SIGGRAPH 2024 Course: Generative Models for Visual Content Editing and Creation

ANYI RAO, Stanford University, USA
YUANBO XIANGLI, Cornell University, USA
YUWEI GUO, Chinese University of Hong Kong, China
MIA TANG, Stanford University, USA
CHENLIN MENG, Stanford University, USA
MANEESH AGRAWALA, Stanford University, USA

Fig. 1. Synthetic images generated with ControlNet [Zhang et al. 2023]

Interest in generative models is surging in academia and industry, with their impressive capabilities and creativity outputs. Crucially, these models are also providing users with a growing degree of control over the generation process via texts or visuals prompts. Concretely, large-scale text-to-image foundation models like Stable Diffusion [Rombach et al. 2021],

Authors' addresses: Anyi Rao, Stanford University, Palo Alto, CA, USA, anyirao@stanford.edu; Yuanbo Xiangli, Cornell University, New York, USA; Yuwei Guo, Chinese University of Hong Kong, Hong Kong, China; Mia Tang, Stanford University, Palo Alto, CA, USA; Chenlin Meng, Stanford University, Palo Alto, CA, USA; Maneesh Agrawala, Stanford University, Palo Alto, CA, USA.

0730-0301/2024/8-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

SDXL [Podell et al. 2023], eDiff-I [Balaji et al. 2022], DALL-E 3 [Betker et al. 2023]; text-to-video foundation models like Imagen Video [Ho et al. 2022] and Make-a-video [Singer et al. 2022], Sora [OpenAI 2024] have boosted the growth of visual content editing and generation. Representatively, works such as AnimateDiff [Guo et al. 2023], ControlNet [Zhang et al. 2023] democratized video creation with diverse user-defined conditions, and have become practical tools for graphic designs and personalized media. There has also been a revolution in 3D asset generation in terms of fidelity and efficiency. Harvesting the powerful prior of 2D diffusion models, works such as DreamFusion [Poole et al. 2022], Magic3D [Lin et al. 2023], Zero123 [Liu et al. 2023], Wonder3D [Long et al. 2023] were enabled high-quality text- and image-to-3D object generation, with plausible geometry and physical properties to support their usage in gaming and simulation tasks. At the meantime, the emergence of high-quality large-scale 3D data [Deitke et al. 2023a,b; Yu et al. 2023] also empowered direct generative model training in 3D space [Hong et al. 2023; Xu et al. 2023]. Inspired by the success of 3D asset generation, scene-level 3D synthesis also gained increasing interest. Work such as GeNVS [Chan et al. 2023], ReconFusion [Wu et al. 2023] also benefit from 2D diffusion priors to achieve high-quality novel view synthesis. Another branch of work, such as AssetField [Xiangli et al. 2023], BlockPlanner [Xu et al. 2021] regard scenes as a composition of 3D assets guided by layouts, that can be generatively modeled in a data-driven manner whilst guarantee user controllability.

This course covers the advances in generative models over the last few years, with a slight shift towards the controllability and creativity tasks enabled by generative models. We will first go over the fundamental machine learning and deep learning techniques relevant to generative models. Next, we will showcase recent representative work in controllable image, video and 3D content generation and compositional representation learning. Finally, we will conclude with a discussion on the future application of this technology, societal impact and open research problems. After the course, the attendees will learn basic knowledge about diffusion models and how such models can be applied to different applications.

P.S. Website: https://cveu.github.io/event/sig2024.html; Twitter: https://twitter.com/cveu_workshop

CCS Concepts: • **Information systems** → Multimedia content creation.

Additional Key Words and Phrases: Generative Models, Creativity Support

ACM Reference Format:

Anyi Rao, Yuanbo Xiangli, Yuwei Guo, Mia Tang, Chenlin Meng, and Maneesh Agrawala. 2024. SIGGRAPH 2024 Course: Generative Models for Visual Content Editing and Creation. *ACM Trans. Graph.* 1, 1 (August 2024), 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

REFERENCES

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. 2022. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324* (2022).

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf* 2, 3 (2023), 8.

Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. 2023. GeNVS: Generative Novel View Synthesis with 3D-Aware Diffusion Models. In *arXiv*.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. 2023a. Objaverse-XL: A Universe of 10M+ 3D Objects. *arXiv preprint arXiv:2307.05663* (2023).

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023b. Objaverse: A Universe of Annotated 3D Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13142–13153.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2023. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. 2023. LRM: Large Reconstruction Model for Single Image to 3D. arXiv:2311.04400 [cs.CV]

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. 2023. Magic3D: High-Resolution Text-to-3D Content Creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot One Image to 3D Object. arXiv:2303.11328 [cs.CV]

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 2023. Wonder3D: Single Image to 3D using Cross-Domain Diffusion. *arXiv preprint arXiv:2310.15008* (2023).

OpenAI. 2024. Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2022. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv* (2022).

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. 2023. ReconFusion: 3D Reconstruction with Diffusion Priors. *arXiv* (2023).

Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Bo Dai, and Dahua Lin. 2023. AssetField: Assets Mining and Reconfiguration in Ground Feature Plane Representation. arXiv:2303.13953 [cs.CV]

Linning Xu, Yuanbo Xiangli, Anyi Rao, Nanxuan Zhao, Bo Dai, Ziwei Liu, and Dahua Lin. 2021. BlockPlanner: City Block Generation With Vectorized Graph Representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5077–5086.

Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, and Kai Zhang. 2023. DMV3D: Denoising Multi-View Diffusion using 3D Large Reconstruction Model. arXiv:2311.09217 [cs.CV]

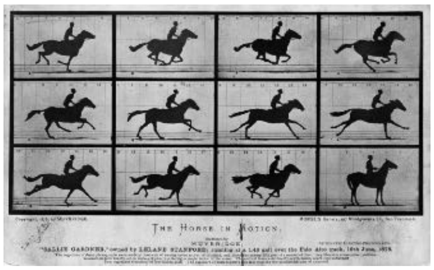
Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. 2023. MVImgNet: A Large-scale Dataset of Multi-view Images. In *CVPR*.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

The Birth of Videos

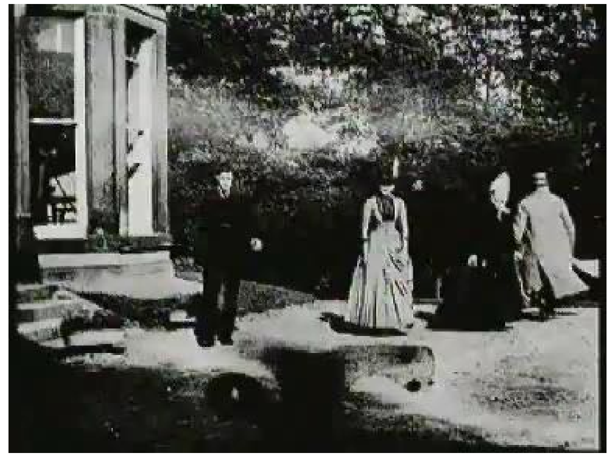
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

The Horse in Motion (1878)



The first motion picture ever made
Eadweard Muybridge

Roundhay Garden Scene (1888)



The first film with 20 frames
Louis Le Prince

https://www.thevintagenews.com/2016/06/27/46591-2/ https://headsup.scoutlife.org/what-was-the-first-movie-ever-made/#:~:text=Roundhay%20Garden%20Scene%20(1888),it%20is%20technically%20a%20movie

Video and Its Origins in Magic

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

The Vanishing Lady (1897)

Alter Time and Space through Editing
George Melies

Un Homme De Tete (1898)

The Father of Visual Effects
George Melies

Rao, Caba, etal, Oragnizing ICCV23, ECCV22, ICCV21 Creative Video Editing and Understanding Workshop

Creative Video and Its Origins in Magic

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

@kassupalen – TikTok 2020

@zachking – TikTok 2019

Rao, Caba, etal, Oragnizing ICCV23, ECCV22, ICCV21 Creative Video Editing and Understanding Workshop

Text to Video Generation: SORA

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

SORA: Prompt: The story of a robot's life in a cyberpunk setting.

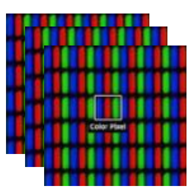
How Visual Content is Created?

Visual Content from Pixels

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

$rgb\ (w, h)$ t

appearance, width, height



HBO Opening 1993

[Paul Trillo https://mobile.twitter.com/paultrillo/status/1572301755710582784](https://mobile.twitter.com/paultrillo/status/1572301755710582784)

Visual Content from a Camera Navigating in the 3D Environment

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

$(x, y, z, \alpha, \beta, \gamma, f)\ t$

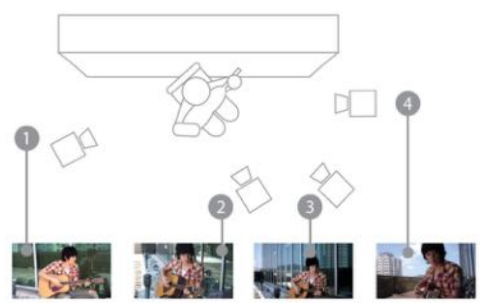
Position, Angle, Focal Length

Visual Content from Multi-view Editing

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

i, t

camera index, time



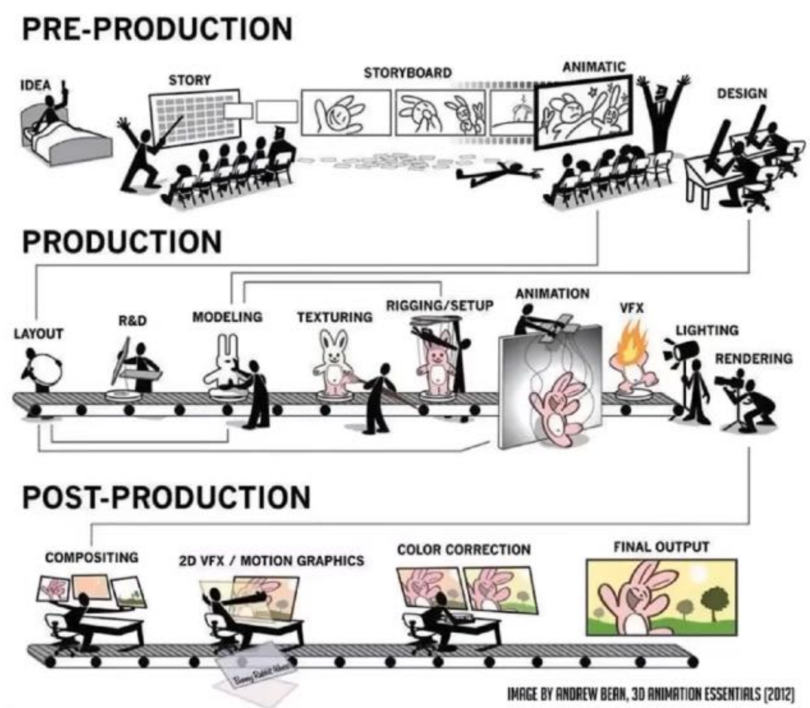
$(\alpha, \beta, f)\ t$

horizontal/vertical angle,focal, time

Visual Content from Professional Pipeline

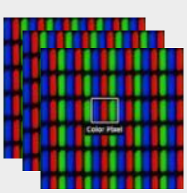
A Comprehensive workflow that combines

- **People's efforts**
- **Natural language processing**
- **Computer graphics**
- **Computer vision**
- **Animation**
- **VFX**
- **Artificial intelligence**
- **More…..**



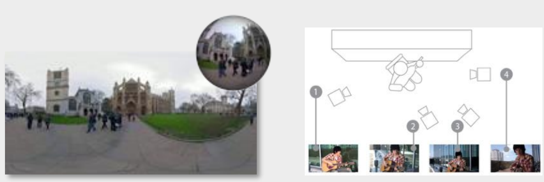
How Visual Content is Created?

