Temporal Image-to-Image Generation

Presenter: Sakura
Contact me: bili_sakura@zju.edu.cn
Date: August 19, 2024

Outline

- Problem Definition
- Image Editing
  - SDEdit
  - CycleDiffusion
  - InstructPix2Pix
- Image-to-Video Generation
  - SVD
  - Video LDM

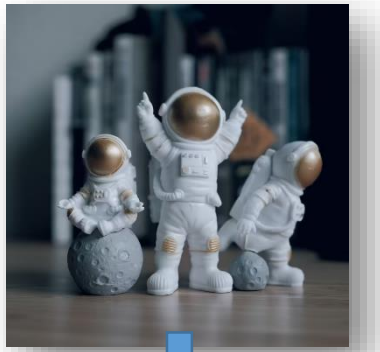
Problem Definition

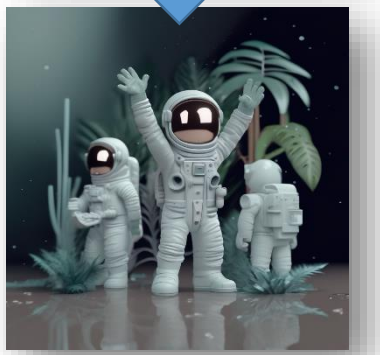
Generative Tasks

Text-to-Image

Astronaut in a jungle, cold color palette, muted colors, detailed, 8k

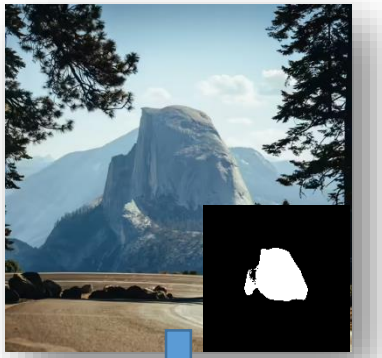
Image-to-Image

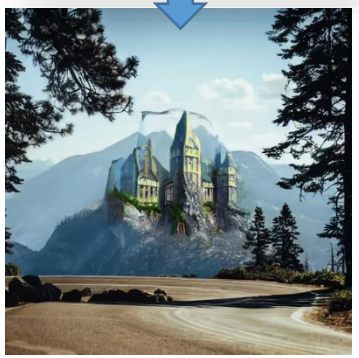




Astronaut in a jungle, cold color palette, muted colors, detailed, 8k

Inpainting





concept art digital painting of an elven castle, inspired by lord of the rings, highly detailed, 8k

