Denoising Diffusion Models
A Generative Learning Big Bang

CVPR 2023 Tutorial

Part III: Applications on other domains

Jiaming Song

Chenlin Meng

Arash Vahdat

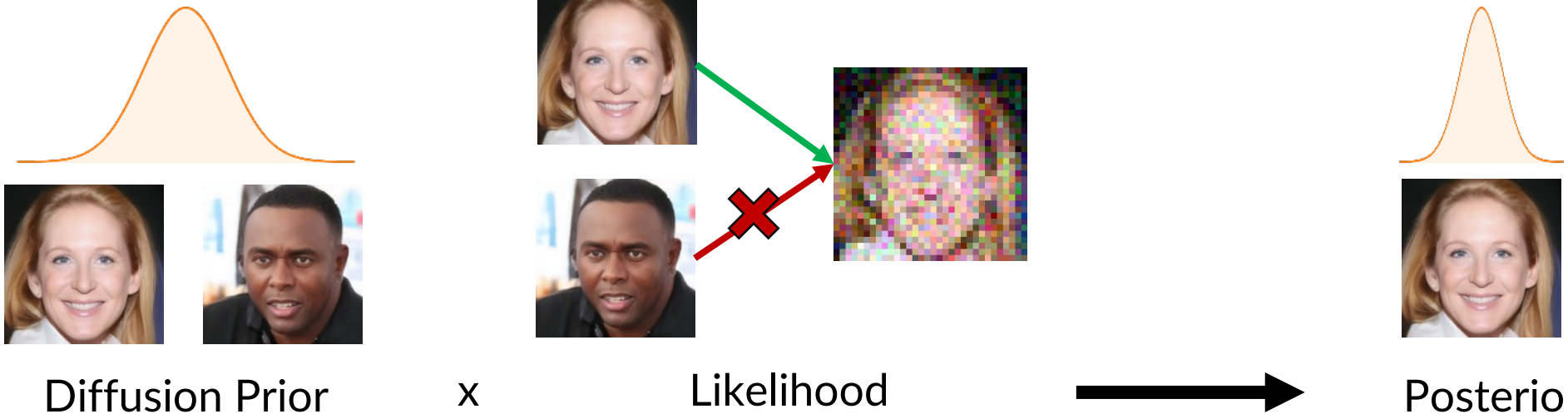
Outline

- Inverse problems
- 3D
- Video
- Miscellaneous

- Setup
- Replacement-based methods
- Reconstruction-based methods

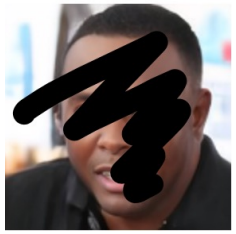
Diffusion Models for Inverse Problems

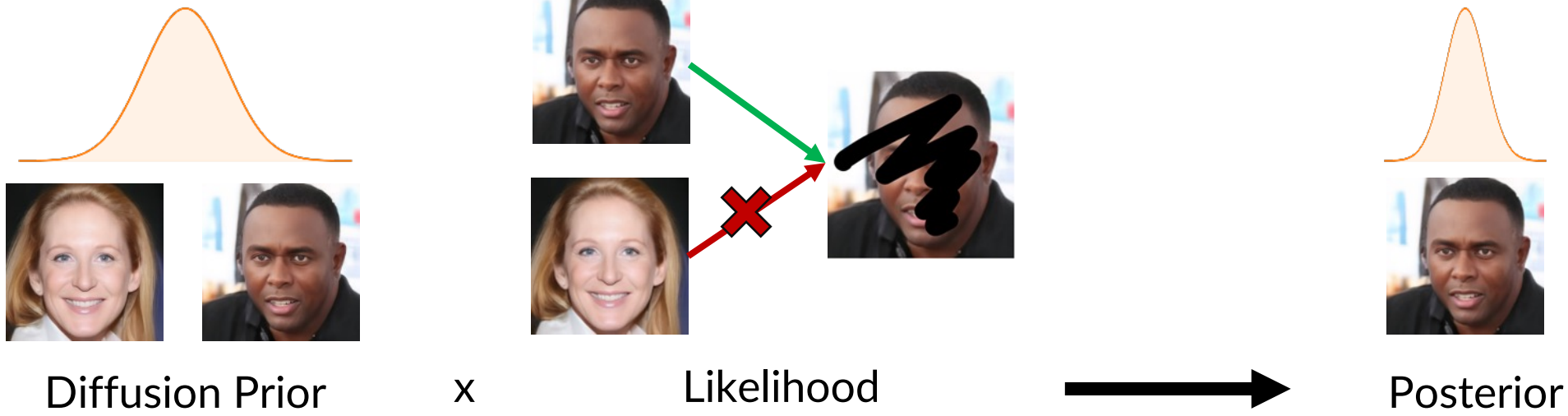
Goal: denoise and super-resolve an image



Diffusion Models for Inverse Problems

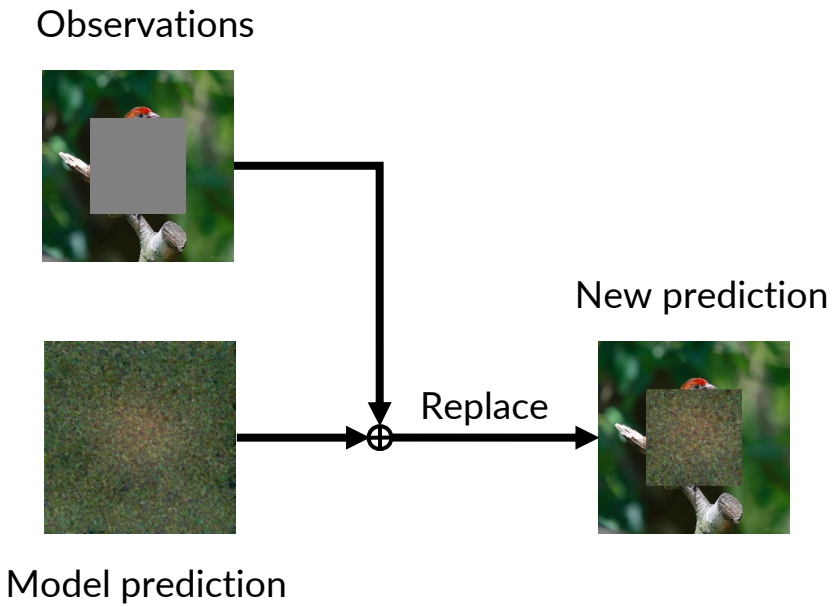
Goal: recover the masked region of an image

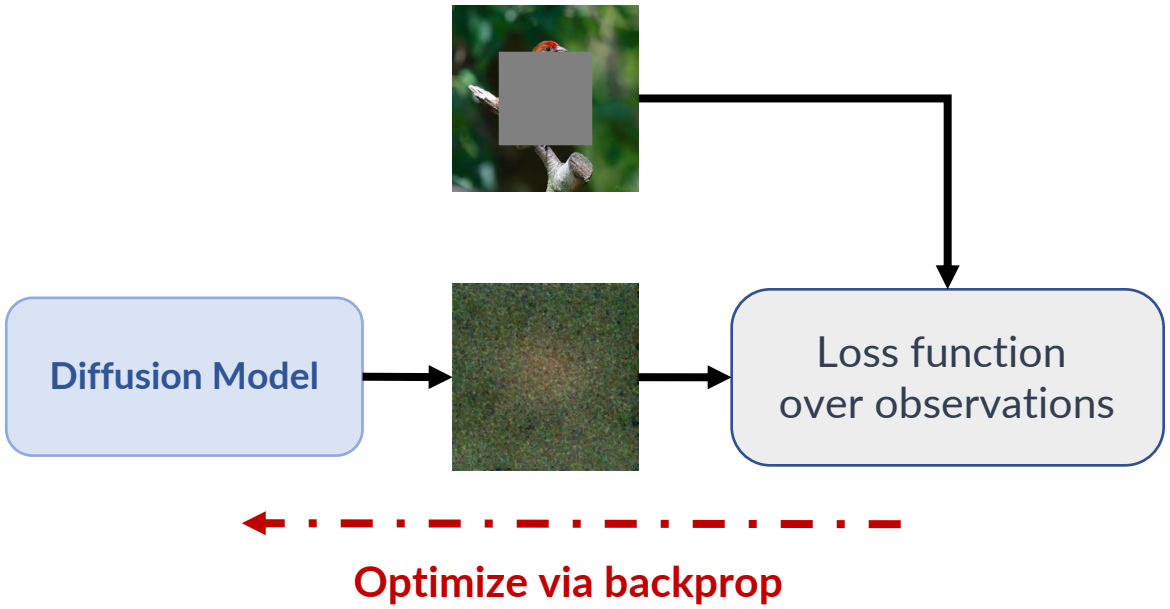




We want to use the same diffusion model for different problems!

Diffusion Models for Inverse Problems: Two Paradigms

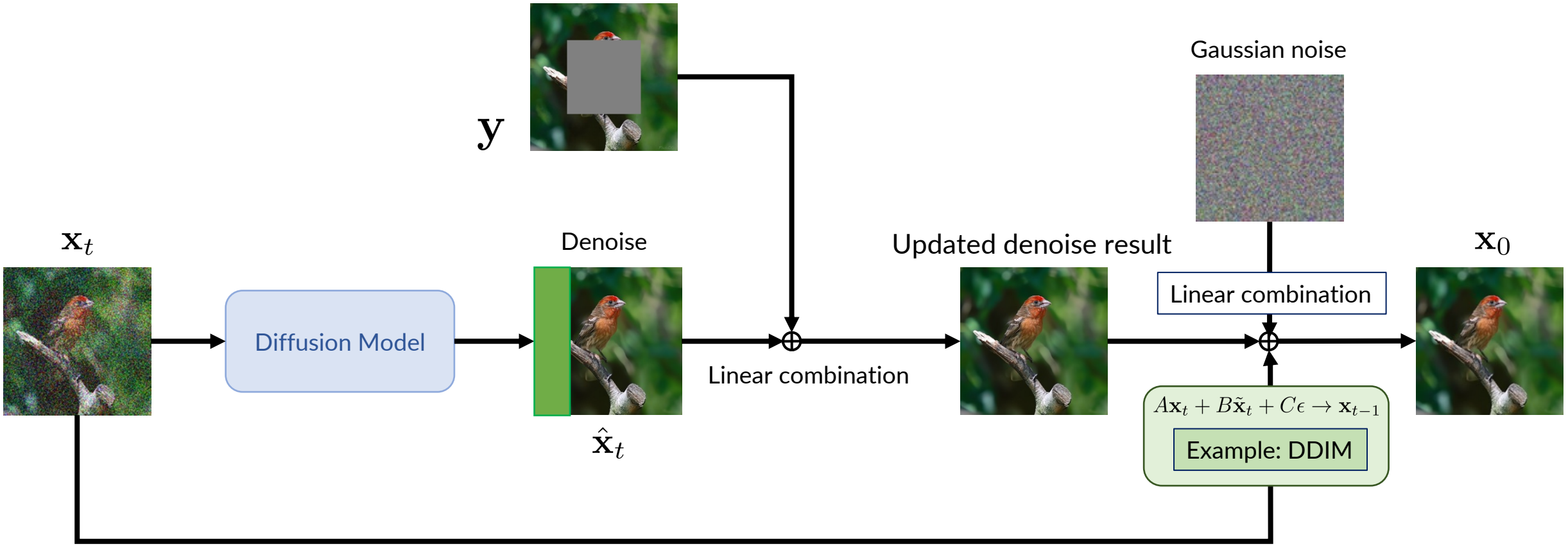




Replacement-based methods
(Overwrites model prediction with known information)

Reconstruction-based methods
(Approximate classifier-free guidance without additional training)

Replacement-based Methods: An Example



Song et al., "Score-Based Generative Modeling through Stochastic Differential Equations", ICLR 2021

Reconstruction-based Methods: An Example

- Use reconstruction-based loss to approximate classifier guidance

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|\mathbf{y}) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t)$$

"Conditional score" — Prior diffusion model — Want to approximate without training!

- Use the denoising model to approximate the classifier guidance term

$$p_t(\mathbf{y}|\mathbf{x}_t) = \mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t)}[p(\mathbf{y}|\mathbf{x}_0)] \approx p(\mathbf{y}|\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t])$$

Denoiser output

- The above equality is from the Markov chain: $\mathbf{y} \leftarrow \mathbf{x}_0 \rightarrow \mathbf{x}_t$

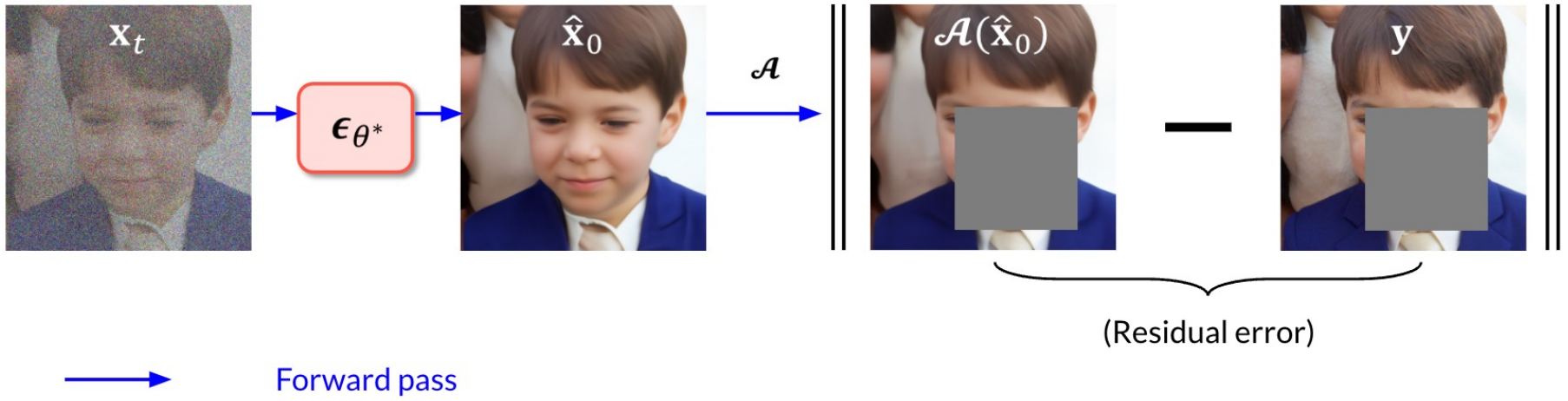
Chung et al., "Diffusion Posterior Sampling for General Noisy Inverse Problems", ICLR 2023

Diffusion Posterior Sampling

In the Gaussian case,

$$p(\mathbf{y}|\mathbb{E}[\mathbf{x}_0|\mathbf{x}_t]) = -c\|\mathcal{A}(\hat{\mathbf{x}}_0) - \mathbf{y}\|_2^2$$

Maximizing the likelihood is minimizing the L2 distance between measured and generated!



Chung et al., "Diffusion Posterior Sampling for General Noisy Inverse Problems", ICLR 2023

Solving Inverse Problems with Diffusion Models

Reconstruction-based methods

- **ScoreSDE:** simple linear problems, e.g., inpainting, colorization; later extended to MRI and CT.
- **ILVR:** more linear problems, e.g., super-resolution.
- **SNIPS:** slow solution for noisy linear problems.
- **CCDF:** better initializations.
- **DDRM:** fast solution for all noisy linear problems, and JPEG.

Choi et al., "ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models", ICCV 2021
Kawar et al., "SNIPS: Solving Noisy Inverse Problems Stochastically", NeurIPS 2021
Chung et al., "Come-Closer-Diffuse-Faster: Accelerating Conditional Diffusion Models for Inverse Problems through Stochastic Contraction", CVPR 2022
Song et al., "Solving Inverse Problems in Medical Imaging with Score-Based Generative Models", ICLR 2022
Kawar et al., "Denoising Diffusion Restoration Models", NeurIPS 2022

Solving Inverse Problems with Diffusion Models

Replacement-based methods

- **Video Diffusion/Pyramid DDPM:** used for super-resolution.
- **Pseudoinverse guidance:** linear and some non-differentiable problems, e.g., JPEG
- **MCG:** combines replacement & reconstruction for linear problems.

Others

- **CSGM:** Posterior sampling with Langevin Dynamics based on the diffusion score model.
- **RED-Diff:** A Regularizing-by-Denoising (RED), variational inference approach.
- **Posterior sampling**: use RealNVP to approximate posterior samples from diffusion models.

Ho et al., "Video Diffusion Models", NeurIPS 2022
Chung et al., "Improving Diffusion Models for Inverse Problems using Manifold Constraints", NeurIPS 2022
Ryu and Ye, "Pyramidal Denoising Diffusion Probabilistic Models", arXiv 2022
Chung et al., "Diffusion Posterior Sampling for General Noisy Inverse Problems", arXiv 2022
Song et al., "Pseudoinverse-Guided Diffusion Models for Inverse Problems", ICLR 2023
Jalal et al., "Robust Compressed Sensing MRI with Deep Generative Priors", NeurIPS 2021
Mardani et al., "A Variational Perspective on Solving Inverse Problems with Diffusion Models", arXiv 2023
Feng et al., "Score-Based Diffusion Models as Principled Priors for Inverse Imaging", arXiv 2023

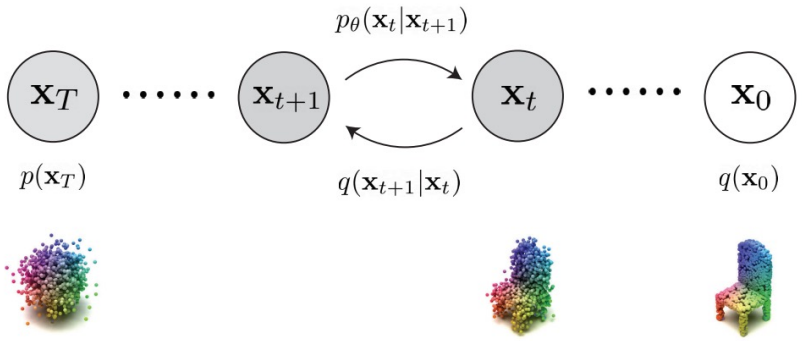
Outline

- Inverse problems
- 3D
- Video
- Miscellaneous

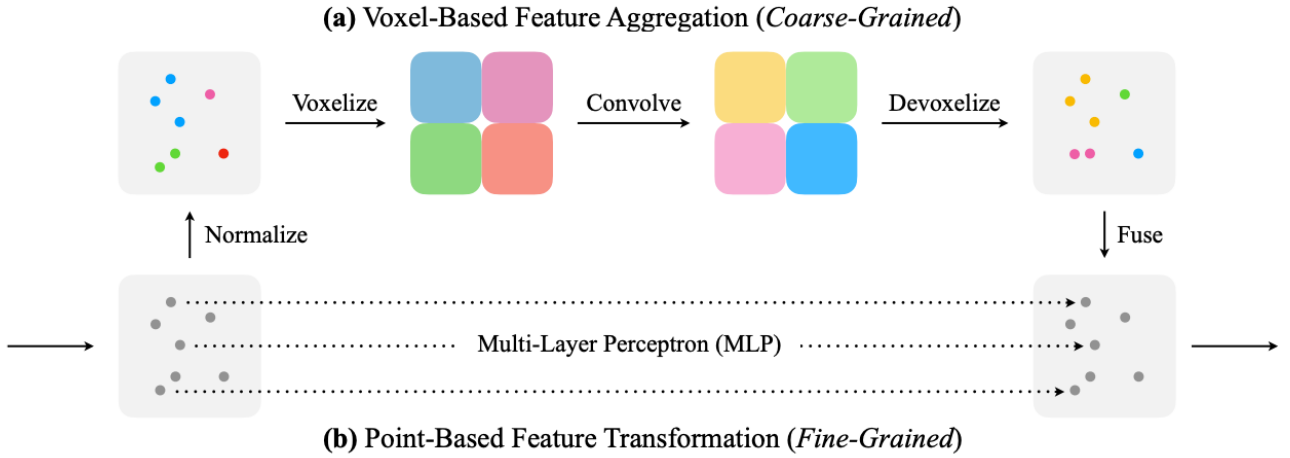
- Diffusion on various 3D representations
- 2D diffusion models for 3D generation
- Diffusion models for view synthesis
- 3D reconstruction
- 3D editing

Diffusion Models for Point Clouds

A set of points with location information.