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

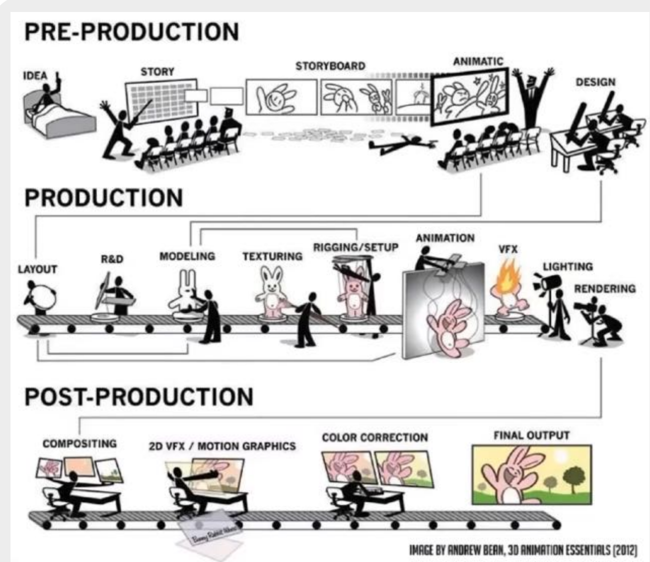


Pixels

Camera in 3D



Editing from Multi view Footage



Professional Video Pipeline

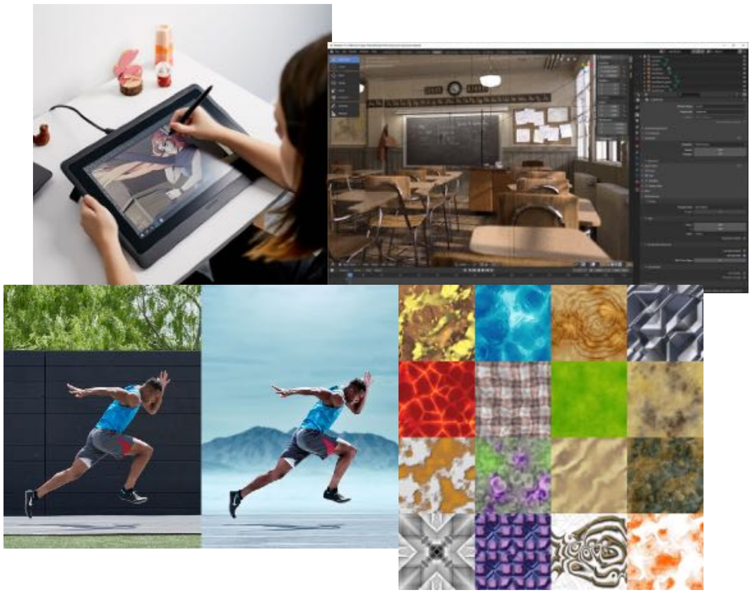
🤔 More e.g., Remixing, Augmenting, Editing ….

Introduction to Generative Models

Agenda

- Introduction to Diffusion
- Conditional generation and guidance
- Implementation Architectures

To make a beautiful synthetic image…

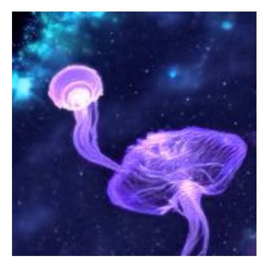


A cute corgi sitting on a beach sipping on a glass of lemonade. 4K, photorealistic.

Past

Now!

Generative AI Applications

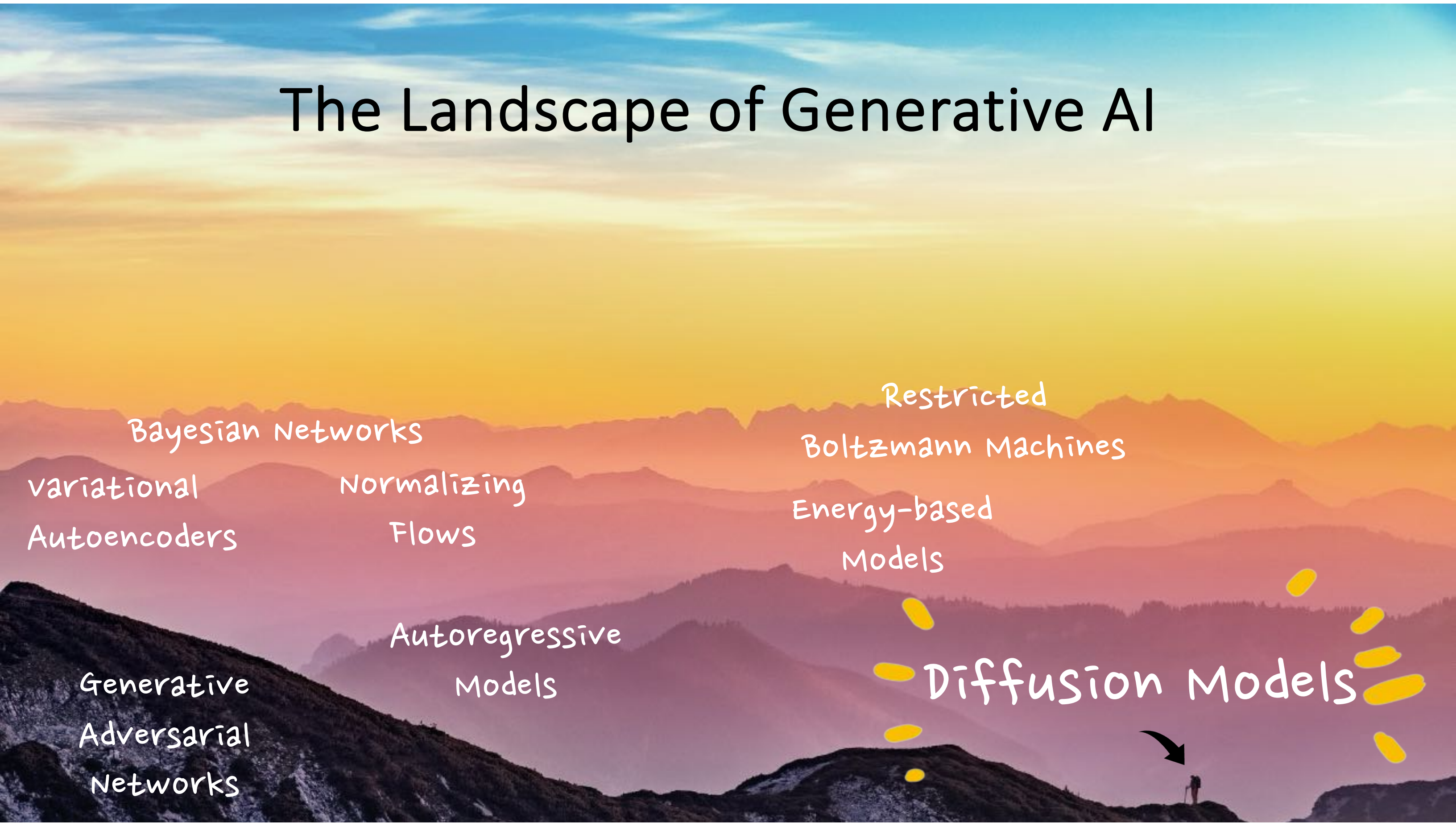


Art & Design

Content Generation

Entertainment

The Landscape of Generative AI



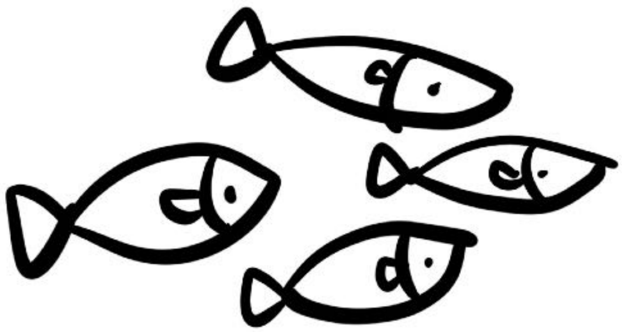
How do we generate new data?

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

our goal:

Generate fish that looks and behaves like it belongs to this river

How do we generate new data?

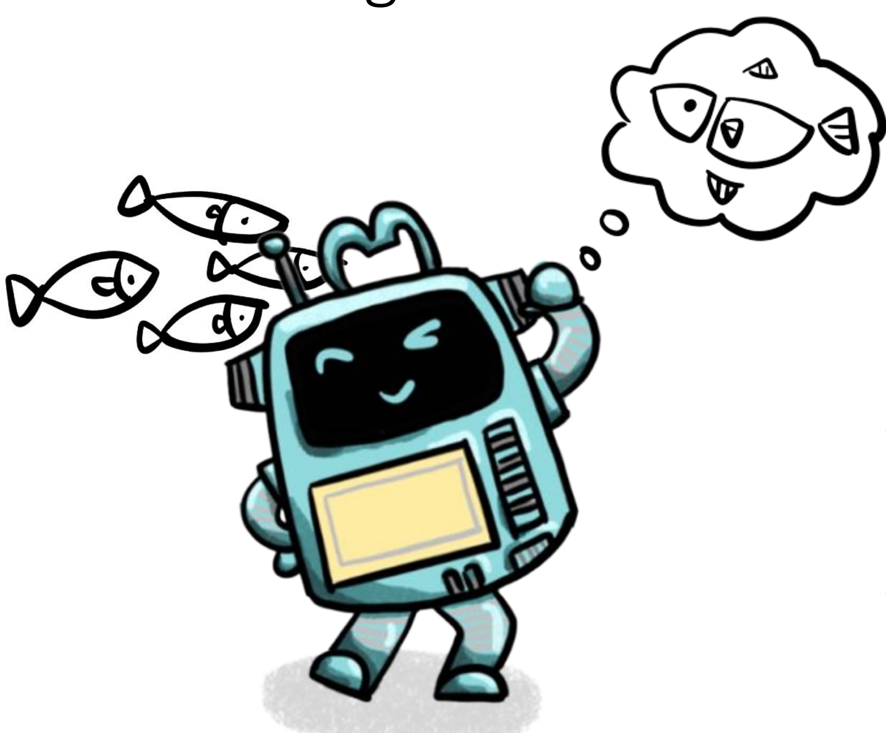


Step 1:

capture fish from the river

a lot of

How do we generate new data?

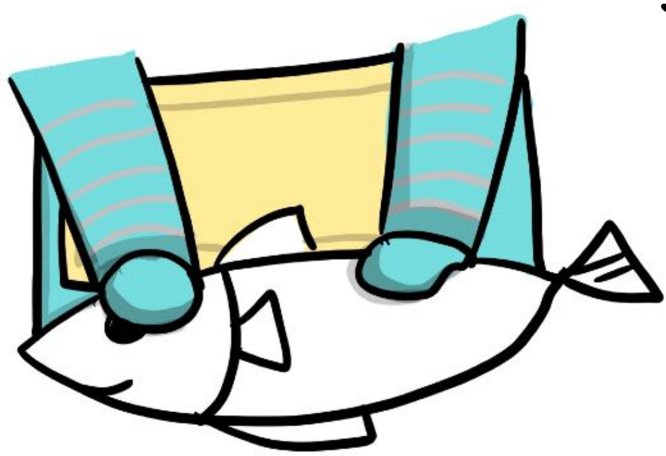


Step 2:

Train a neural network to learn the fish distribution by analyzing the captured fish!

How do we generate new data?

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

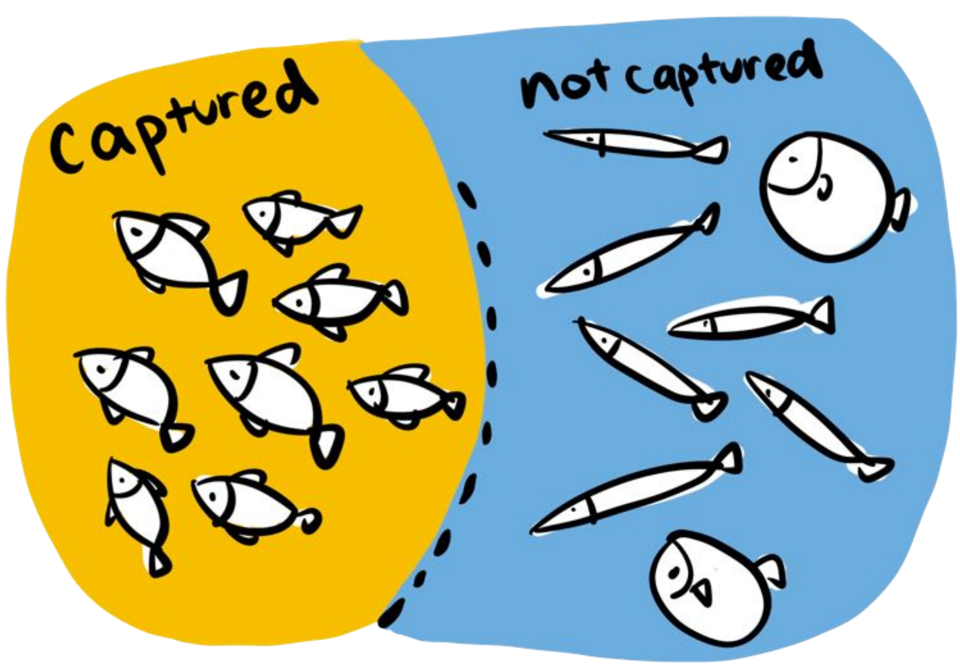


Step 3:

