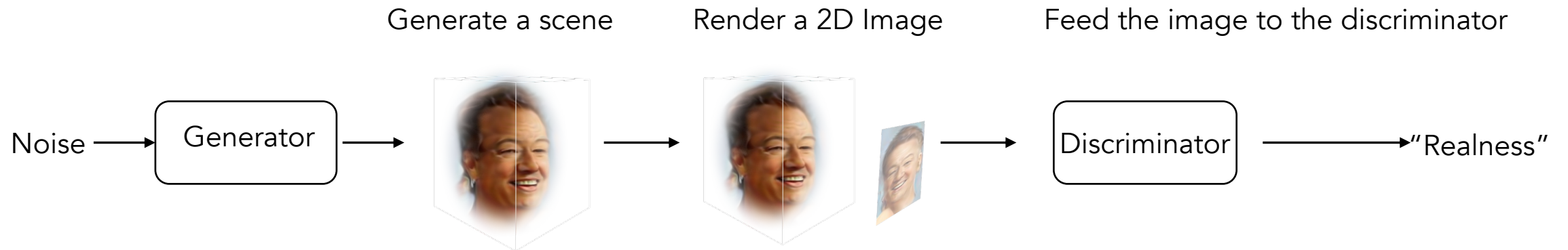


3D Generation with 2D Generative Models

Training Steps

1. Generate a representation of a *scene*
2. Render the scene from a random camera pose
3. Feed the image to a 2D discriminator
4. Backpropagate through the discriminator and differentiable rendering



Slide credit: Eric Chan



Video Synthesis and Editing

Jun-Yan Zhu

16-726 Learning-based Image Synthesis, Spring 2025

Image Editing and Synthesis

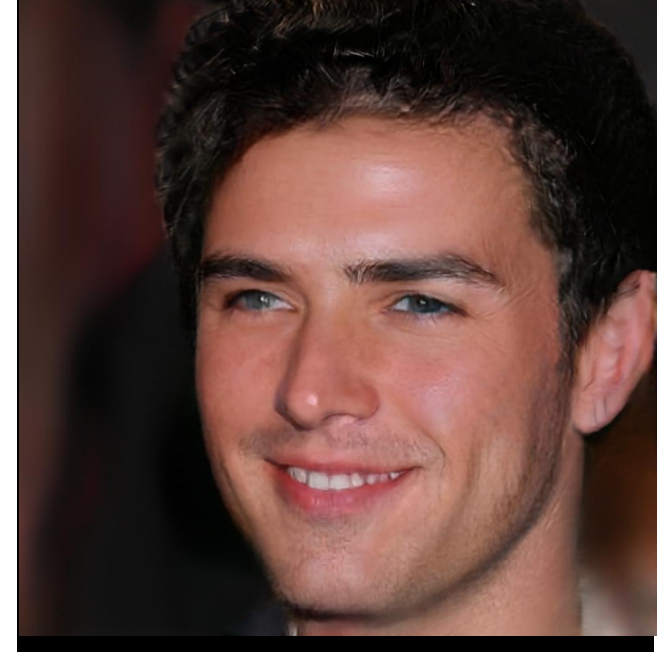


Image Editing and Synthesis



Image Editing and Synthesis



“Emma Stone”

“Mohawk hairstyle”

“Without makeup”

“Cute cat”

“Lion”

“Gothic church”

Images vs. Videos

- What are the major differences?
 - 2D vs. 3D
 - Spatial vs. spatial + temporal
- Differences between 3D vs. videos?
 - (x, y, z) vs. (x, y, t)
- Why are videos much more challenging?
 - Understanding motions/actions/intensions.
 - Long-term dependence.
 - Hard to annotate.
 - Computationally-expensive (e.g., memory, training time)

Why not apply it to video?

Someone gave you an image synthesis model

asked you to build a video synthesis application

How to Generalize to Videos

- Idea 1: Frame-by-Frame

Frame-by-Frame Result (pix2pixHD)



Frame-by-Frame Result (CycleGAN)



How to Generalize to Videos

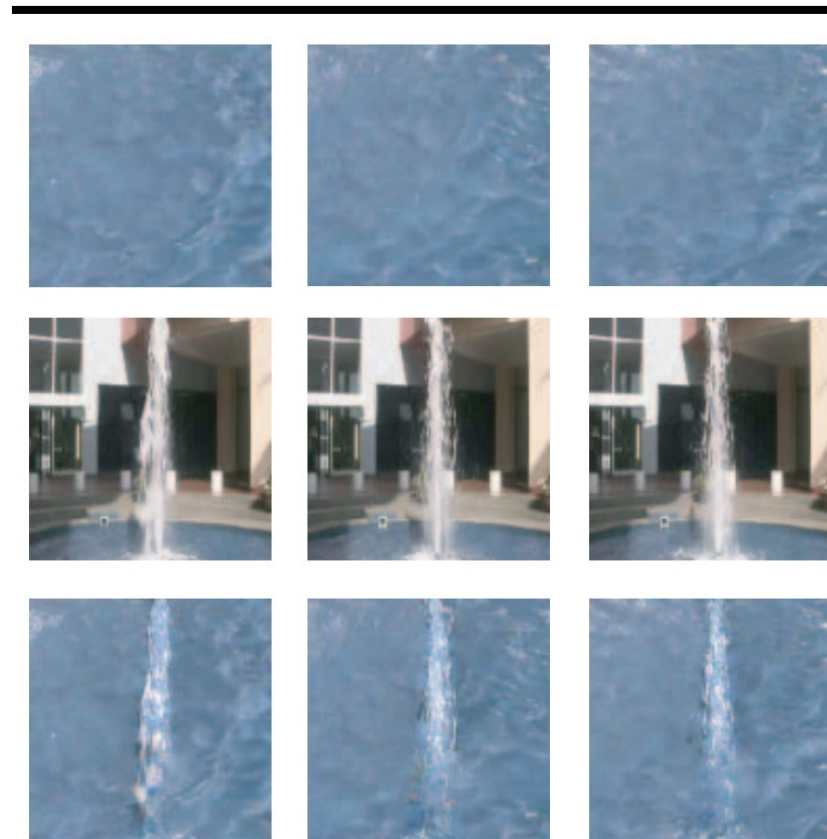
- Idea 1: Frame-by-Frame
 - Temporal inconsistency (flicking, color drift)
- Idea 2: Video as 3D data (height x width x time)

Case Study: 3D Poisson blending

Spatial-temporal Constraints

Background

Time



$$F(\nabla I, G) = \|\nabla I - G\|^2$$

$$= \left(\frac{\partial I}{\partial x} - G_x\right)^2 + \left(\frac{\partial I}{\partial y} - G_y\right)^2 + \left(\frac{\partial I}{\partial t} - G_t\right)^2$$

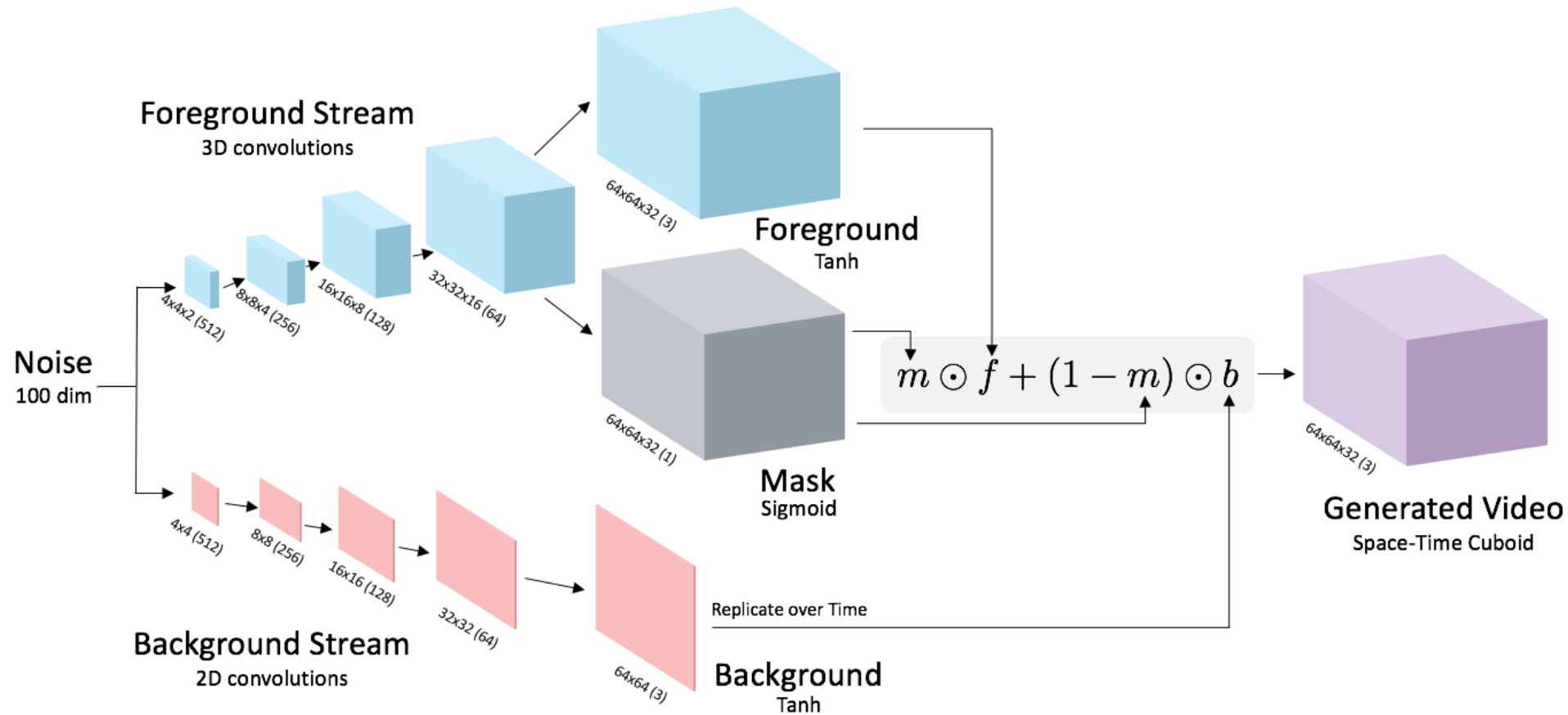
Foreground

Output
Image

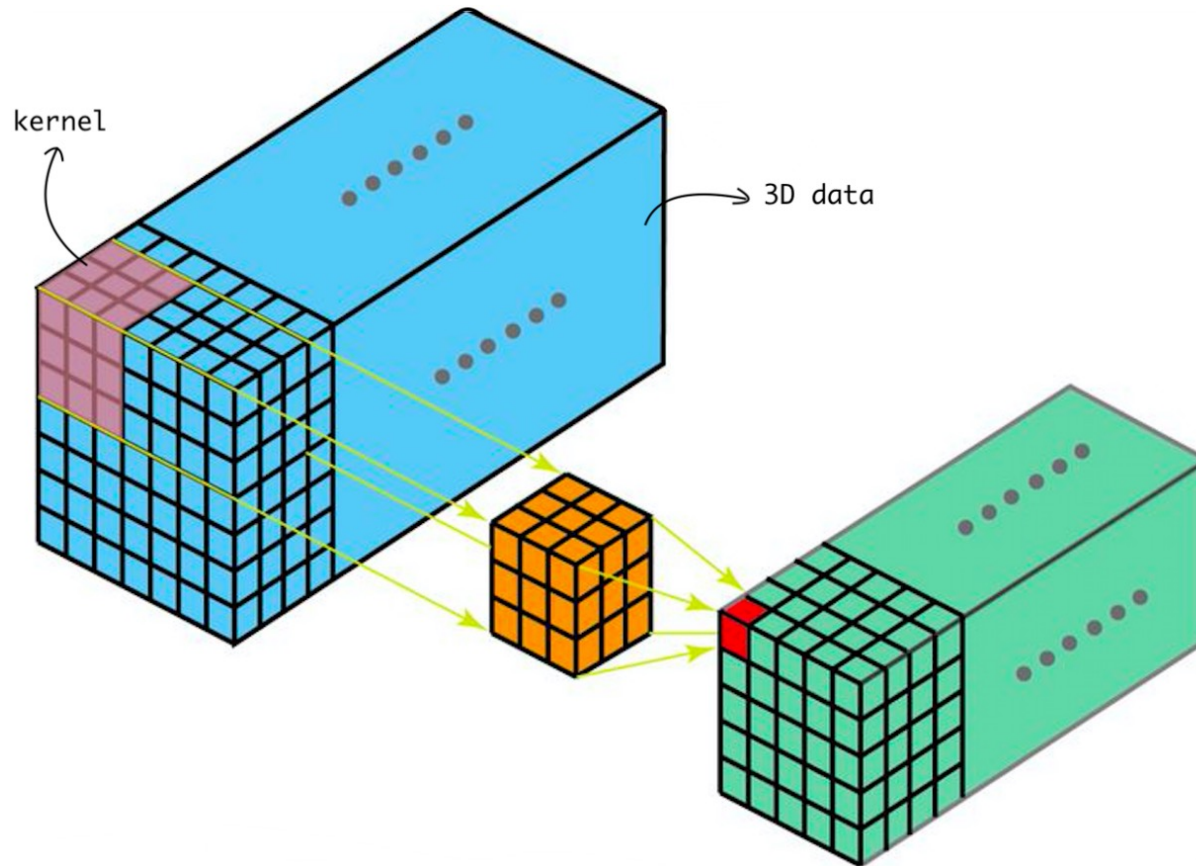
Guidance Gradient

Output

Case Study: Video GANs



Recap: 3D Conv



Easy to implement:

