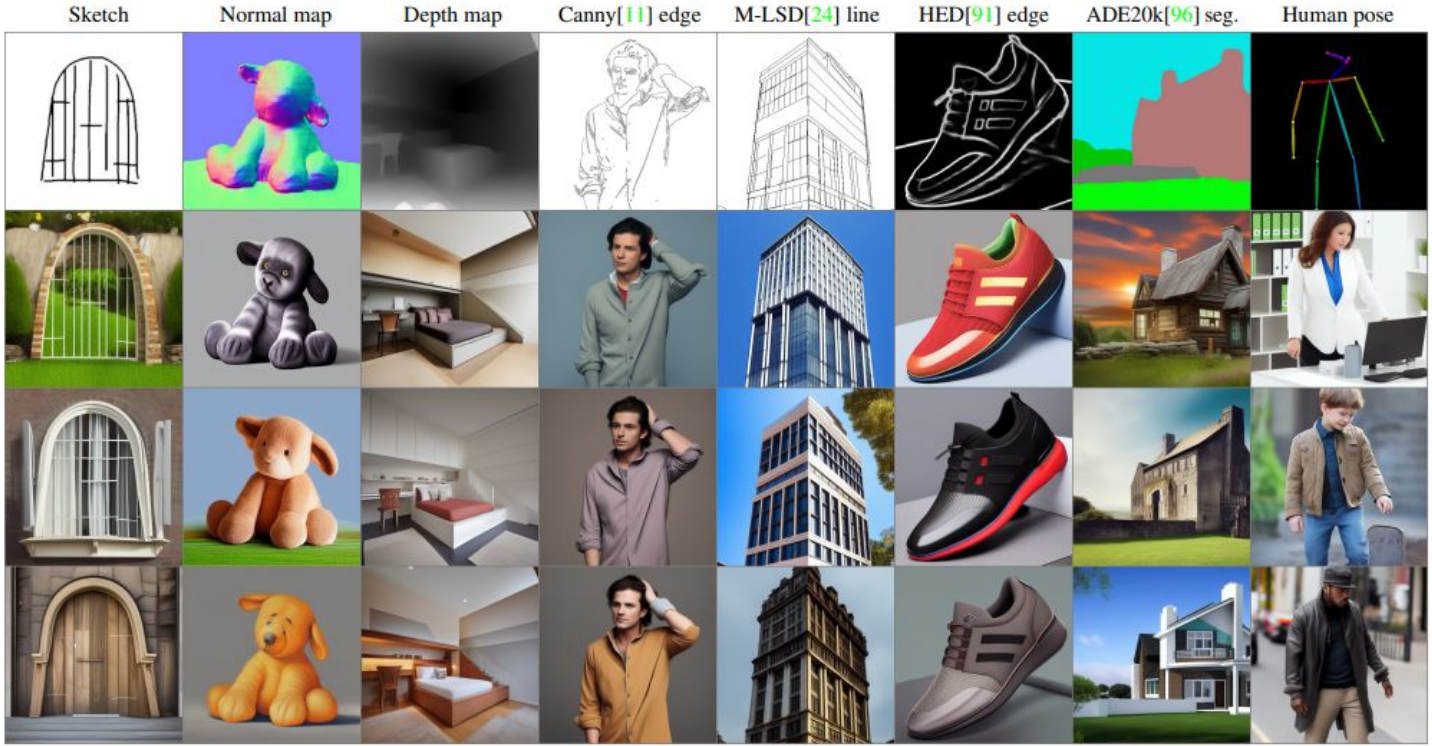
Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