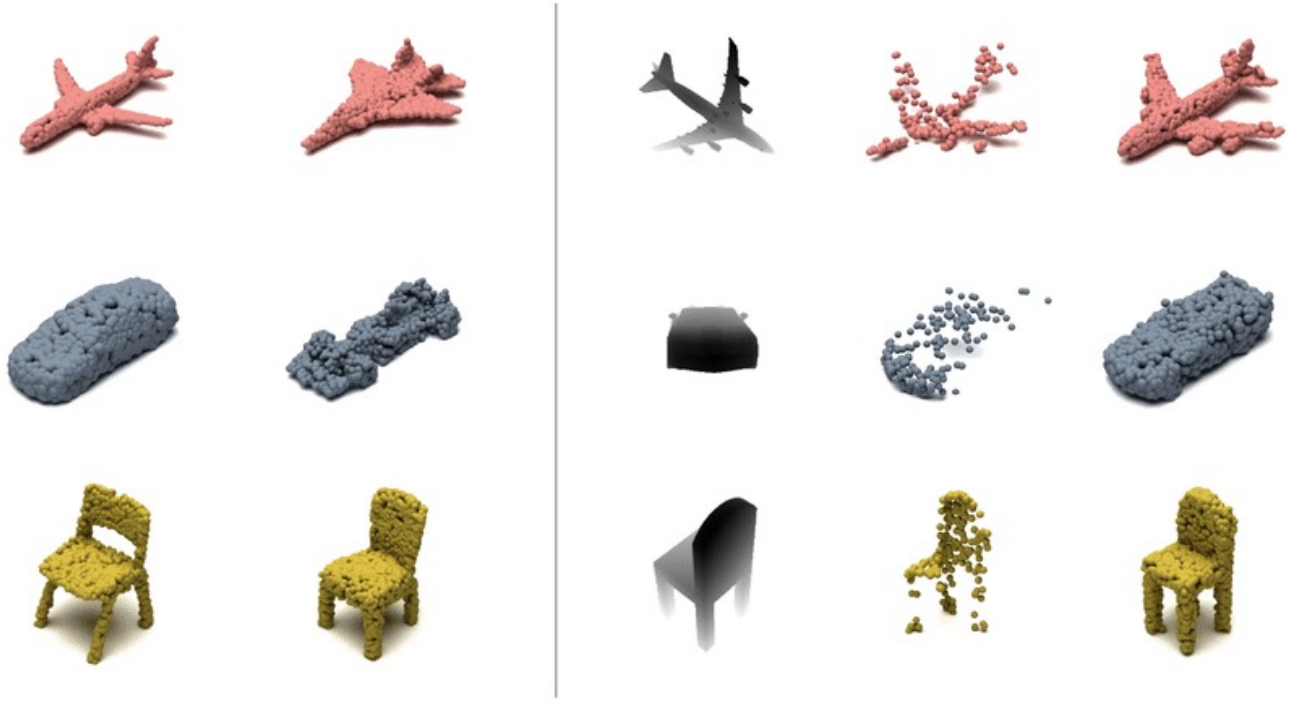
Procedure



Point-Voxel CNN architecture

Zhou et al., "3D Shape Generation and Completion through Point-Voxel Diffusion", ICCV 2021
Liu et al, "Point-Voxel CNN for Efficient 3D Deep Learning", NeurIPS 2019

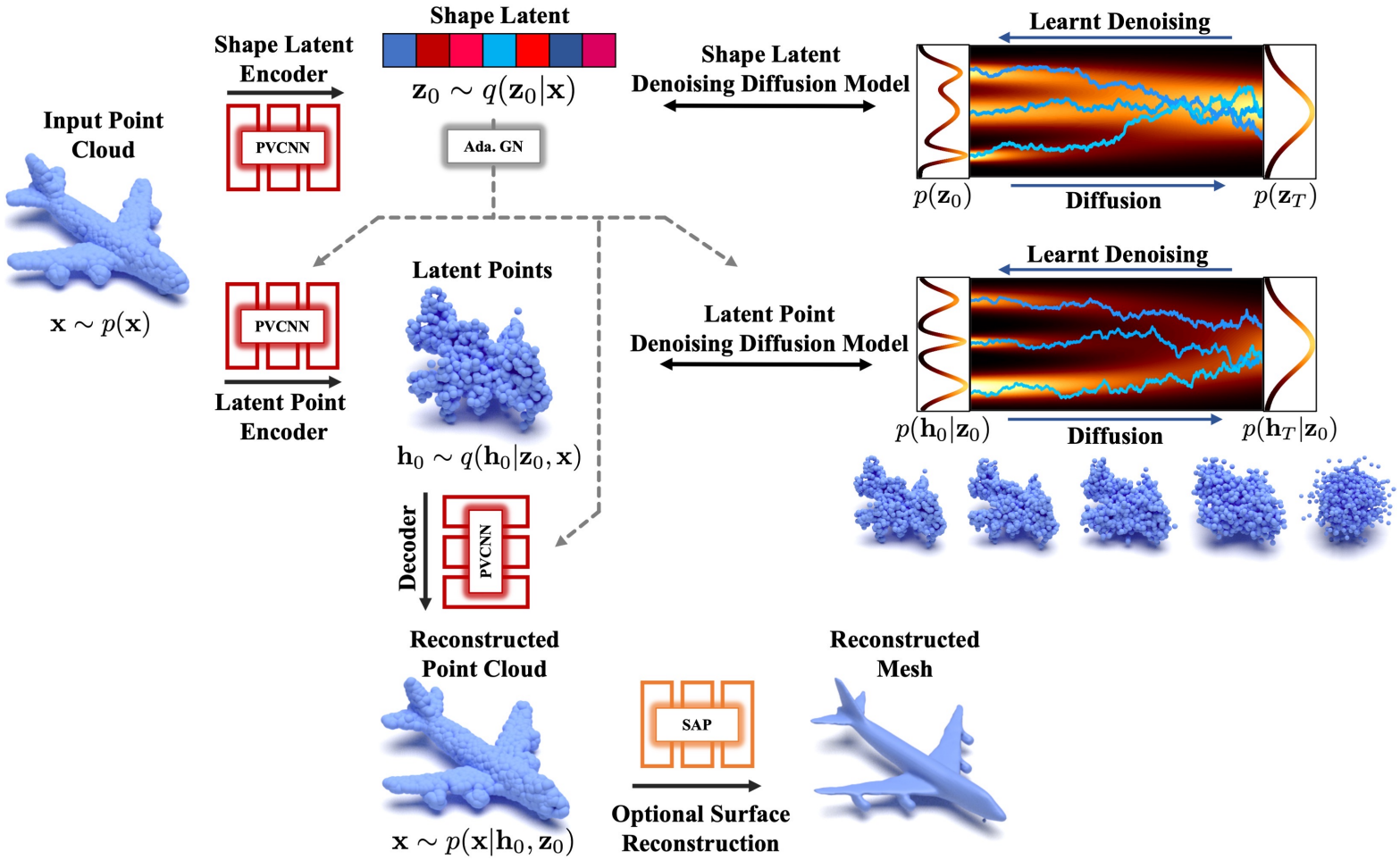
Diffusion Models for Point Clouds



Generation Completion

Zhou et al., "3D Shape Generation and Completion through Point-Voxel Diffusion", ICCV 2021

Diffusion Models for Point Clouds

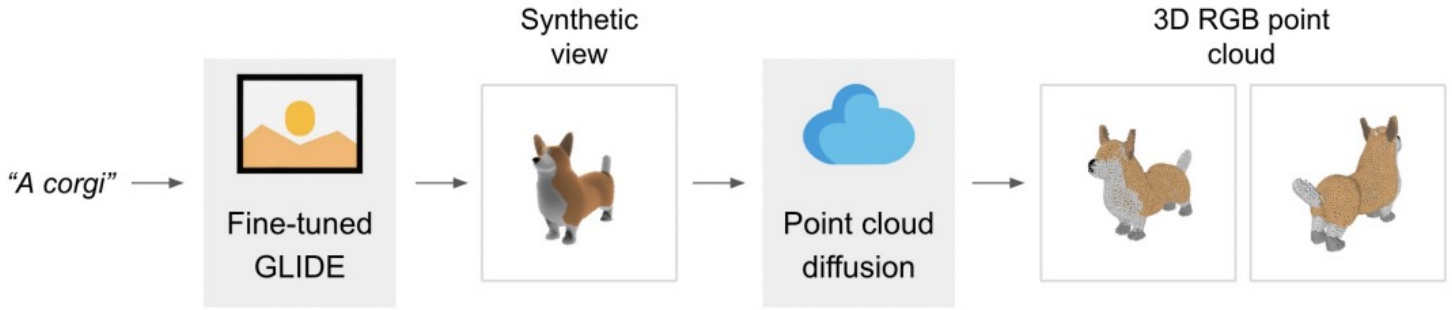


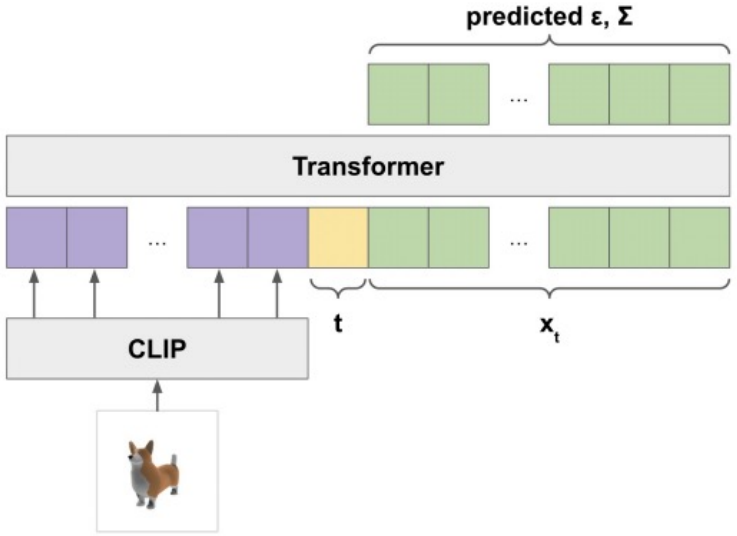
Point cloud diffusion in the latent space

Zeng et al., "LION: Latent Point Diffusion Models for 3D Shape Generation", NeurIPS 2022

Diffusion Models for Point Clouds

Point-E uses a synthetic view from fine-tuned GLIDE, and then ”lifts” the image to a 3d point cloud.





A transformer-based architecture

Nichol et al., "Point-E: A System for Generating 3D Point Clouds from Complex Prompts", arXiv 2022

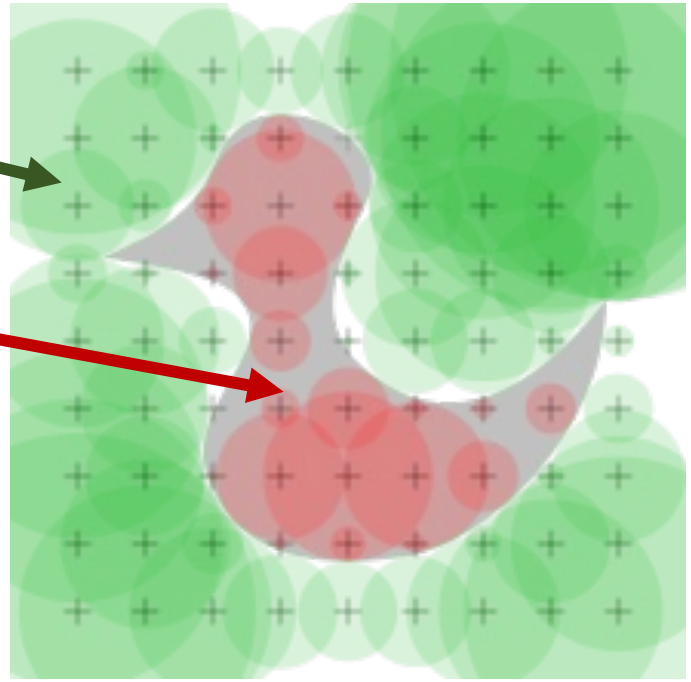
Diffusion Models for Signed Distance Functions

SDF is a function representation of a surface.

For each location x, |SDF(x)| = smallest distance to any point on the surface.

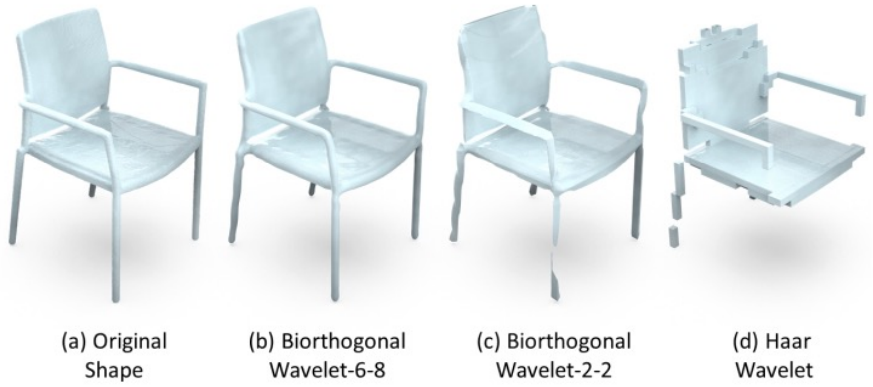
Green: outside of surface, positive

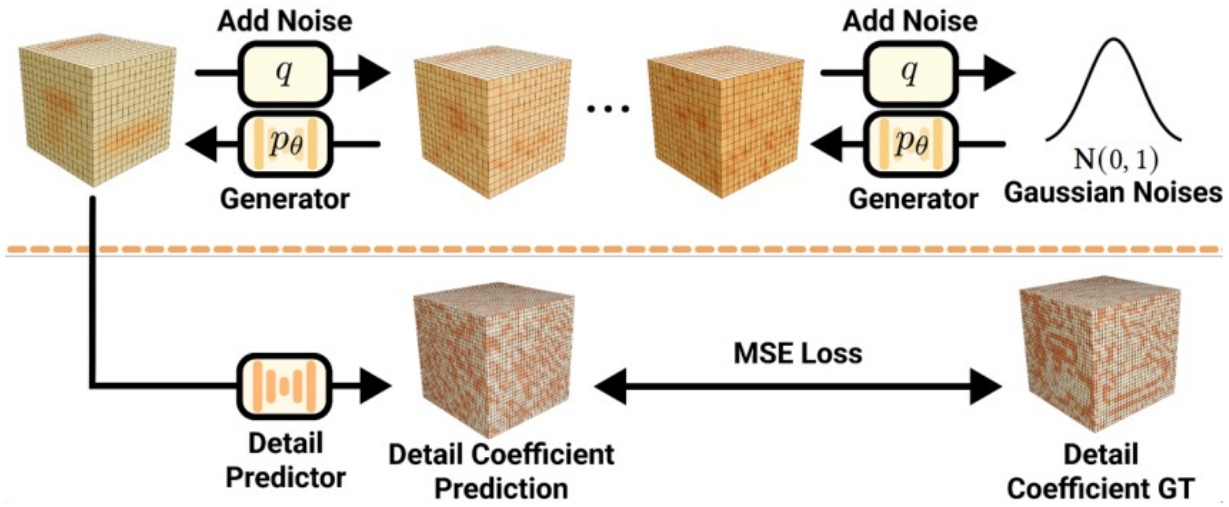
Red: inside of surface, negative



Diffusion Models for Signed Distance Functions

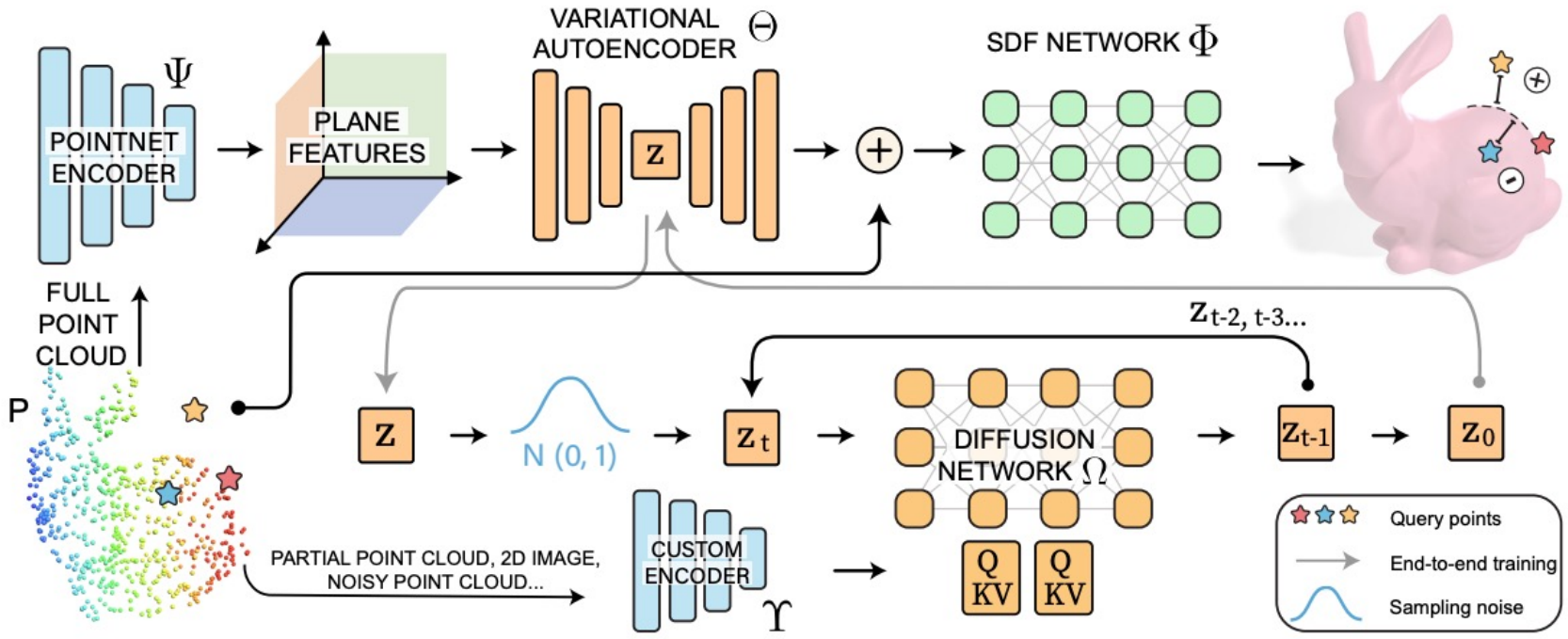
- Memory of SDF grows cubically with resolution
- Wavelets can be used for compression!
- Diffusion for coarse coefficients, then predict detailed ones.





Hui et al., "Neural Wavelet-domain Diffusion for 3D Shape Generation", arXiv 2022

Diffusion Models for Signed Distance Functions

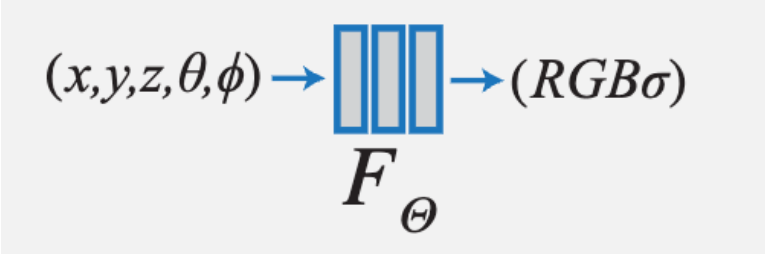


Latent space diffusion for SDFs, where conditioning can be provided with cross attention

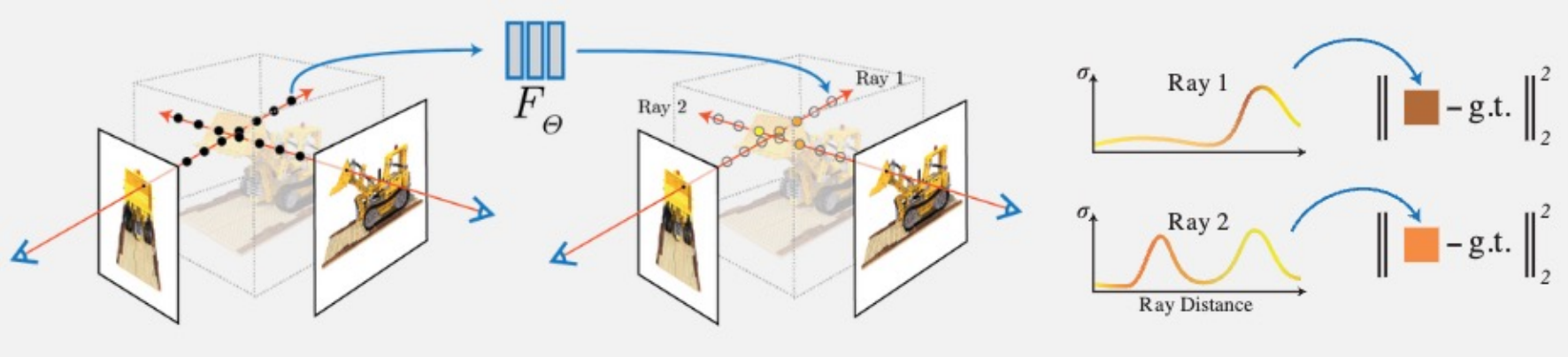
Chou et al., "DiffusionSDF: Conditional Generative Modeling of Signed Distance Functions", arXiv 2022

Diffusion Models for Other 3D Representations

Neural Radiance Fields (NeRF) is another representation of a 3D object.