- Replace 2D by 3D in your code

e.g., Conv2D -> Conv3D

ConvTranspose2d->ConvTranspose3d

MaxPool2d -> MaxPool3d

```
CLASS torch.nn.Conv3d(in_channels, out_channels, kernel_size, stride=1, padding=0,  
                        dilation=1, groups=1, bias=True, padding_mode='zeros', device=None, dtype=None) [SOURCE]
```


Case Study: Video GANs

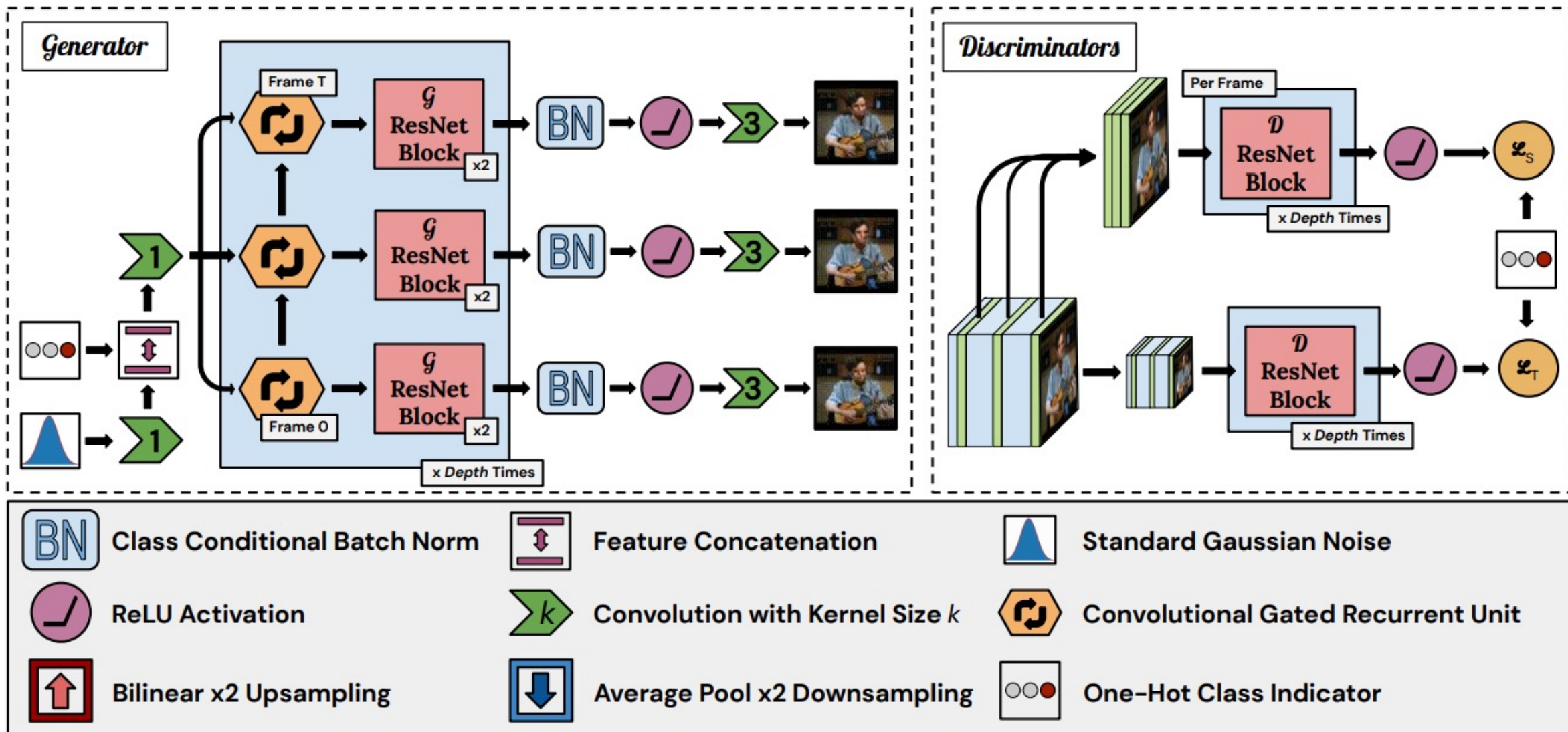


Case Study: DVDGANs



BigGAN-based generator + Spatial Discriminator + Temporal Discriminator

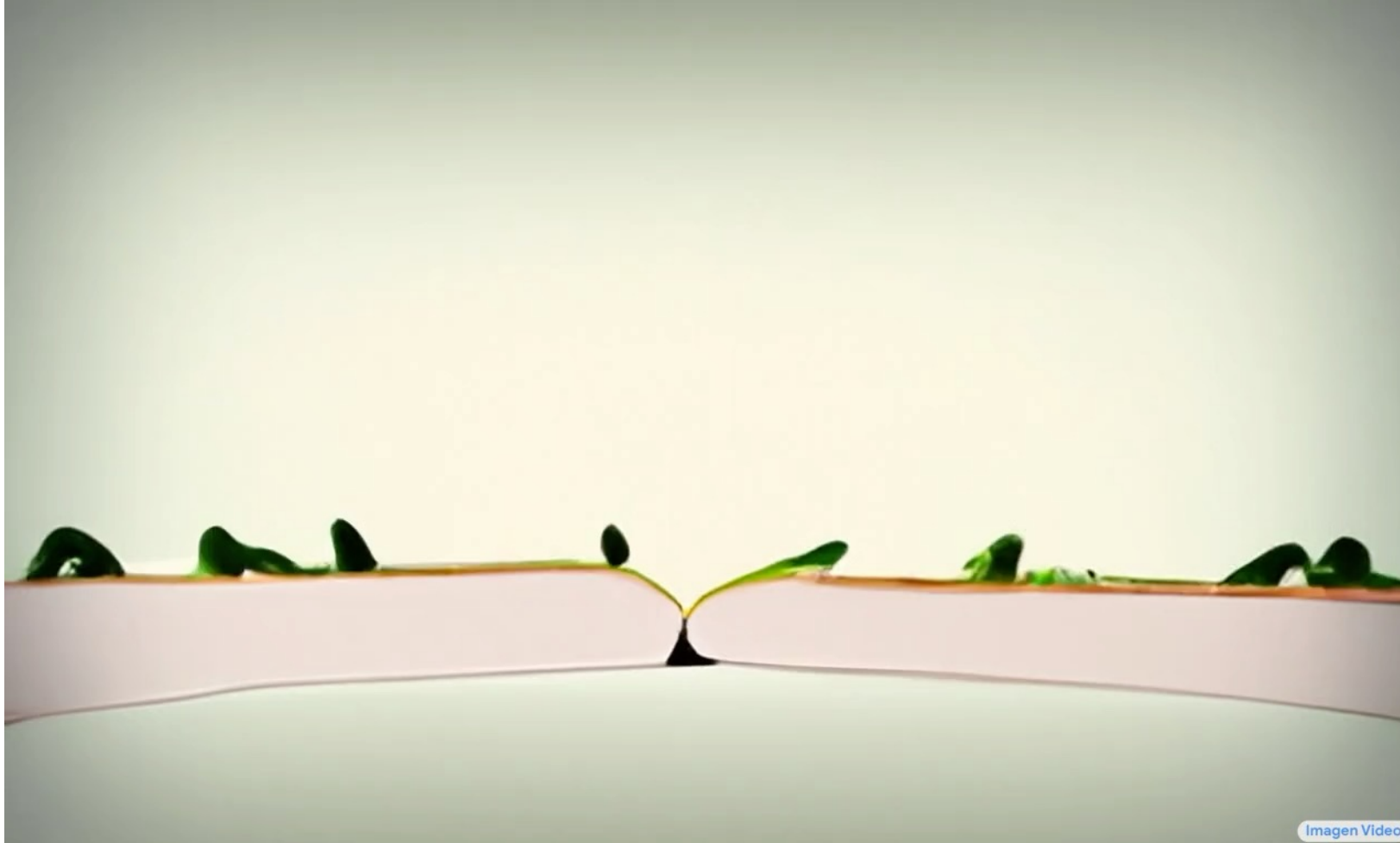
Case Study: DVDGANs



Case Study: DVDGANs



Case Study: Imagen Video



Case Study: Imagen Video



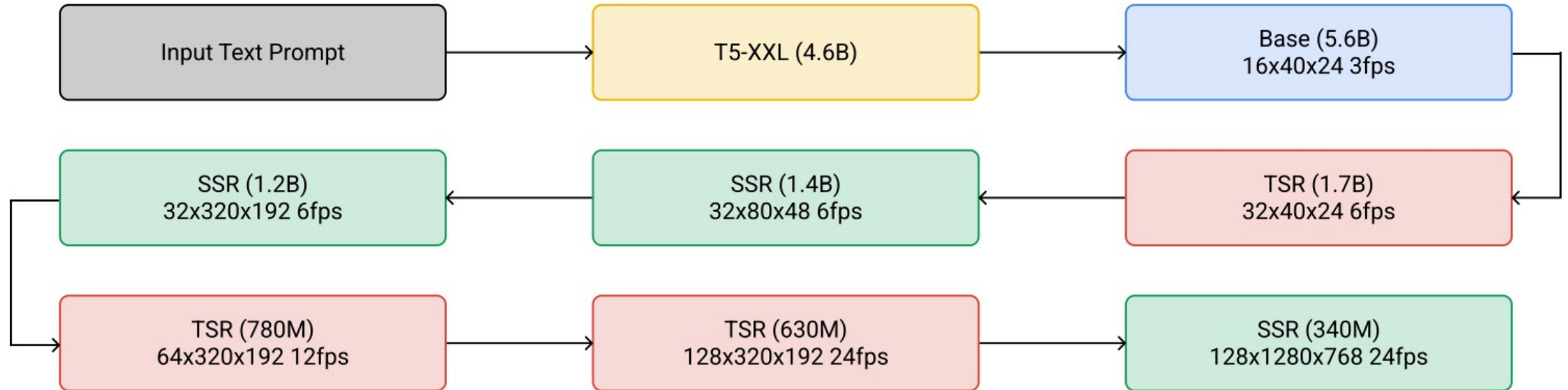
An astronaut riding a horse

Case Study: Imagen Video

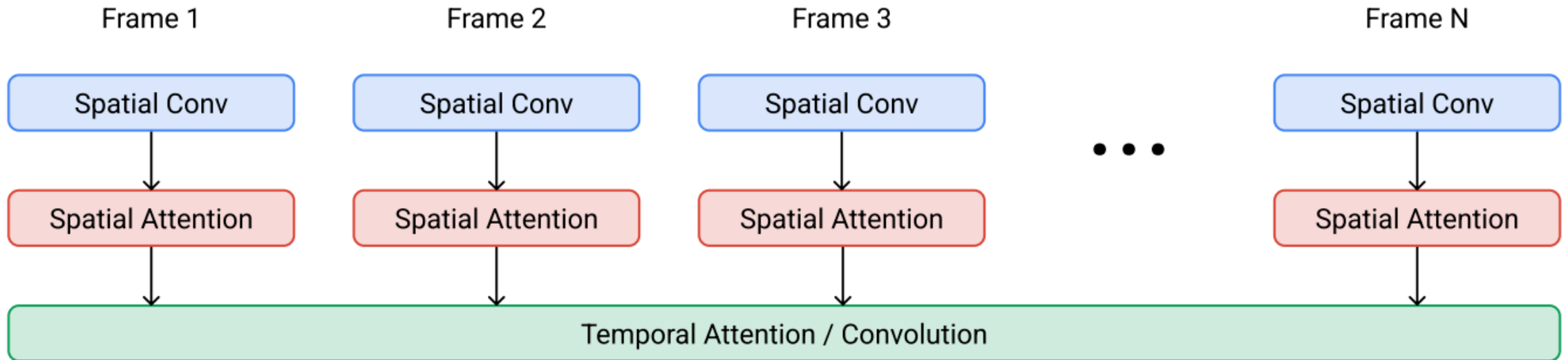


Wooden figure surfing on a surfboard in space

Case Study: Imagen Video



Case Study: Imagen Video



How to Generalize to Videos

- Idea 1: Frame-by-Frame
 - Temporal inconsistency (flicking, color drift)
- Idea 2: Video as 3D data (height x width x time)
 - memory-intensive and time-consuming
 - only work for a short video at low resolution
- Idea 3: recurrent (autoregressive) synthesis
 - Generate 1st frame, generate 2nd frame based on 1st one, ...
 - Using optical flow (optional): warp 1st frame to 2nd frame.