ControlNet

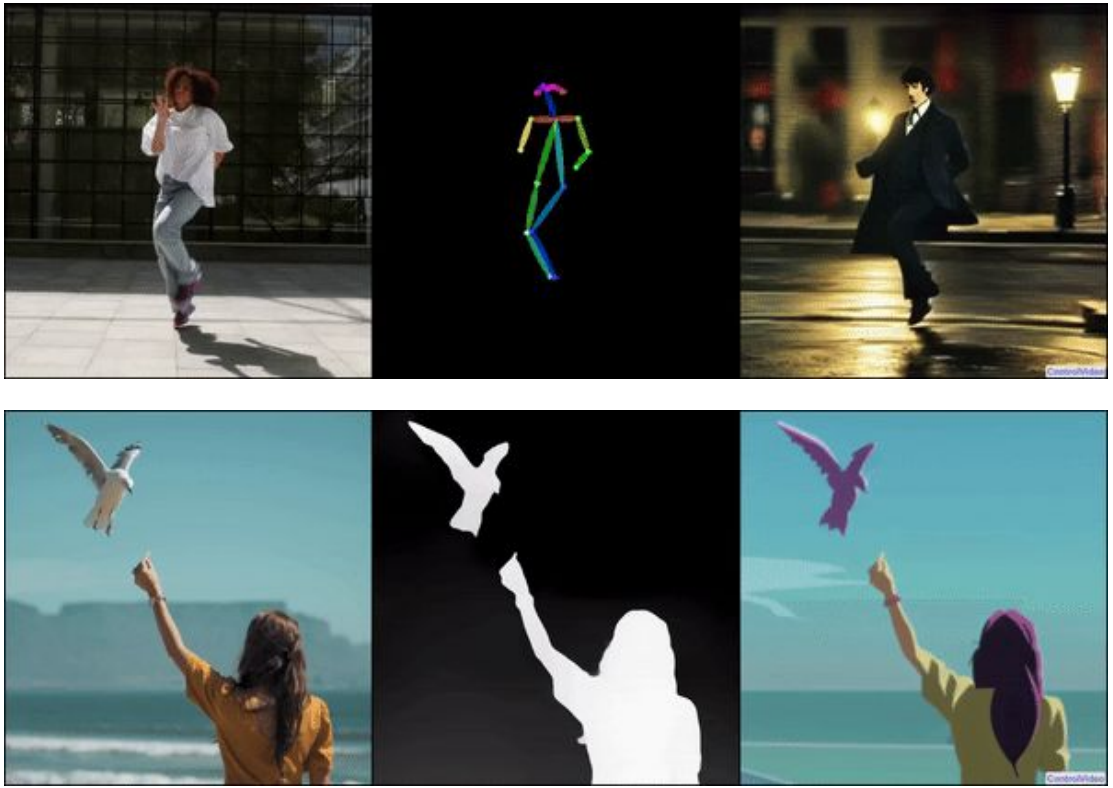

x
neural network block
y
(a) Before
c
zero convolution
x
neural network block (locked)
trainable copy
zero convolution
y_c
ControlNet
(b) After


Prompt c_t
Time t
Text Encoder
Time Encoder
Condition c_f
zero convolution
Input z_t
Prompt&Time
SD Encoder Block A 64×64
SD Encoder Block A 64×64 (trainable copy)
SD Encoder Block B 32×32
SD Encoder Block B 32×32 (trainable copy)
SD Encoder Block C 16×16
SD Encoder Block C 16×16 (trainable copy)
SD Encoder Block D 8×8
SD Encoder Block D 8×8 (trainable copy)
SD Middle Block 8×8
SD Middle Block 8×8 (trainable copy)
zero convolution
SD Decoder Block D 8×8
SD Decoder Block C 16×16
SD Decoder Block B 32×32
SD Decoder Block A 64×64
×3
Output $\epsilon_\theta(z_t, t, c_t, c_f)$

ControlNet


Sketch
Normal map
Depth map
Canny[11] edge
M-LSD[24] line
HED[91] edge
ADE20k[96] seg.
Human pose