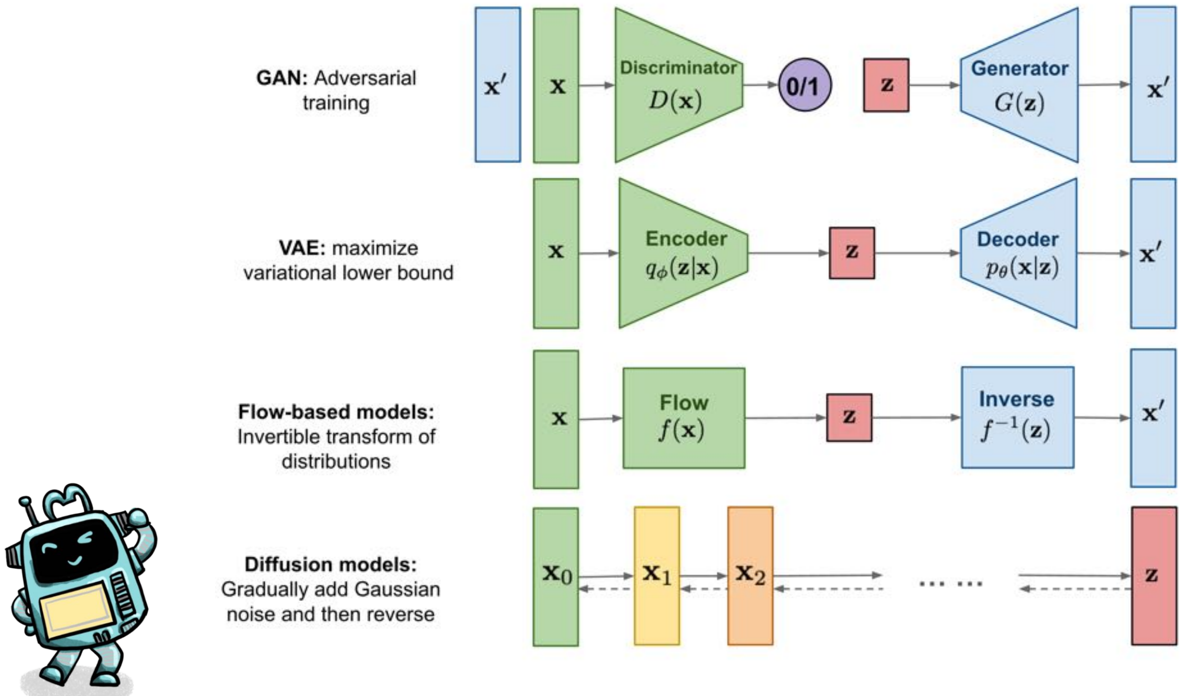
- Use the trained neural network to generate new fish.
- Ensure the generated fish is good: have characteristics learned from the dataset.

Are we modeling the actual distribution?

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG



Models

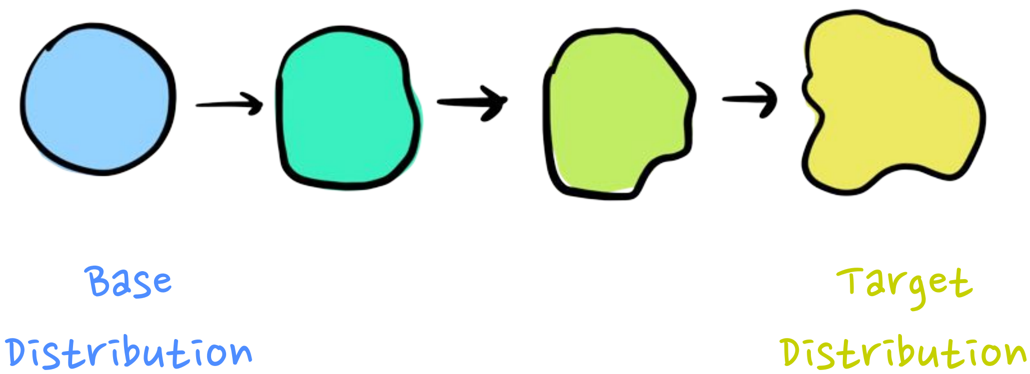


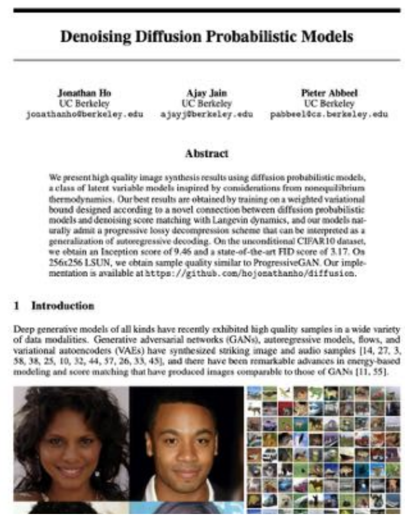
https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Diffusion Models:
A Generative AI Big Bang

Diffusion models

- Main idea: iteratively convert a base distribution to the target distribution via Markov chain





Denoising Diffusion Probabilistic Models

Jonathan Ho, UC Berkeley, jonathanho@berkeley.edu
Ajay Jain, UC Berkeley, ajayj@berkeley.edu
Pieter Abbeel, UC Berkeley, pabbeel@cs.berkeley.edu

Abstract

We present high quality image synthesis results using diffusion probabilistic models, a class of latent variable models inspired by considerations from nonequilibrium thermodynamics. Our best results are obtained by training on a weighted variational bound designed according to a novel connection between diffusion probabilistic models and denoising score matching with Langevin dynamics, and our models naturally admit a progressive lossy decompression scheme that can be interpreted as a generalization of autoregressive decoding. On the unconditional CIFAR10 dataset, we obtain an Inception score of 9.46 and a state-of-the-art FID score of 3.17. On 256x256 LSUN, we obtain sample quality similar to ProgressiveGAN. Our implementation is available at https://github.com/hojonathanho/diffusion.

1 Introduction

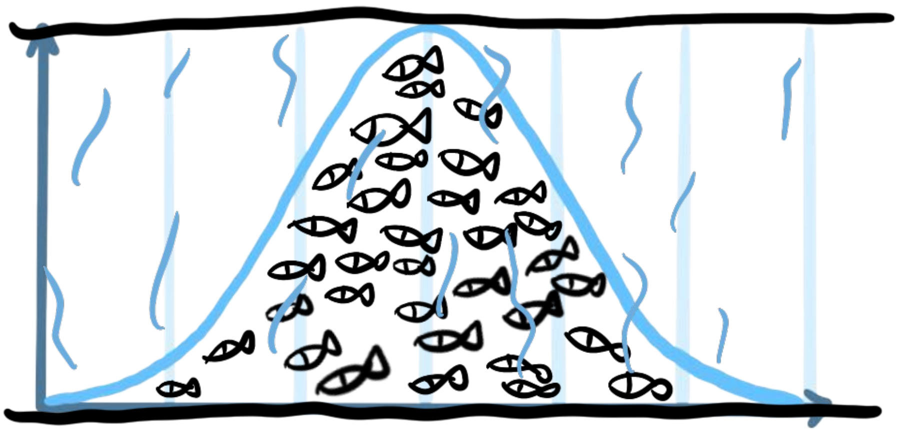
Deep generative models of all kinds have recently exhibited high quality samples in a wide variety of data modalities. Generative adversarial networks (GANs), autoregressive models, flows, and variational autoencoders (VAEs) have synthesized striking image and audio samples [14, 27, 3, 58, 38, 25, 10, 32, 44, 57, 26, 33, 45], and there have been remarkable advances in energy-based modeling and score matching that have produced images comparable to those of GANs [11, 55].

Basics of diffusion models

Forward Diffusion Process	Reverse Diffusion Process	Training & Sampling

Refresh on distributions

Gaussian Distribution



Defined by:

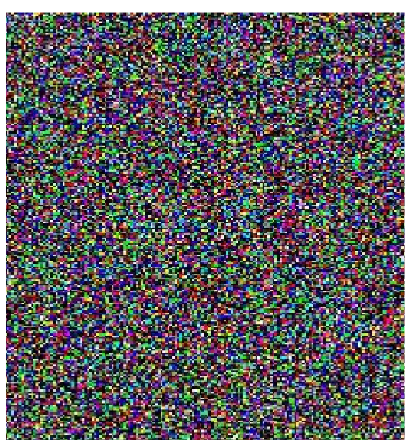
- Mean
- Std. deviation

$=\sqrt{\text{variance}}$

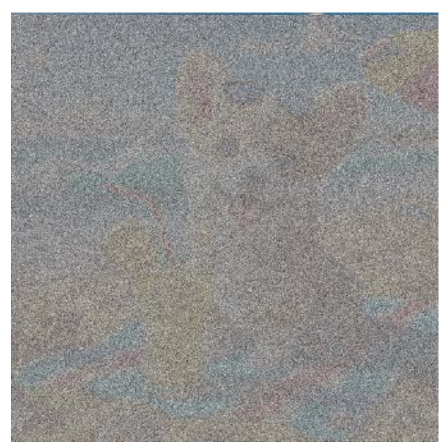
Gaussian noise

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

+



=

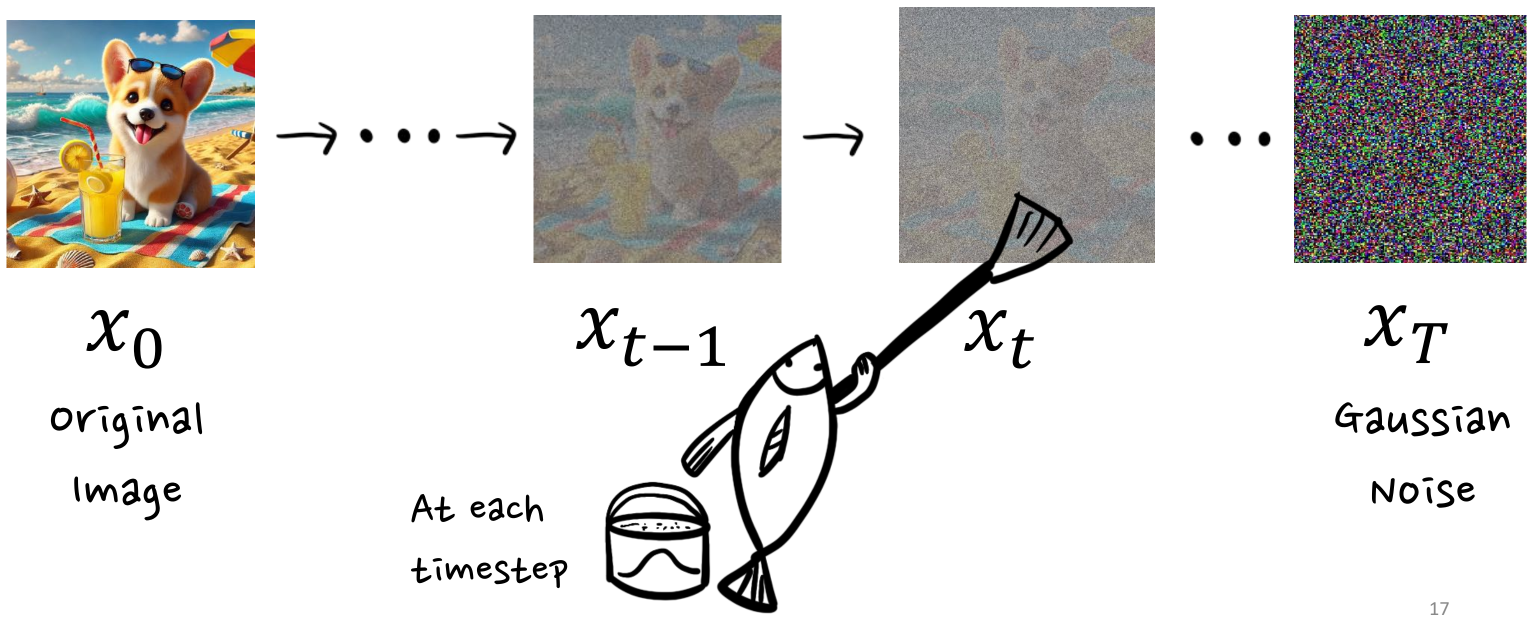


Original Image

Gaussian Noise

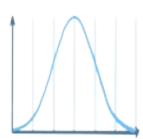
Noised Image

Forward diffusion process



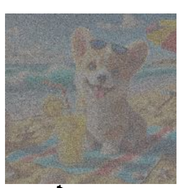
Forward diffusion process

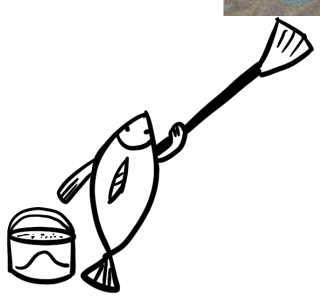
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG



mean var.

$$q(x_t | \ldots) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\ \ldots, \beta_t \mathrm{I})$$

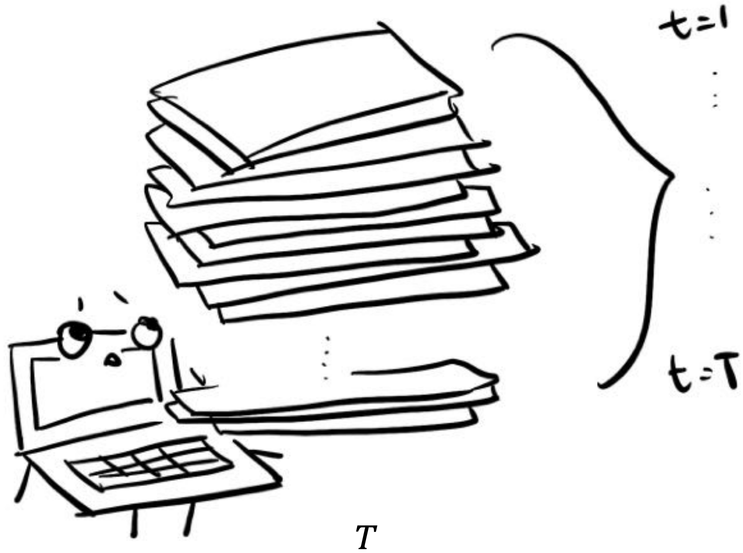




variance Schedule

Apply forward process one by one?