Text Synthesis

- [Shannon,'48] proposed a way to generate English-looking text using N-grams:
 - Assume a generalized Markov model
 - Use a large text to compute prob. distributions of each letter given N-1 previous letters
 - Starting from a seed repeatedly sample this Markov chain to generate new letters
 - Also works for whole words

WE NEED TO EAT CAKE

Case Study: Video Textures



Still image



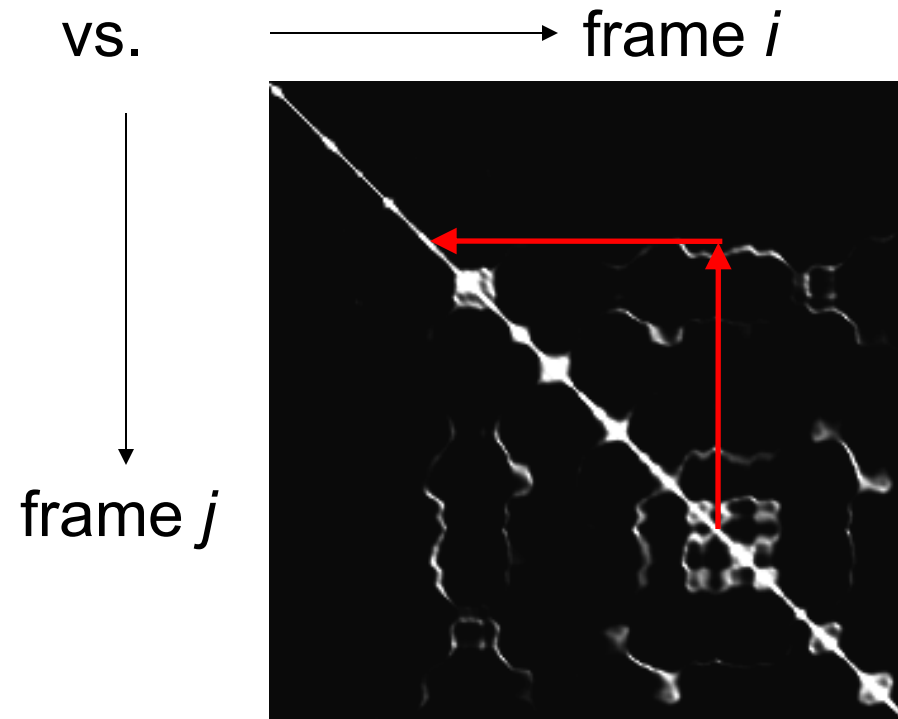
Loopy video

Case Study: Video Textures



Finding good transitions

- Compute L_2 distance $D_{i,j}$ between all frames

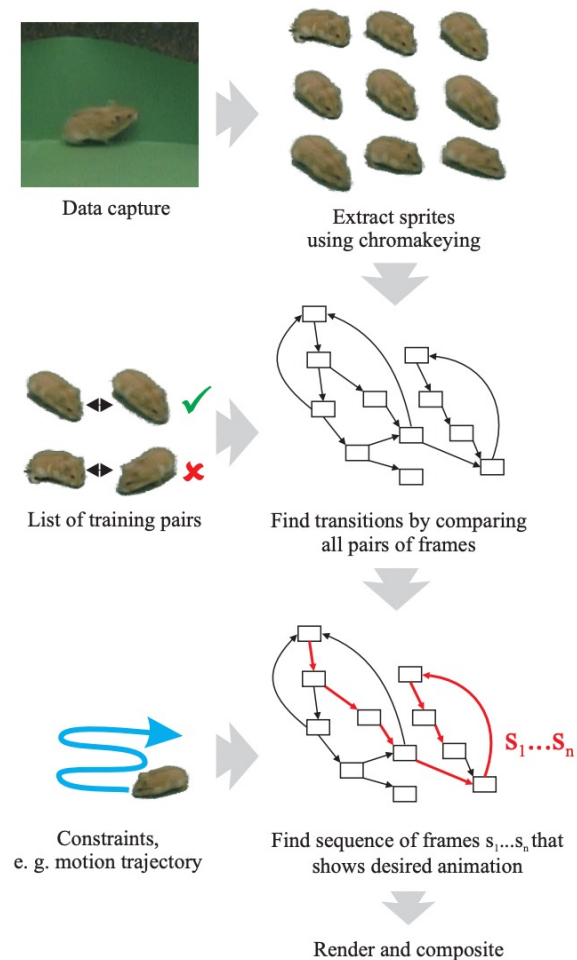


Similar frames make good transitions

Case Study: Video Textures



Case Study: Controlled Animation of Video Sprites



Case Study: Video-to-Video Translation



T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, B. Catanzaro,
“Video-to-Video Synthesis,” NeurIPS 2018.

<https://github.com/NVIDIA/vid2vid>

Previous Work: Frame-by-Frame Result

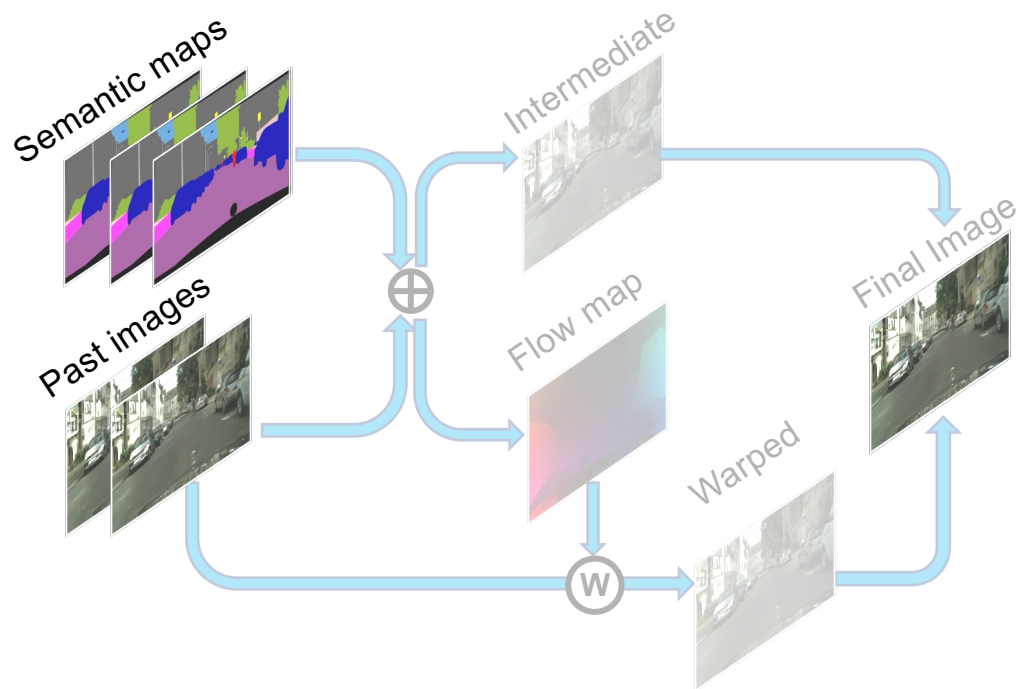


vid2vid

- Sequential generator
- Multi-scale temporal discriminator
- Spatio-temporal progressive training procedure