Text/Image-to-Video

fps=7

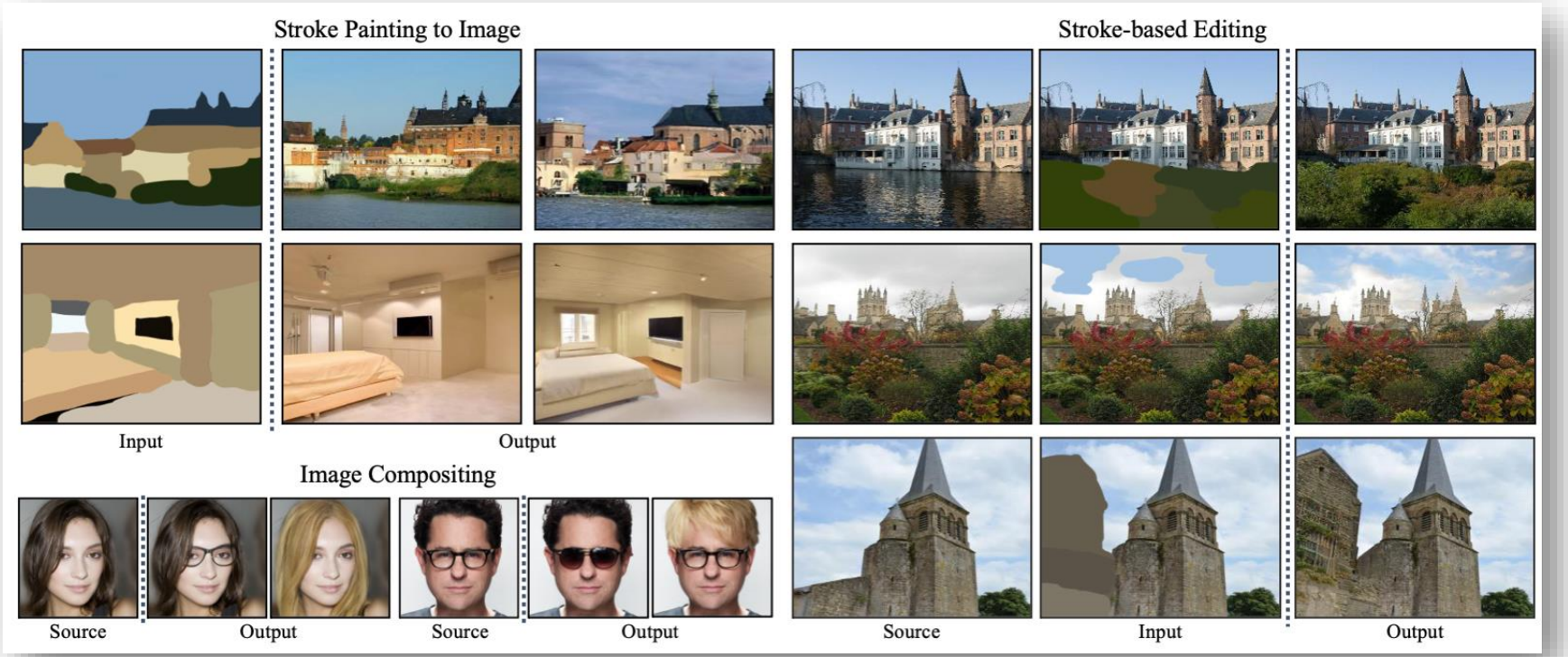
All examples derive from huggingface*. From left to right, images are generated by SD 1.5 [2], SD 1.5, SD 1.5 Inpainting and SVD XT respectively.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, 'High-Resolution Image Synthesis With Latent Diffusion Models', presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.

Image Editing

Stochastic Differential Editing (SDEdit)

- Training-free
- Integrated with SDE-based generative model (e.g. Stable Diffusion)



Stochastic Differential Editing (SDEdit) is a **unified** image synthesis and editing framework based on stochastic differential equations. SDEdit allows stroke painting to image, image compositing, and stroke-based editing **without** task-specific model training and loss functions [1].

[1] C. Meng et al., 'SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations', presented at the International Conference on Learning Representations, Oct. 2021.

Stochastic Differential Editing (SDEdit)

- Training-free
- Integrated with SDE-based generative model (e.g. Stable Diffusion)

initial image

generated image

Stable Diffusion v1.5 [2] with SDEdit [1] technique for (better?) image editing.