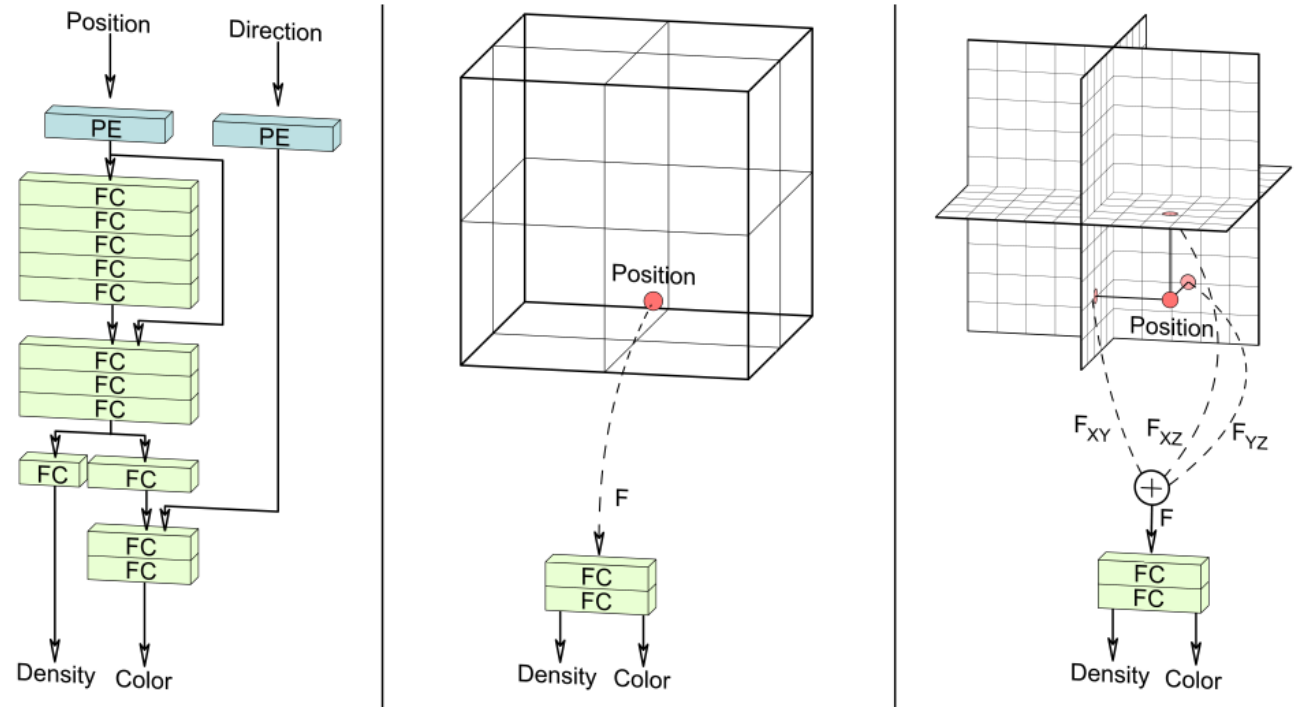


World coordinates + viewing direction -> RGB + density



Volume rendering of NeRFs

Diffusion Models for Other 3D Representations



NeRF
(Fully implicit)

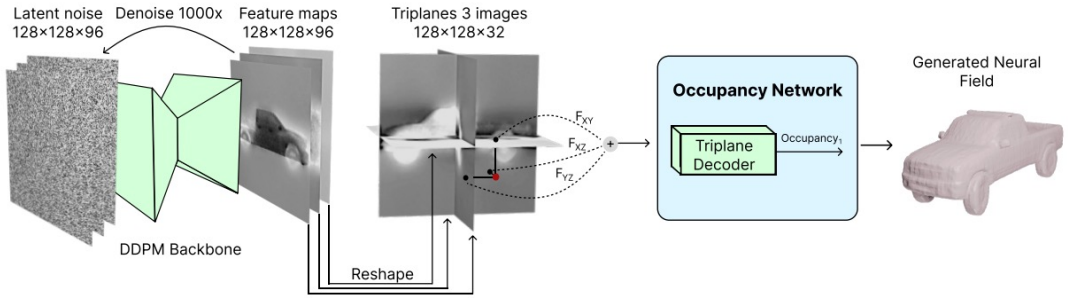
Voxels
(Explicit / hybrid)

Triplanes
(Factorized, hybrid)

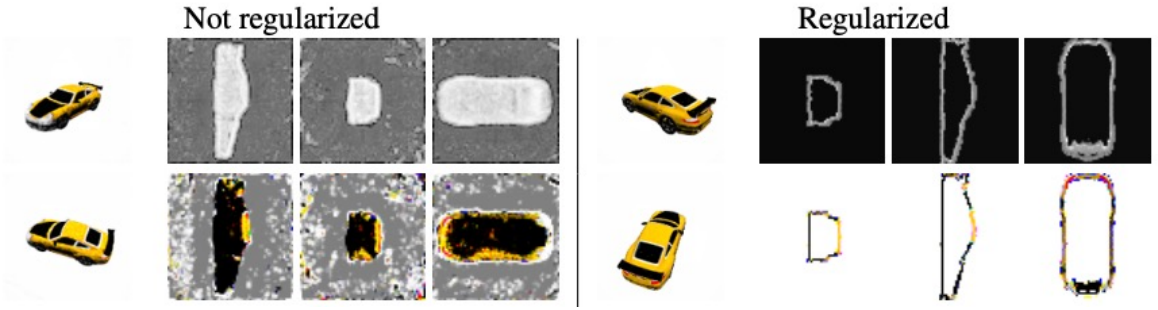
Image from EG3D paper.

Diffusion Models for Other 3D Representations

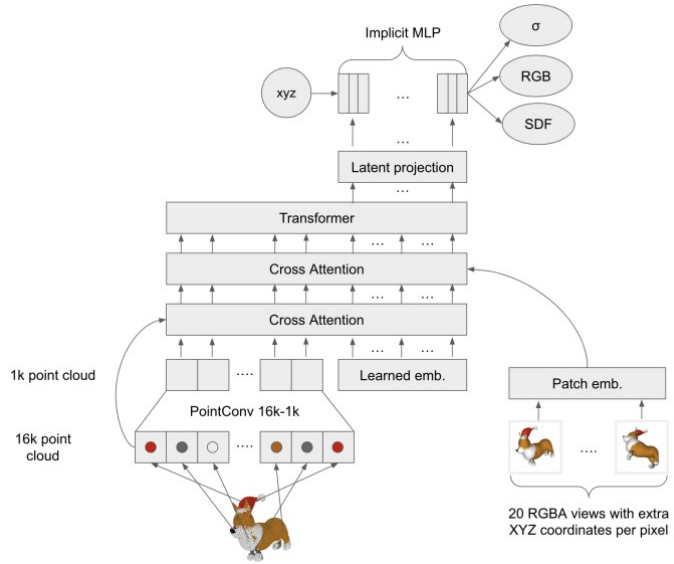
- Triplanes, regularized ReLU Fields, the MLP of NeRFs...
- A good representation is important!



Triplane diffusion



Regularized ReLU Fields



Implicit MLP of NeRFs

Shue et al., "3D Neural Field Generation using Triplane Diffusion", arXiv 2022
Yang et al., "Learning a Diffusion Prior for NeRFs", ICLR Workshop 2023
Jun and Nichol, "Shap-E: Generating Conditional 3D Implicit Functions", arXiv 2023

Outline

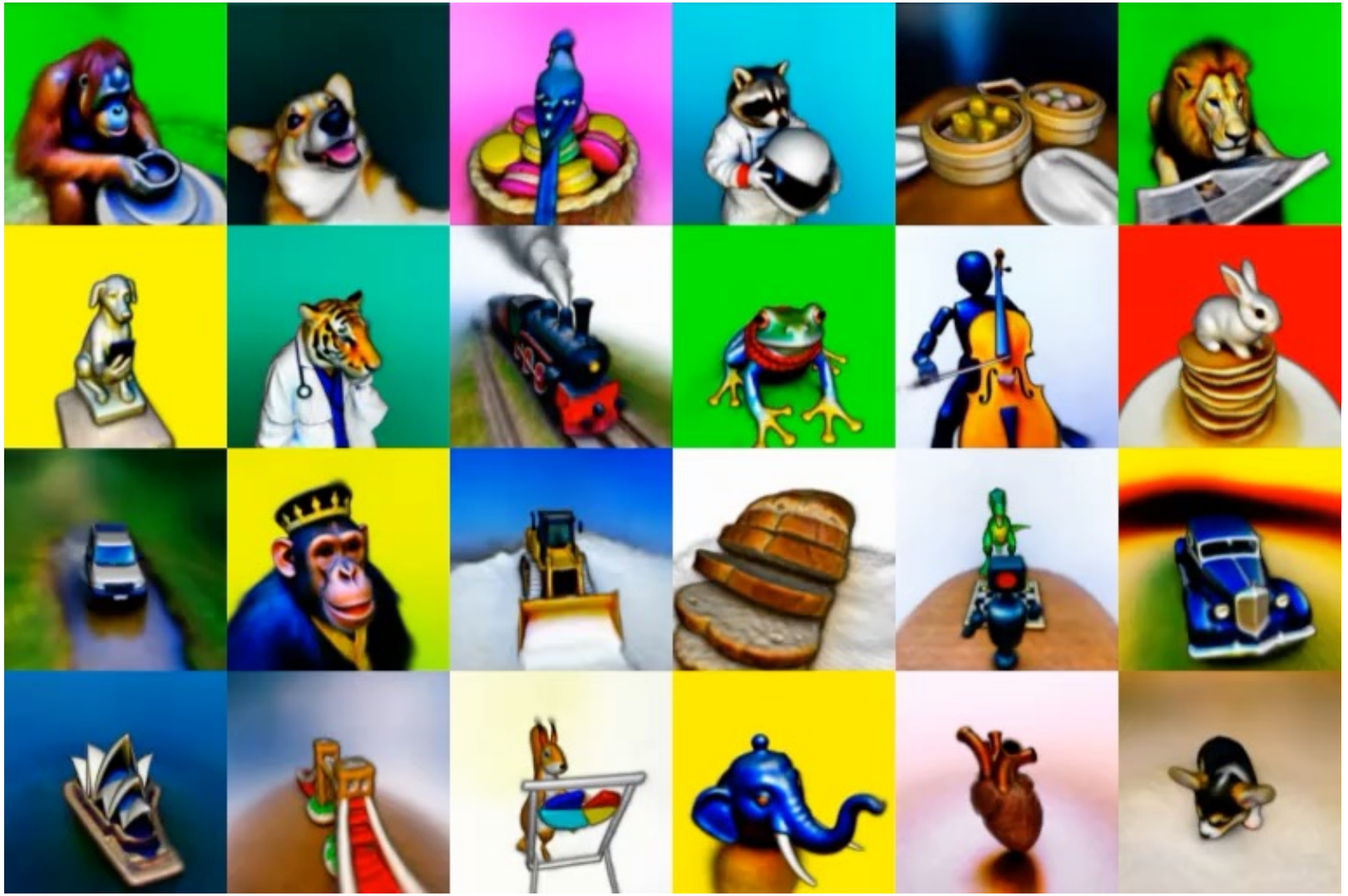
- Inverse problems
- 3D
- Video
- Miscellaneous

- Diffusion on various 3D representations
- 2D diffusion models for 3D generation
- Diffusion models for view synthesis
- 3D reconstruction
- 3D editing

2D Diffusion Models for 3D Generation

- Just now, we discussed diffusion models directly on 3d.
- However, there are a lot fewer 3d data than 2d.
  - A lot of experiments are based on ShapeNet!
- Can we use 2d diffusion models as a “prior” for 3d?

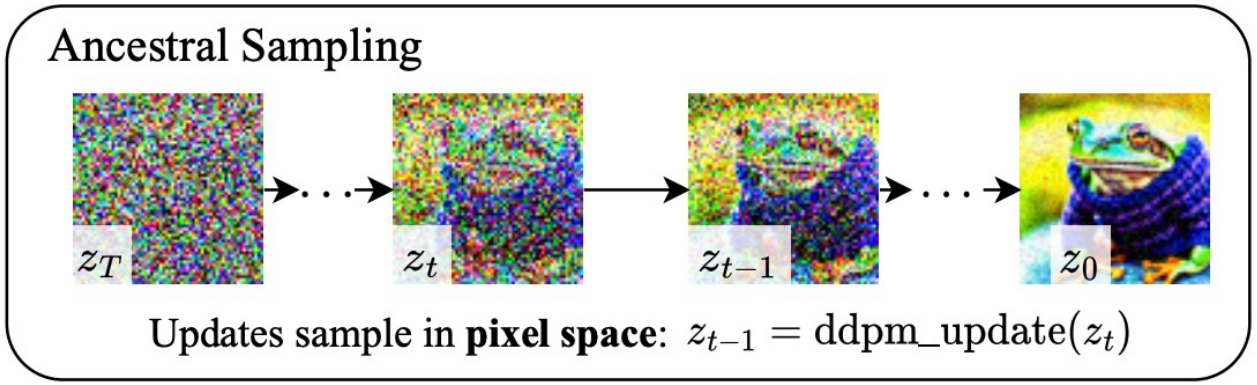
DreamFusion: where it all started

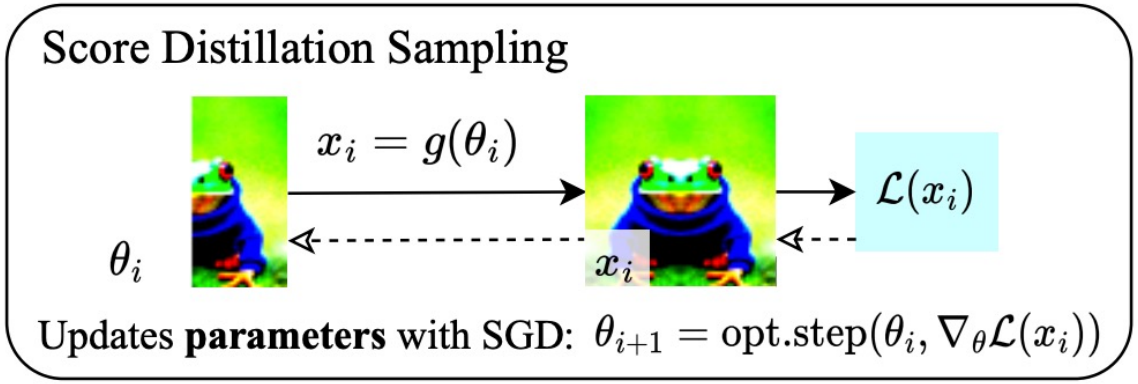


Poole et al., "DreamFusion: Text-to-3D using 2D Diffusion", ICLR 2023

DreamFusion: Setup

- Suppose there is a text-to-image diffusion model.
- Goal: optimize NeRF parameter such that each angle “looks good” from the text-to-image model.
- Unlike ancestral sampling (e.g., DDIM), the underlying parameters are being optimized over some loss function.





Poole et al., "DreamFusion: Text-to-3D using 2D Diffusion", ICLR 2023

DreamFusion: Score Distillation Sampling

- Consider the diffusion model objective for a sample x:

$$\mathcal{L}_{\text{Diff}}(\phi, \mathbf{x}) = \mathbb{E}_{t\sim\mathcal{U}(0,1),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[w(t)\|\epsilon_\phi(\alpha_t\mathbf{x}+\sigma_t\epsilon;t)-\epsilon\|_2^2\right],$$

- Directly computing the gradient leads to a Jacobian term over the U-Net:

$$\nabla_\theta\mathcal{L}_{\text{Diff}}(\phi, \mathbf{x}=g(\theta)) = \mathbb{E}_{t,\epsilon}\Bigg[w(t)\underbrace{\left(\hat{\epsilon}_\phi(\mathbf{z}_t;y,t)-\epsilon\right)}_{\text{Noise Residual}}\underbrace{\frac{\partial\hat{\epsilon}_\phi(\mathbf{z}_t;y,t)}{\mathbf{z}_t}}_{\text{U-Net Jacobian}}\underbrace{\frac{\partial\mathbf{x}}{\partial\theta}}_{\text{Generator Jacobian}}\Bigg]$$

$$\mathbf{z}_t = \alpha_t\mathbf{x}+\sigma_t\epsilon$$

- However, it turns out we can consider removing the U-Net Jacobian!

$$\nabla_\theta\mathcal{L}_{\text{SDS}}(\phi, \mathbf{x}=g(\theta)) \triangleq \mathbb{E}_{t,\epsilon}\left[w(t)\left(\hat{\epsilon}_\phi(\mathbf{z}_t;y,t)-\epsilon\right)\frac{\partial\mathbf{x}}{\partial\theta}\right]$$

Poole et al., "DreamFusion: Text-to-3D using 2D Diffusion", ICLR 2023

DreamFusion: Score Distillation Sampling

Consider the KL term to minimize (given t):

$$\mathrm{KL}(q(\mathbf{z}_t|g(\theta);y,t)\|p_\phi(\mathbf{z}_t;y,t))$$

KL between noisy real image distribution and generated image distributions, conditioned on y!

KL and its gradient is defined as:

$$\mathrm{KL}(q(\mathbf{z}_t|\mathbf{x}=g(\theta))\|p_\phi(\mathbf{z}_t|y)) = \mathbb{E}_\epsilon\left[\log q(\mathbf{z}_t|\mathbf{x}=g(\theta)) - \log p_\phi(\mathbf{z}_t|y)\right]$$

$$\nabla_\theta \mathrm{KL}(q(\mathbf{z}_t|\mathbf{x}=g(\theta))\|p_\phi(\mathbf{z}_t|y)) = \mathbb{E}_\epsilon\Big[\underbrace{\nabla_\theta \log q(\mathbf{z}_t|\mathbf{x}=g(\theta))}_{\text{(A)}} - \underbrace{\nabla_\theta \log p_\phi(\mathbf{z}_t|y)}_{\text{(B)}}\Big]$$

Approx. with diffusion model

Score of sample perturbed with noise

(B) can be derived from chain rule

$$\nabla_\theta \log p_\phi(\mathbf{z}_t|y) = s_\phi(\mathbf{z}_t|y)\frac{\partial \mathbf{z}_t}{\partial \theta} = \alpha_t s_\phi(\mathbf{z}_t|y)\frac{\partial \mathbf{x}}{\partial \theta} = -\frac{\alpha_t}{\sigma_t}\hat{\epsilon}_\phi(\mathbf{z}_t|y)\frac{\partial \mathbf{x}}{\partial \theta}$$

(A) is the gradient of the entropy of the forward process with fixed variance = 0.

Poole et al., "DreamFusion: Text-to-3D using 2D Diffusion", ICLR 2023

DreamFusion: Score Distillation Sampling

(A) + (B) = $\frac{\alpha_t}{\sigma_t}\hat{\epsilon}_\phi(\mathbf{z}_t|y)\frac{\partial \mathbf{x}}{\partial \theta}$

However, this objective can be quite noisy.