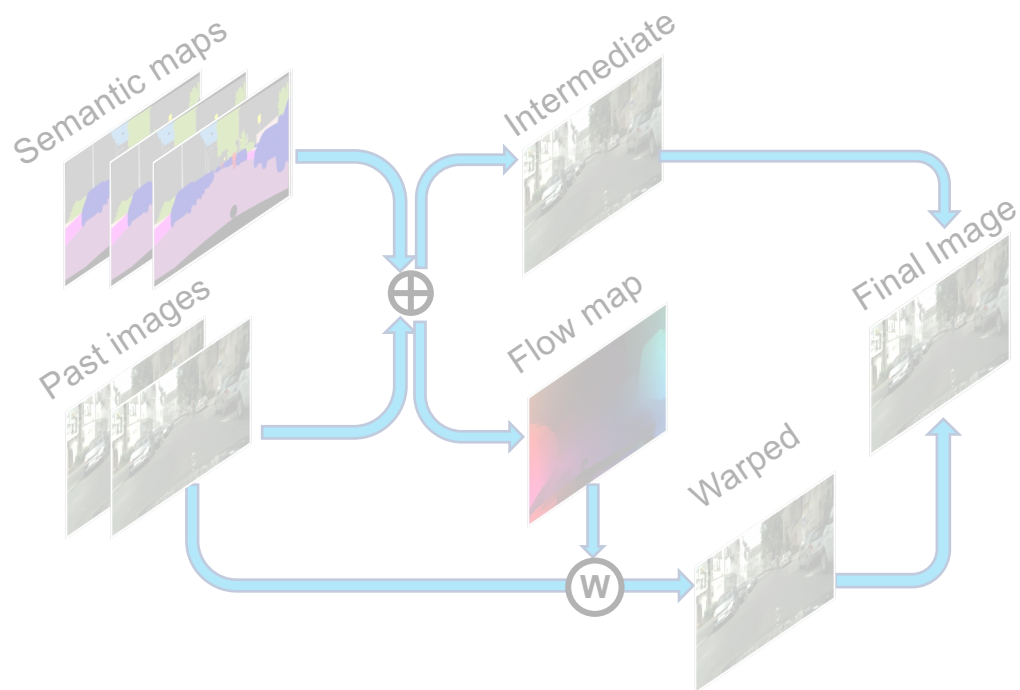
vid2vid

Sequential Generator



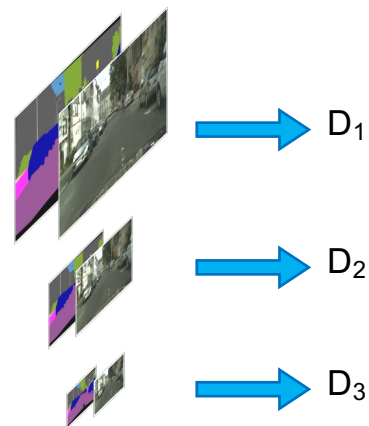
vid2vid

Sequential Generator

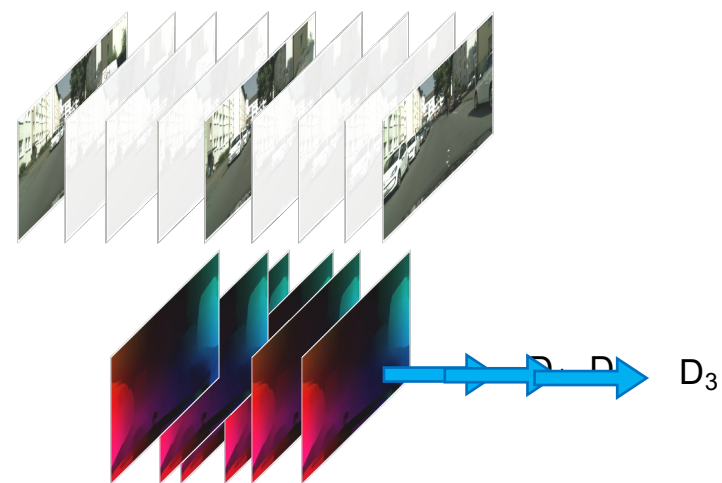


Multi-scale Discriminators

Image Discriminator



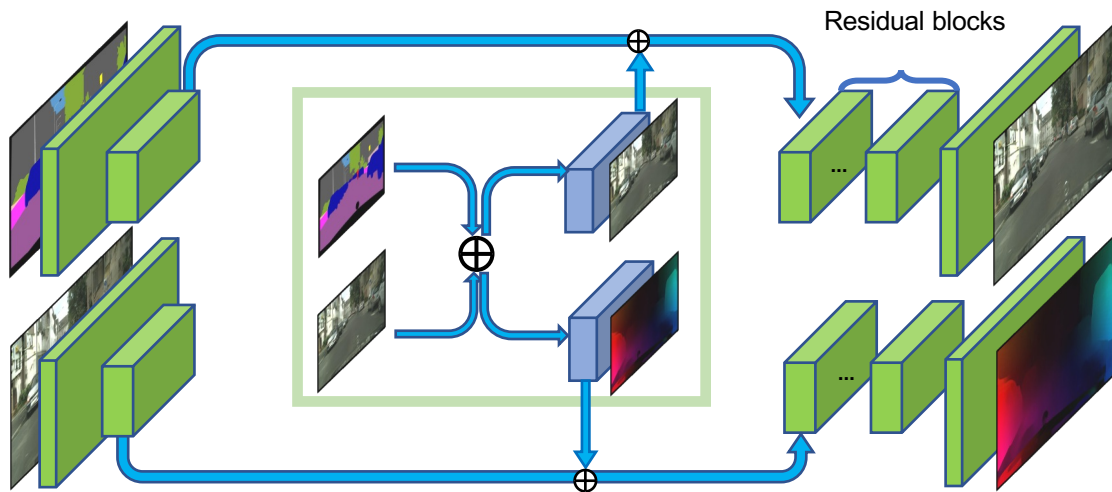
Video Discriminator



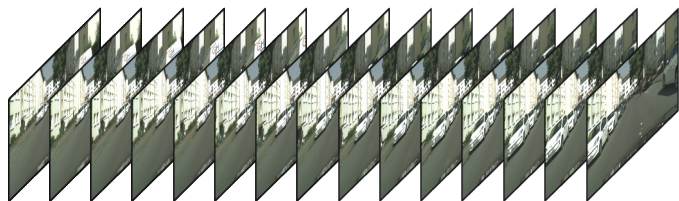
vid2vid

Spatio-temporally Progressive Training

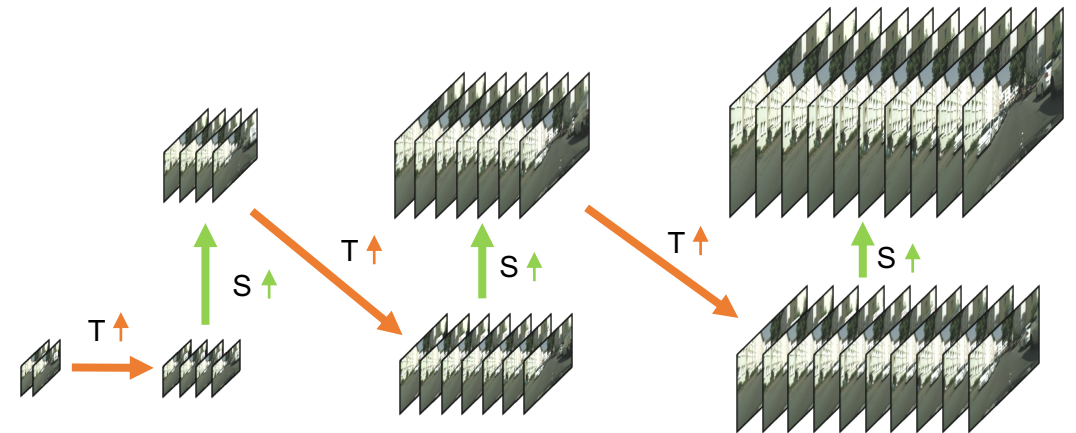
Spatially progressive



Temporally progressive



Alternating training



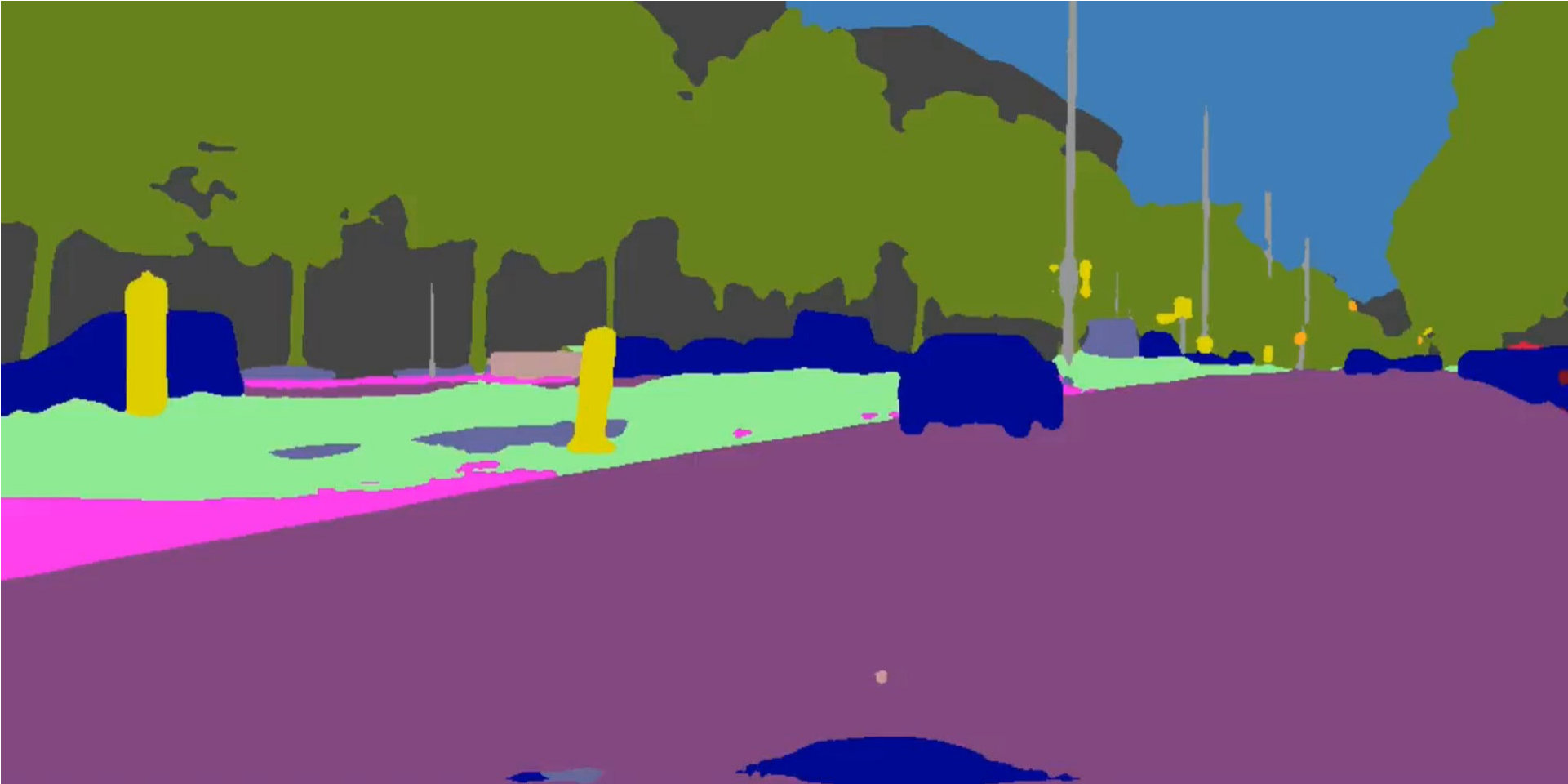
vid2vid Results

- Semantic → Street view scenes
- Edges → Human faces
- Poses → Human bodies

vid2vid Results

- Semantic → Street view scenes

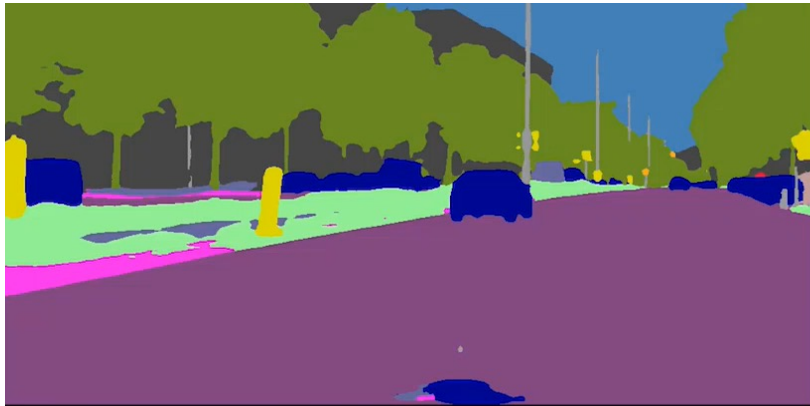
Street View: Cityscapes



Street View: Cityscapes



Street View: Cityscapes



Labels



pix2pixHD

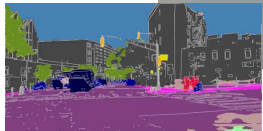


COVST



Ours

Street View: Boston



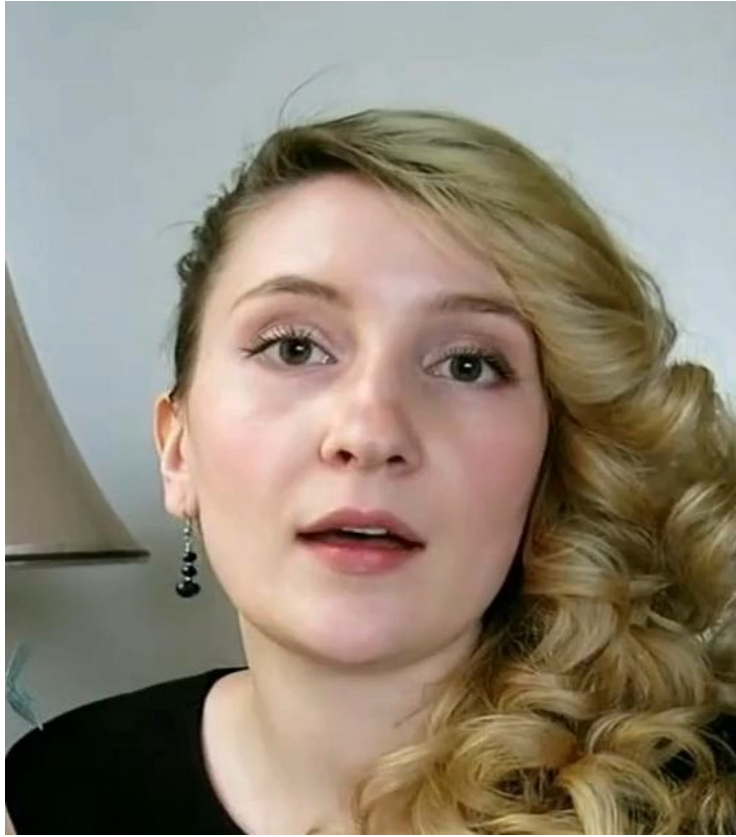
Street View: NYC



Results

- Edges → Human faces

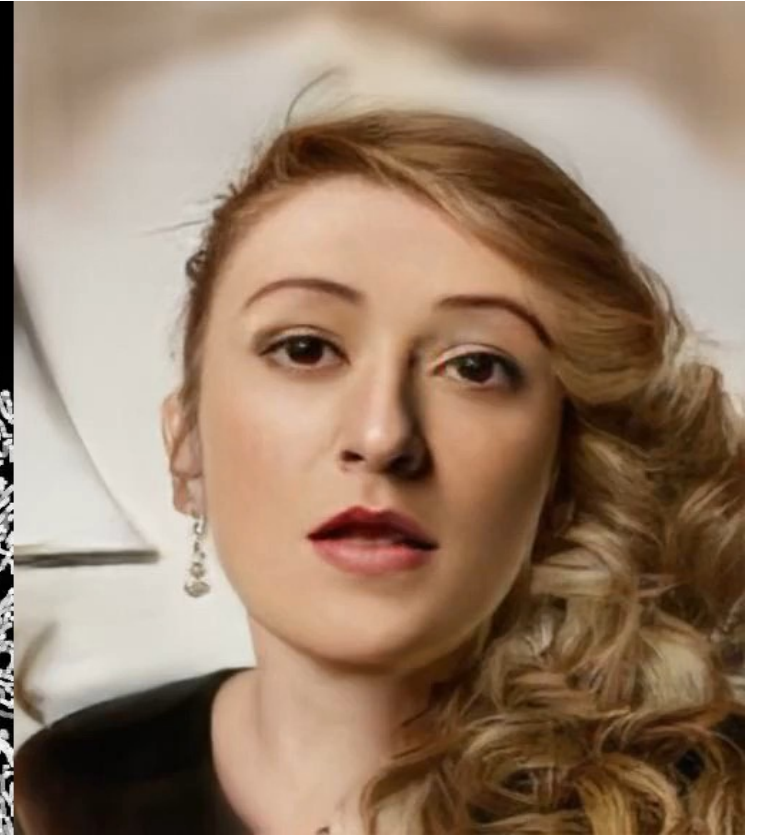
Face Swapping (face $\rightarrow$ edge $\rightarrow$ face)



input

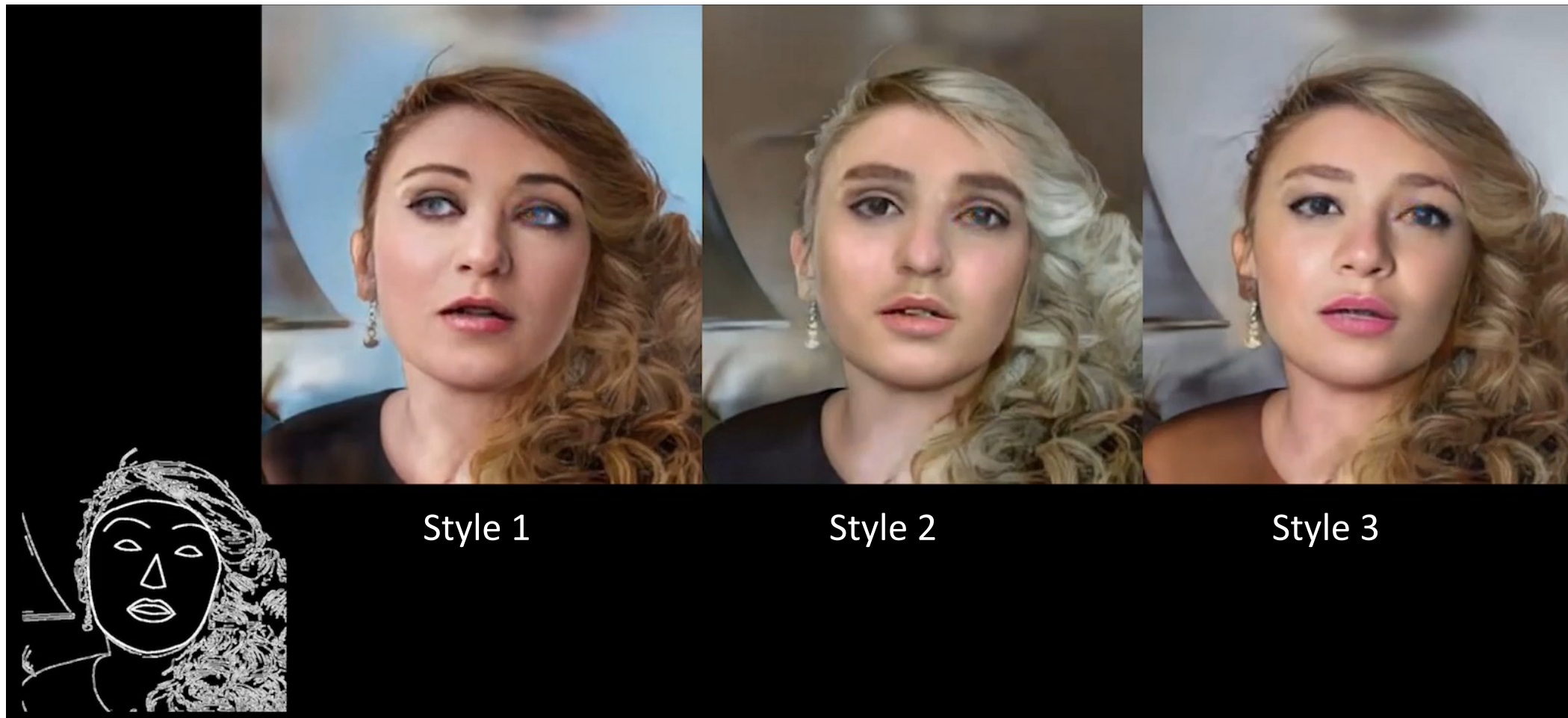


edges



output

Multi-modal Edge $\rightarrow$ Face



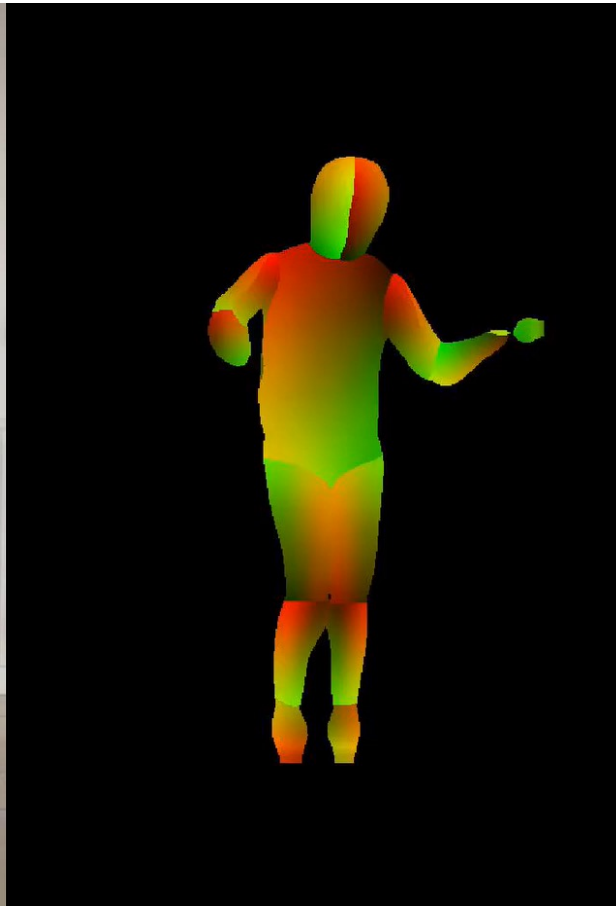
Results

- Poses → Human bodies

Motion Transfer (body $\rightarrow$ pose $\rightarrow$ body)