ControlVideo

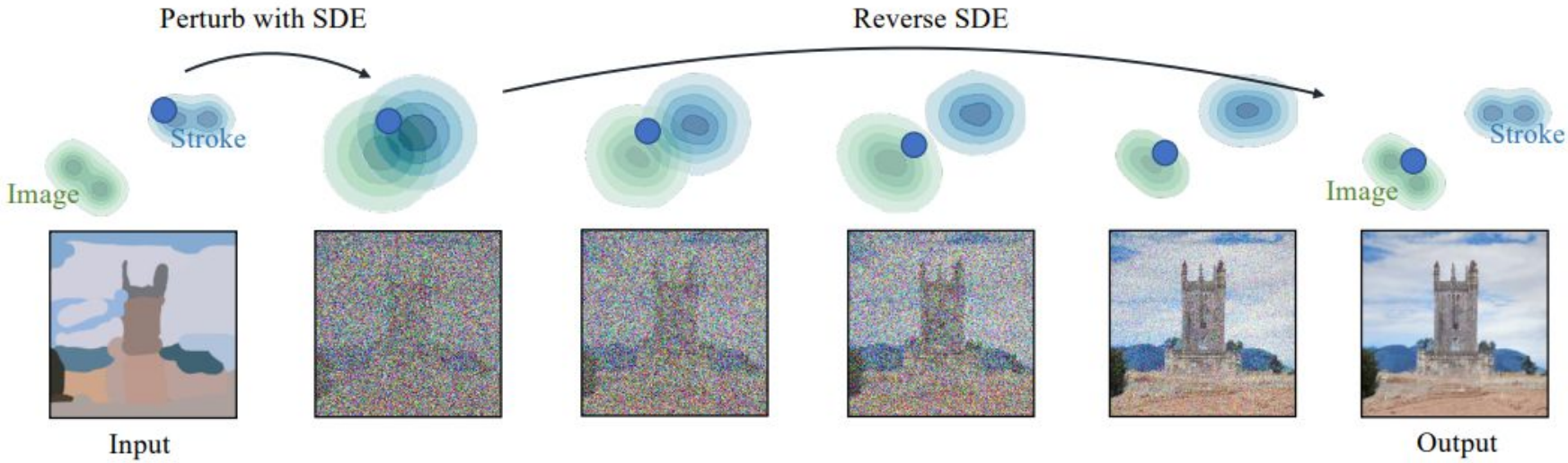


ControlVideo

Zhao, Min, et al. "Controlvideo: Adding conditional control for one shot text-to-video editing." *arXiv preprint arXiv:2305.17098* (2023).

SDEdit

Goal: High-level, editing of semantic image features while retaining global structure



Meng, Chenlin, et al. "Sdedit: Guided image synthesis and editing with stochastic differential equations." *arXiv preprint arXiv:2108.01073* (2021).

SDEdit

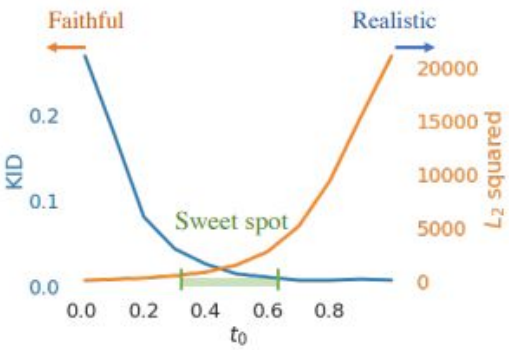
Given some diffusion timestep t, image x_0, edit caption c:

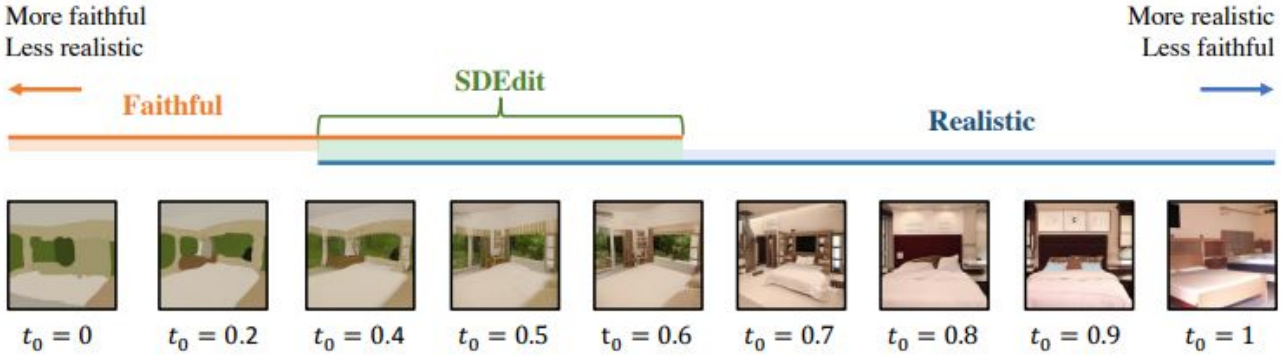
1) Apply the forward process to t: $q(x_t \mid x_0)$
2) Apply the reverse process to 0: $p(x_{t-1} \mid x_t, c)$ (t times)