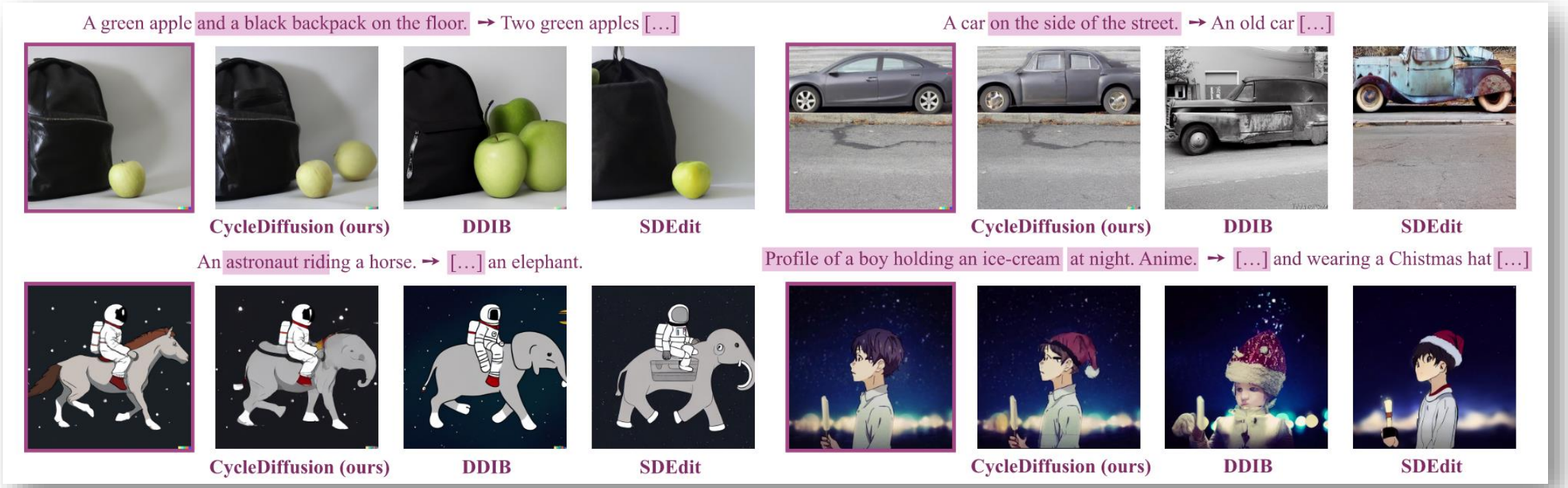
(Example is accessed from huggingface.co/docs/diffusers.*)

[1] C. Meng et al., 'SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations', presented at the International Conference on Learning Representations, Oct. 2021

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, 'High-Resolution Image Synthesis With Latent Diffusion Models', presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.

CycleDiffusion

- Training-free
- Integrated with SDE-based generative model (e.g. Stable Diffusion)



Examples of CycleDiffusion [3] for zero-shot image editing. Within each pair of source and target texts, overlapping text spans are marked in purple in the source text and abbreviated as [. . .] in the target text. Visual comparison to the baselines, DDIB [4] and SDEdit [1].

[1] C. Meng et al., 'SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations', presented at the International Conference on Learning Representations, Oct. 2021
[3] C. H. Wu and F. De La Torre, 'A Latent Space of Stochastic Diffusion Models for Zero-Shot Image Editing and Guidance', in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France: IEEE, Oct. 2023, pp. 7344–7353. doi: 10.1109/ICCV51070.2023.00678.
[4] X. Su, J. Song, C. Meng, and S. Ermon, 'Dual Diffusion Implicit Bridges for Image-to-Image Translation', presented at the The Eleventh International Conference on Learning Representations, Sep. 2022.