input



poses

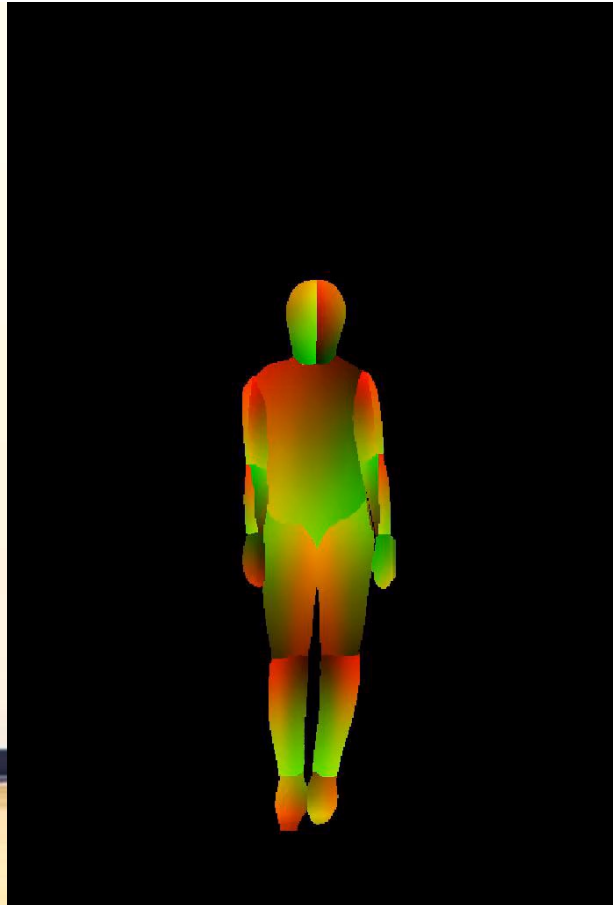


output

Motion Transfer (body $\rightarrow$ pose $\rightarrow$ body)



input



poses

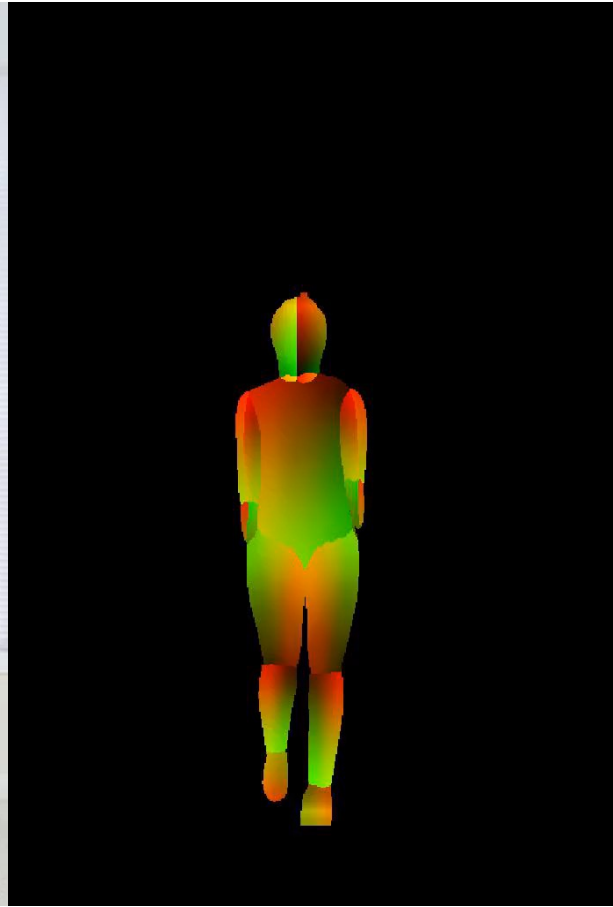


output

Motion Transfer (body $\rightarrow$ pose $\rightarrow$ body)



input



poses

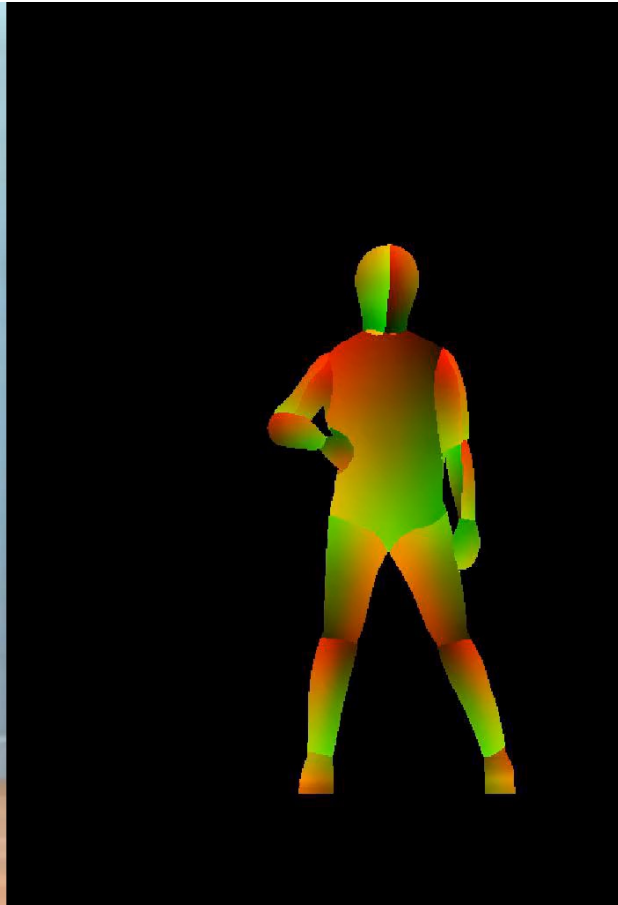


output

Motion Transfer (body $\rightarrow$ pose $\rightarrow$ body)



input



poses



output

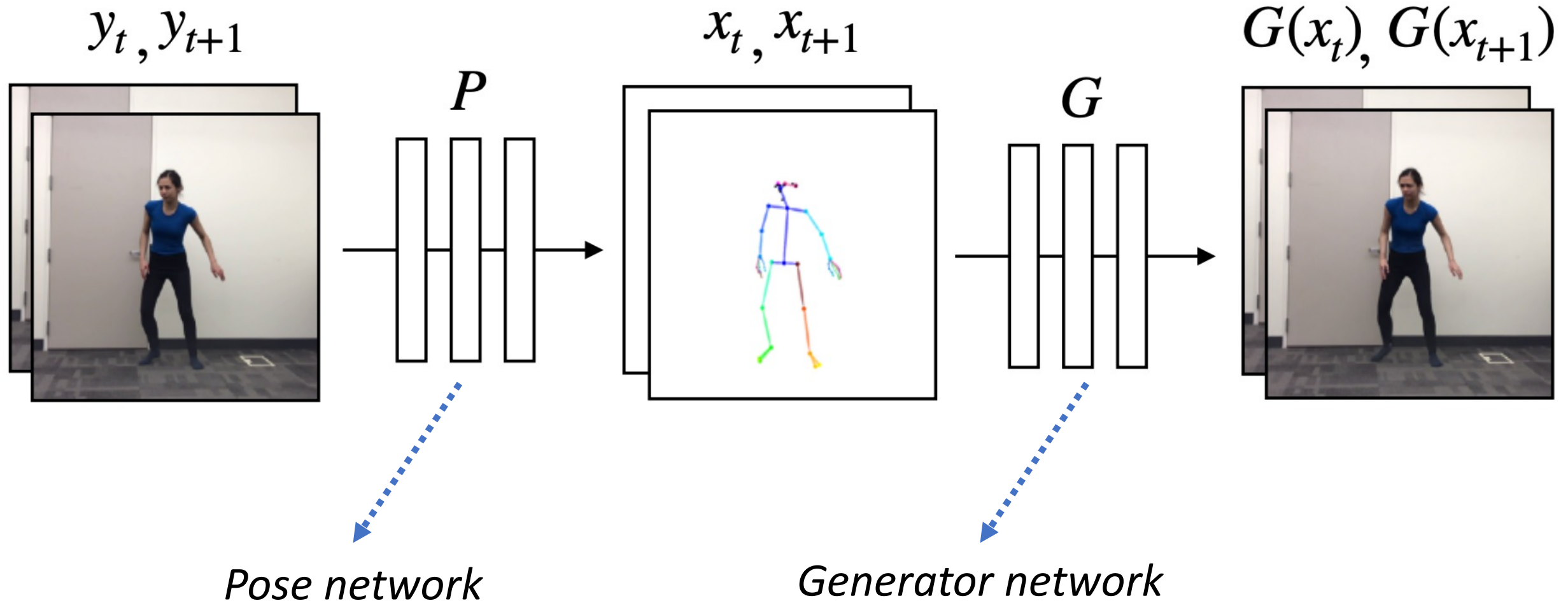
More Dancing ...