SDEdit

T is a hyperparameter

- Trade-off between faithfulness to original image, and alignment with target edit





Using SDEdit

Source Video

Made of wooden blocks

Using SDEdit

Source Video

Make of colorful toy bricks

Dreamix

How about more general video editing?

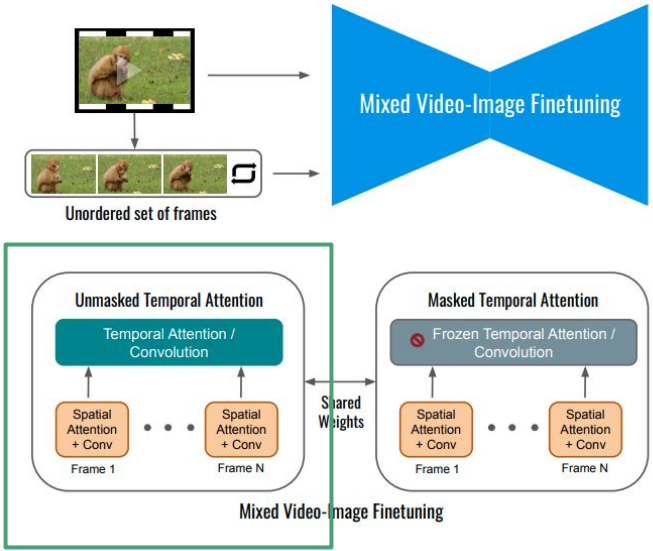
Molad, Eyal, et al. "Dreamix: Video diffusion models are general video editors." *arXiv preprint arXiv:2302.01329* (2023).

Dreamix

Mixed Video-Image Finetuning (Video)

- Reconstruct the original video **v** conditioned on noised version $\mathbf{z}_s$ and rare token **t***

$$\mathcal{L}_\theta^{vid}(v) = \mathbb{E}_{\epsilon \sim N(0,\mathbf{I}), s \in \mathcal{U}(0,1)} \| D_{\theta'}(z_s, s, t^*, c) - v \|^2$$



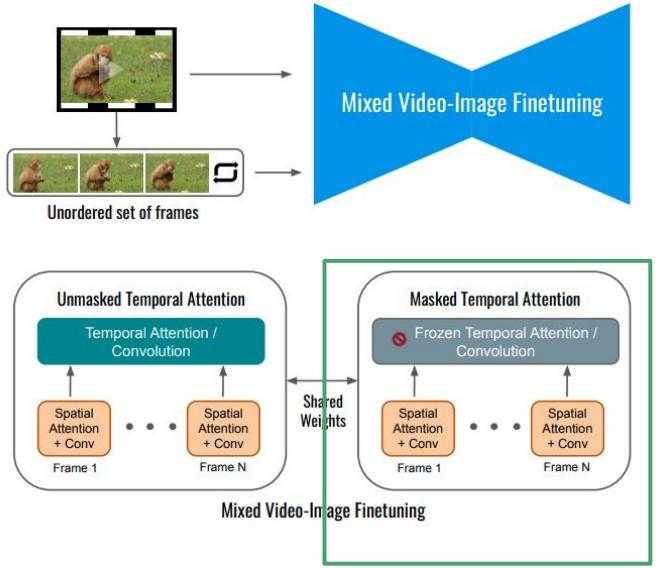
Dreamix

Mixed Video-Image Finetuning (Image)

- Reconstruct the original video **v** conditioned on noised version $\mathbf{z}_s$ and rare token **t***

Mask out any temporal operations for only frame-based denoising

$$\mathcal{L}_\theta^{frame}(u) = \mathbb{E}_{\epsilon\sim N(0,\mathbf{I}),s\in\mathcal{U}(0,1)}\left\|D_{\theta'}^a(z_s, s, t^*, c) - u\right\|^2$$