InstructPix2Pix

- Fine-tuned on SD 1.5
- InstructPix2Pix checkpoint is accessible here

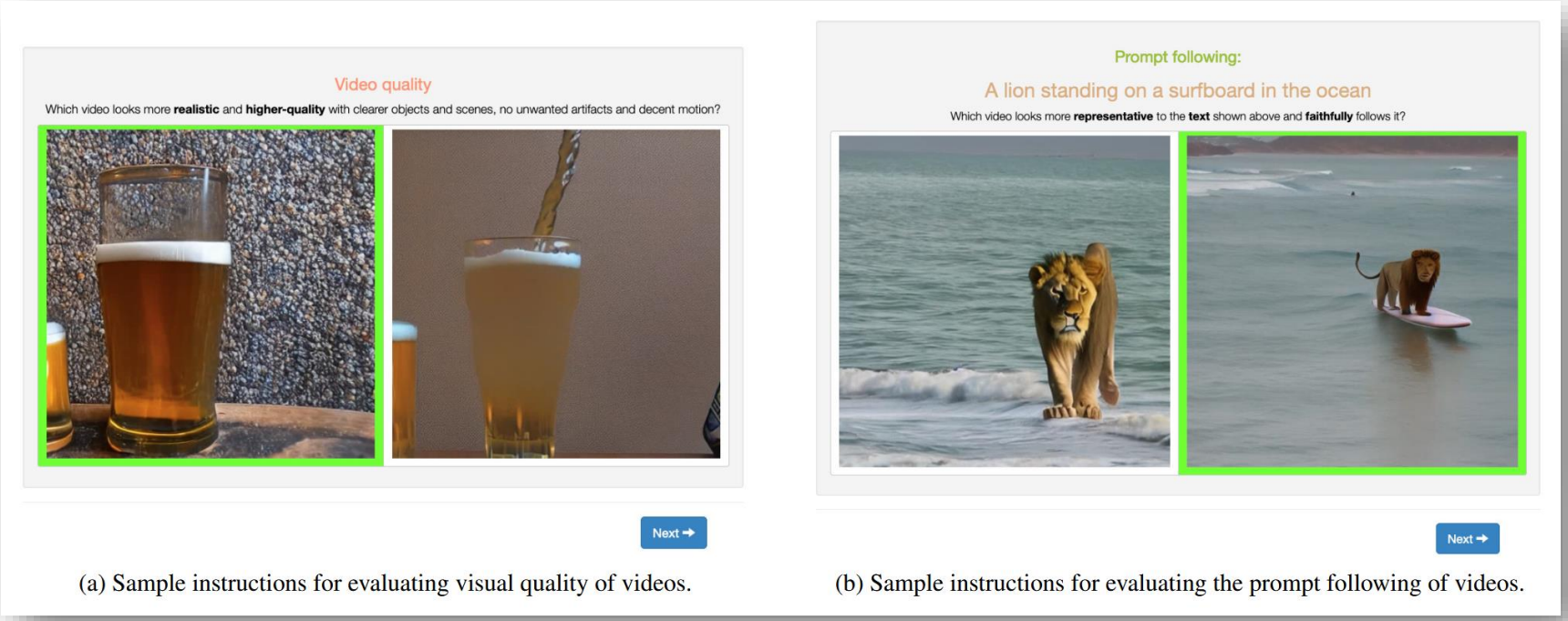
Given an image and an instruction for how to edit that image, our model [5] performs the appropriate edit. Our model does not require full descriptions for the input or output image, and edits images in the forward pass without per-example inversion or fine-tuning.

[5] T. Brooks, A. Holynski, and A. A. Efros, 'InstructPix2Pix: Learning To Follow Image Editing Instructions', presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18392–18402.

Image-to-Video Generation

Stable Video Diffusion (SVD)

- Developed by Stability AI
- Start from SD 2.1 [8] checkpoint



(a) Sample instructions for evaluating visual quality of videos.

(b) Sample instructions for evaluating the prompt following of videos.