Too much to store in my memory! or disk!



$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

Reparameterization trick

Forward process $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathrm{I})$

Rewriting Def. $x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathrm{I})$

Define variables $\alpha_t = 1 - \beta_t \qquad \overline{\alpha_t} = \prod_{i=1}^{t} \alpha_i$

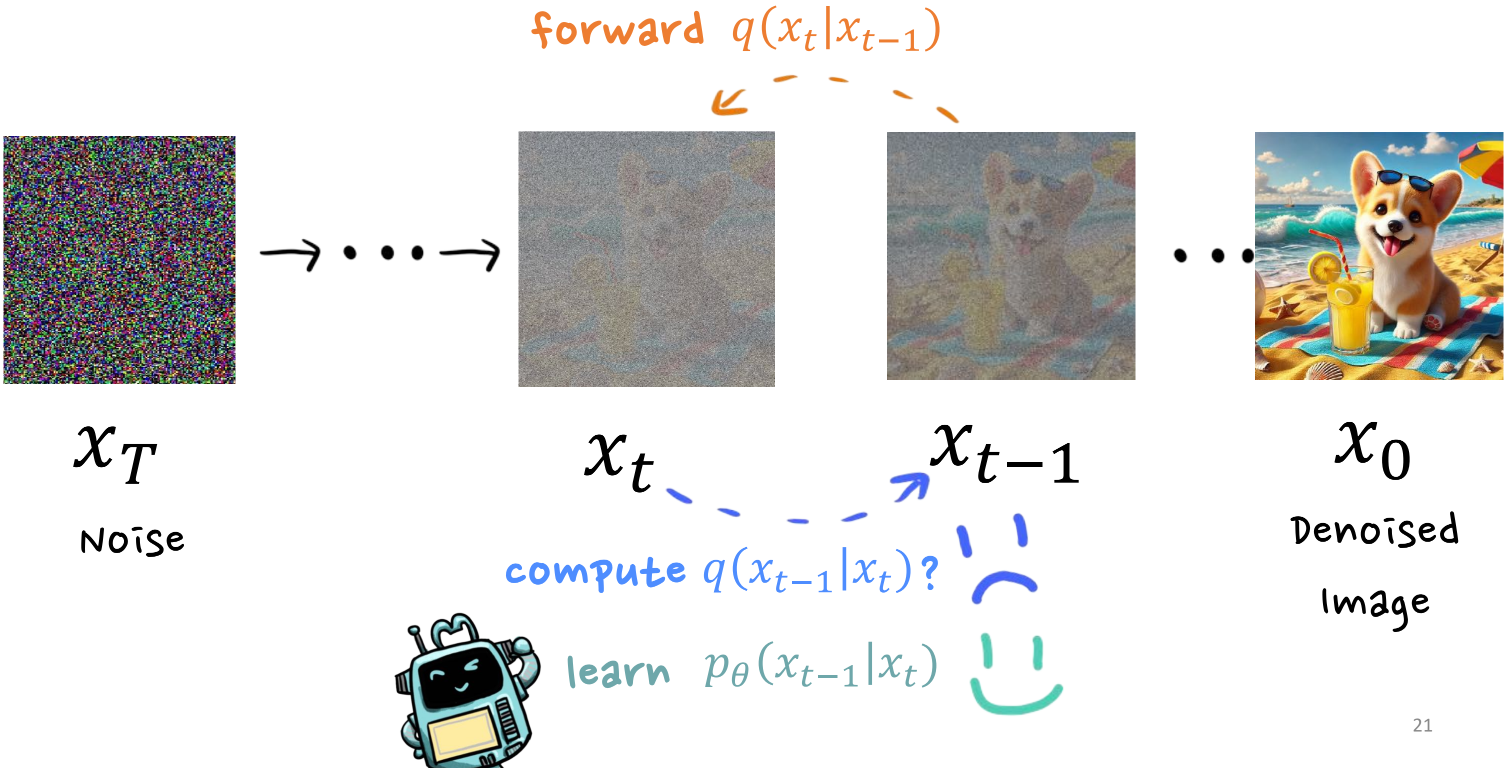
Gaussian Recap We can merge gaussians with different variances. $\mathcal{N}(0, \sigma_1^2 \mathrm{I}), \mathcal{N}(0, \sigma_2^2 \mathrm{I}) \rightarrow \mathcal{N}(0, (\sigma_1^2 + \sigma_2^2)\mathrm{I})$

Plug in definitions!

$$\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1} \\
&= \sqrt{\alpha_t \alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_t \alpha_{t-1}}\bar{\boldsymbol{\epsilon}}_{t-2} \\
&= \dots \\
&= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}
\end{aligned}$$

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

Reverse diffusion process



Training diffusion models

Model: $\epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon\right)$

Loss: $MSE\left[\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon\right)\right]$

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \mathrm{Uniform}(\{1,\ldots,T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$
5: Take gradient descent step on
$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$
6: **until** converged

Loss calculation

DDPM Training Loop

Sampling diffusion models

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

Algorithm 2 Sampling

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \dots, 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$

DDPM Sampling Loop

Basics of diffusion models

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

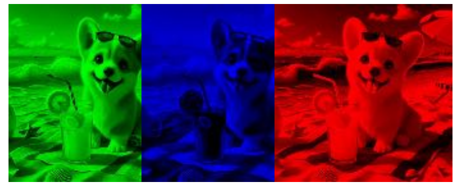
Forward Diffusion Process

Reverse Diffusion Process

Training & Sampling

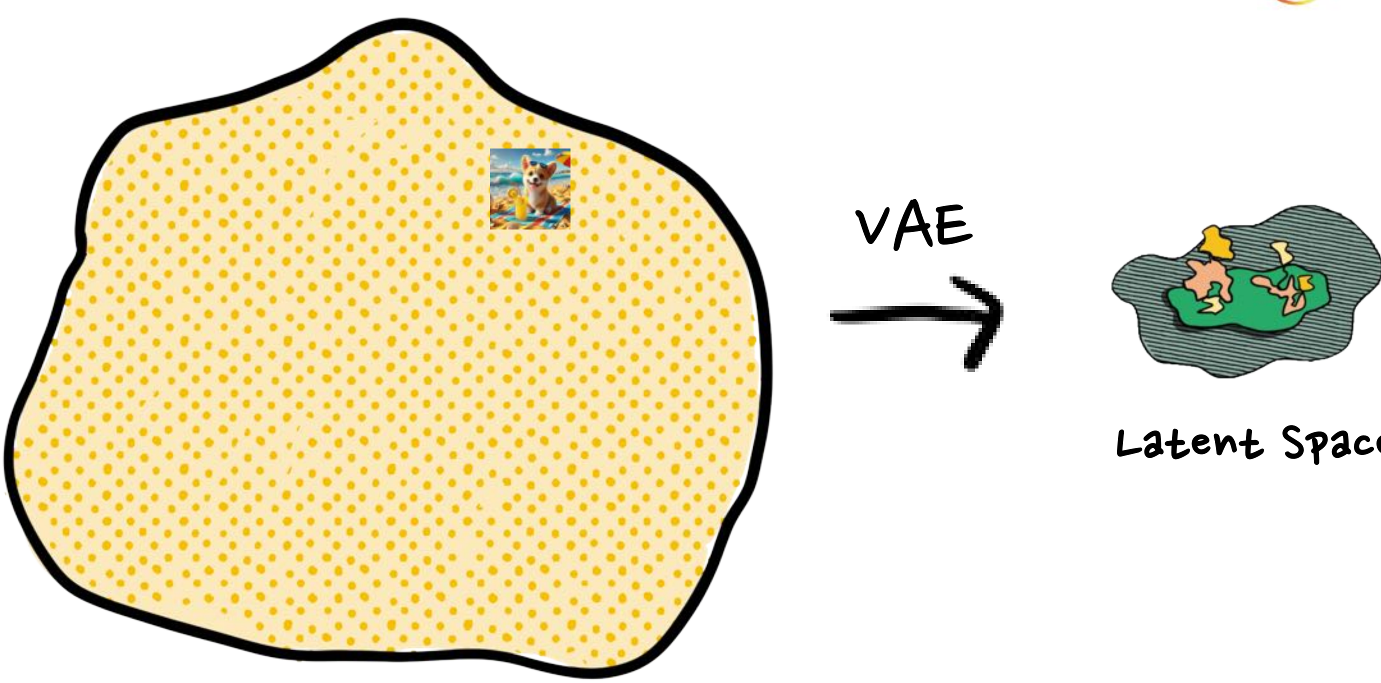
Pixels are expensive!

High Quality Image



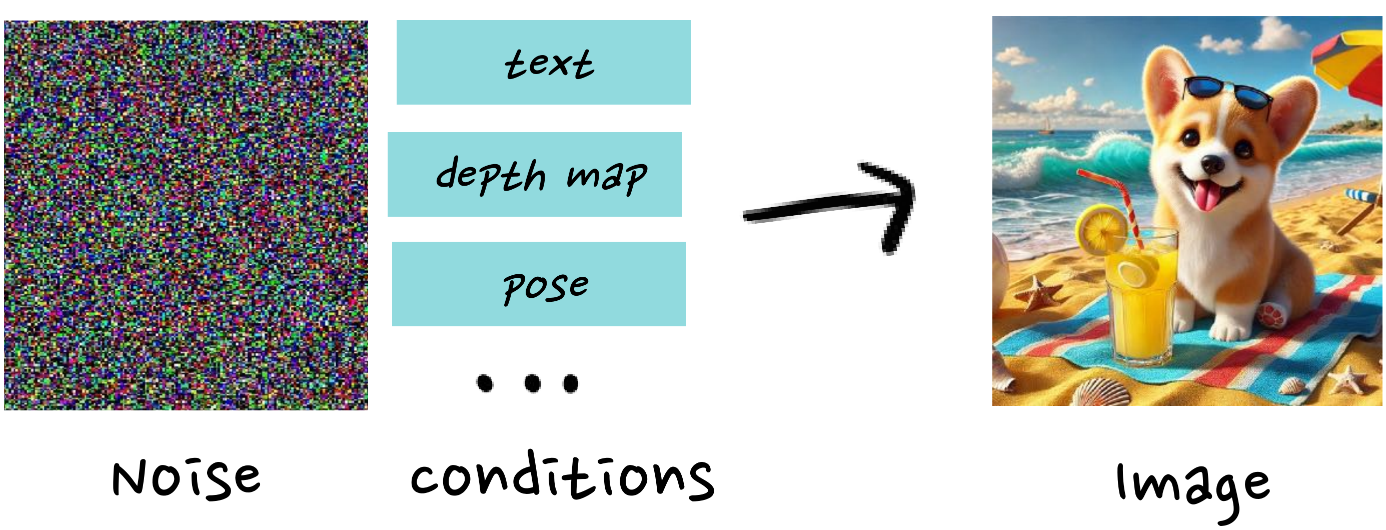
Need to store a lot of data!