Source Subject

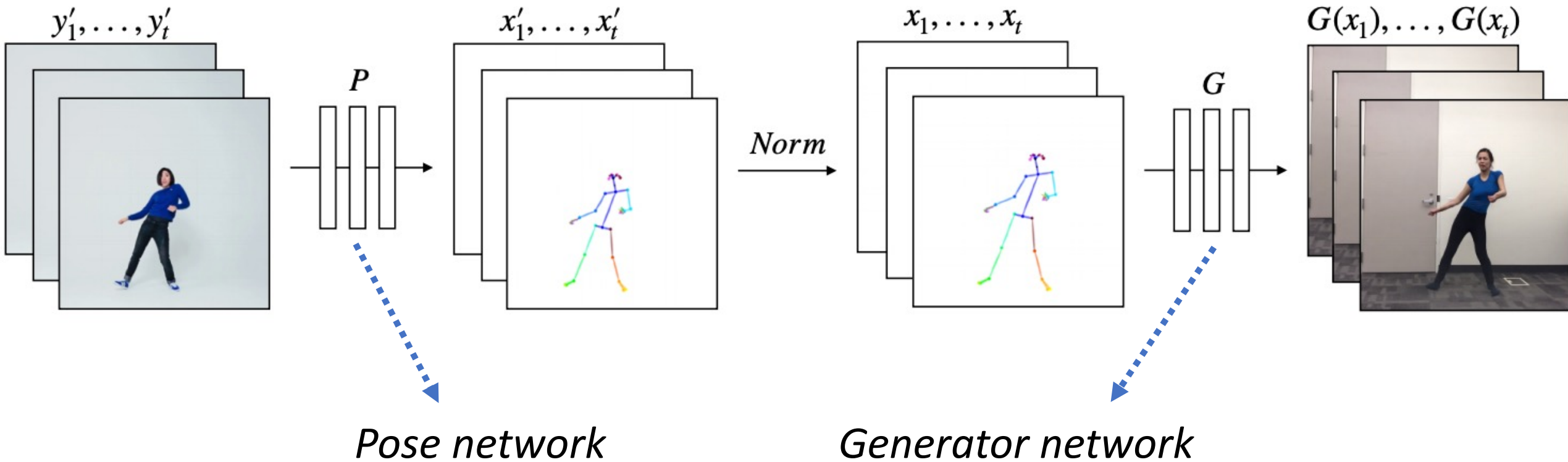
*Challenging due to missed detections



Pose-guided Synthesis

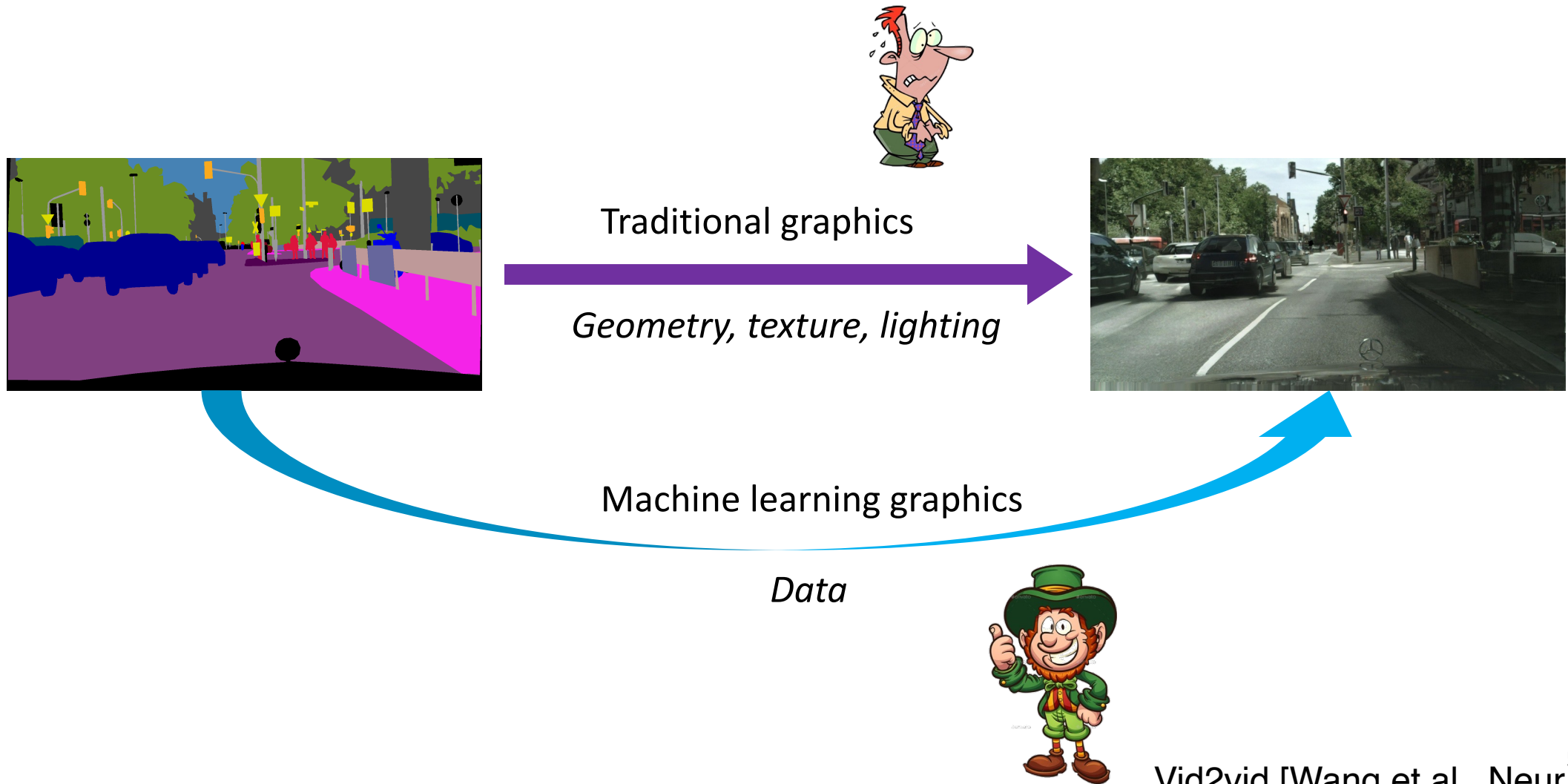


Transfer Phase





ML-based Rendering



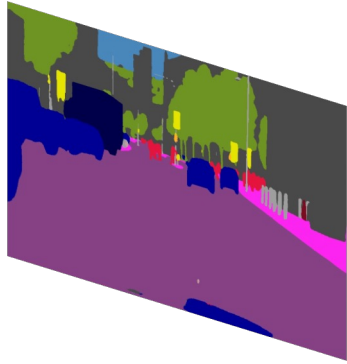
vid2vid Extensions: Interactive Graphics

User control

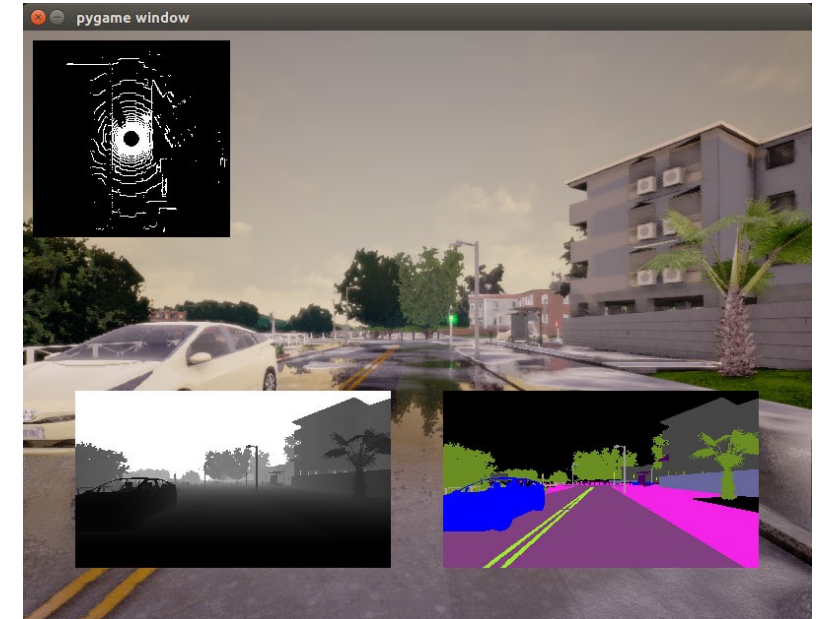
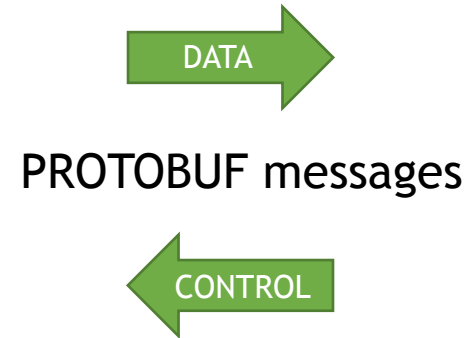
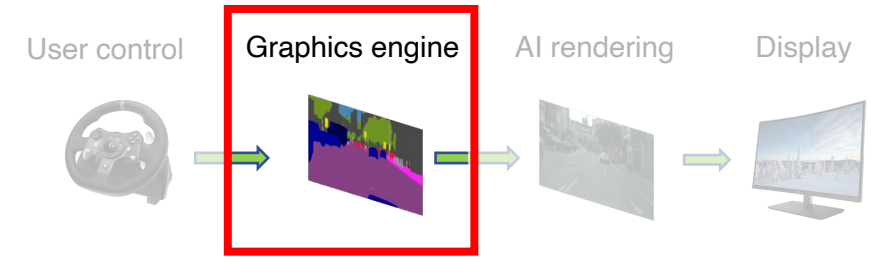
Graphics engine

AI rendering

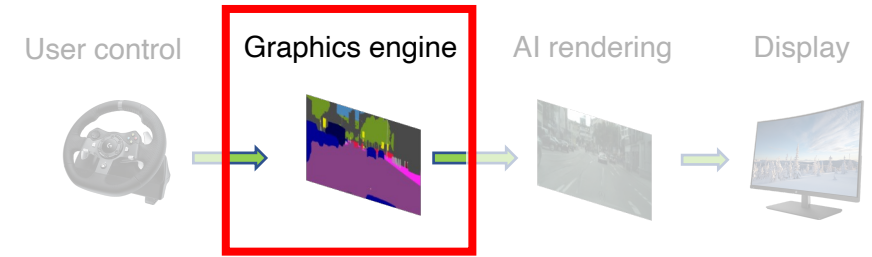
Display



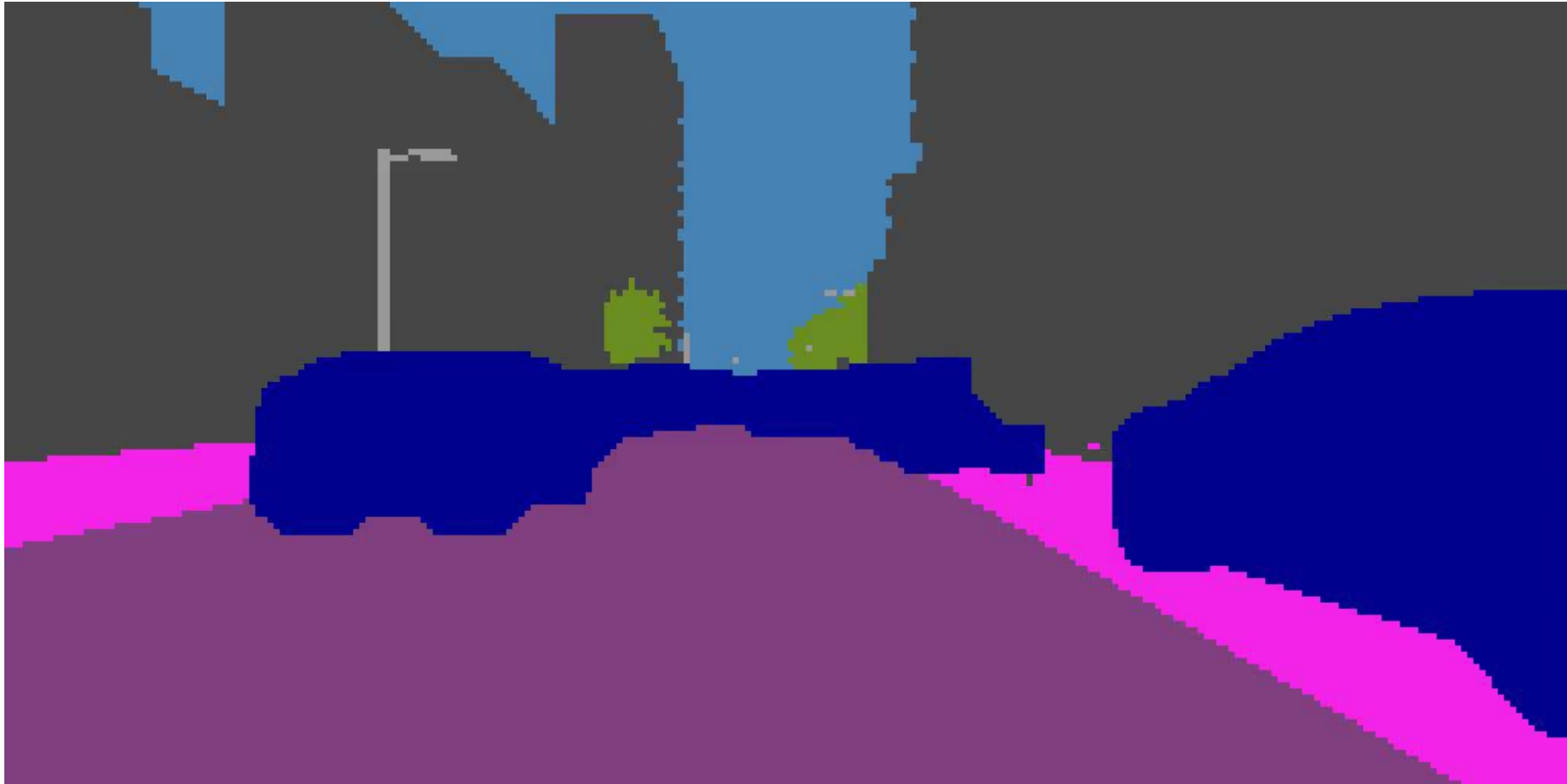
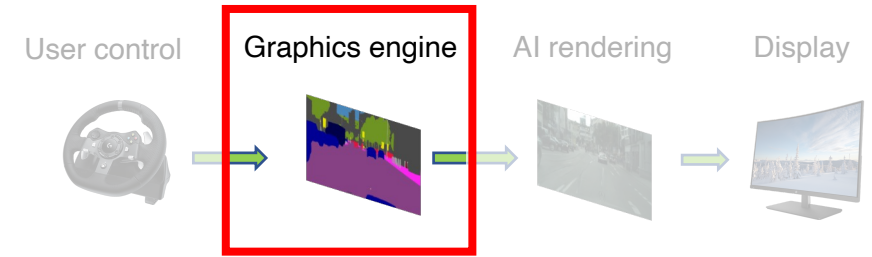
Graphics Engine: CARLA