Alternatively, we can consider a "baseline" approach in reinforcement learning: add a component that has zero mean but reduces variance. Writing out (A) again:

$$\nabla_\theta \log q(\mathbf{z}_t|\mathbf{x}) = \Big(\underbrace{\frac{\partial \log q(\mathbf{z}_t|\mathbf{x})}{\partial \mathbf{x}}}_{\text{parameter score}} + \underbrace{\frac{\partial \log q(\mathbf{z}_t|\mathbf{x})}{\partial \mathbf{z}_t}\frac{\partial \mathbf{z}_t}{\partial \mathbf{x}}}_{\text{path derivative}}\Big)\alpha_t\frac{\partial \mathbf{x}}{\partial \theta} = \left(\frac{\alpha_t}{\sigma_t}\epsilon - \frac{\alpha_t}{\sigma_t}\epsilon\right)\alpha_t\frac{\partial \mathbf{x}}{\partial \theta} = 0.$$

Thus, we have:

$$\mathbb{E}_{t,\epsilon}\left[(A)+(B)\right] = \mathbb{E}_{t,\epsilon}\left[\frac{\alpha_t}{\sigma_t}(\hat{\epsilon}_\phi(\mathbf{z}_t|y)+\epsilon-\epsilon)\frac{\partial \mathbf{x}}{\partial \theta}\right]$$

$$= \mathbb{E}_{t,\epsilon}\left[\frac{\alpha_t}{\sigma_t}(\hat{\epsilon}_\phi(\mathbf{z}_t|y)-\epsilon)\frac{\partial \mathbf{x}}{\partial \theta}\right]$$

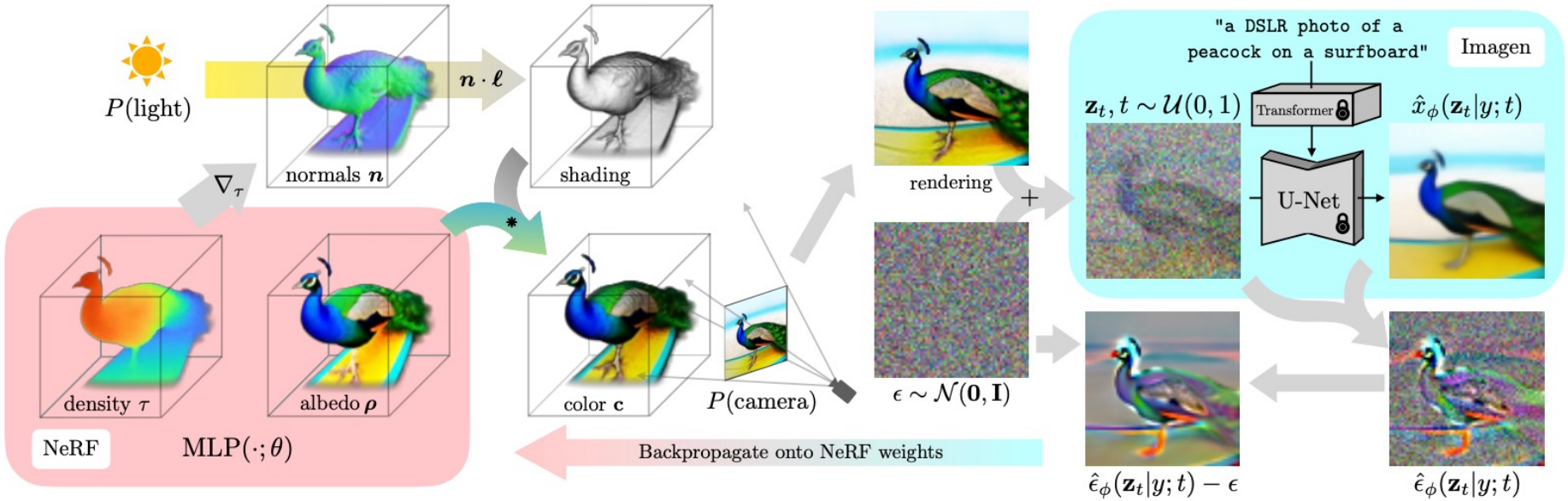
Gaussian noise, independent of NeRF, image, time, mean=0!

This has the same mean, but reduced variance, as we train $\hat{\epsilon}_\phi$ to predict ϵ

Poole et al., "DreamFusion: Text-to-3D using 2D Diffusion", ICLR 2023

DreamFusion in Text-to-3D

- SDS can be used to optimize a 3D representation, like NeRF.

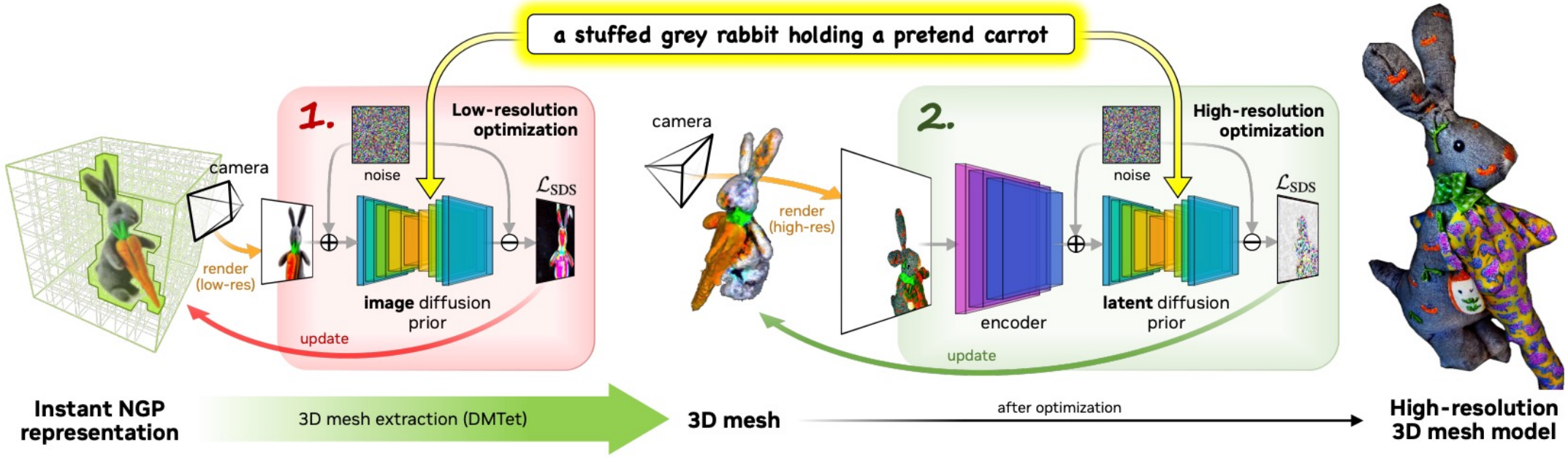


Poole et al., "DreamFusion: Text-to-3D using 2D Diffusion", ICLR 2023

Extensions to SDS: Magic3D

2x speed and higher resolution

- Accelerate NeRF with Instant-NGP, for coarse representations.
- Optimize a fine mesh model with differentiable renderer.



Lin et al., "Magic3D: High-Resolution Text-to-3D Content Creation", CVPR 2023

Alternative to SDS: Score Jacobian Chaining

A different formulation, motivated from approximating 3D score.

$$\nabla_{\boldsymbol{\theta}} \log \tilde{p}_\sigma(\boldsymbol{\theta}) = \mathbb{E}_\pi \left[\nabla_{\boldsymbol{\theta}} \log p_\sigma(\boldsymbol{x}_\pi) \right]$$

$$\frac{\partial \log \tilde{p}_\sigma(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbb{E}_\pi \left[\frac{\partial \log p_\sigma(\boldsymbol{x}_\pi)}{\partial \boldsymbol{x}_\pi} \cdot \frac{\partial \boldsymbol{x}_\pi}{\partial \boldsymbol{\theta}} \right]$$

$$\underbrace{\nabla_{\boldsymbol{\theta}} \log \tilde{p}_\sigma(\boldsymbol{\theta})}_{\text{3D score}} = \mathbb{E}_\pi \Big[\underbrace{\nabla_{\boldsymbol{x}_\pi} \log p_\sigma(\boldsymbol{x}_\pi)}_{\text{2D score; pretrained}} \cdot \underbrace{\boldsymbol{J}_\pi}_{\text{renderer Jacobian}} \Big].$$

In principle, the diffusion model is the noisy 2D score (over clean images),
but in practice, the diffusion model suffers from out-of-distribution (OOD) issues!

For diffusion model on noisy images, the non-noisy images are OOD!

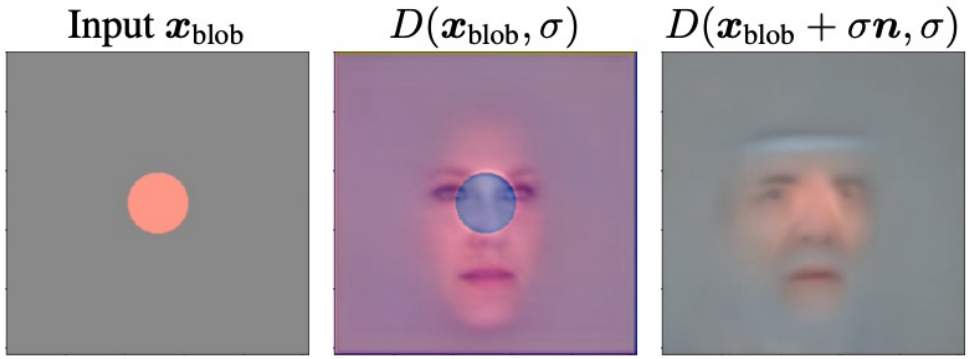
Wang et al., "Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation", CVPR 2023.

Score Jacobian Chaining

SJC approximates noisy score with “Perturb-and-Average Scoring”, which is not present in SDS.

- Use score model on multiple noise-perturbed data, then average it.

$$
\begin{aligned}
&\mathrm{PAAS}(\boldsymbol{x}_\pi, \sqrt{2}\sigma) \\
\triangleq &\mathbb{E}_{\boldsymbol{n}\sim\mathcal{N}(0,\mathbf{I})}\left[\mathrm{score}(\boldsymbol{x}_\pi + \sigma\boldsymbol{n}, \sigma)\right] \\
= &\mathbb{E}_{\boldsymbol{n}}\left[\frac{D(\boldsymbol{x}_\pi + \sigma\boldsymbol{n}, \sigma) - (\boldsymbol{x}_\pi + \sigma\boldsymbol{n})}{\sigma^2}\right] \\
= &\mathbb{E}_{\boldsymbol{n}}\left[\frac{D(\boldsymbol{x}_\pi + \sigma\boldsymbol{n}, \sigma) - \boldsymbol{x}_\pi}{\sigma^2}\right] - \underbrace{\mathbb{E}_{\boldsymbol{n}}\left[\frac{\boldsymbol{n}}{\sigma}\right]}_{=0}. \\
&\mathrm{PAAS}(\boldsymbol{x}_\pi, \sqrt{2}\sigma) \approx \nabla_{\boldsymbol{x}_\pi} \log p_{\sqrt{2}\sigma}(\boldsymbol{x}_\pi).
\end{aligned}
$$

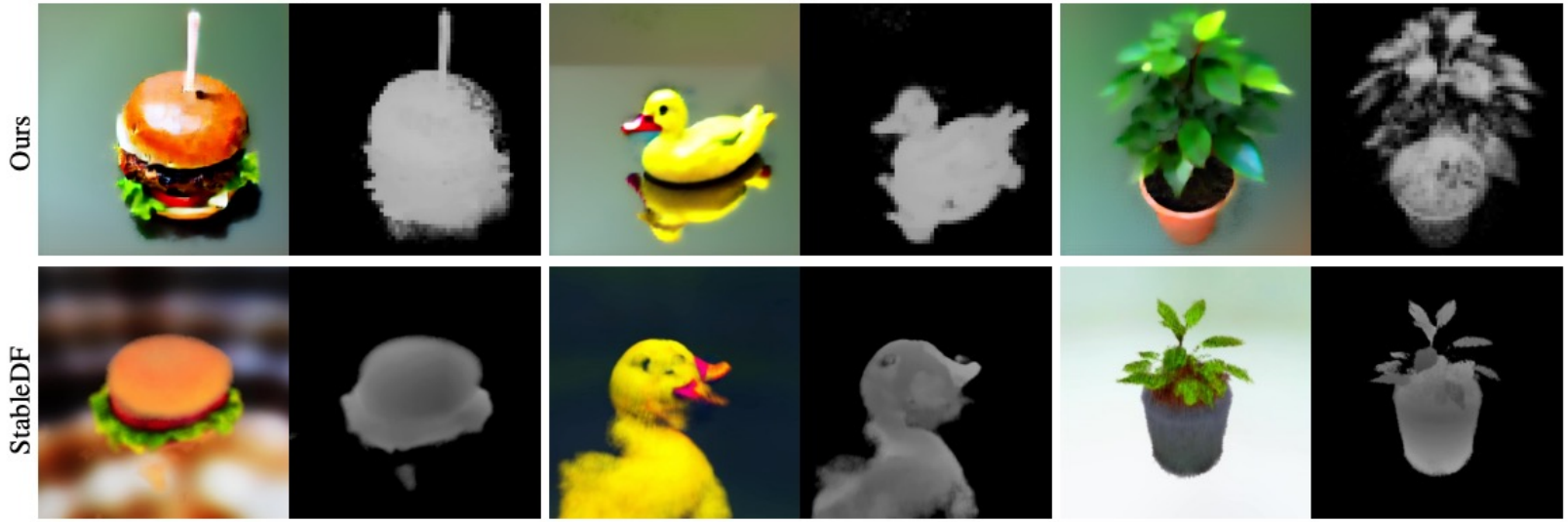


PAAS helps guide updates with a better score.

Wang et al., "Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation", CVPR 2023.

SJC and SDS

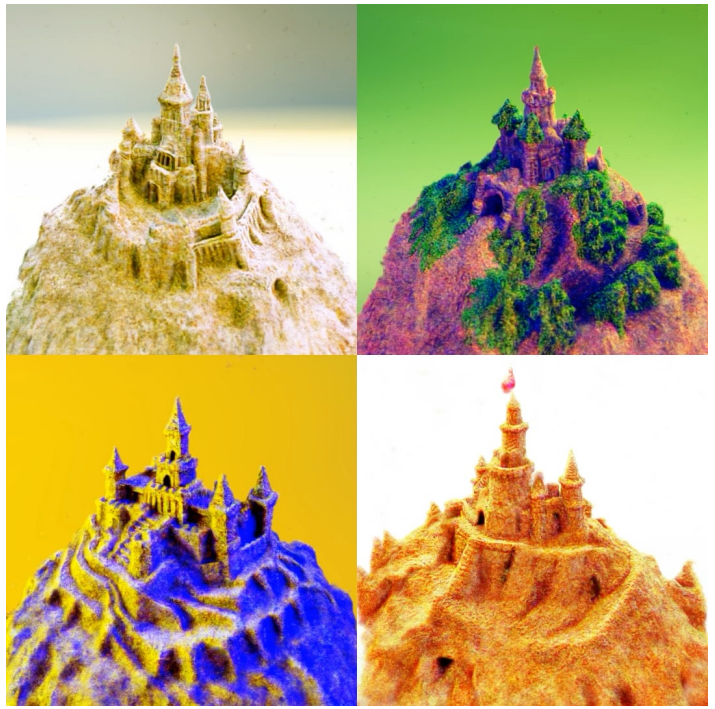
SJC is a competitive alternative to SDS.



Wang et al., "Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation", CVPR 2023.

Alternative to SDS: ProlificDreamer

- SDS-based method often set classifier-guidance weight to 100, which limits the “diversity” of the generated samples.
- ProlificDreamer reduces this to 7.5, leading to diverse samples.



Wang et al., “ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation”, arXiv 2023

ProlificDreamer and Variational Score Distillation

Instead of maximizing the likelihood under diffusion model, VSD minimizes the KL divergence via variational inference.

$$\min_{\mu} D_{\mathrm{KL}}(q_0^{\mu}(\boldsymbol{x}_0|y) \parallel p_0(\boldsymbol{x}_0|y)).$$

μ is the distribution of NeRFs.

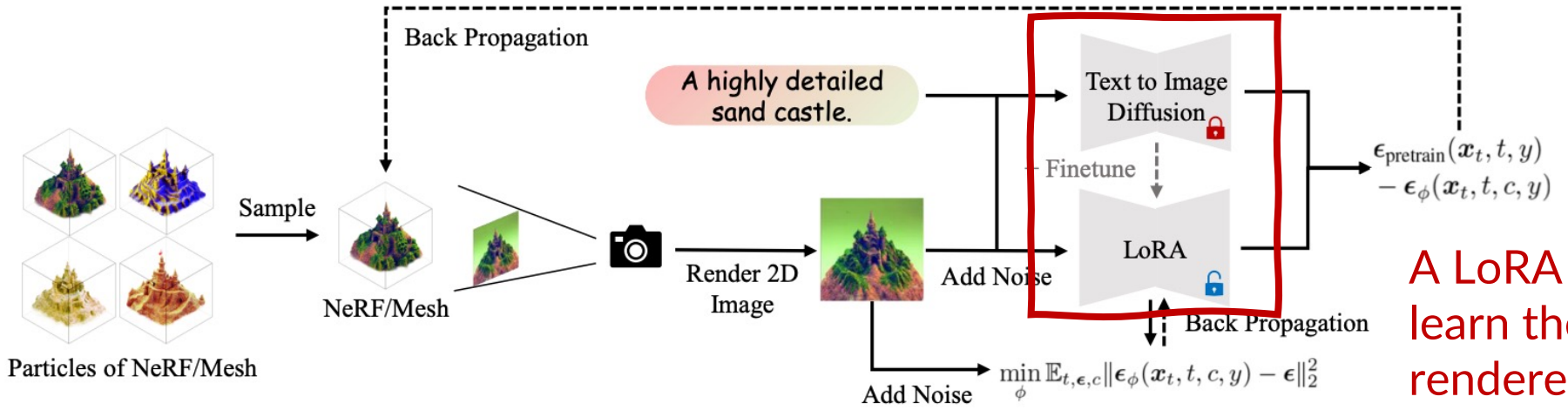
Suppose $\theta_\tau \sim \mu$ is a NeRF sample, then VSD simulates this ODE:

$$\frac{\mathrm{d}\theta_\tau}{\mathrm{d}\tau} = -\mathbb{E}_{t,\boldsymbol{\epsilon},c}\Big[\omega(t)\Big(\underbrace{-\sigma_t \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|y)}_{\textit{score of noisy real images}} - \underbrace{(-\sigma_t \nabla_{\boldsymbol{x}_t} \log q_t^{\mu_\tau}(\boldsymbol{x}_t|c,y))}_{\textit{score of noisy rendered images}}\Big)\frac{\partial \boldsymbol{g}(\theta_\tau, c)}{\partial \theta_\tau}\Big],$$

- Diffusion model can be used to approximate score of noisy real images.
- How about noisy rendered images?

ProlificDreamer and Variational Score Distillation

- Learn another diffusion model to approximate the score of noisy rendered images!



A LoRA step is added to learn the score of noisy rendered images!

Wang et al., "ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation", arXiv 2023

Why does VSD work in practice?

- The valid text-to-image NeRFs form a distribution with infinite possibilities!
- In SDS, epsilon is the score of noisy “dirac distribution” over finite renders, which converges to the true score with infinite renders!
- In VSD, the LoRA model aims to represent the (true) score of noisy distribution over infinite number of renders!
- If the generated NeRF distribution is only one point and LoRA overfits perfectly, then VSD = SDS!
- But LoRA has good generalization (and learns from a trajectory of NeRFs), so closer to the true score!