Latent diffusion

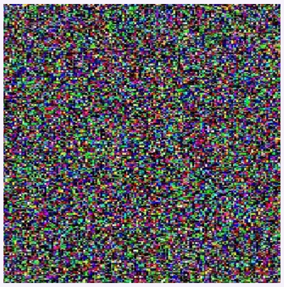


Conditional generation and guidance

So far, we've seen noise to image



Score function

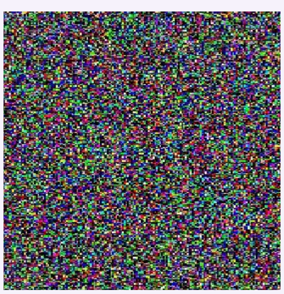


Noise

Image

$$p(x) \approx \nabla_x \log p(x)$$

Conditional diffusion models



"A corgi on a beach"

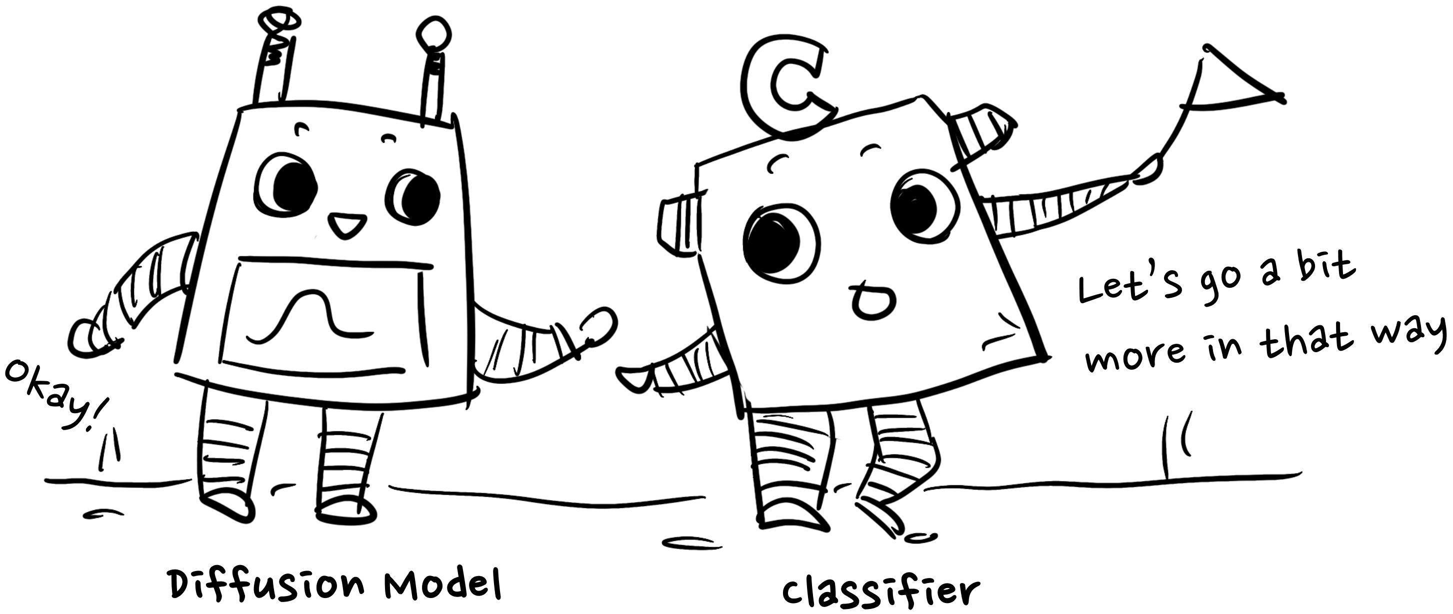
Noise

condition

Image

$$p(x|y) \approx \nabla_x \log p(x|y) = \underbrace{\nabla_x \log p(y|x)}_{\text{conditioning term}} + \underbrace{\nabla_x \log p(x)}_{\text{unconditional score function}}$$

Classifier guidance



Classifier guidance

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Guidance Strength

$$\nabla_x \log p(x|y) = \boxed{w}\, \nabla_x \log p(y|x) + \nabla_x \log p(x)$$

w = 1.0

w = 10.0

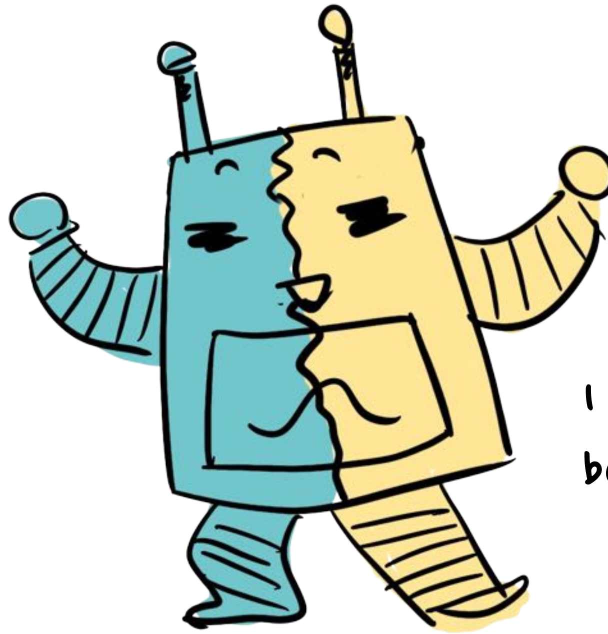
[Diffusion Models Beat GANs on Image Synthesis]

Classifier-Free Guidance (CFG)

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

Conditioning Dropout:

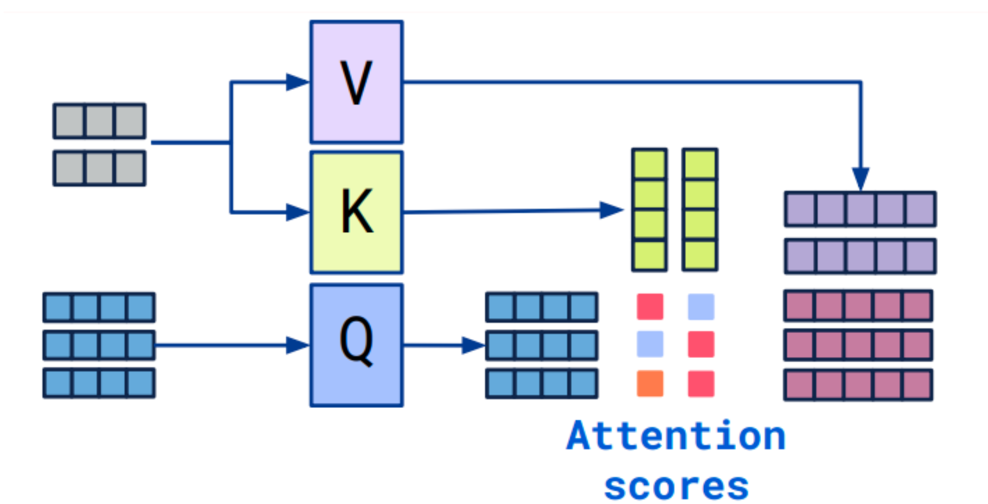
10 to 20 percentage of the time, the conditioning information is removed.



I can do both!

How do we feed conditioning signals?

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG



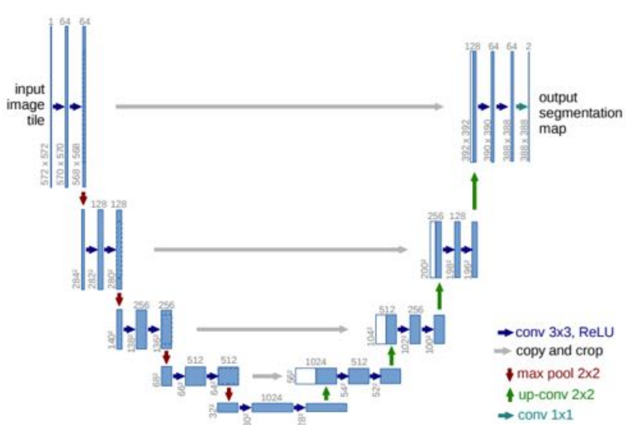
cross-attention layers!

Figure from https://vaclavkosar.com/ml/cross-attention-in-transformer-architecture

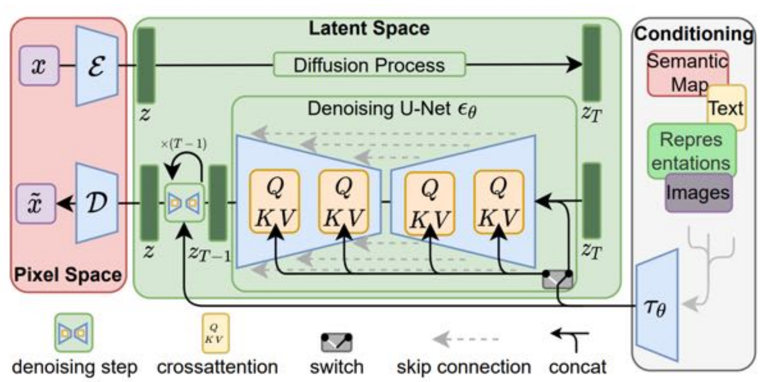
Implementation Architectures

U-Net Architecture

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG



U-Net architecture



U-Net based diffusion architecture

Ronneberger et al., "U-Net: Convolutional Networks for Biomedical Image Segmentation", MICCAI 2015
Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models", CVPR 2022

U-Net Architecture

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Imagen.

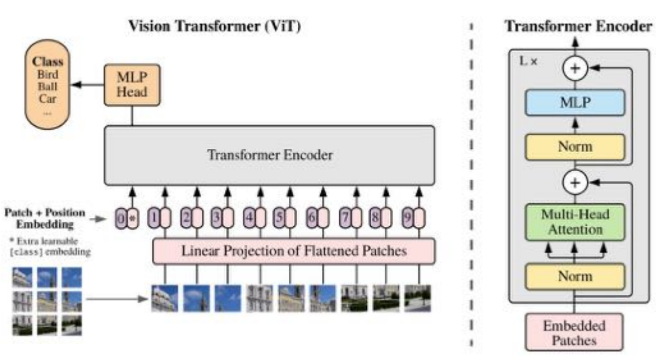
Stable Diffusion

eDiff-I

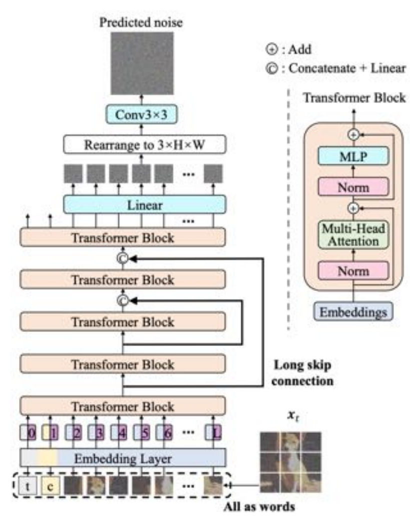
Saharia et al. “Photorealistic text-to-image diffusion models with deep language understanding”, NeurIPS 2022
Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models", CVPR 2022
Balaji et al.,” ediffi: Text-to-image diffusion models with an ensemble of expert denoisers”, arXiv 2022

Transformer Architecture

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG



Vision transformer.



Transformer based diffusion model

Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale”, ICLR 2021
Bao et al., "All are Worth Words: a ViT Backbone for Score-based Diffusion Models", arXiv 2022

Transformer Architecture

Scalable Diffusion Models with Transformers

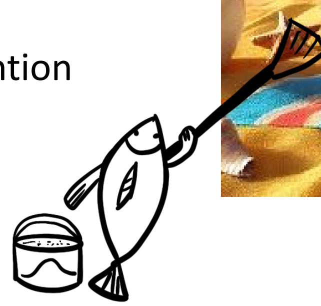
One Transformer Fits All Distributions in Multi-Modal Diffusion at Scale

simple diffusion: End-to-end diffusion for high resolution images

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
Hoogeboom et al., "simple diffusion: End-to-end diffusion for high resolution images", arXiv 2023

Summary

- Basics of diffusion models
  - Forward & reverse diffusion process
  - Sampling and training
  - Latent diffusion
- Conditional generation
  - Classifier guidance
  - Classifier-free guidance (CFG)
  - Adding condition using cross-attention
- Implementation architectures
  - U-net
  - Vision Transformers