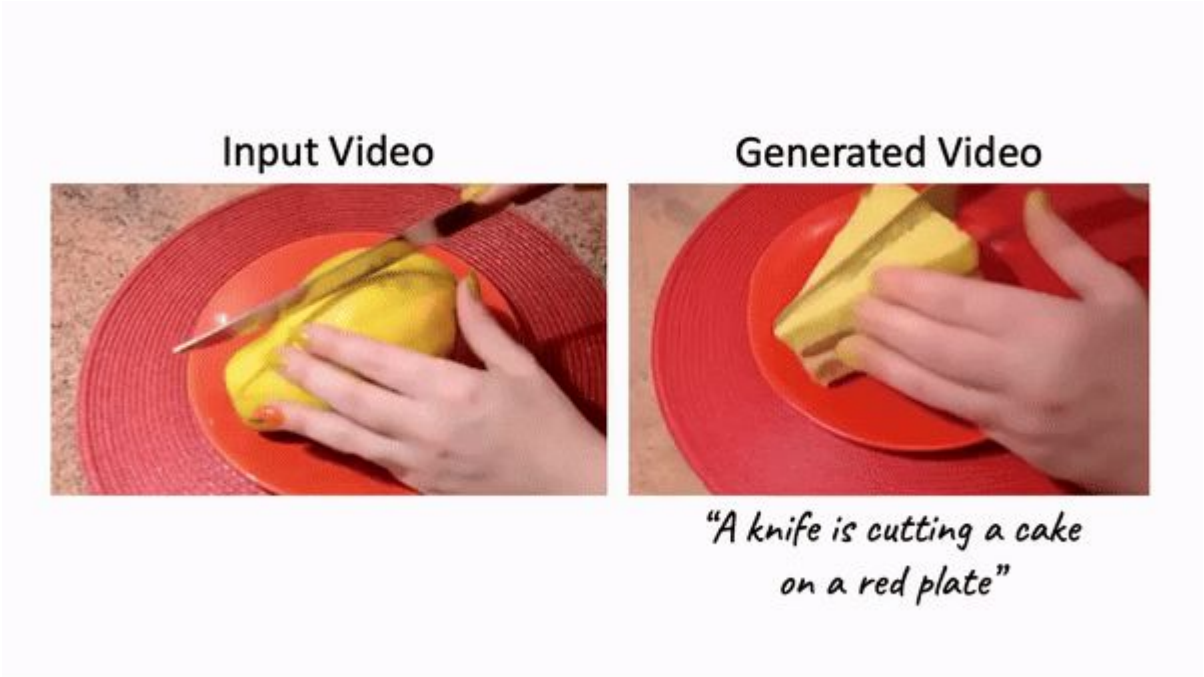


Dreamix

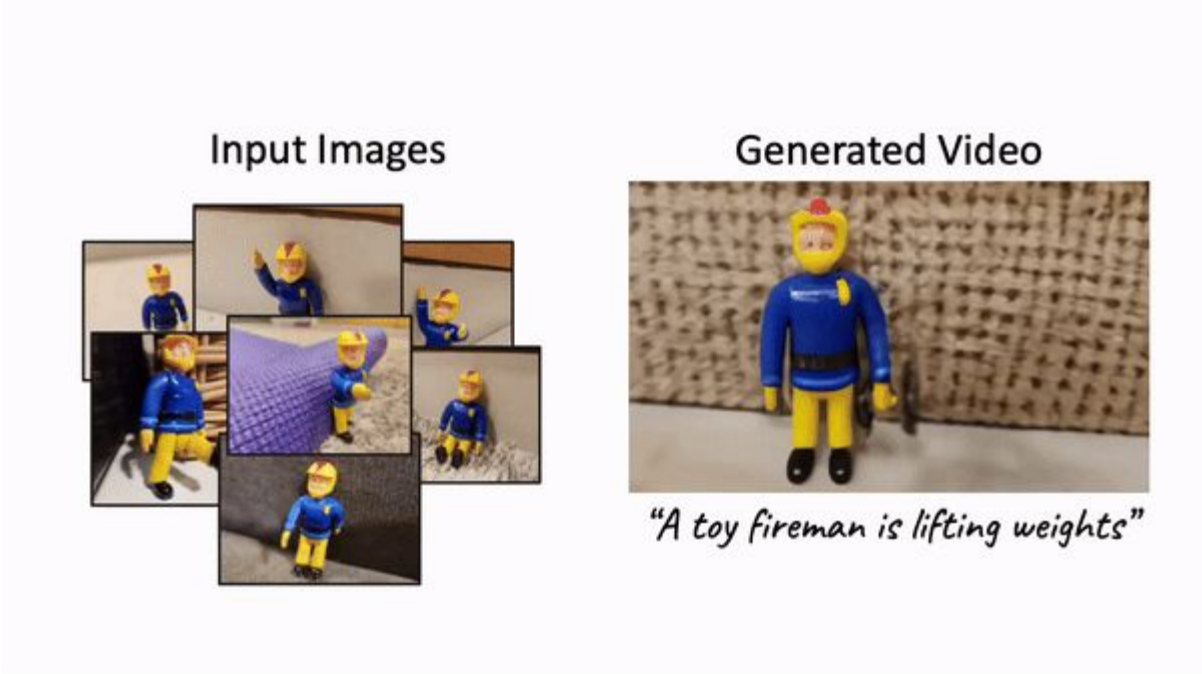
Final Fine-tuning Loss

$$\theta = arg \min_{\theta'} \alpha \mathcal{L}_{\theta'}^{vid}(v) + (1-\alpha)\mathcal{L}_{\theta'}^{frame}(u)$$

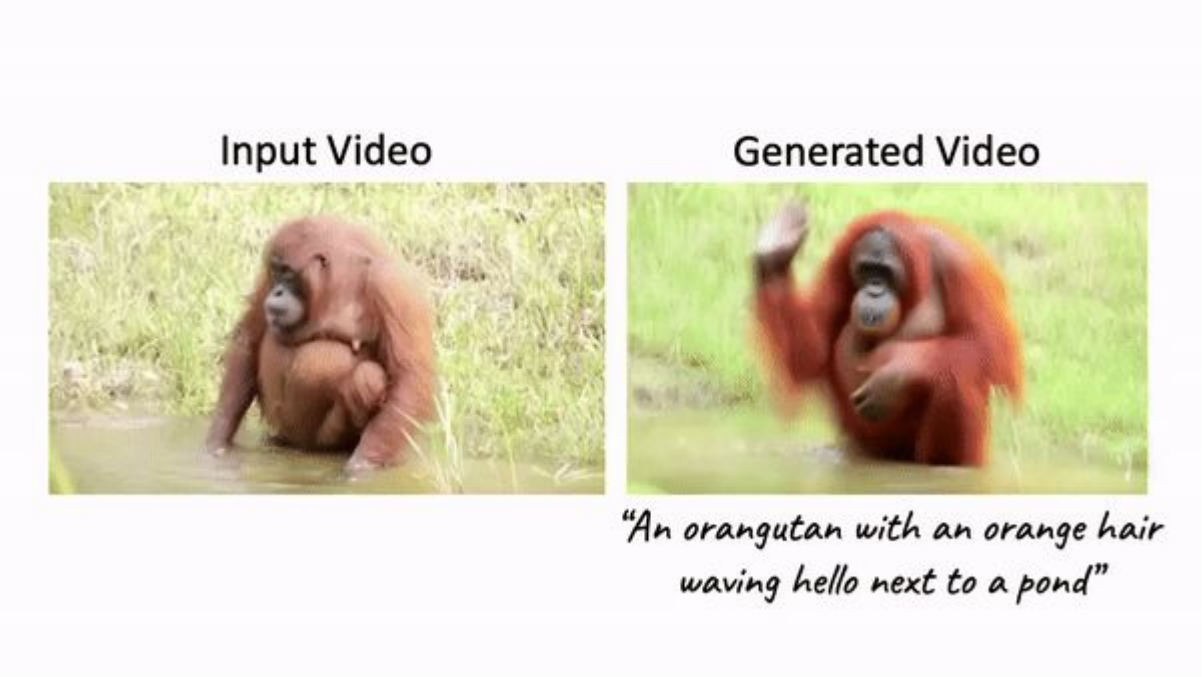
Dreamix


Input Video
Generated Video
"A knife is cutting a cake on a red plate"