- This is analogous to
  - Representing the dataset score via mixture of Gaussians on the dataset (SDS), versus
  - Representing the dataset score via the LoRA UNet (VSD)

Wang et al., “ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation”, arXiv 2023

Outline

- Inverse problems
- 3D
- Video
- Miscellaneous

- Diffusion on various 3D representations
- 2D diffusion models for 3D generation
- Diffusion models for view synthesis
- 3D reconstruction
- 3D editing

Novel-view Synthesis with Diffusion Models

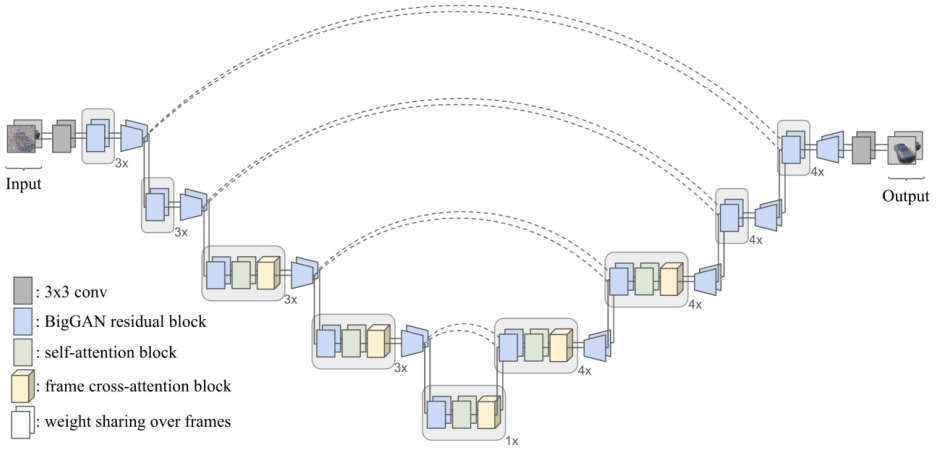
- These do not produce 3D as output, but synthesis the view at different angles.



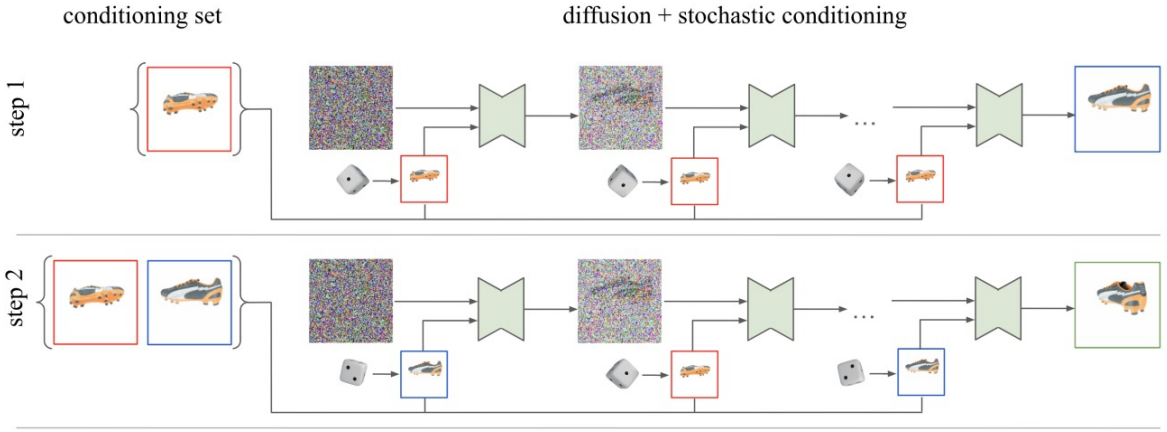

Watson et al., "Novel View Synthesis with Diffusion Models", ICLR 2023

3DiM

- Condition on a frame and two poses, predict another frame.



UNet with frame cross-attention

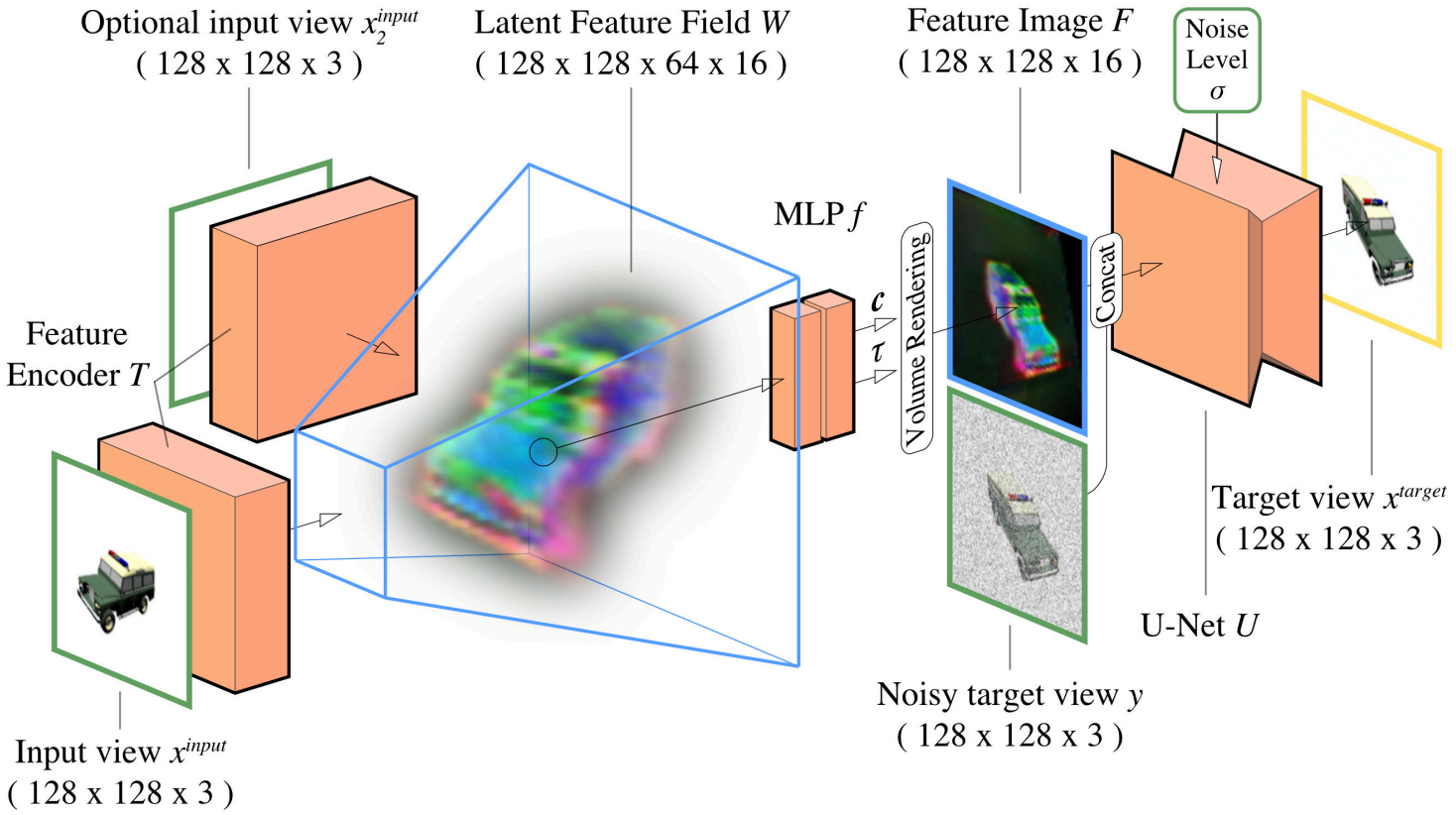


Sample based on stochastic conditions, allowing the use of multiple conditional frames.

Watson et al., "Novel View Synthesis with Diffusion Models", ICLR 2023

GenVS

- 3D-aware architecture with latent feature field.
- Use diffusion model to improve render quality based on structure.



Chan et al., "Generative Novel View Synthesis with 3D-Aware Diffusion Models", arXiv 2023

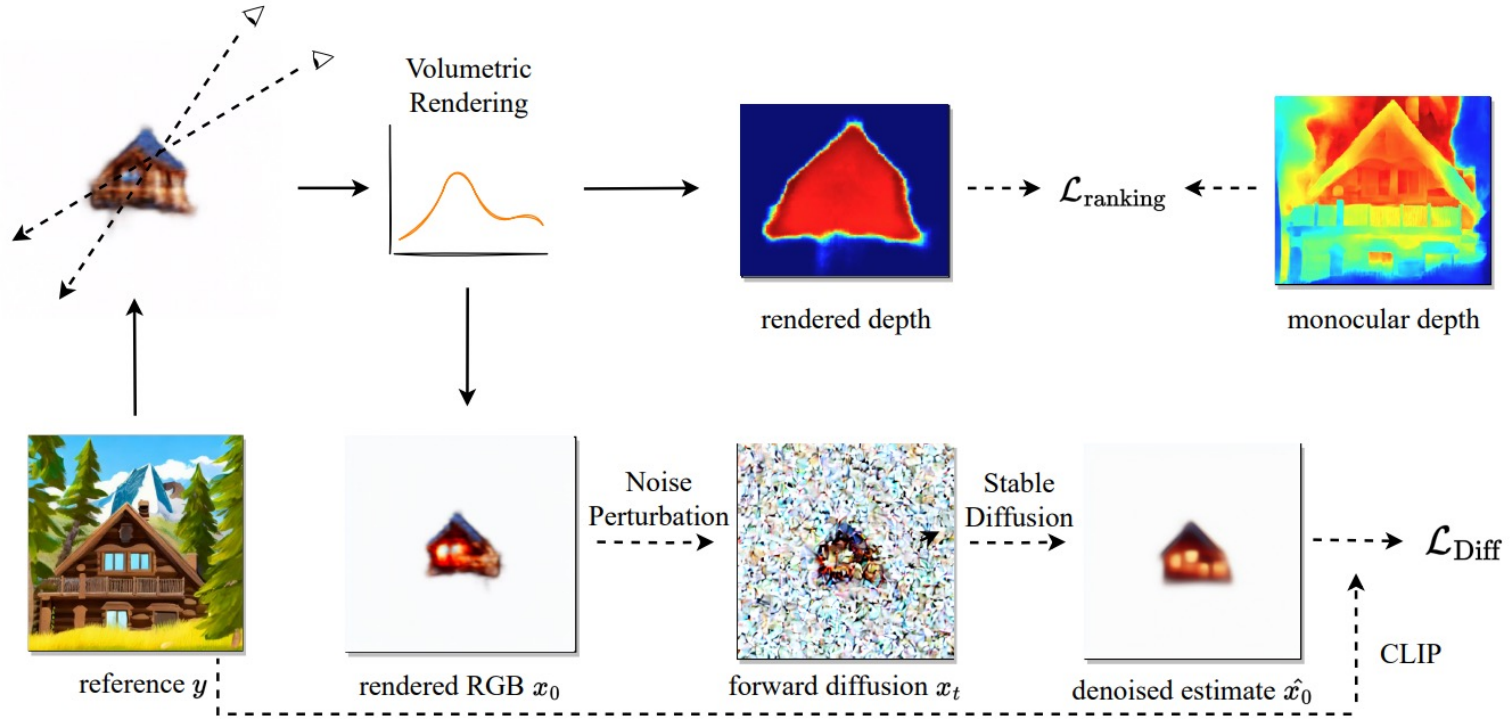
Outline

- Inverse problems
- 3D
- Video
- Miscellaneous

- Diffusion on various 3D representations
- 2D diffusion models for 3D generation
- Diffusion models for view synthesis
- 3D reconstruction
- 3D editing

NeuralLift-360 for 3D reconstruction

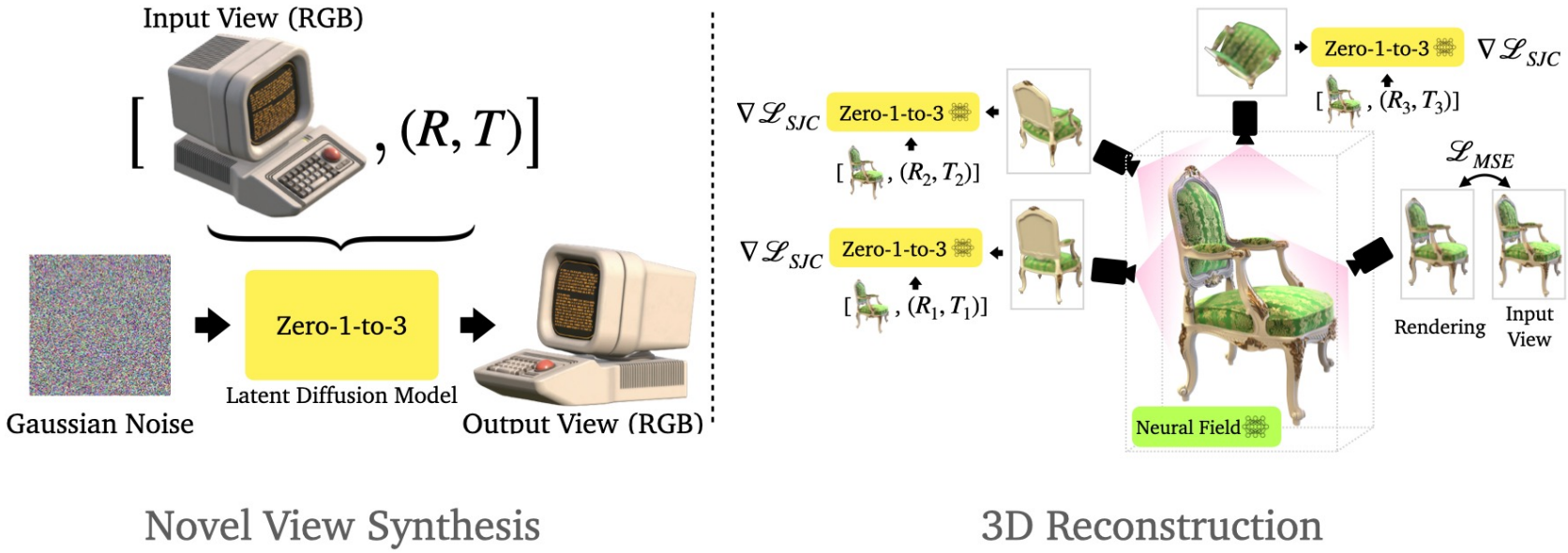
- SDS + Fine-tuned CLIP text embedding + Depth supervision



Xu et al., "NeuralLift-360: Lifting An In-the-wild 2D Photo to A 3D Object with 360° Views", CVPR 2023

Zero 1-to-3

- Generate novel view from 1 view and pose, with 2d model.
- Then, run SJC / SDS-like optimizations with view-conditioned model.



Liu et al., [“Zero-1-to-3: Zero-shot One Image to 3D Object”](), arXiv 2023

Outline

- Inverse problems
- 3D
- Video
- Miscellaneous

- Diffusion on various 3D representations
- 2D diffusion models for 3D generation
- Diffusion models for view synthesis
- 3D reconstruction
- 3D editing

Instruct NeRF2NeRF

Edit a 3D scene with text instructions

Haque et al., "Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions", arXiv 2023

Instruct NeRF2NeRF

Edit a 3D scene with text instructions

- Given existing scene, use Instruct Pix2Pix to edit image at different viewpoints.
- Continue to train the NeRF and repeat the above process

Haque et al., "Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions", arXiv 2023

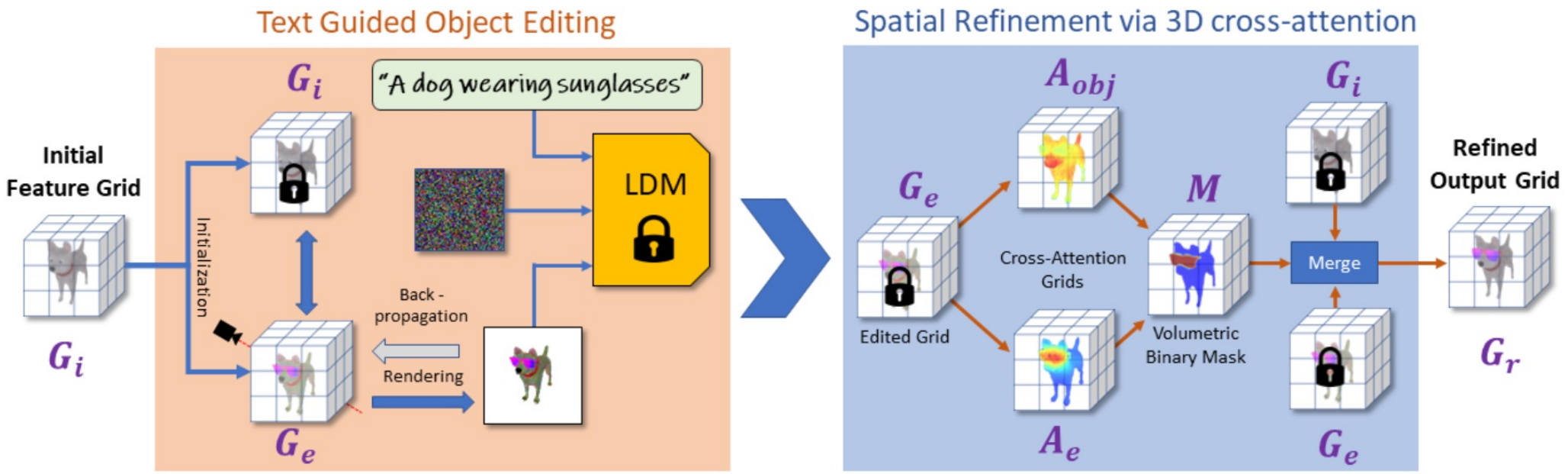
Instruct NeRF2NeRF

With each iteration, the edits become more consistent.

Haque et al., "Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions", arXiv 2023

Vox-E: Text-guided Voxel Editing of 3D Objects

- Text-guided object editing with SDS
- Regularize the structure of the new voxel grid.



Sella et al., "Vox-E: Text-guided Voxel Editing of 3D Objects", arXiv 2023

Outline

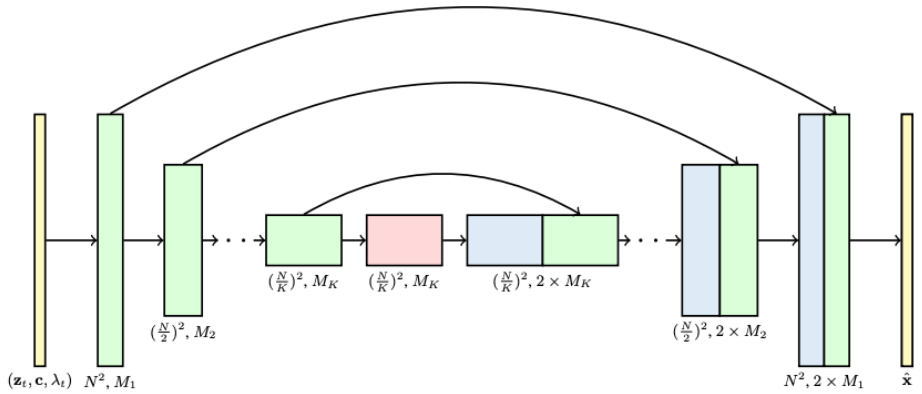
- Inverse problems
- 3D
- Video
- Miscellaneous

- Video generative models
- Video style transfer methods
- Video editing methods

Video Diffusion Models

3D UNet from a 2D UNet.

- 3x3 2d conv to 1x3x3 3d conv.
- Factorized spatial and temporal attentions.