Original CARLA Sequence

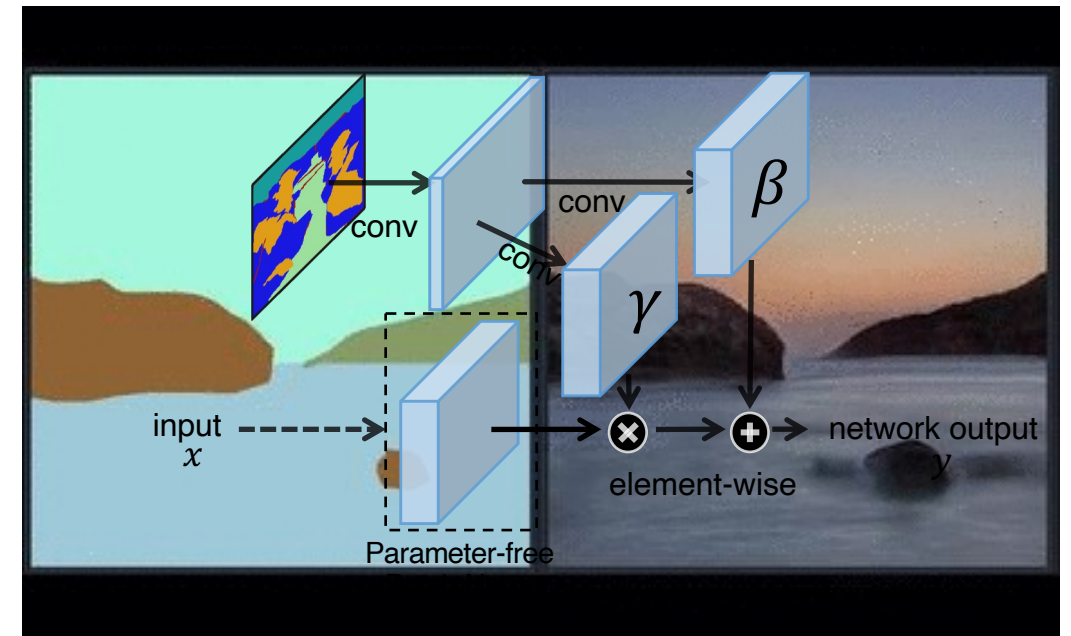
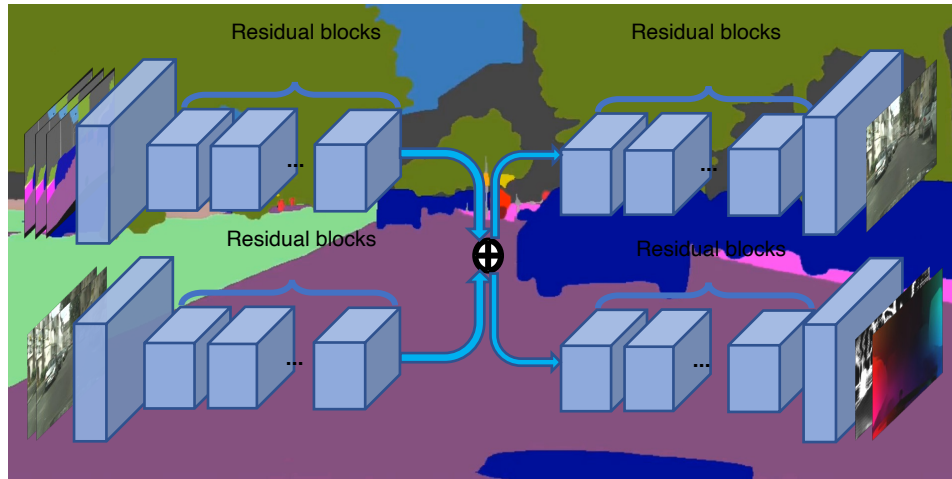
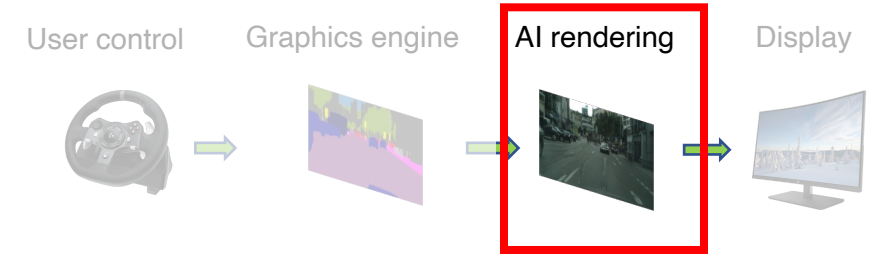


CARLA Semantic Maps

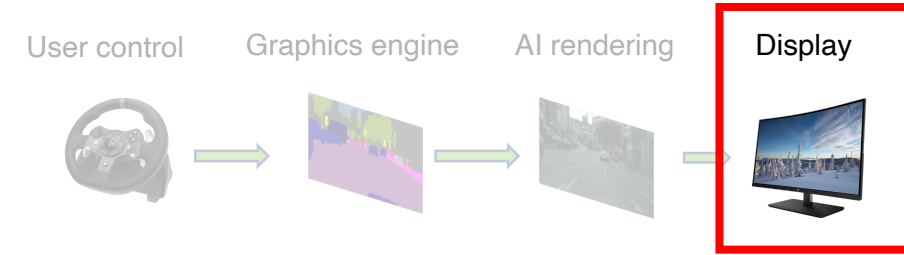


Methodology

- Combine vid2vid with SPADE



Demo Result



Driving Game



vid2game (playable videos)

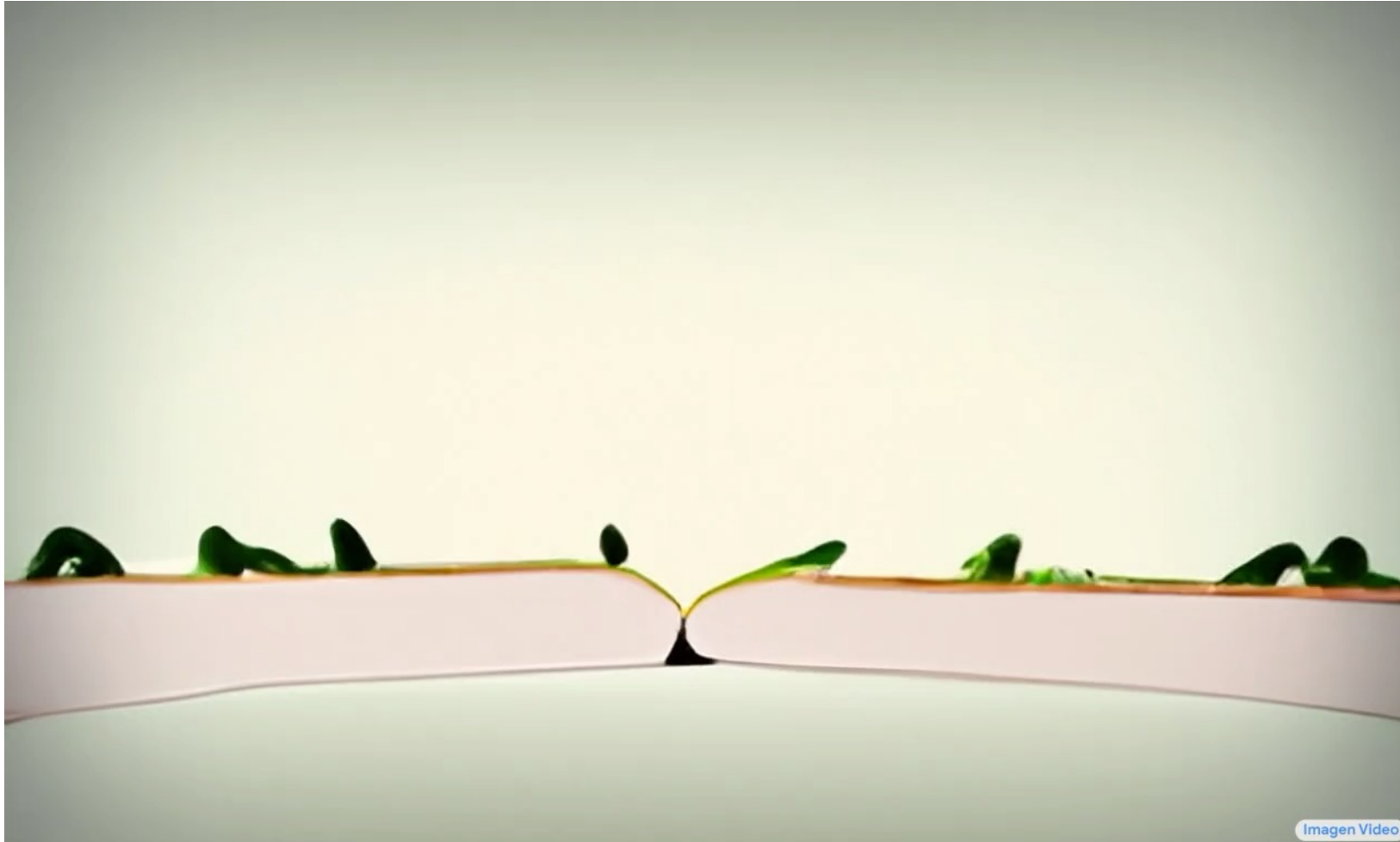
O. Gafni, L. Wolf, Y. Taigman. "Vid2Game: Controllable Characters Extracted from Real-World Videos," 2019

Target Video



Video Diffusion Models

Imagen Video (Pixel Diffusion Videos)



Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

Imagen Video (Pixel Diffusion Videos)



An astronaut riding a horse

Imagen Video (Pixel Diffusion Videos)

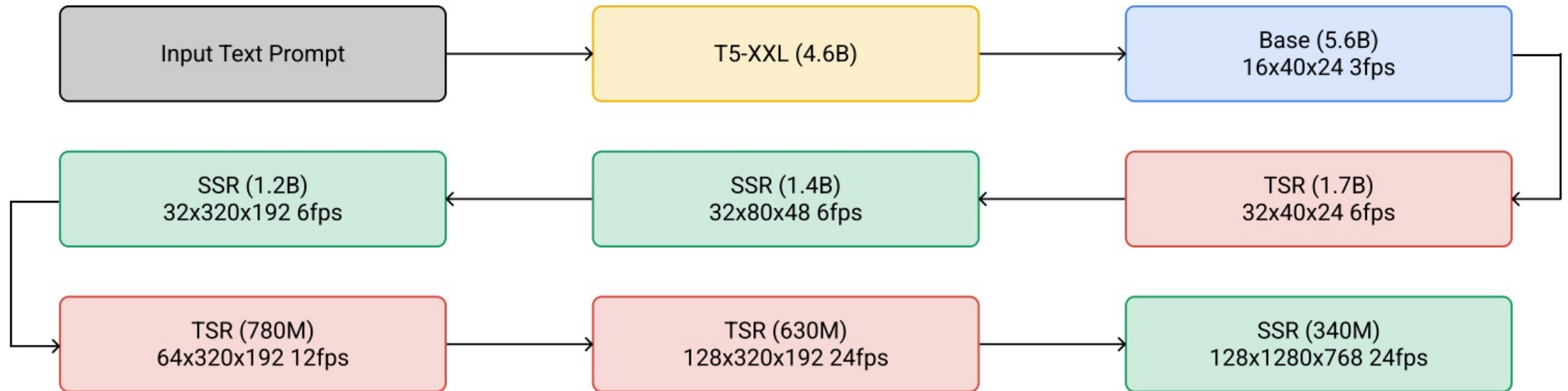


Imagen Video: High Definition Video Generation with Diffusion Models

Jonathan Ho et al., 2022

Imagen Video (Pixel Diffusion Videos)

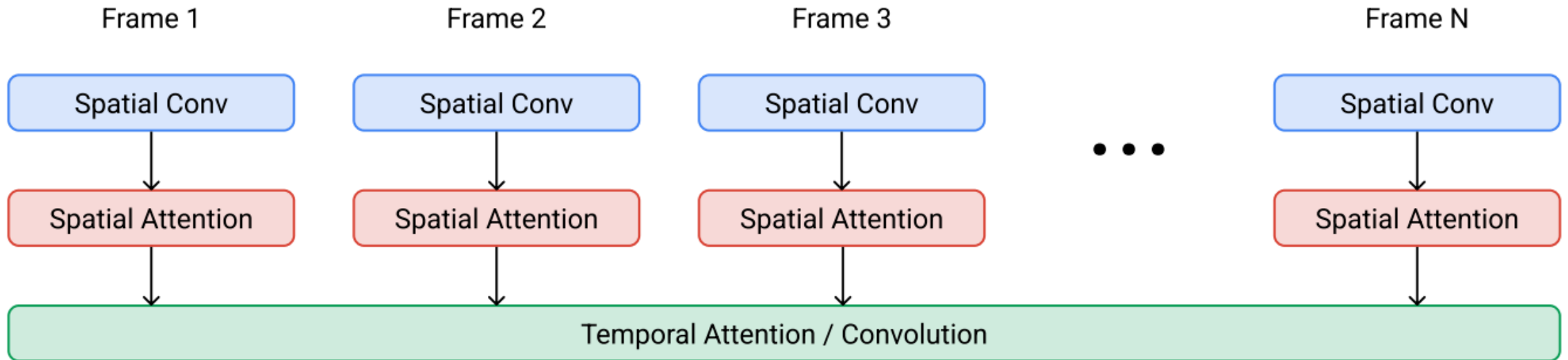
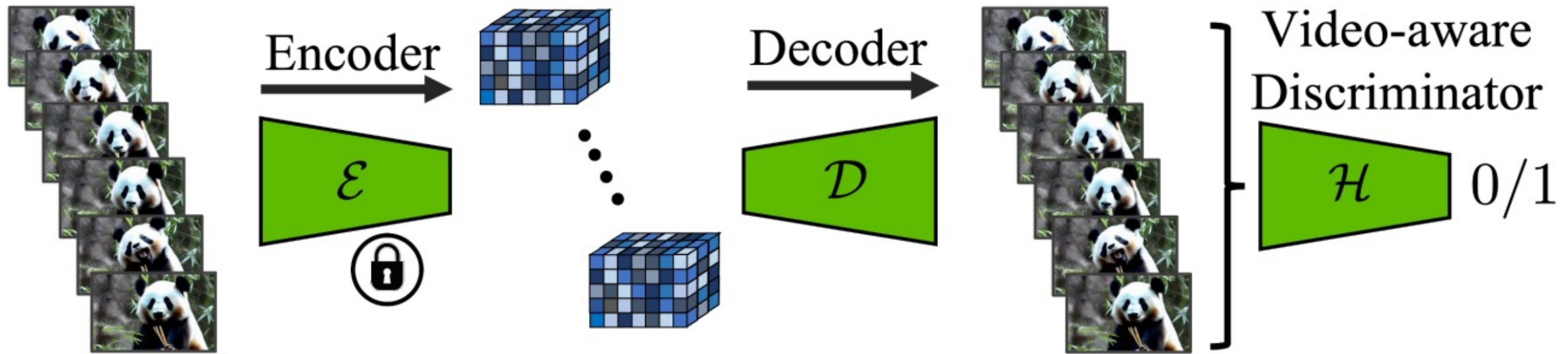