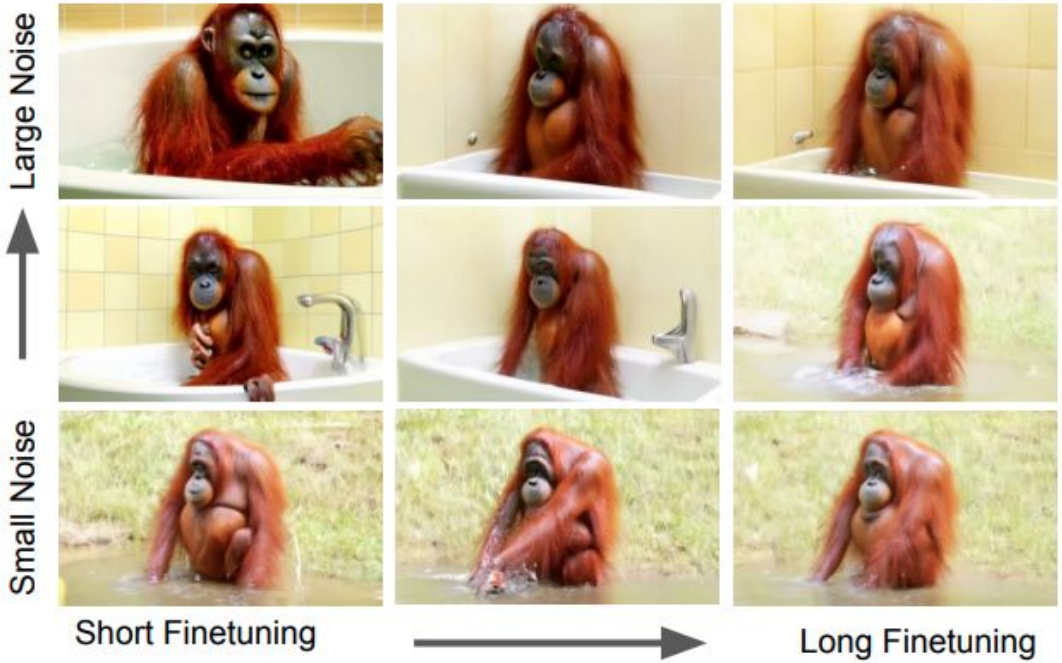
Dreamix


Input Images
Generated Video
"A toy fireman is lifting weights"

Dreamix

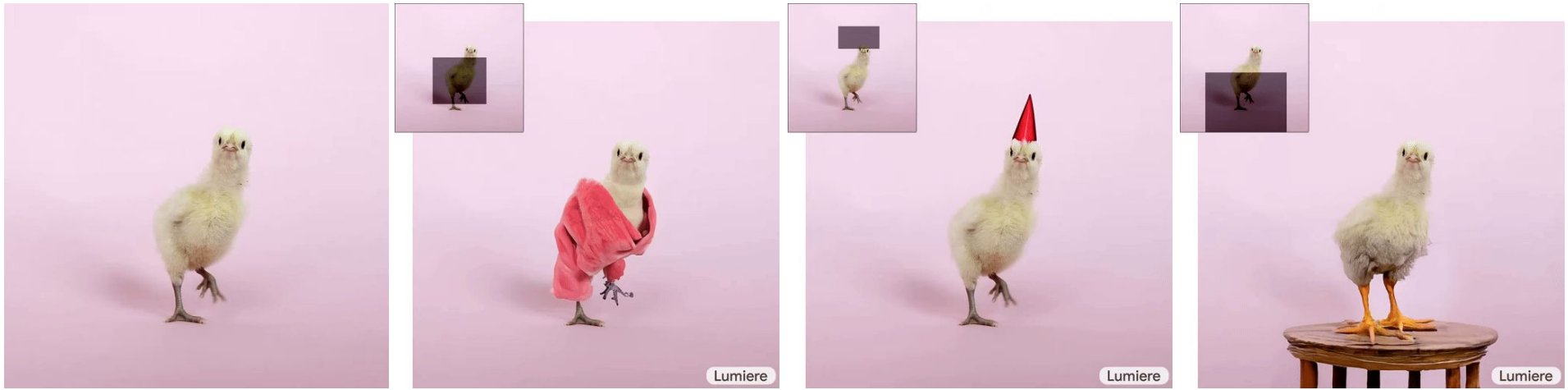

Input Video
Generated Video
"An orangutan with an orange hair waving hello next to a pond"