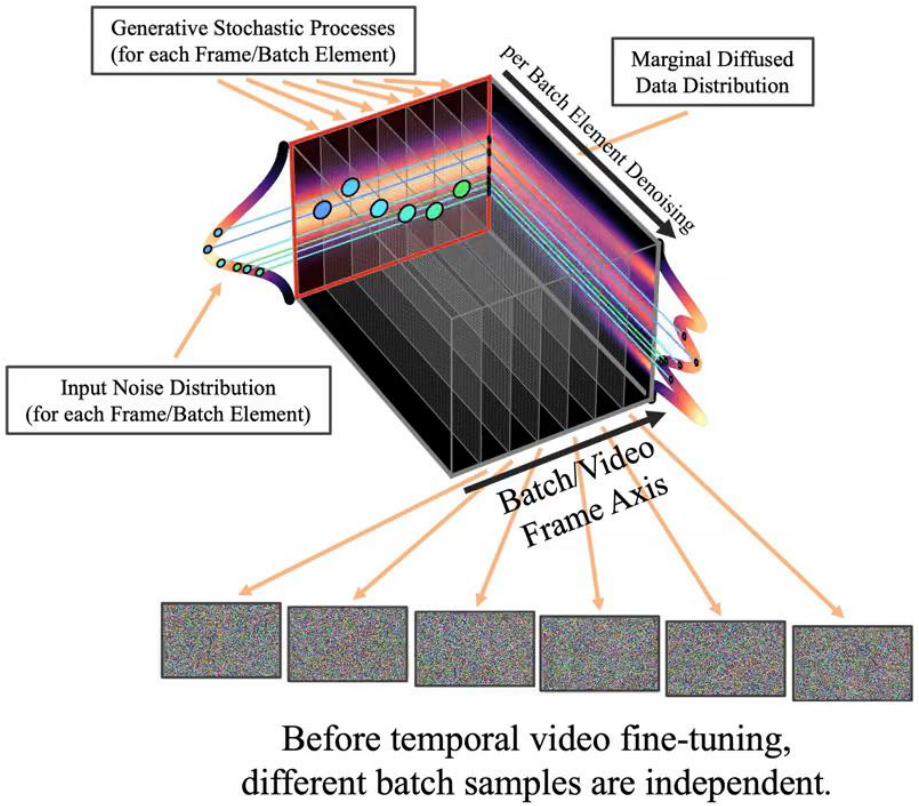
Left: Our human evaluation framework, as seen by the annotators. The prompt & task order and model choices are fully randomized. Right: Image-to-video examples. [6]

[6] A. Blattmann et al., 'Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets', Nov. 