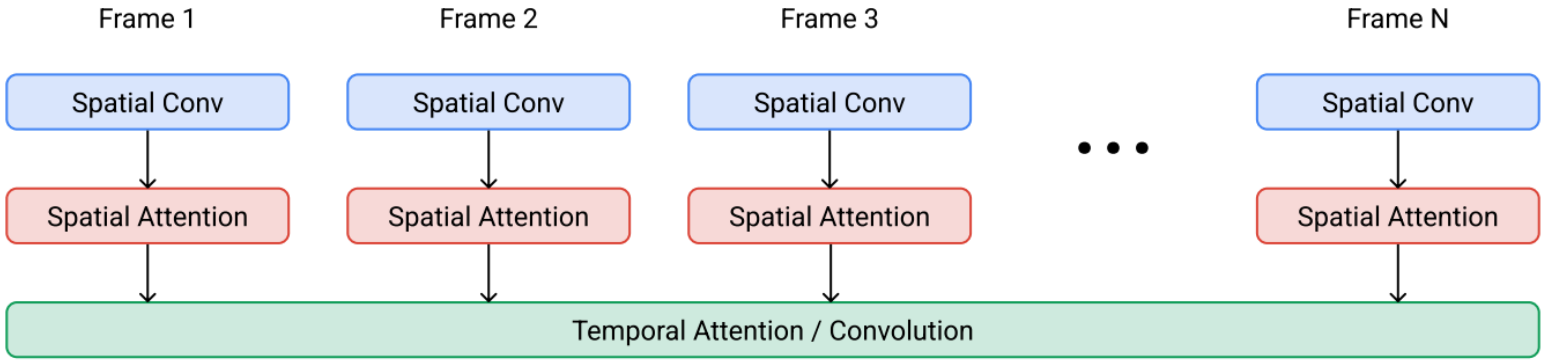
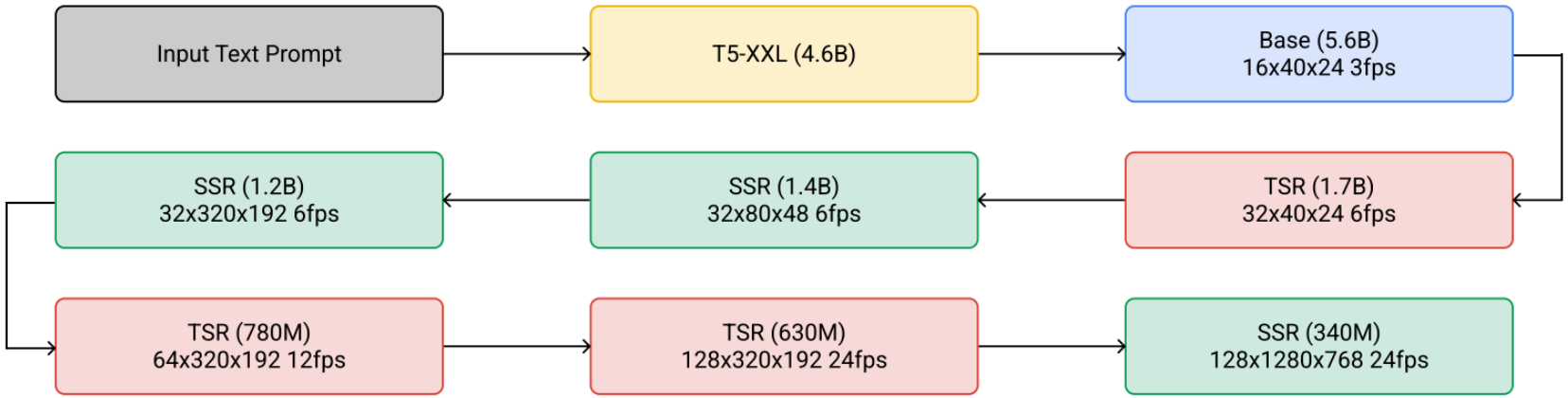


Illustration on how the 3d attention is factorized (from Imagen video)

Ho et al., "Video Diffusion Models", NeurIPS 2022

Imagen Video: Large Scale Text-to-Video

- 7 cascade models in total.
- 1 Base model (16x40x24)
- 3 Temporal super-resolution models.
- 3 Spatial super-resolution models.



Ho et al., "Imagen Video: High Definition Video Generation with Diffusion Models", 2022

Make-a-Video

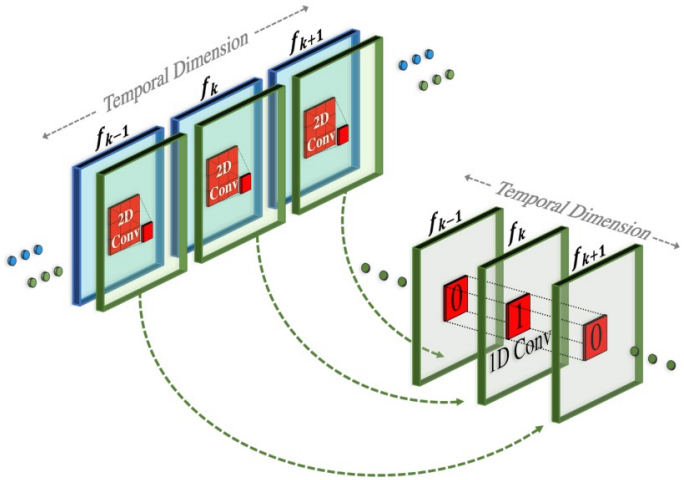
- Start with an unCLIP (DALL-E 2) base network.

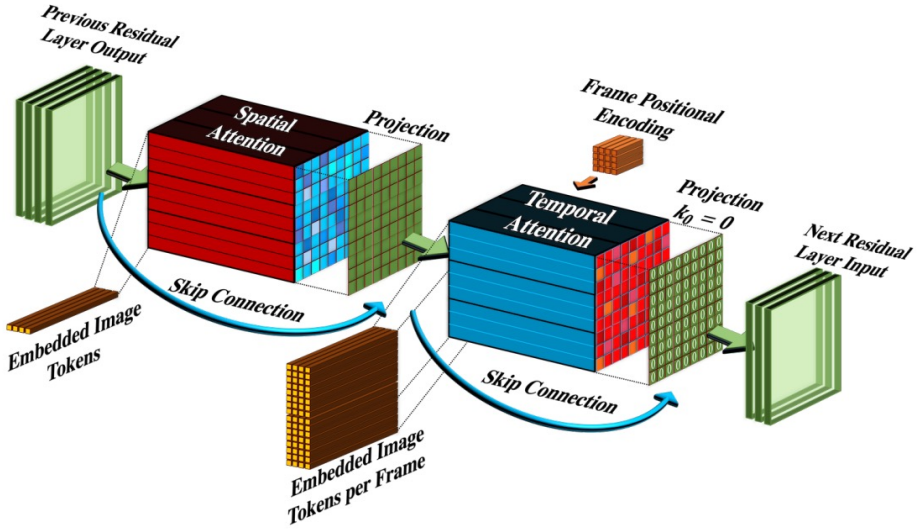
$$\hat{y}_t = \mathrm{SR}_h \circ \mathrm{SR}_l^t \circ \uparrow_F \circ \mathrm{D}^t \circ \mathrm{P} \circ (\hat{x}, \mathrm{C}_x(x)),$$

SR_h	Spatial SR
SR_l^t	Spatialtemporal SR
$\uparrow_F$	Frame interpolation
D^t	Spatialtemporal decoder
P	Prior

Singer et al., "Make-A-Video: Text-to-Video Generation without Text-Video Data", ICLR 2023

Make-a-Video





3D Conv from Spatial Conv + Temporal Conv

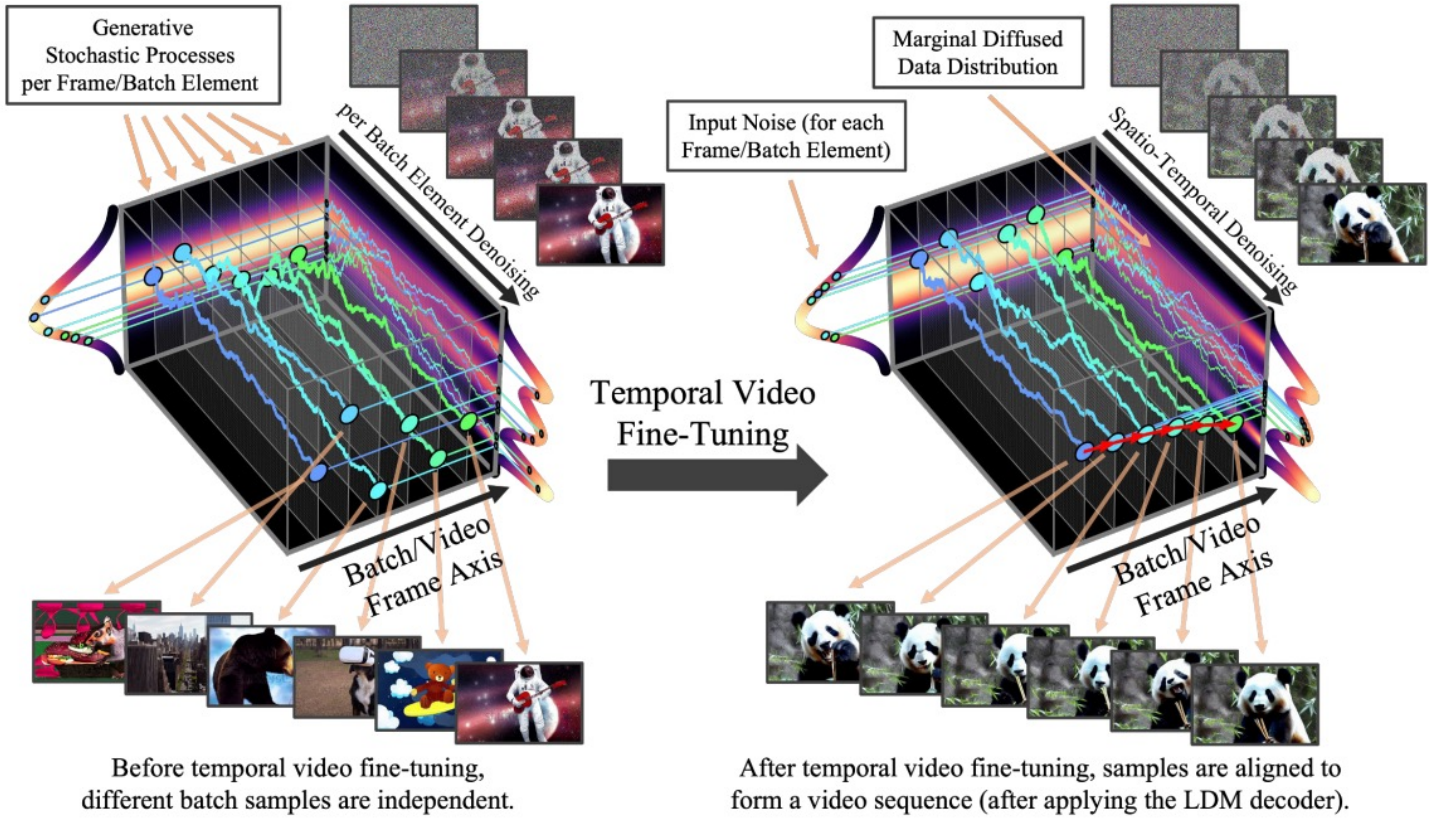
3D Attn from Spatial Attn + Temporal Attn

Different from Imagen Video, only the image prior takes text as input!

Singer et al., "Make-A-Video: Text-to-Video Generation without Text-Video Data", ICLR 2023

Video LDM

- Similarly, fine-tune a text-to-video model from text-to-image model.



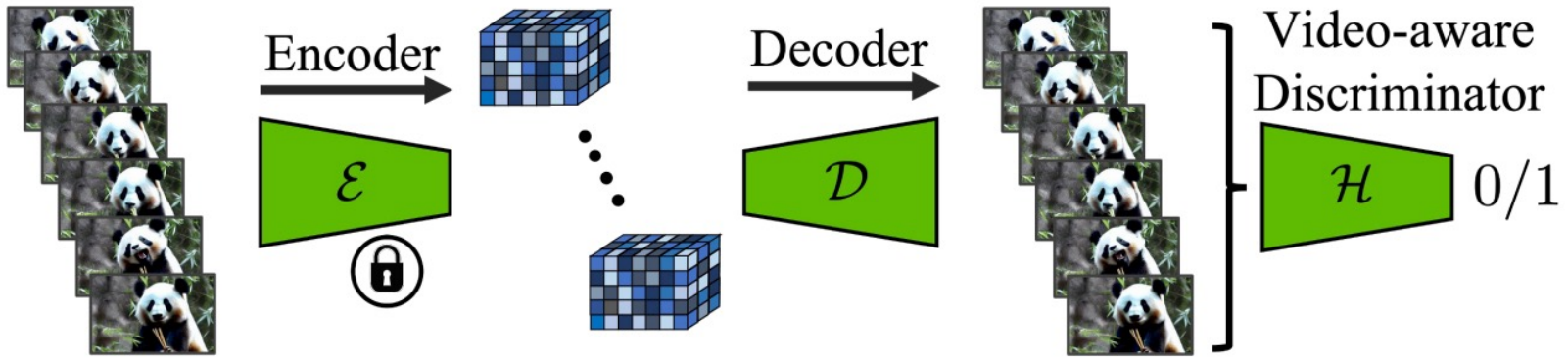
Before temporal video fine-tuning, different batch samples are independent.

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

Blattmann et al., "Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models", CVPR 2023

Video LDM: Decoder Fine-tuning

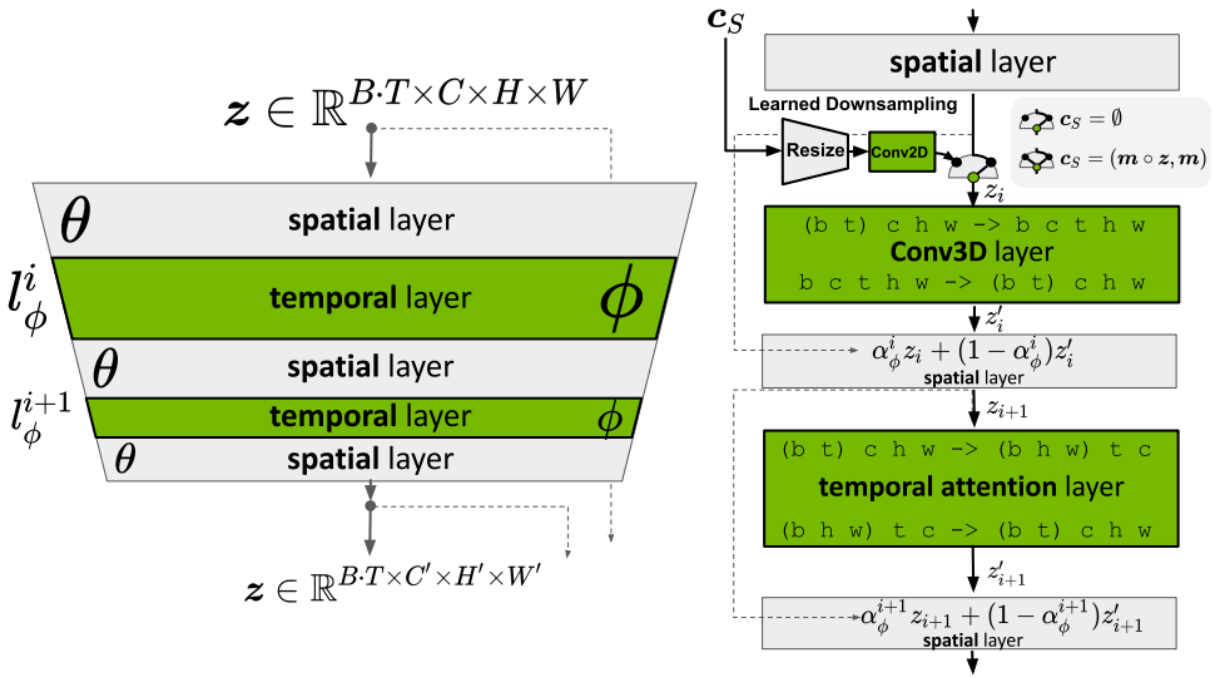
- Fine-tune the decoder to be video-aware, keeping encoder frozen.



Blattmann et al., "Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models", CVPR 2023

Video LDM: LDM Fine-tuning

- Interleave spatial layers and temporal layers.
- The spatial layers are frozen, whereas temporal layers are trained.
- Temporal layers can be Conv3D or Temporal attentions.
- Context can be added for autoregressive generation.



Optional context via learned down-sampling operation.

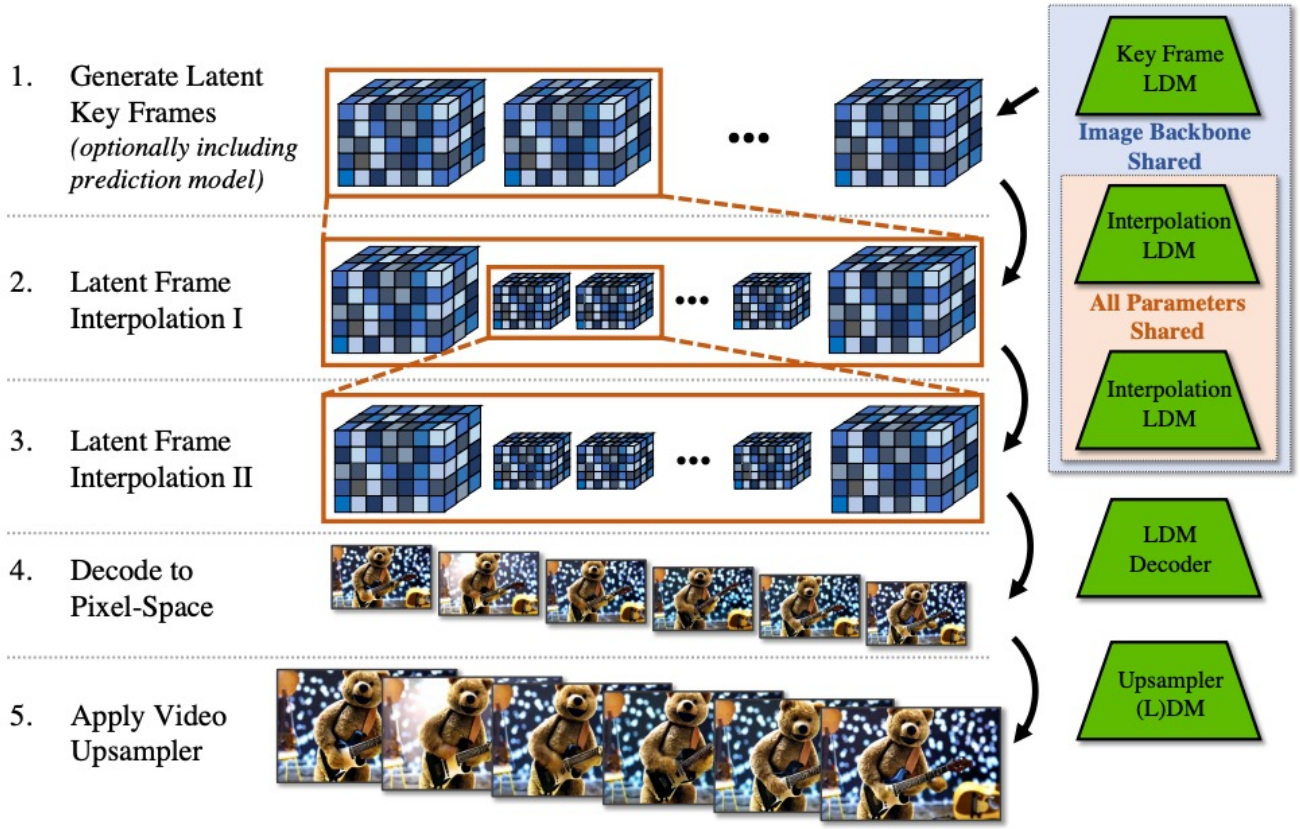
For **Conv3D**,
shape is [batch, time, channel, height, width]

For **Temporal attention**,
shape is [batch * height * width, time, channel]

Blattmann et al., "Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models", CVPR 2023

Video LDM: Upsampling

- After key latent frames are generated, the latent frames go through temporal interpolation.
- Then, they are decoded to pixel space and optionally upsampled.



Blattmann et al., "Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models", CVPR 2023

Outline

- Inverse problems
- 3D
- Video
- Miscellaneous

- Video generative models
- Video style transfer / editing methods

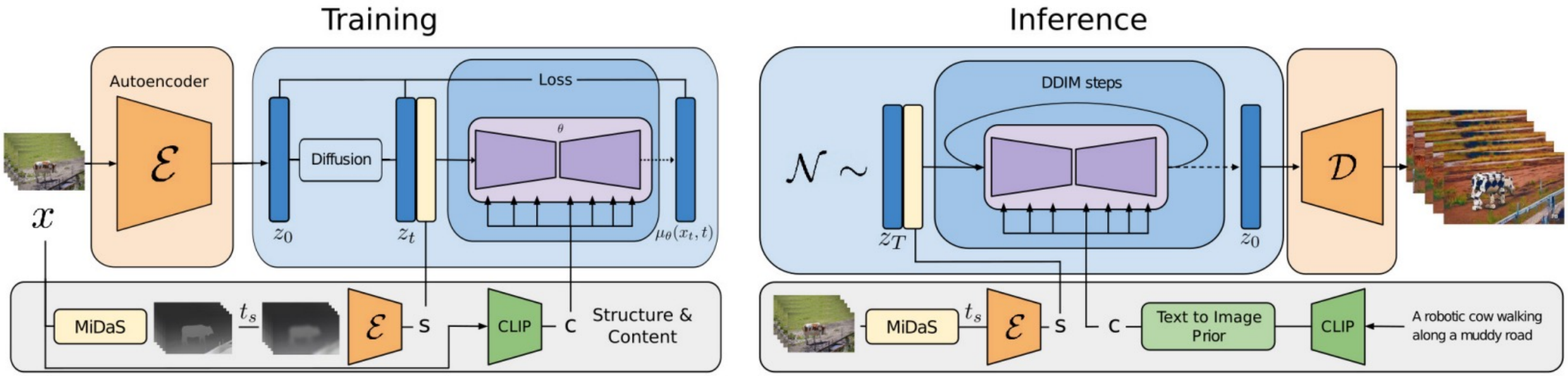
Gen-1

- Transfer the style of a video using text prompts given a “driving video”

Esser et al., "Structure and Content-Guided Video Synthesis with Diffusion Models", arXiv 2023

Gen-1

- Condition on structure (depth) and content (CLIP) information.
- Depth maps are passed with latents as input conditions.
- CLIP image embeddings are provided via cross-attention blocks.
- During inference, CLIP text embeddings are converted to CLIP image embeddings.