Generated with DALL-E

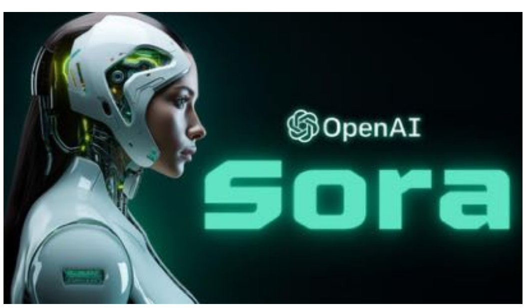
Start from Text-to-Image Large Models

DALL·E 3

Stable Diffusion

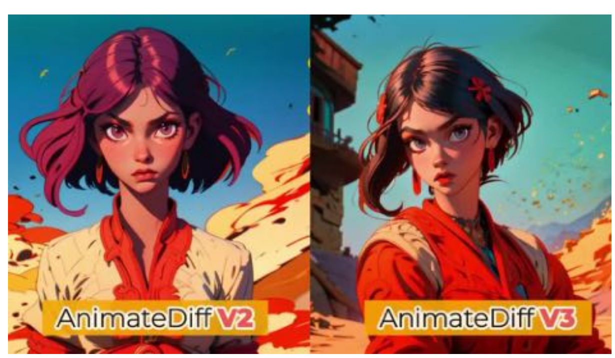
Stable Diffusion XL

Start from Text-to-Video Large Models



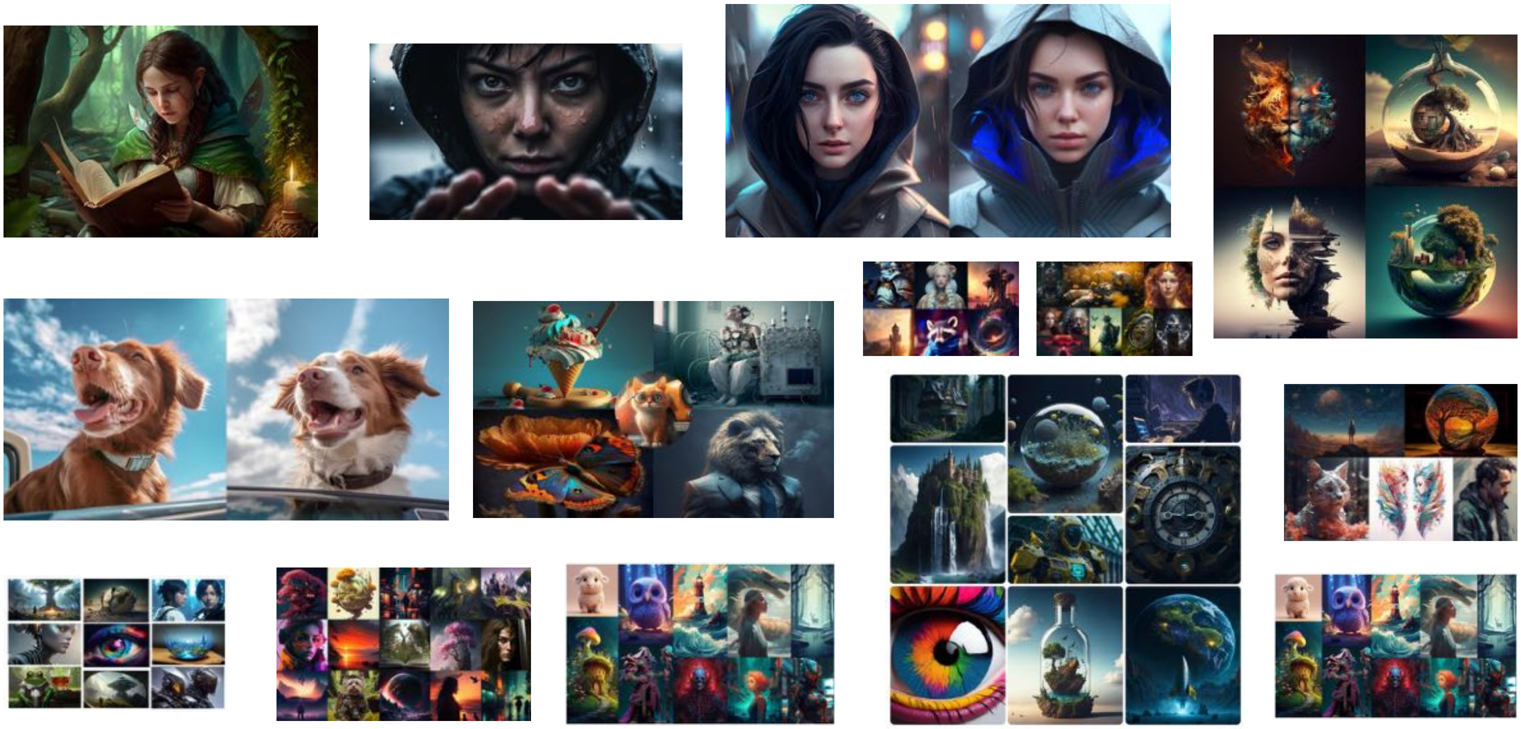
Gen-2: The Next Step Forward for Generative AI

Google Research
LUMIERE



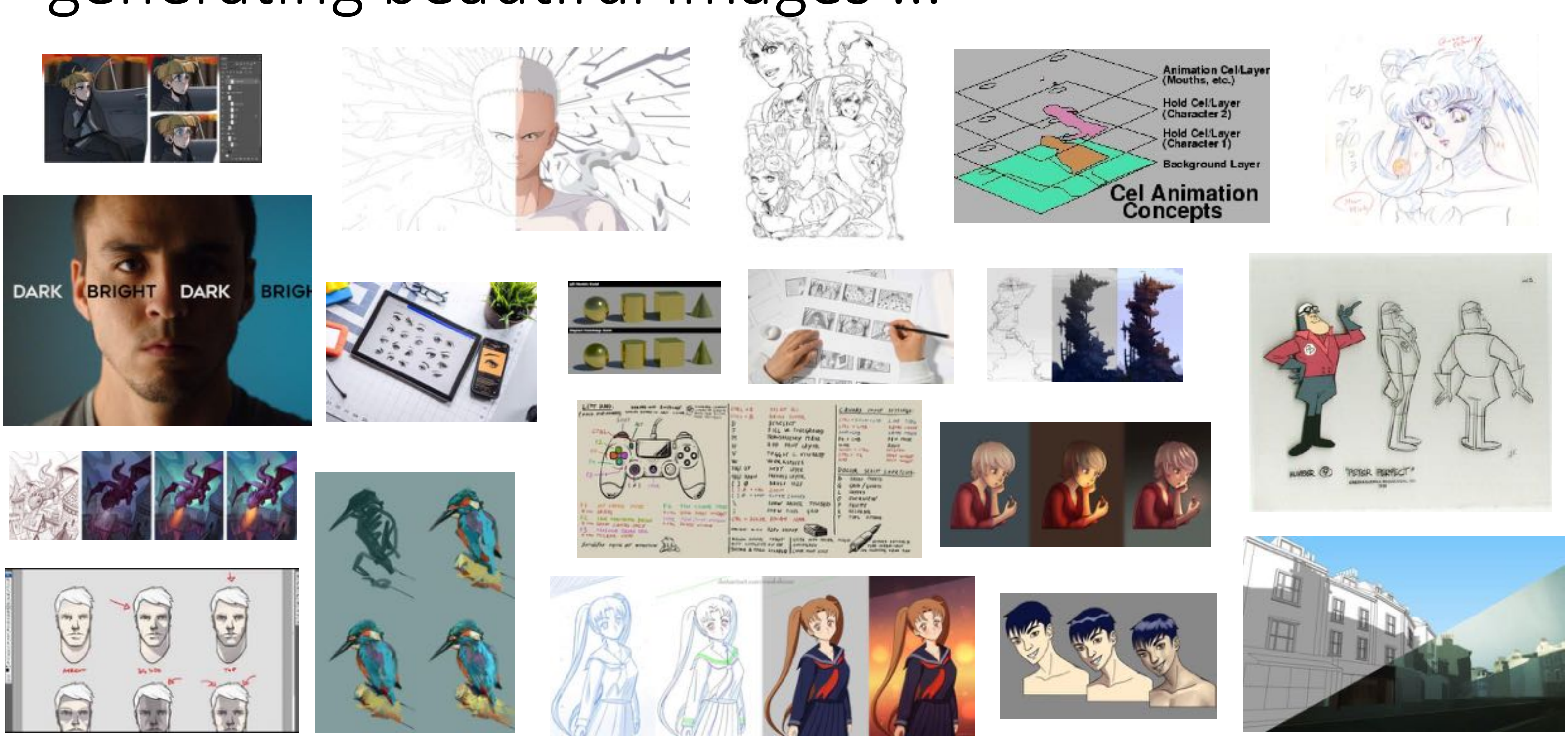
Generating High-quality Images …

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG



But visual creation is more than just generating beautiful images …

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG



More Control Other Than Texts?

stanford memorial church with neon signage in the style of bladerunner

Iteration 1

stanford memorial church **and main quad with palm trees** in the style of bladerunner

Iteration 3

nighttime rain stanford memorial church and main quad with palm trees, **night market food stalls and neon signs** in the style of bladerunner

Iteration 8

nighttime rain stanford memorial church and main quad with palm trees, night market food stalls and neon signs **like downtown tokyo**

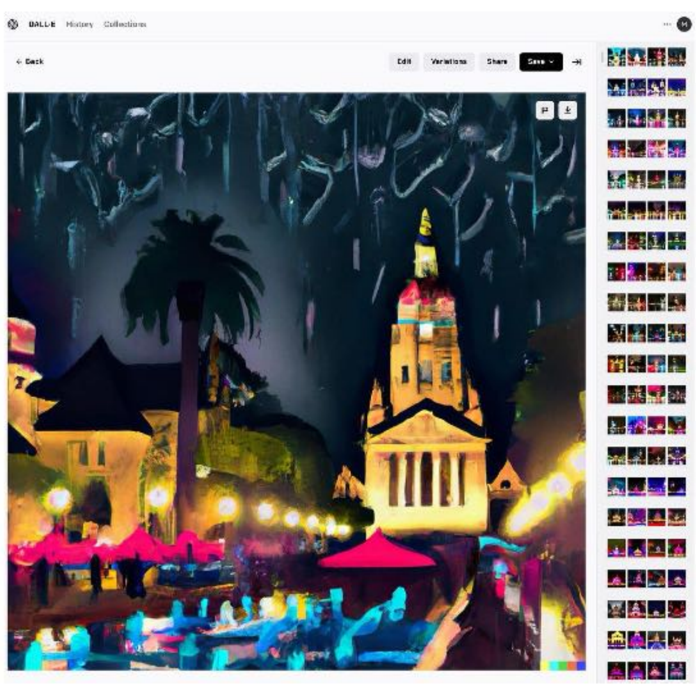
Iteration 17

More Control Other Than Texts?

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

nighttime rain stanford memorial church and main quad with palm trees, night market **japadog** food stalls and neon signs, **neo** tokyo **bladerunner** style **film still illustration**

Iteration 21

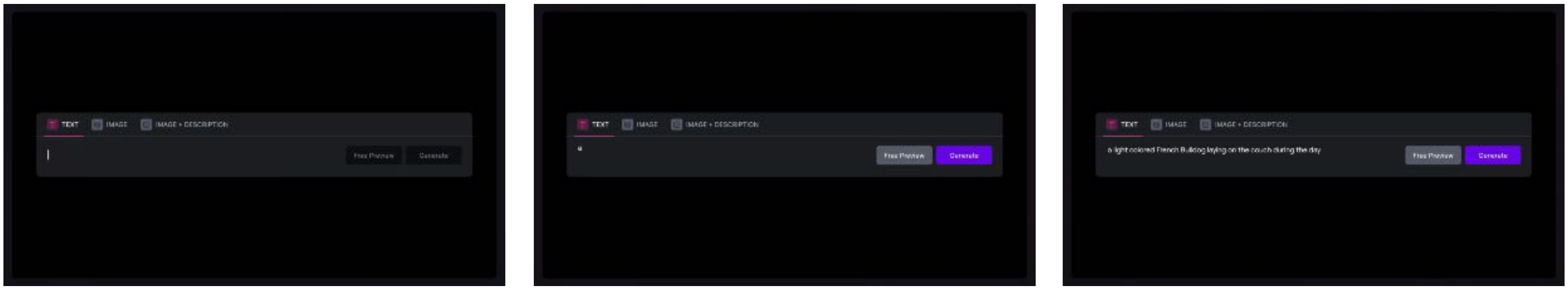


Lots of trial-and-error!

https://magrawala.substack.com/p/unpredictable-black-boxes-are-terrible

Text Control is Limited in Creation

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG



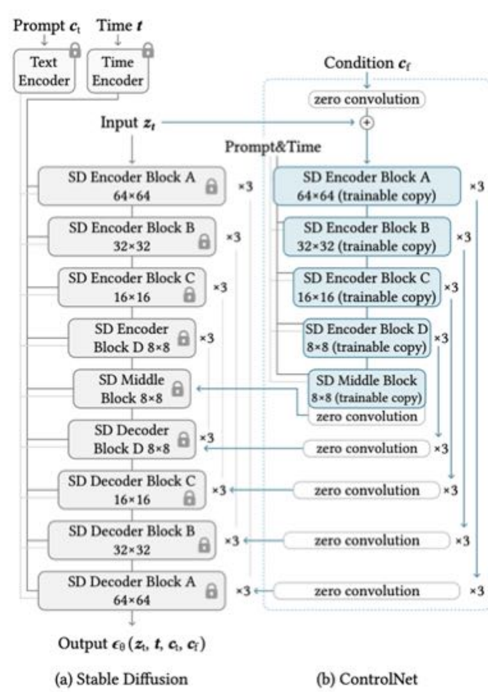
Iteration 1 | Iteration 5 | Iteration 17

Text does not match user's mental representation, which leads to lots of trial-and-error!