25, 2023, arXiv: arXiv:2311.15127. doi: 10.48550/arXiv.2311.15127.

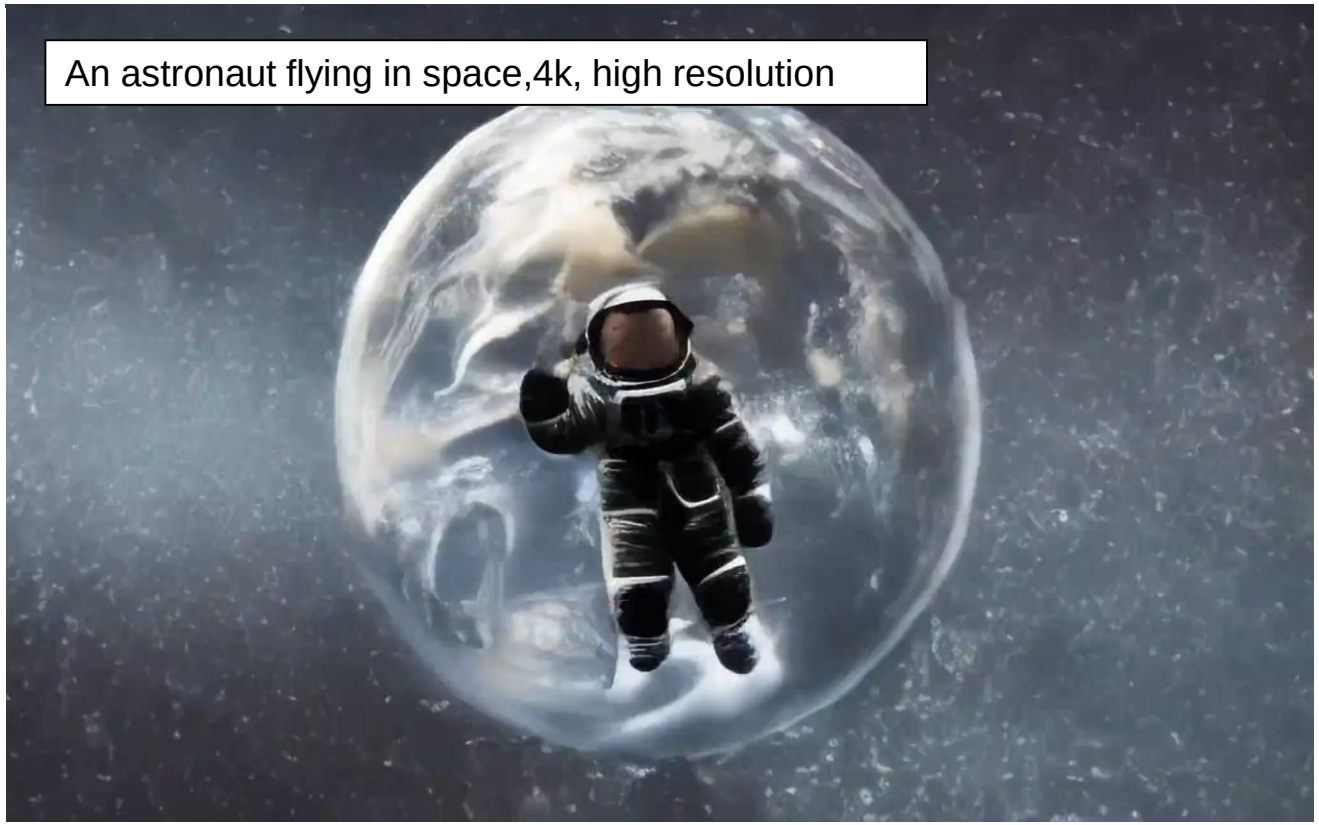
[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, 'High-Resolution Image Synthesis With Latent Diffusion Models', presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.