Esser et al., "Structure and Content-Guided Video Synthesis with Diffusion Models", arXiv 2023

Pix2Video: Video Editing Using Image Diffusion

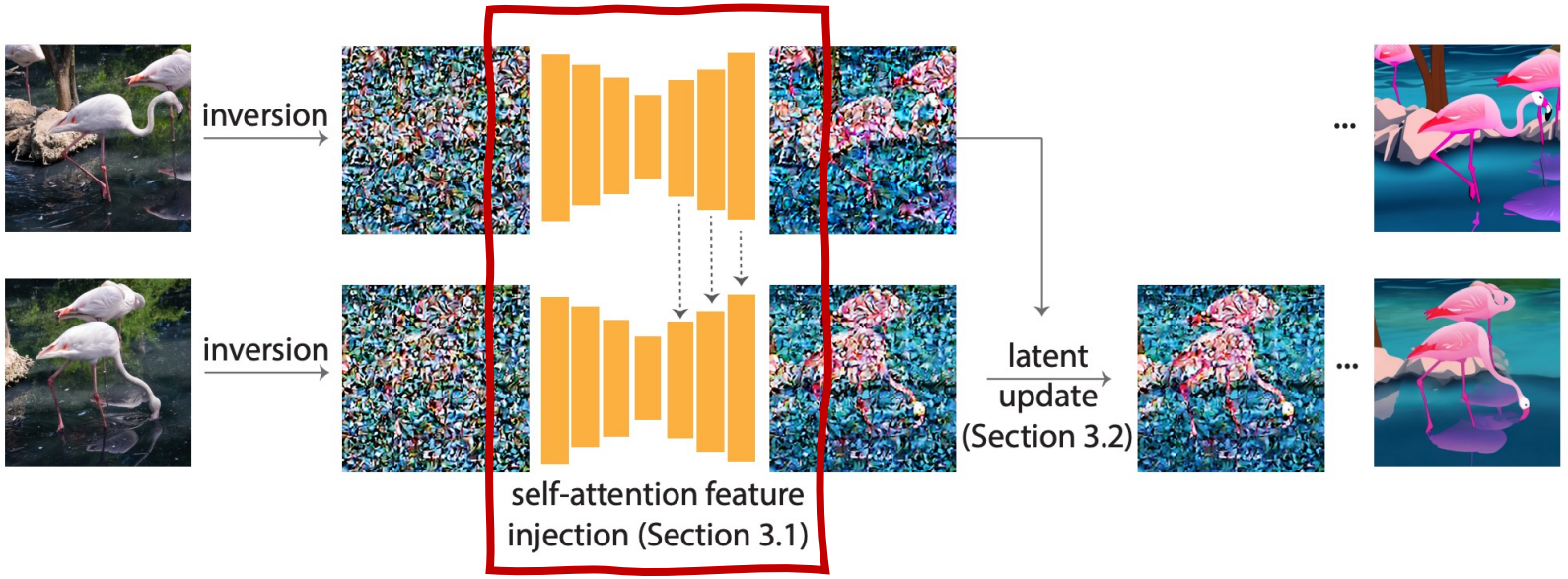
- Given a sequence of frames, generate a new set of images that reflects an edit.
- Editing methods on individual images fail to preserve temporal information.

a group of 8-bit pixelated fish swimming in an aquarium

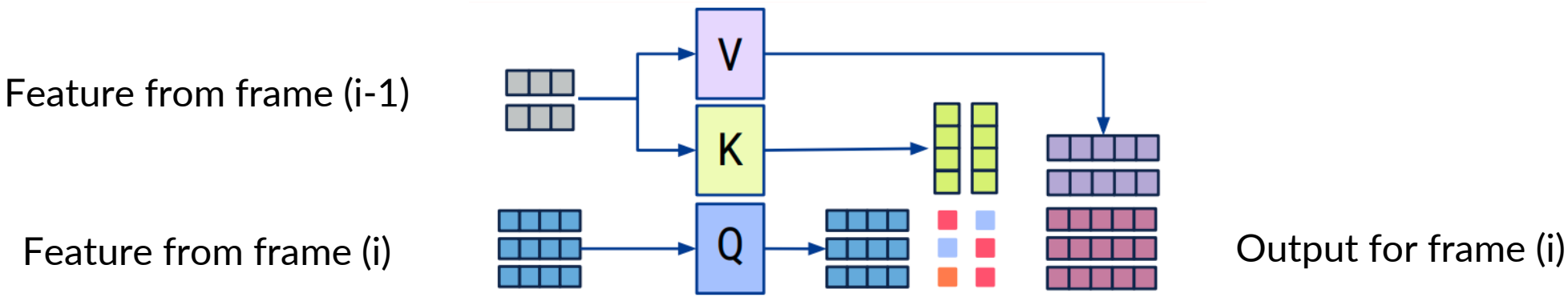
Inconsistent frames!

Ceylan et al., "Pix2Video: Video Editing using Image Diffusion", arXiv 2023

Pix2Video: Video Editing Using Image Diffusion

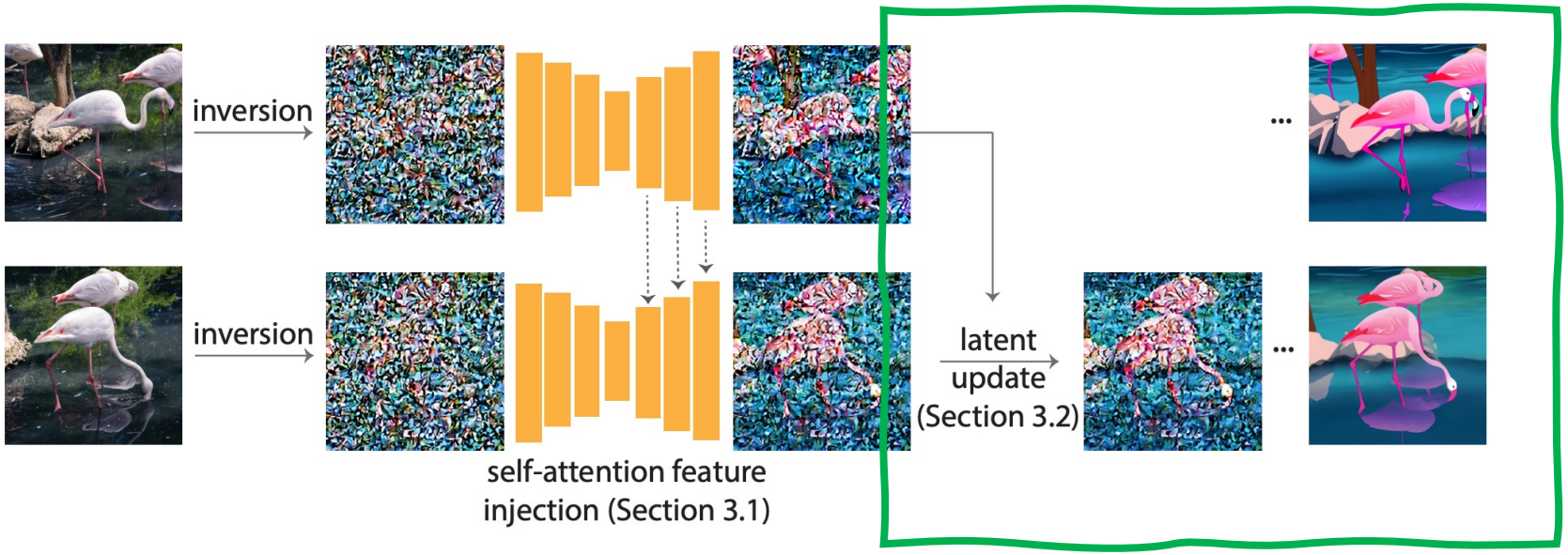


- Self-Attention injection: use the features of previous frame for Key and Values.



Ceylan et al., "Pix2Video: Video Editing using Image Diffusion", arXiv 2023

Pix2Video: Video Editing Using Image Diffusion



- Guided latent update: use “reconstruction guidance” for x0 prediction at frame (i)

$$g(\hat{x}_0^{i,t}, \hat{x}_0^{i-1,t}) = \|\hat{x}_0^{i,t} - \hat{x}_0^{i-1,t}\|_2^2$$

$$x_{t-1}^i \leftarrow x_{t-1}^i - \delta_{t-1} \nabla_{x_t^i} g(\hat{x}_0^{t,i-1}, \hat{x}_0^{t,i}),$$

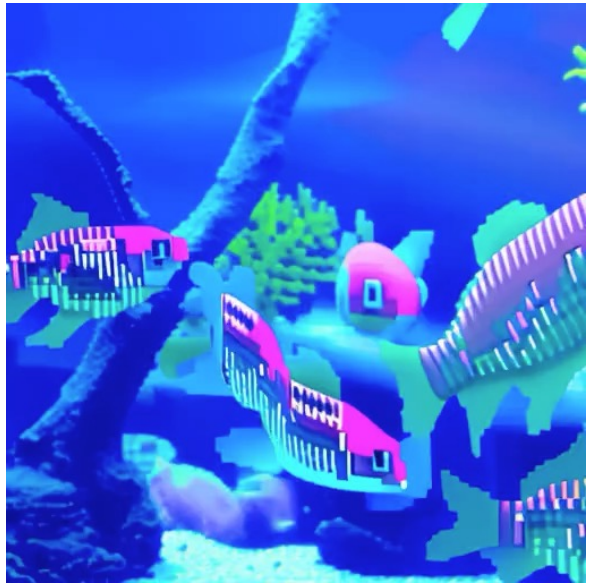
- This make frames (i) and (i-1) similar.

Ceylan et al., "Pix2Video: Video Editing using Image Diffusion", arXiv 2023

Pix2Video: Video Editing Using Image Diffusion

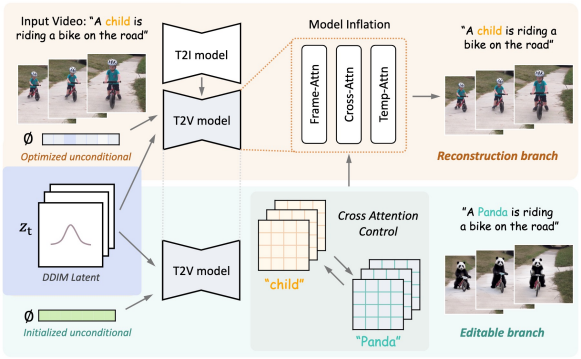
- The two methods improve the temporal consistency of the final video!

a group of 8-bit pixelated fish swimming in an aquarium

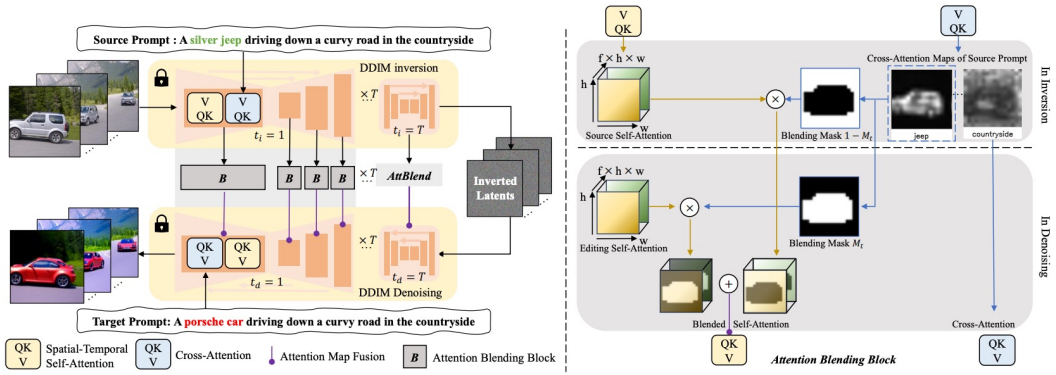


Ceylan et al., "Pix2Video: Video Editing using Image Diffusion", arXiv 2023

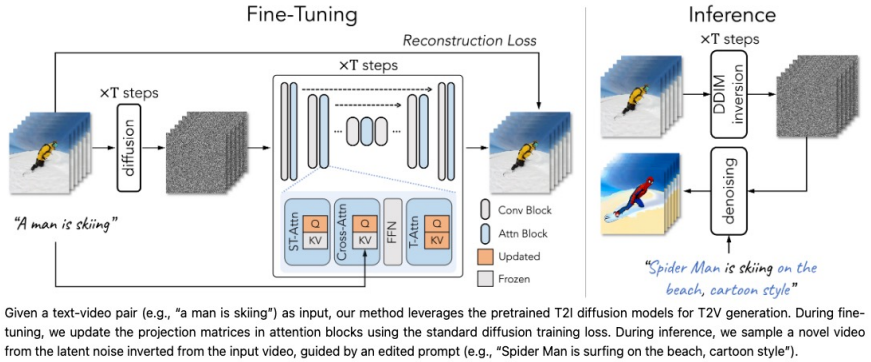
Concurrent / Related Works



Video-P2P: Cross-Attention Control on text-to-video model

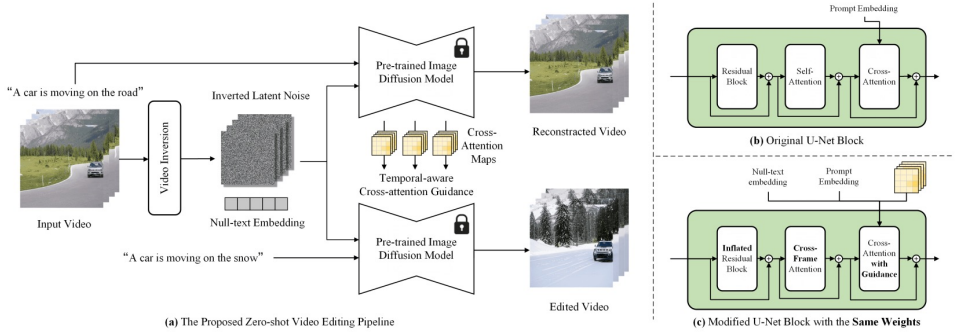


FateZero: Store attention maps from DDIM inversion for later use



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").

Tune-A-Video: Fine-tune projection matrices of the attention layers, from text2image model to text2video model.



Vid2vid-zero: Learn a null-text embedding for inversion, then use cross-frame attention with original weights.

Outline

- Inverse problems
- 3D
- Video
- Miscellaneous

- Diffusion models for large contents
- Safety and limitations of diffusion models

Diffusion Models for Large Contents

- Suppose model is trained on small, squared images, how to extend it to larger images?
- Outpainting is always a solution, but not a very efficient one!

Let us generate this image with a diffusion model only trained on squared regions:

1. Generate the center region $q(\mathbf{x}_1, \mathbf{x}_2)$
2. Generate the surrounding region conditioned on parts of the center image $q(\mathbf{x}_3|\mathbf{x}_2)$

$\mathbf{x}_1$ $\mathbf{x}_2$

$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$

Latency scales linearly with the content size!

Diffusion Models for Large Contents

- Unlike autoregressive models, diffusion models can generate large contents in parallel!

1. Generate the center region $q(\mathbf{x}_1, \mathbf{x}_2)$
2. Generate the surrounding region conditioned on parts of the center image $q(\mathbf{x}_3|\mathbf{x}_2)$

$\mathbf{x}_1$ $\mathbf{x}_2$

$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$

$$q(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) = q(\mathbf{x}_1, \mathbf{x}_2)q(\mathbf{x}_3|\mathbf{x}_2) = \frac{q(\mathbf{x}_1, \mathbf{x}_2)q(\mathbf{x}_2, \mathbf{x}_3)}{q(\mathbf{x}_2)}$$

$$\log q(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) = \log q(\mathbf{x}_1, \mathbf{x}_2) + \log q(\mathbf{x}_2, \mathbf{x}_3) - \log q(\mathbf{x}_2)$$

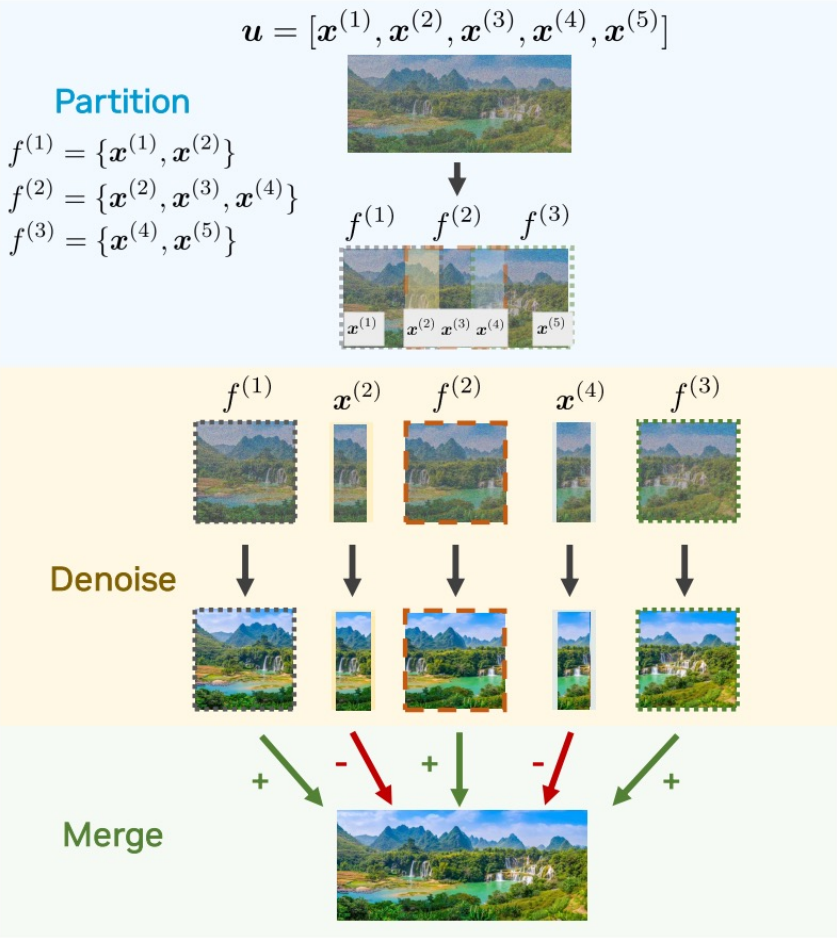
(Factor node) (Variable node)

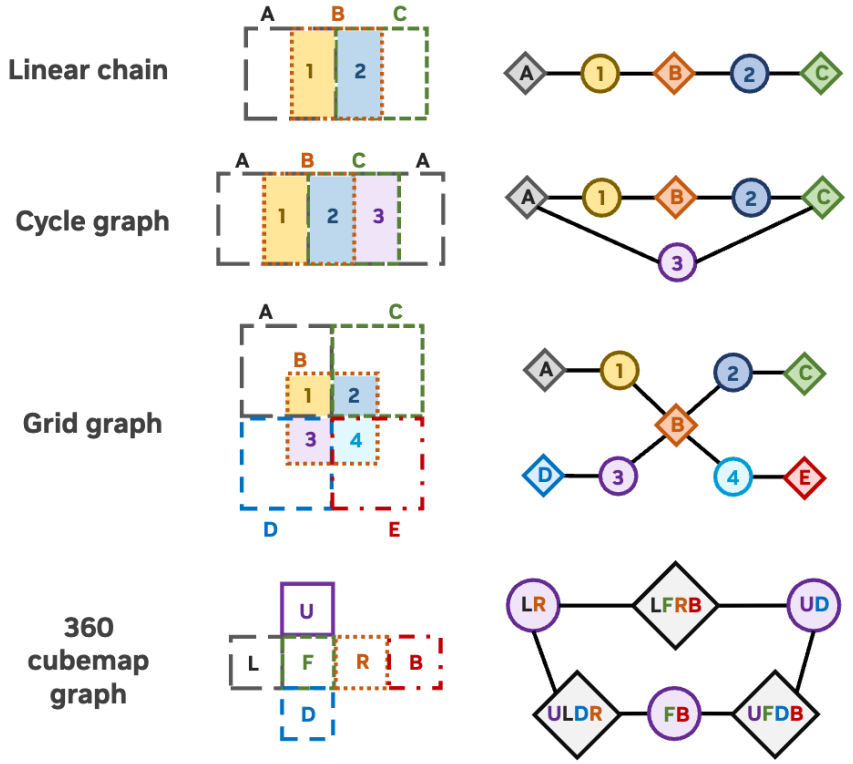
Scores of $q(\mathbf{x}_1, \mathbf{x}_2), q(\mathbf{x}_2, \mathbf{x}_3), q(\mathbf{x}_2)$ can be run in parallel with diffusion model!