Imagen Video: High Definition Video Generation with Diffusion Models

Jonathan Ho et al., 2022

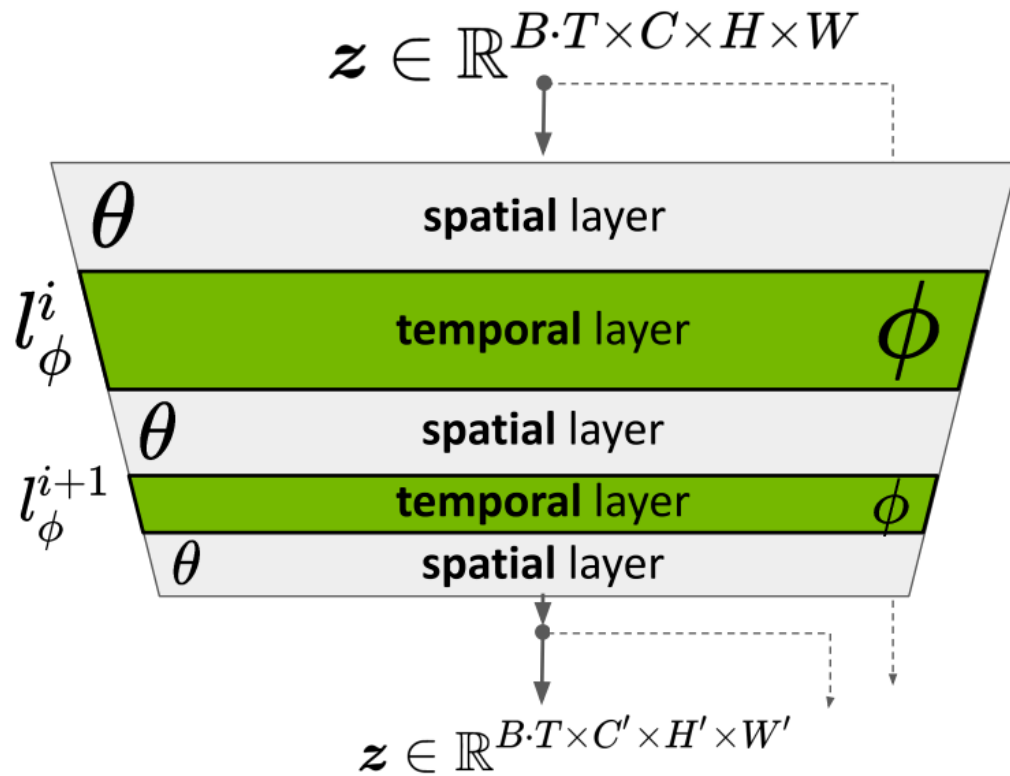
Video LDM (Stable Diffusion Video)



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

Andreas Blattmann, et al., 2023

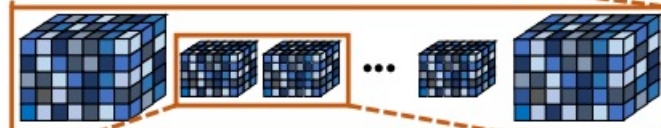
Video LDM (Stable Diffusion Video)



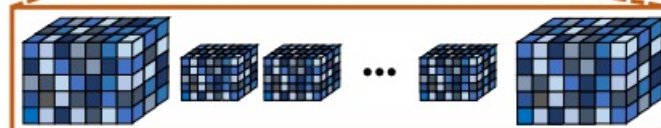
1. Generate Latent Key Frames
(optionally including prediction model)



2. Latent Frame Interpolation I



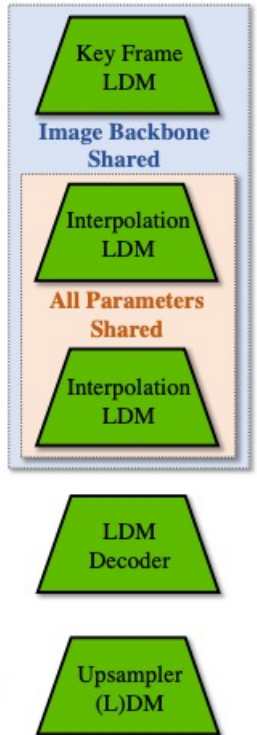
3. Latent Frame Interpolation II



4. Decode to Pixel-Space



5. Apply Video Upsampler



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

Andreas Blattmann, et al., 2023

Video LDM (Stable Diffusion Video)



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

Andreas Blattmann, et al., 2023

Video LDM (Stable Diffusion Video)

Pros:

- Reuse the image models (encoder, decoder, and Diffusion)
- Generate both images and videos

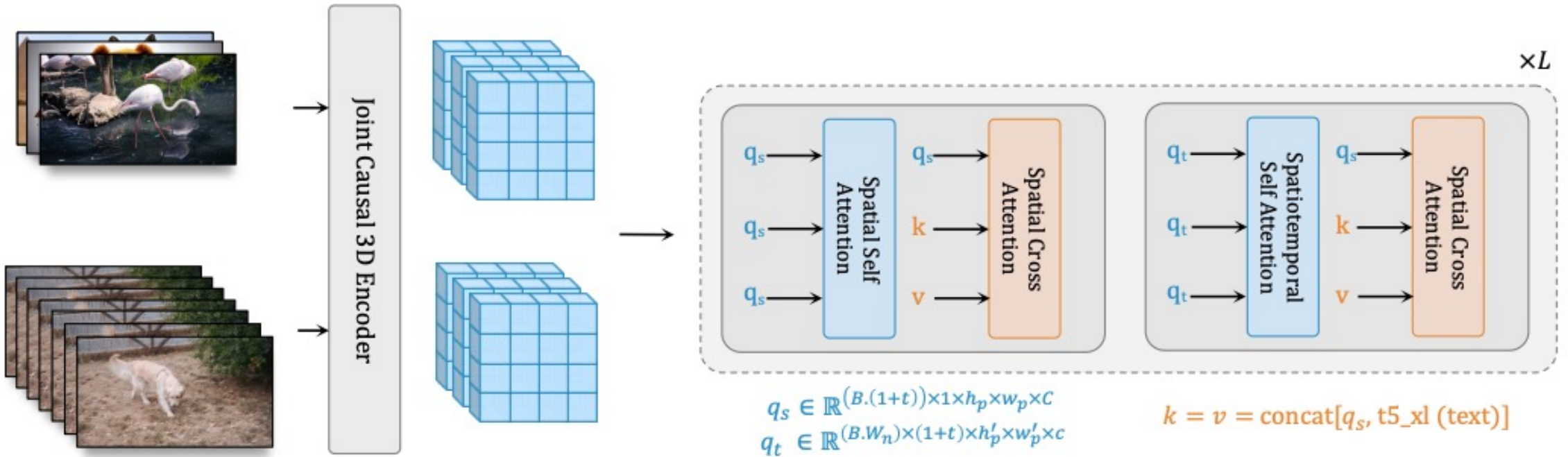
Cons:

- Per-frame Image VAE Encoder (no temporal redundancy)
- Causing flickering artifacts.

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

Andreas Blattmann, et al., 2023

W.A.L.T: 3D Casual CNN Tokenizer + DiT

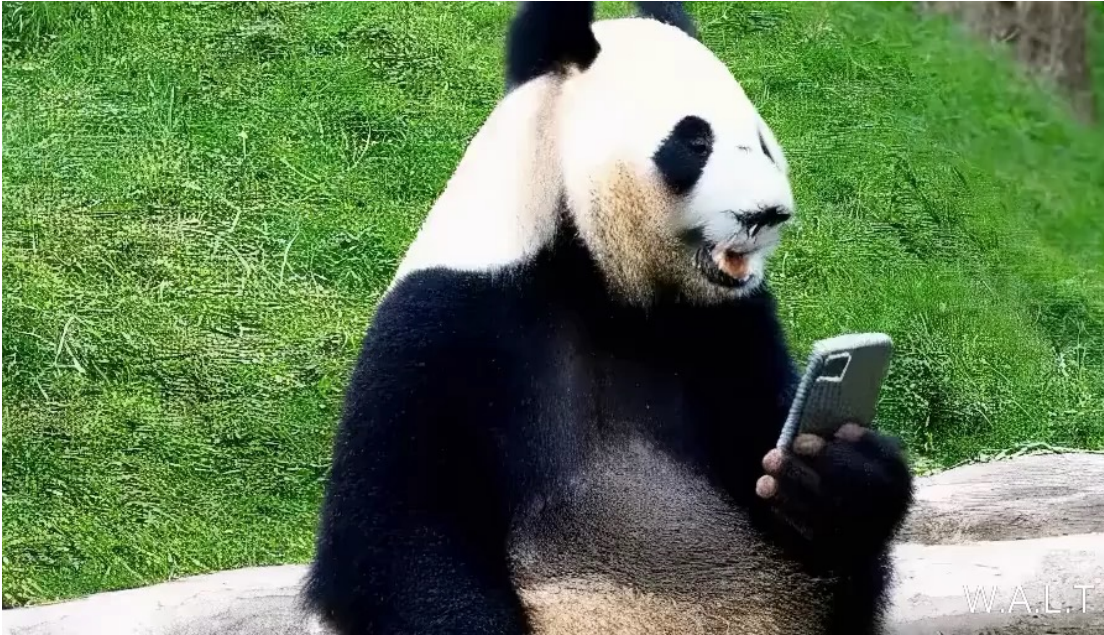


- Causal 3D CNN layers: CNN kernel operates on only the past frames
- Trained on videos (UCF-101 and Kinetics-600) and image (ImageNet)

Photorealistic Video Generation with Diffusion Models

Agrim Gupta, et al., 2023

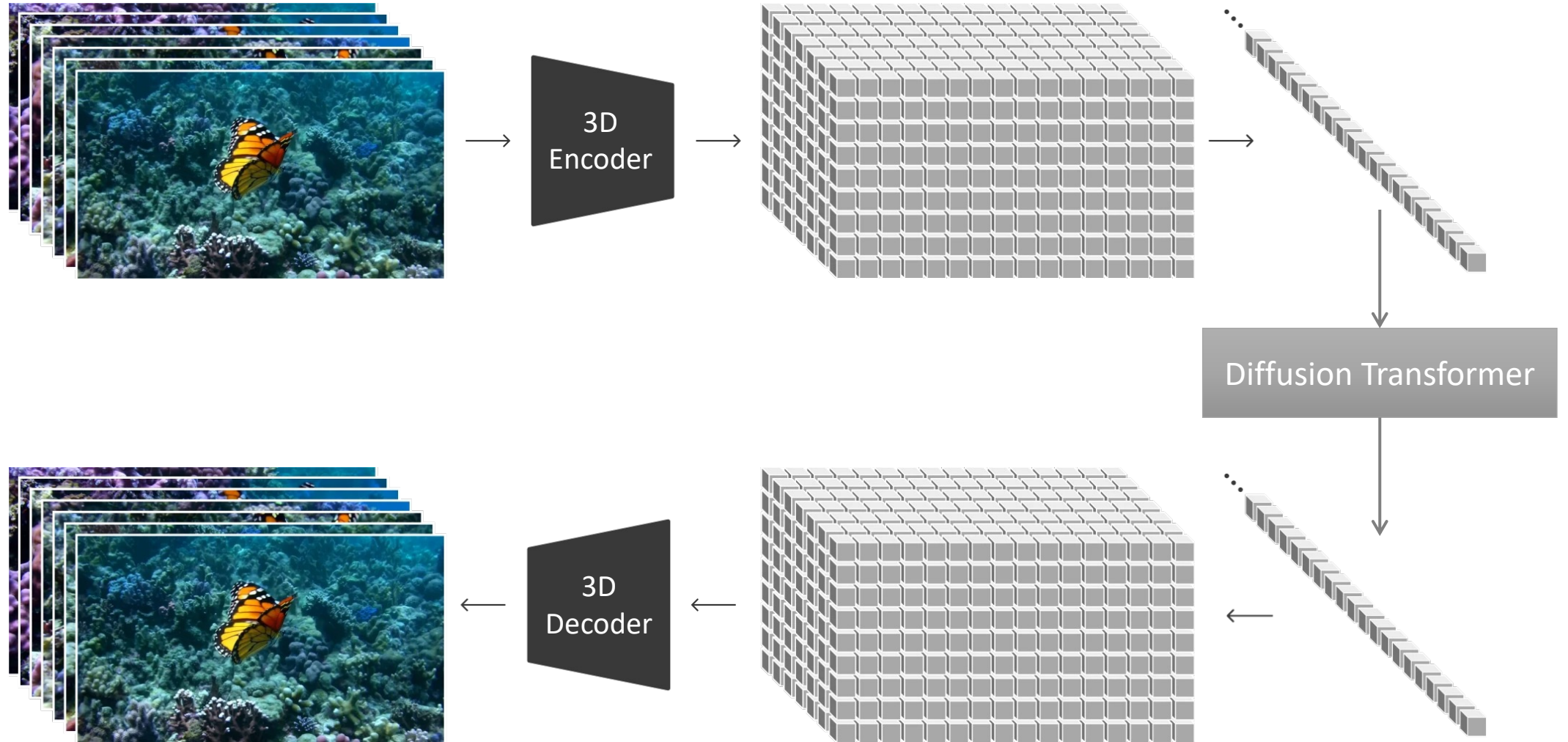
W.A.L.T: 3D Casual CNN Tokenizer + DiT



Photorealistic Video Generation with Diffusion Models

Agrim Gupta, et al., 2023

SORA: 3D Encoder/decoder + DiT



Video Generation Models as World Simulators

Tim Brooks*, Bill Peebles*, et al., 2023

SORA: 3D Encoder/decoder + DiT



SORA: 3D Encoder/decoder + DiT



A Shiba Inu dog wearing a beret and black turtleneck.

SORA: 3D Encoder/decoder + DiT



An image of a realistic cloud that spells “SORA”.

Thank You!



16-726, Spring 2025

<https://learning-image-synthesis.github.io/>