Video LDM

- Developed by NVIDIA



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

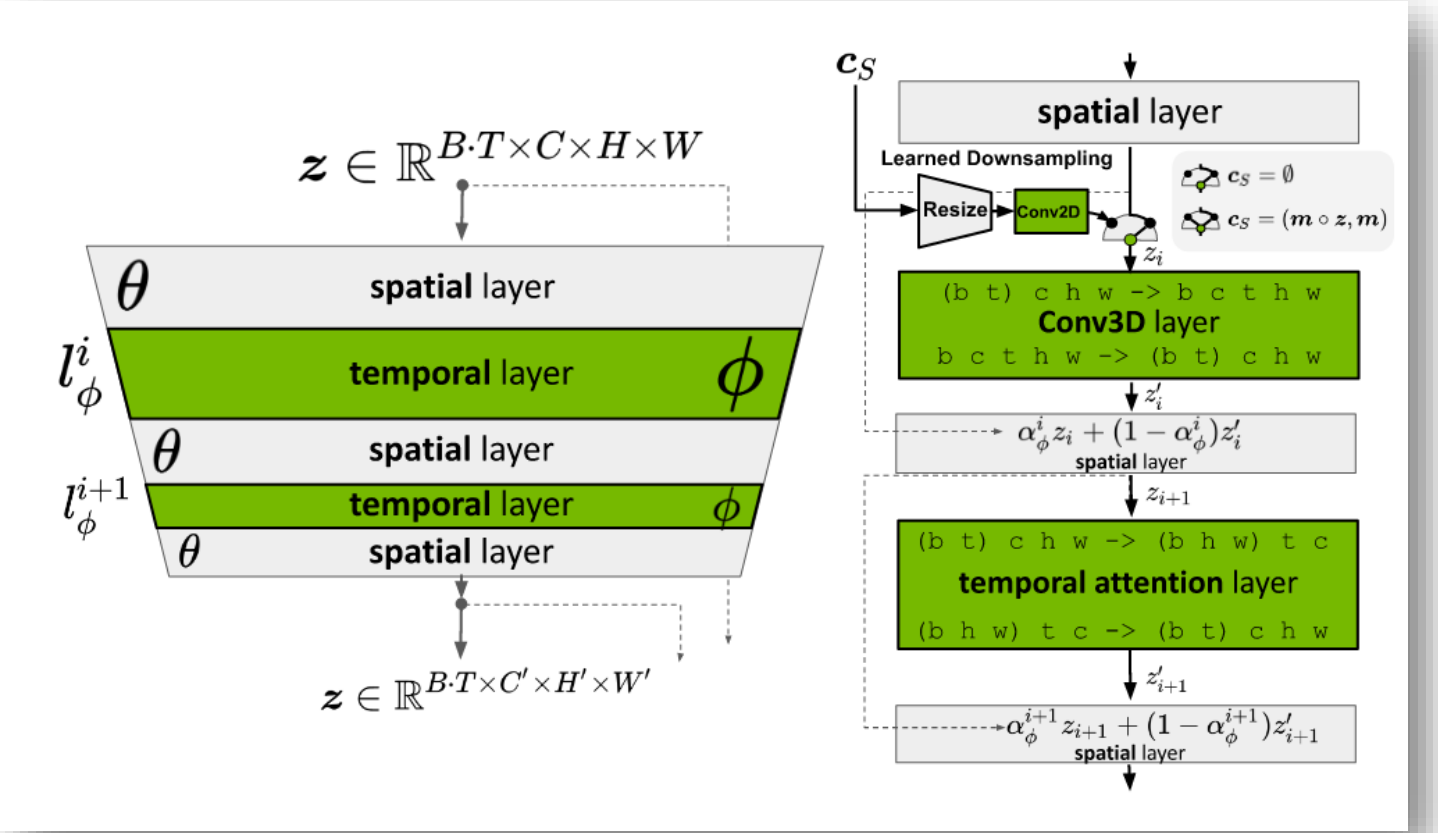


Left: Animation of temporal video fine-tuning in our Video Latent Diffusion Models (Video LDMs). Right: A example from Video LDMs.

[7] A. Blattmann et al., 'Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models', in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 22563–22575. doi: 10.1109/CVPR52729.2023.02161.

Video LDM

- Developed by NVIDIA



Temporal block design in Video LDM.

Left: We 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 l^i_ϕ are trained. Right: During training, the base model θ interprets the input sequence of length T as a batch of images. For the temporal layers l^i_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 (a^i_ϕ =1) yields the original image model. For illustration purposes, only a single U-Net Block is shown. B denotes batch size, T sequence length, C input channels and H and W the spatial dimensions of the input. c_s is optional context frame conditioning, when training prediction models.

[7] A. Blattmann et al., 'Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models', in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 22563–22575. doi: 10.1109/CVPR52729.2023.02161.

References and Related Works

[1] C. Meng et al., ‘SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations’, presented at the International Conference on Learning Representations, Oct. 2021.
[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, ‘High-Resolution Image Synthesis With Latent Diffusion Models’, presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.
[3] C. H. Wu and F. De La Torre, ‘A Latent Space of Stochastic Diffusion Models for Zero-Shot Image Editing and Guidance’, in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France: IEEE, Oct. 2023, pp. 7344–7353. doi: 10.1109/ICCV51070.2023.00678.
[4] X. Su, J. Song, C. Meng, and S. Ermon, ‘Dual Diffusion Implicit Bridges for Image-to-Image Translation’, presented at the The Eleventh International Conference on Learning Representations, Sep. 2022.
[5] T. Brooks, A. Holynski, and A. A. Efros, ‘InstructPix2Pix: Learning To Follow Image Editing Instructions’, presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18392–18402.
[6] A. Blattmann et al., ‘Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets’, Nov. 25, 2023, arXiv: arXiv:2311.15127. doi: 10.48550/arXiv.2311.15127.
[7] A. Blattmann et al., ‘Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models’, in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 22563–22575. doi: 10.1109/CVPR52729.2023.02161.
[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, ‘High-Resolution Image Synthesis With Latent Diffusion Models’, presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.