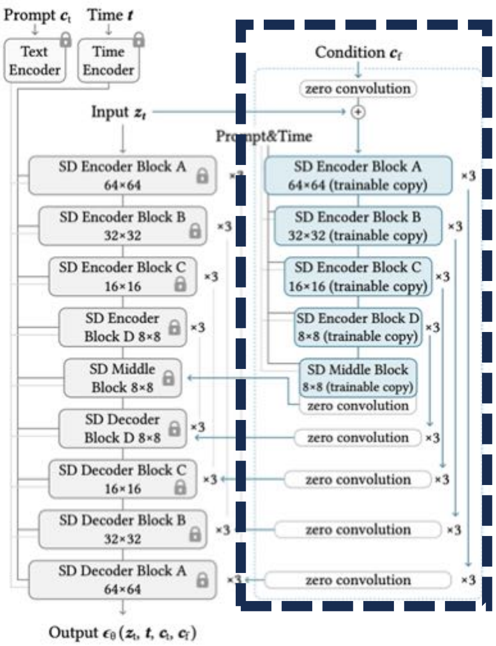
https://magrawala.substack.com/p/unpredictable-black-boxes-are-terrible

ControlNet

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG



Architecture of ControlNet

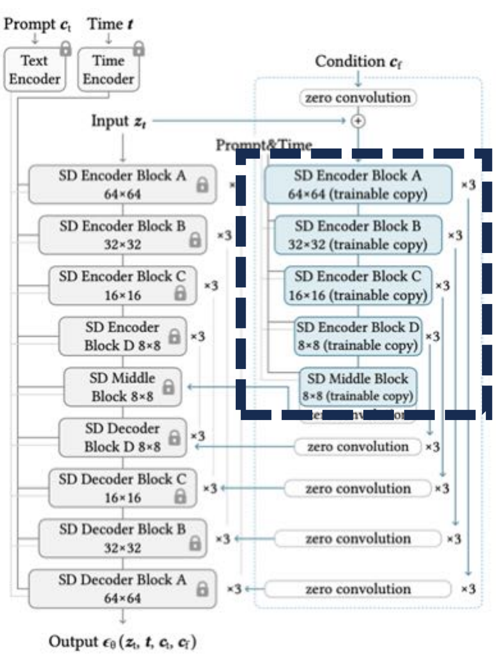


(a) Stable Diffusion (b) ControlNet

- Using **external model** to process control signals.
- **Re-using pretrained weights** as the backbone of control model.
- Connecting with **zero-initialized layers** to reduce initial noise.

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

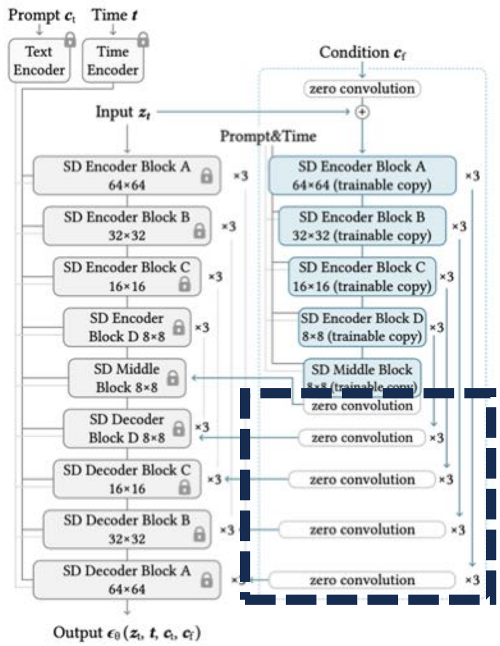
Architecture of ControlNet



(a) Stable Diffusion (b) ControlNet

- Using **external model** to process control signals.
- **Re-using pretrained weights** as the backbone of control model.
- Connecting with **zero-initialized layers** to reduce initial noise.

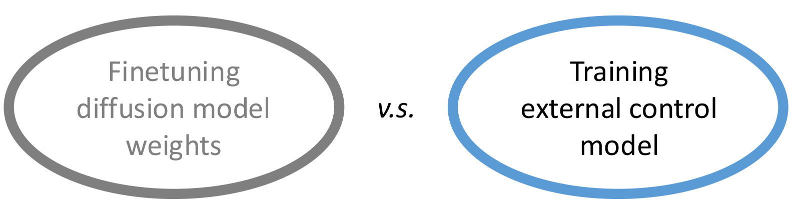
Architecture of ControlNet



(a) Stable Diffusion (b) ControlNet

- Using **external model** to process control signals.
- **Re-using pretrained weights** as the backbone of control model.
- Connecting with **zero-initialized layers** to reduce initial noise.

External model to process control signals

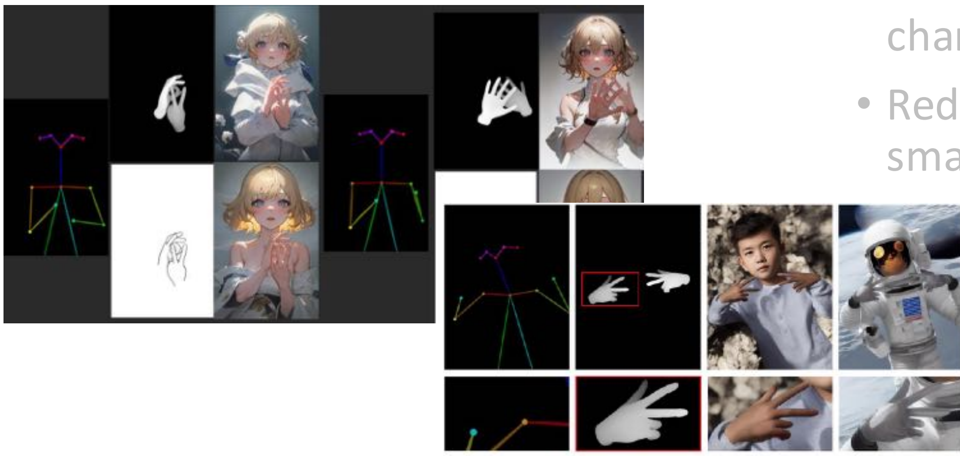


- Compossible control (multiple conditions)
- Minimal influence to the base model (the base model can be changed)
- Reduced overfitting risk (training with small dataset becomes easier)

External model to process control signals

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

Finetuning diffusion model weights *v.s.* Training external control model



Multiple condition (pose&depth) "boy" "astronaut"

- Compossible control (multiple conditions)
- Minimal influence to the base model (the base model can be changed)
- Reduce overfitting (training with small dataset becomes easier)

External model to process control signals

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

Finetuning diffusion model weights *v.s.* Training external control model

- Compossible control (multiple conditions)
- Minimal influence to the base model (the base model can be changed)
- Reduce overfitting (training with small dataset becomes easier)

"house" SD 1.5 Comic Diffusion Protogen 3.4

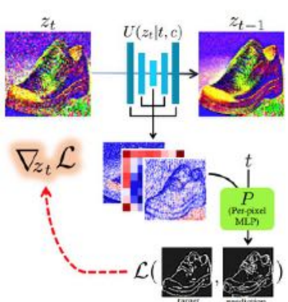
External model to process control signals

without ControlNet
(using Stability's "official" method to add the channels to input layer, same as their depth-to-image structure)

SD + ControlNet

- Compossible control (multiple conditions)
- Minimal influence to the base model (the base model can be changed)
- Reduce overfitting (training with small dataset becomes easier)

Reusing pretrained backbone



Some insights from some previous works …

In the paper "Sketch-Guided Text-to-Image Diffusion Models" (from 2022 November), Voynov *et.al.* discussed that one of the major challenge of "sketches" guided diffusion is **the difficult alignment of complex scenes with mixed and ambiguous semantics.**

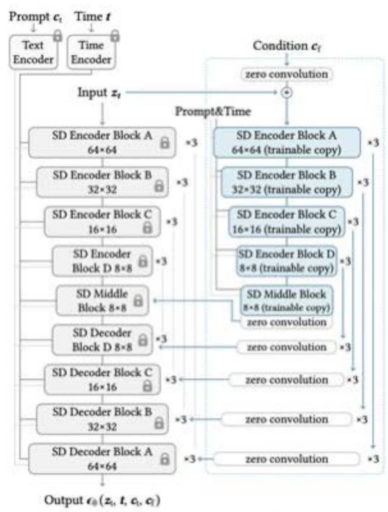
Figure 14. **Failure cases.** The quality of the results may drop for different initialization, and on complex scenes with mixed and ambiguous semantics.

This motivates us to find a **stronger backbone** to solve the semantic alignment and understanding problem …

By the way, this is the result from ControlNet 1.1.

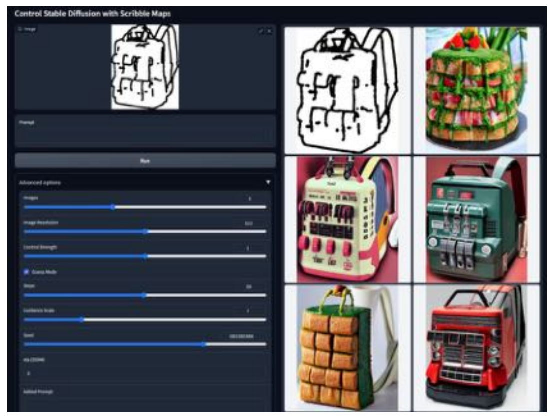
Reusing pretrained backbone

The ability to “guess” contexts without accurate prompts …



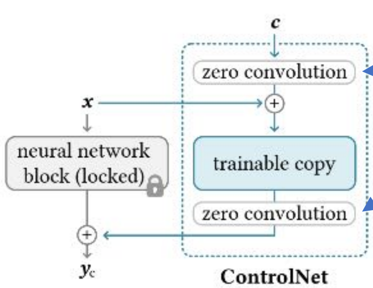
(a) Stable Diffusion (b) ControlNet

Input “a high-quality and extremely detailed image”



Using zero-initialized layers

The ability to “guess” contexts without accurate prompts …



Zero-initialized connection layers
- Reduce initial harmful noise
- Protect the trainable copy

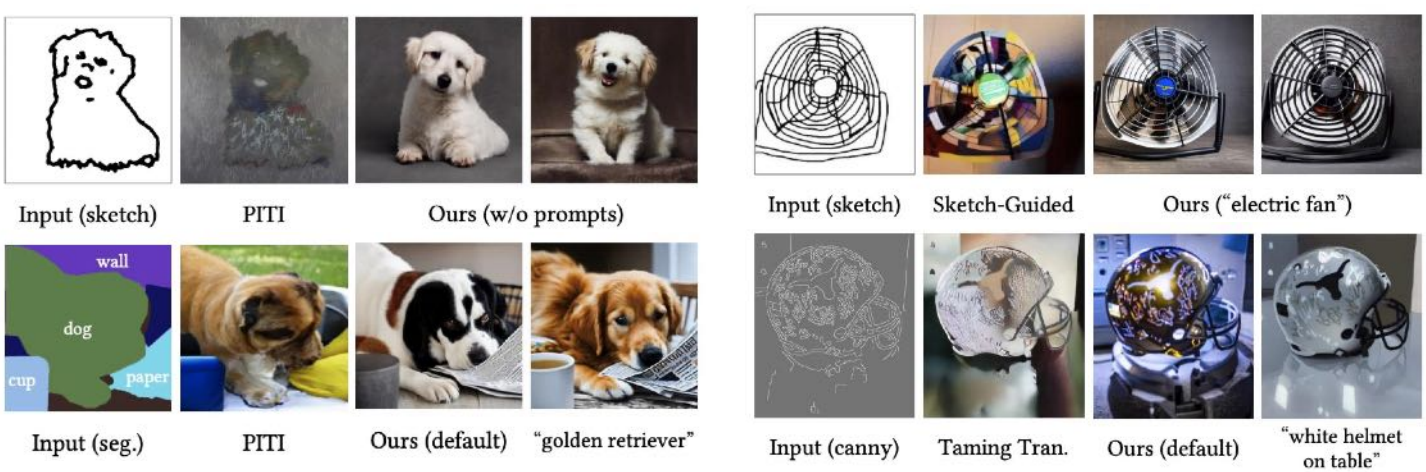
$$y_c = \mathcal{F}(x;\Theta) + \mathcal{Z}(\mathcal{F}(x + \mathcal{Z}(c;\Theta_{z1});\Theta_c);\Theta_{z2}),$$

In the first training step, $y_c = y$.

All experiments are conducted with Stable Diffusion 1.5

Comparisons

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG



Unleash Human Creativity

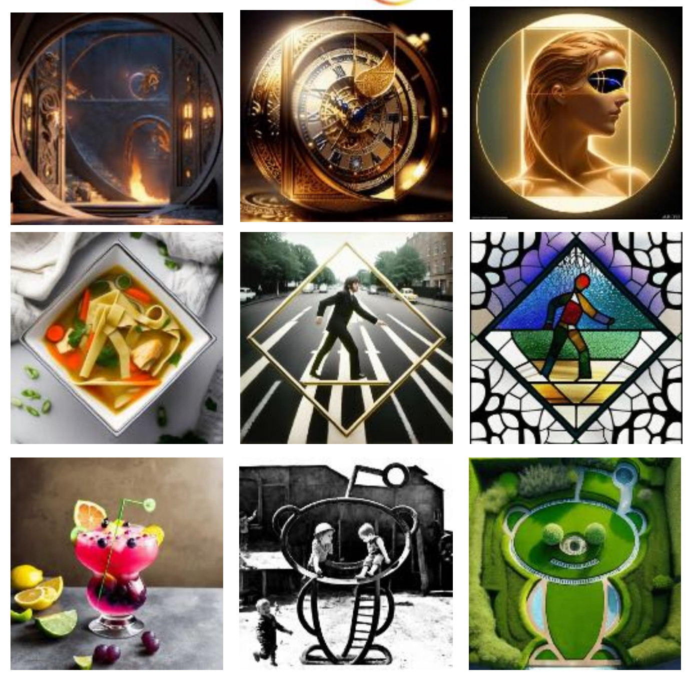
Reddit
https://www.reddit.com › StableDiffusion › comments
SDBattle: Week 5 - ControlNet Cross Walk Challenge! Use ...
Mar 20, 2023 — Welcome back to the weekly Stable Diffusion **Battle** Challenge! Excited to see what you all make! Join us for more **battles** over at r/SDBattles. If ...