Dreamix



Video Inpainting

Easier with text-video diffusion models



References

Ho, Jonathan, et al. "Imagen video: High definition video generation with diffusion models." *arXiv preprint arXiv:2210.02303* (2022).

Ge, Songwei, et al. "Preserve your own correlation: A noise prior for video diffusion models." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

Singer, Uriel, et al. "Make-a-video: Text-to-video generation without text-video data." *arXiv preprint arXiv:2209.14792* (2022).

Bar-Tal, Omer, et al. "Lumiere: A space-time diffusion model for video generation." *arXiv preprint arXiv:2401.12945* (2024).

Blattmann, Andreas, et al. "Align your latents: High-resolution video synthesis with latent diffusion models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

Girdhar, Rohit, et al. "Emu video: Factorizing text-to-video generation by explicit image conditioning." *arXiv preprint arXiv:2311.10709* (2023).

Blattmann, Andreas, et al. "Stable video diffusion: Scaling latent video diffusion models to large datasets." *arXiv preprint arXiv:2311.15127* (2023).

Yan, Wilson, et al. "Videogpt: Video generation using vq-vae and transformers." *arXiv preprint arXiv:2104.10157* (2021).

Ge, Songwei, et al. "Long video generation with time-agnostic vqgan and time-sensitive transformer." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.

He, Yingqing, et al. "Latent video diffusion models for high-fidelity long video generation." *arXiv preprint arXiv:2211.13221* (2022).

Villegas, Ruben, et al. "Phenaki: Variable length video generation from open domain textual descriptions." *International Conference on Learning Representations*. 2022.

Gupta, Agrim, et al. "Photorealistic video generation with diffusion models." *arXiv preprint arXiv:2312.06662* (2023).

Kondratyuk, Dan, et al. "Videopoet: A large language model for zero-shot video generation." *arXiv preprint arXiv:2312.14125* (2023).

Yu, Lijun, et al. "Language Model Beats Diffusion--Tokenizer is Key to Visual Generation." *arXiv preprint arXiv:2310.05737* (2023).

Mentzer, Fabian, et al. "Finite scalar quantization: Vq-vae made simple." *arXiv preprint arXiv:2309.15505* (2023).

Finn, Chelsea, Ian Goodfellow, and Sergey Levine. "Unsupervised learning for physical interaction through video prediction." *Advances in neural information processing systems* 29 (2016).

Yang, Mengjiao, et al. "Learning interactive real-world simulators." *arXiv preprint arXiv:2310.06114* (2023).

Hu, Anthony, et al. "Gaia-1: A generative world model for autonomous driving." *arXiv preprint arXiv:2309.17080* (2023).

Gupta, Agrim, et al. "Maskvit: Masked visual pre-training for video prediction." *arXiv preprint arXiv:2206.11894* (2022).

Du, Yilun, et al. "Learning universal policies via text-guided video generation." *Advances in Neural Information Processing Systems* 36 (2024).

Escontrela, Alejandro, et al. "Video prediction models as rewards for reinforcement learning." *Advances in Neural Information Processing Systems* 36 (2024).
Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
Zhao, Min, et al. "Controlvideo: Adding conditional control for one shot text-to-video editing." *arXiv preprint arXiv:2305.17098* (2023).
Meng, Chenlin, et al. "Sdedit: Guided image synthesis and editing with stochastic differential equations." *arXiv preprint arXiv:2108.01073* (2021).
Molad, Eyal, et al. "Dreamix: Video diffusion models are general video editors." *arXiv preprint arXiv:2302.01329* (2023).