Zhang et al., "DiffCollage: Parallel Generation of Large Content with Diffusion Models", CVPR 2023

Diffusion Models for Large Contents

- A “large” diffusion model from “small” diffusion models!





Example factor graphs

Zhang et al., "DiffCollage: Parallel Generation of Large Content with Diffusion Models", CVPR 2023

Diffusion Models for Large Contents

- Applications such as long images, looped motion, 360 images…

*Prompt: A person **runs** forward, then **kicks** his legs, then **skips** rope, then **bends** down to pick something up off the ground.*

*Prompt: A person **runs** forward, then **skips** rope, then **bends** down to pick something up off the ground, then **kicks** his legs.*

Zhang et al., "DiffCollage: Parallel Generation of Large Content with Diffusion Models", CVPR 2023

Related Works

- Based on similar ideas but differ in how overlapping regions are mixed.

Jiménez, "Mixture of Diffusers for scene composition and high resolution image generation", arXiv 2023
Bar-Tal et al., "MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation", ICML 2023

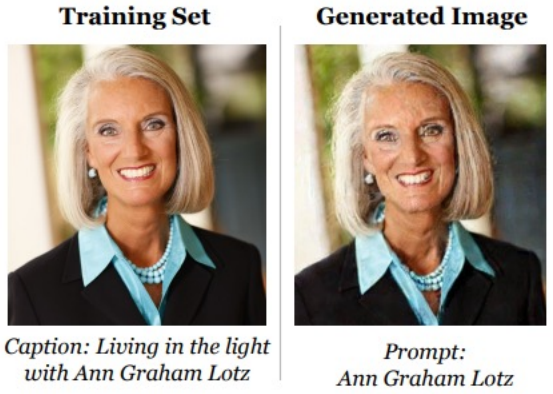
Outline

- Inverse problems
- 3D
- Video
- Miscellaneous

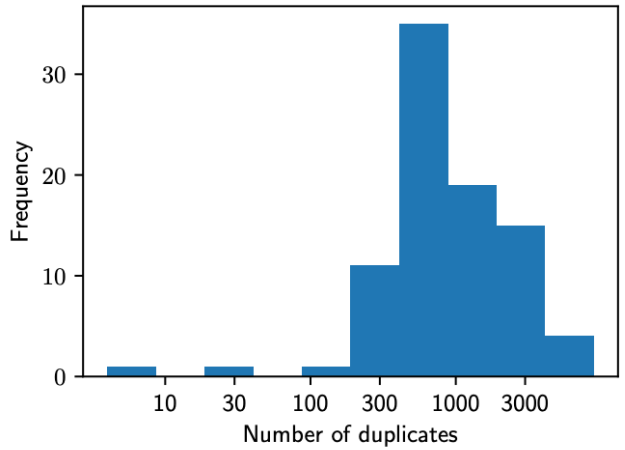
- Diffusion models for large contents
- Safety and limitations of diffusion models

Data Memorization in Diffusion Models

- Due to the likelihood-base objective function, diffusion models can ”memorize” data.
- And with a higher chance than GANs!
- Nevertheless, a lot of “memorized images” are highly-duplicated in the dataset.



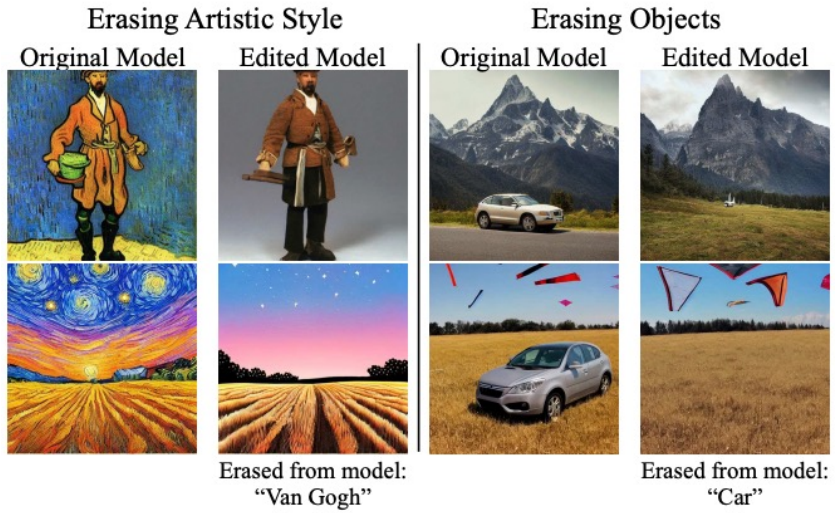
Architecture		Images Extracted	FID
GANs	StyleGAN-ADA [43]	**150**	**2.9**
	DiffBigGAN [82]	57	4.6
	E2GAN [69]	95	11.3
	NDA [63]	70	12.6
	WGAN-ALP [68]	49	13.0
DDPMs	OpenAI-DDPM [52]	**301**	**2.9**
	DDPM [33]	232	3.2

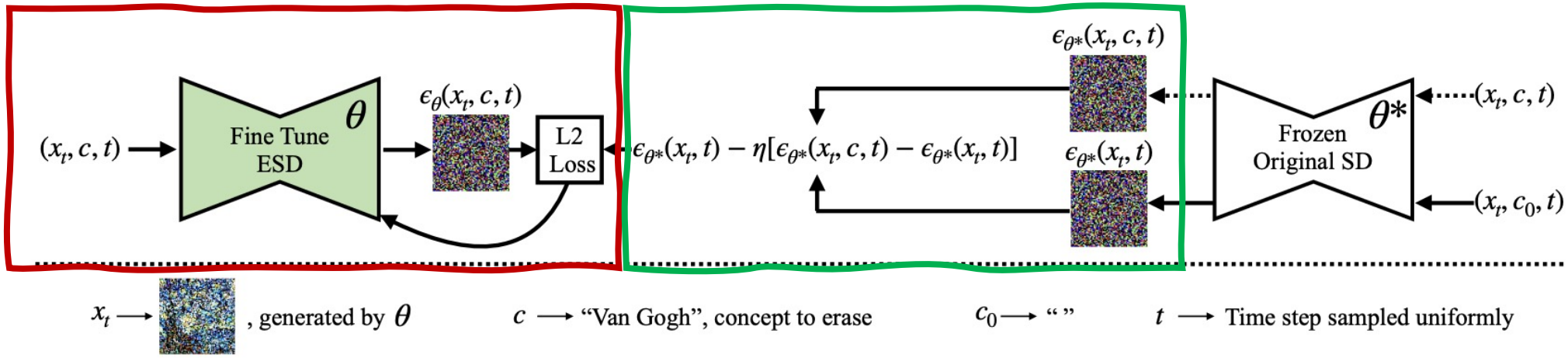


Carlini et al., "Extracting Training Data from Diffusion Models", arXiv 2023

Erasing Concepts in Diffusion Models

- Fine-tune a model to remove unwanted concepts.
- From original model, obtain score via negative CFG.
- A new model is fine-tuned from the new score function.





Gandikota et al., "Erasing Concepts from Diffusion Models", arXiv 2023

Summary

We covered a lot of topics: fundamentals, image applications, other applications.

There are many good papers on the topic, but we can only cover some of them!

We thank the community effort to give us a curated list of papers:
https://github.com/cvpr2023-tutorial-diffusion-models/papers/graphs/contributors

Slides and recording will be available at:

https://cvpr2023-tutorial-diffusion-models.github.io/

Thank you!

Finally, the list of contributed papers to our repo.

Part I

Ho et al., "Denoising Diffusion Probabilistic Models", NeurIPS 2020
Song et al., "Score-Based Generative Modeling through Stochastic Differential Equations", ICLR 2021
Kingma et al., "Variational Diffusion Models", arXiv 2021
Karras et al., "Elucidating the Design Space of Diffusion-Based Generative Models", NeurIPS 2022
Song et al., "Denoising Diffusion Implicit Models", ICLR 2021
Jolicoeur-Martineau et al., "Gotta Go Fast When Generating Data with Score-Based Models", arXiv 2021
Liu et al., "Pseudo Numerical Methods for Diffusion Models on Manifolds", ICLR 2022
Lu et al., "DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps", NeurIPS 2022
Lu et al., "DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models", NeurIPS 2022
Zhang and Chen, "Fast Sampling of Diffusion Models with Exponential Integrator", arXiv 2022
Zhang et al., "gDDIM: Generalized denoising diffusion implicit models", arXiv 2022
Zhao et al., "UniPC: A Unified Predictor-Corrector Framework for Fast Sampling of Diffusion Models", arXiv 2023
Shih et al., "Parallel Sampling of Diffusion Models", arxiv 2023
Chen et al., "A Geometric Perspective on Diffusion Models", arXiv 2023
Xiao et al., "Tackling the Generative Learning Trilemma with Denoising Diffusion GANs", arXiv 2021
Salimans and Ho, "Progressive Distillation for Fast Sampling of Diffusion Models", ICLR 2022
Meng et al., "On Distillation of Guided Diffusion Models", arXiv 2022
Dockhorn et al., "GENIE: Higher-Order Denoising Diffusion Solvers", NeurIPS 2022
Watson et al., "Learning Fast Samplers for Diffusion Models by Differentiating Through Sample Quality", ICLR 2022
Phung et al., "Wavelet Diffusion Models Are Fast and Scalable Image Generators", CVPR 2023
Dhariwal and Nichol, "Diffusion Models Beat GANs on Image Synthesis", arXiv 2021
Ho and Salimans, "Classifier-Free Diffusion Guidance", NeurIPS Workshop 2021
Automatic1111, "Negative Prompt", GitHub
Hong et al., "Improving Sample Quality of Diffusion Models Using Self-Attention Guidance", arXiv 2022
Saharia et al., "Image Super-Resolution via Iterative Refinement", arXiv 2021
Ho et al., "Cascaded Diffusion Models for High Fidelity Image Generation", JMLR 2021
Sinha et al., "D2C: Diffusion-Denoising Models for Few-shot Conditional Generation", NeurIPS 2021
Vahdat et al., "Score-based Generative Modeling in Latent Space", arXiv 2021
Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models", CVPR 2022
Daras et al., "Score-Guided Intermediate Layer Optimization: Fast Langevin Mixing for Inverse Problems", ICML 2022

Part I (cont'd)

Bortoli et al., "Diffusion Schrödinger Bridge", NeurIPS 2021
Bortoli et al., "Riemannian Score-Based Generative Modelling", NeurIPS 2022
Neklyudov et al., "Action Matching: Learning Stochastic Dynamics from Samples", ICML 2023
Bansal et al., "Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise", arXiv 2022
Daras et al., "Soft Diffusion: Score Matching for General Corruptions", TMLR 2023
Delbracio and Milanfar, "Inversion by Direct Iteration: An Alternative to Denoising Diffusion for Image Restoration", arXiv 2023
Luo et al., "Image Restoration with Mean-Reverting Stochastic Differential Equations", ICML 2023

Part II

Bao et al., "All are Worth Words: a ViT Backbone for Score-based Diffusion Models", arXiv 2022
Peebles and Xie, "Scalable Diffusion Models with Transformers", arXiv 2022
Bao et al., "One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale", arXiv 2023
Jabri et al., "Scalable Adaptive Computation for Iterative Generation", arXiv 2022
Hoogeboom et al., "simple diffusion: End-to-end diffusion for high resolution images", arXiv 2023
Meng et al., "SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations", ICLR 2022
Li et al., "Efficient Spatially Sparse Inference for Conditional GANs and Diffusion Models", NeurIPS 2022
Avrahami et al., "Blended Diffusion for Text-driven Editing of Natural Images", CVPR 2022
Hertz et al., "Prompt-to-Prompt Image Editing with Cross-Attention Control", ICLR 2023
Kawar et al., "Imagic: Text-Based Real Image Editing with Diffusion Models", CVPR 2023
Couairon et al., "DiffEdit: Diffusion-based semantic image editing with mask guidance", ICLR 2023
Sarukkai et al., "Collage Diffusion", arXiv 2023
Bar-Tal et al., "MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation", ICML 2023
Gal et al., "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion", ICLR 2023
Ruiz et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation", CVPR 2023
Kumari et al., "Multi-Concept Customization of Text-to-Image Diffusion", CVPR 2023
Tewel et al., "Key-Locked Rank One Editing for Text-to-Image Personalization", SIGGRAPH 2023
Zhao et al., "A Recipe for Watermarking Diffusion Models", arXiv 2023
Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models", ICLR 2022
Li et al., "GLIGEN: Open-Set Grounded Text-to-Image Generation", CVPR 2023

Part II (cont'd)

Avrahami et al., "SpaText: Spatio-Textual Representation for Controllable Image Generation", CVPR 2023
Zhang and Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models", arXiv 2023
Mou et al., "T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models", arXiv 2023
Orgad et al., "Editing Implicit Assumptions in Text-to-Image Diffusion Models", arXiv 2023
Han et al., "SVDiff: Compact Parameter Space for Diffusion Fine-Tuning", arXiv 2023
Xie et al., "DiffFit: Unlocking Transferability of Large Diffusion Models via Simple Parameter-Efficient Fine-Tuning", arXiv 2023
Saharia et al., "Palette: Image-to-Image Diffusion Models", SIGGRAPH 2022
Whang et al., "Deblurring via Stochastic Refinement", CVPR 2022
Xu et al., "Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models", arXiv 2023
Saxena et al., "Monocular Depth Estimation using Diffusion Models", arXiv 2023
Li et al., "Your Diffusion Model is Secretly a Zero-Shot Classifier", arXiv 2023
Gowal et al., "Improving Robustness using Generated Data", NeurIPS 2021
Wang et al., "Better Diffusion Models Further Improve Adversarial Training", ICML 2023

Part III

Jalal et al., "Robust Compressed Sensing MRI with Deep Generative Priors", NeurIPS 2021
Song et al., "Solving Inverse Problems in Medical Imaging with Score-Based Generative Models", ICLR 2022
Kawar et al., "Denoising Diffusion Restoration Models", NeurIPS 2022
Chung et al., "Improving Diffusion Models for Inverse Problems using Manifold Constraints", NeurIPS 2022
Ryu and Ye, "Pyramidal Denoising Diffusion Probabilistic Models", arXiv 2022
Chung et al., "Diffusion Posterior Sampling for General Noisy Inverse Problems", arXiv 2022
Feng et al., "Score-Based Diffusion Models as Principled Priors for Inverse Imaging", arXiv 2023
Song et al., "Pseudoinverse-Guided Diffusion Models for Inverse Problems", ICLR 2023
Mardani et al., "A Variational Perspective on Solving Inverse Problems with Diffusion Models", arXiv 2023
Delbracio and Milanfar, "Inversion by Direct Iteration: An Alternative to Denoising Diffusion for Image Restoration", arxiv 2023
Stevens et al., "Removing Structured Noise with Diffusion Models", arxiv 2023
Wang et al., "Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model", ICLR 2023
Zhou et al., "3D Shape Generation and Completion through Point-Voxel Diffusion", ICCV 2021
Zeng et al., "LION: Latent Point Diffusion Models for 3D Shape Generation", NeurIPS 2022
Nichol et al., "Point-E: A System for Generating 3D Point Clouds from Complex Prompts", arXiv 2022
Chou et al., "DiffusionSDF: Conditional Generative Modeling of Signed Distance Functions", arXiv 2022
Cheng et al., "SDFusion: Multimodal 3D Shape Completion, Reconstruction, and Generation", arXiv 2022
Hui et al., "Neural Wavelet-domain Diffusion for 3D Shape Generation", arXiv 2022
Shue et al., "3D Neural Field Generation using Triplane Diffusion", arXiv 2022
Yang et al., "Learning a Diffusion Prior for NeRFs", ICLR Workshop 2023
Jun and Nichol, "Shap-E: Generating Conditional 3D Implicit Functions", arXiv 2023
Poole et al., "DreamFusion: Text-to-3D using 2D Diffusion", arXiv 2022
Lin et al., "Magic3D: High-Resolution Text-to-3D Content Creation", arXiv 2022
Wang et al., "Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation", arXiv 2022
Metzer et al., "Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures", arXiv 2022
Hong et al., "Debiasing Scores and Prompts of 2D Diffusion for Robust Text-to-3D Generation", CVPR Workshop 2023
Watson et al., "Novel View Synthesis with Diffusion Models", arXiv 2022
Chan et al., "Generative Novel View Synthesis with 3D-Aware Diffusion Models", arXiv 2023
Xu et al., "NeuralLift-360: Lifting An In-the-wild 2D Photo to A 3D Object with 360° Views", arXiv 2022
Zhou and Tulsiani, "SparseFusion: Distilling View-conditioned Diffusion for 3D Reconstruction", arXiv 2022

Part III (cont'd)

Seo et al., "DITTO-NeRF: Diffusion-based Iterative Text To Omni-directional 3D Model", arXiv 2023
Haque et al., "Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions", arXiv 2023
Sella et al., "Vox-E: Text-guided Voxel Editing of 3D Objects", arXiv 2023
Ho et al., "Video Diffusion Models", NeurIPS 2022
Harvey et al., "Flexible Diffusion Modeling of Long Videos", arXiv 2022
Voleti et al., "MCVD: Masked Conditional Video Diffusion for Prediction, Generation, and Interpolation", NeurIPS 2022
Ho et al., "Imagen Video: High Definition Video Generation with Diffusion Models", Oct 2022
Singer et al., "Make-A-Video: Text-to-Video Generation without Text-Video Data", arXiv 2022
Mei and Patel, "VIDM: Video Implicit Diffusion Models", arXiv 2022
Blattmann et al., "Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models", CVPR 2023
Wang et al., "Zero-Shot Video Editing Using Off-The-Shelf Image Diffusion Models", arXiv 2023
Ceylan et al., "Pix2Video: Video Editing using Image Diffusion", arXiv 2023
Esser et al., "Structure and Content-Guided Video Synthesis with Diffusion Models", arXiv 2023
Jiménez, "Mixture of Diffusers for scene composition and high resolution image generation", arXiv 2023
Bar-Tal et al., "MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation", arXiv 2023
Zhang et al., "DiffCollage: Parallel Generation of Large Content with Diffusion Models", CVPR 2023
Zhang et al., "MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model", arXiv 2022
Tevet et al., "Human Motion Diffusion Model", arXiv 2022
Chen et al., "Executing your Commands via Motion Diffusion in Latent Space", CVPR 2023
Du et al., "Avatars Grow Legs: Generating Smooth Human Motion from Sparse Tracking Inputs with Diffusion Model", CVPR 2023
Shafir et al., "Human Motion Diffusion as a Generative Prior", arXiv 2023
Somepalli et al., "Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models", CVPR 2023
Carlini et al., "Extracting Training Data from Diffusion Models", arXiv 2023
Gandikota et al., "Erasing Concepts from Diffusion Models", arXiv 2023
Kumari et al., "Ablating Concepts in Text-to-Image Diffusion Models", arXiv 2023
Somepalli et al., "Understanding and Mitigating Copying in Diffusion Models", arXiv 2023

Questions?

Slides and recording will be available at:

https://cvpr2023-tutorial-diffusion-models.github.io/

Jiaming Song

Chenlin Meng

Arash Vahdat

Thank you!