Reddit
https://www.reddit.com › StableDiffusion › comments
SDBattle: Week 3 - ControlNet Fibonacci Challenge! Use ...
Mar 6, 2023 — I think I'll pass this **battle** and just see the results and if they disproof my statement. If I would go into photoshop and add some gradients or ...

Reddit
https://www.reddit.com › StableDiffusion › comments
SDBattle: Week 8 - ControlNet The Thinker Challenge! Use ...
Apr 11, 2023 — In this week's SD **Battle** I hand made a pose based on The Thinker statue since **ControlNet** was having a hard time generating one.

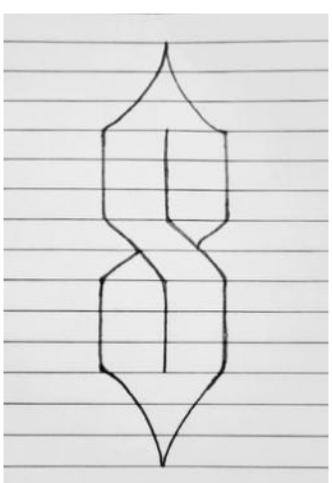
Reddit
https://www.reddit.com › StableDiffusion › comments
SDBattle: Week 6 - ControlNet Dog Paw Challenge! Use ...
Mar 27, 2023 — Join us for more thrilling **battles** over at r/SDBattles, where the action never stops and the stakes are always high! Don't miss this chance to ...

Input



Results Conditioned on the Canny Map from Input

Unleash Human Creativity



Input

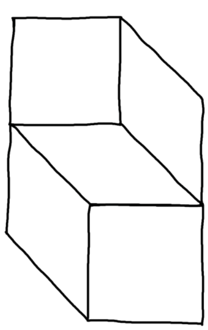
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

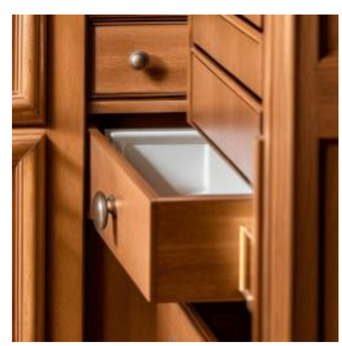
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Unleash Human Creativity

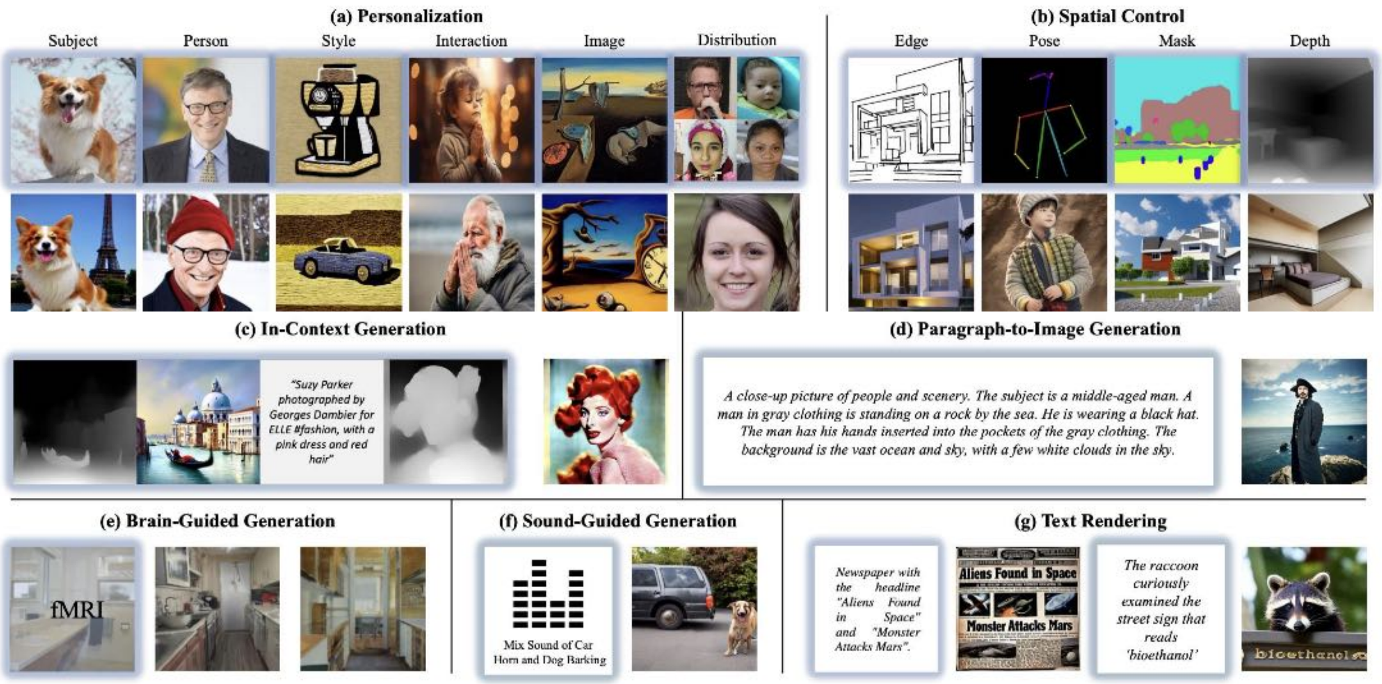
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG



Input



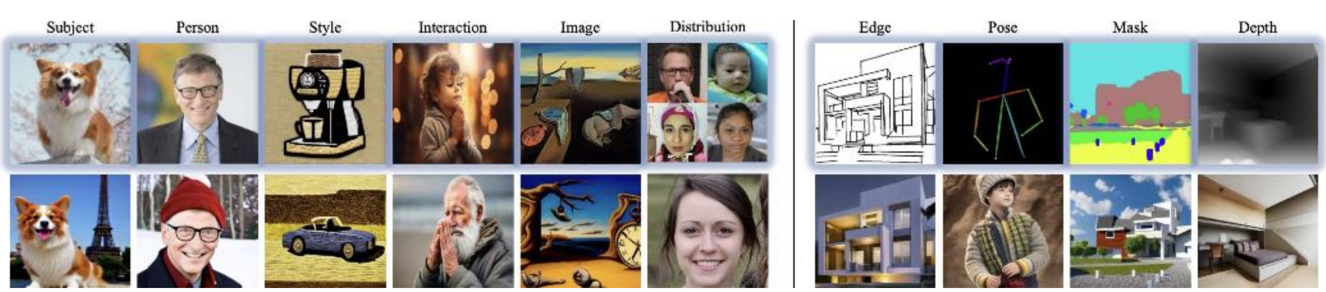
Controllable Generation in General



Cao, Pu, et al. "Controllable generation with text-to-image diffusion models: A survey." *arXiv preprint arXiv:2403.04279* (2024).

Take Away

- Text control is limited
- Better control leads to higher quality



Extend Image Diffusion Models for Videos

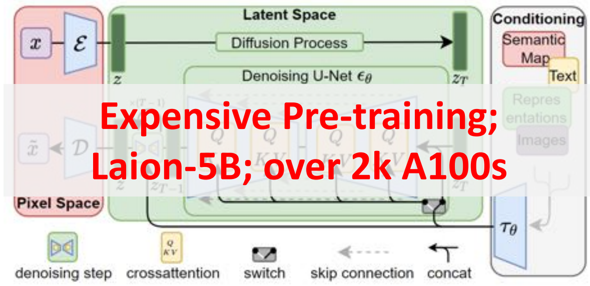
1. Data Format

image

video

2. Dataset availability

LAION (image, 5B) vs. WebVid (video, 10M)

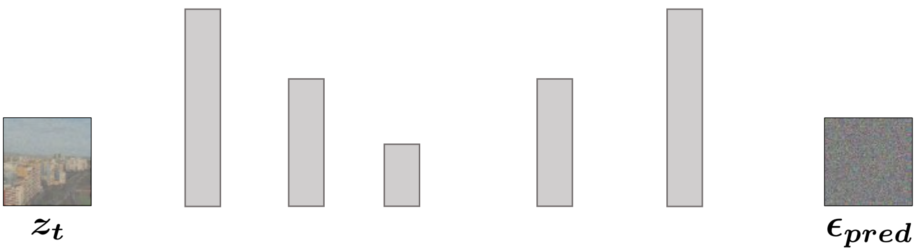


Extend Image Diffusion Models for Videos

Goal: leveraging powerful T2I prior knowledge
Reasons:
(1) better initialization than from scratch
(2) dataset scale (5B vs. 10M)

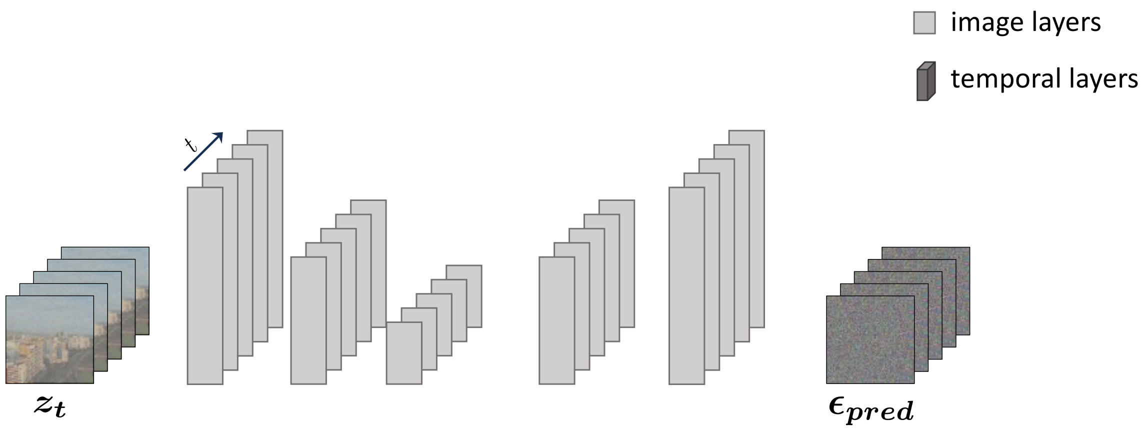
image layers

temporal layers



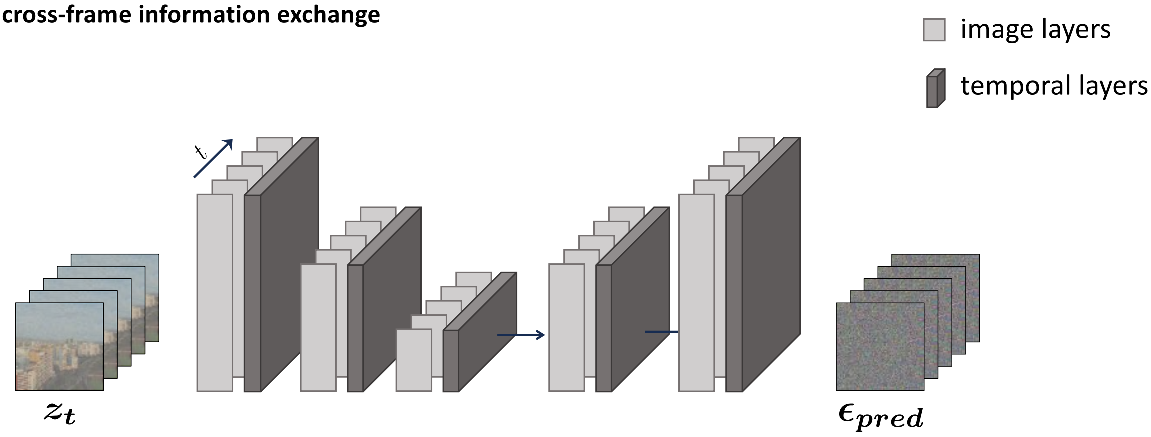
Extend Image Diffusion Models for Videos

1. Repeat the image generator along the time axis (e.g., 16/24 frames)



Extend Image Diffusion Models for Videos

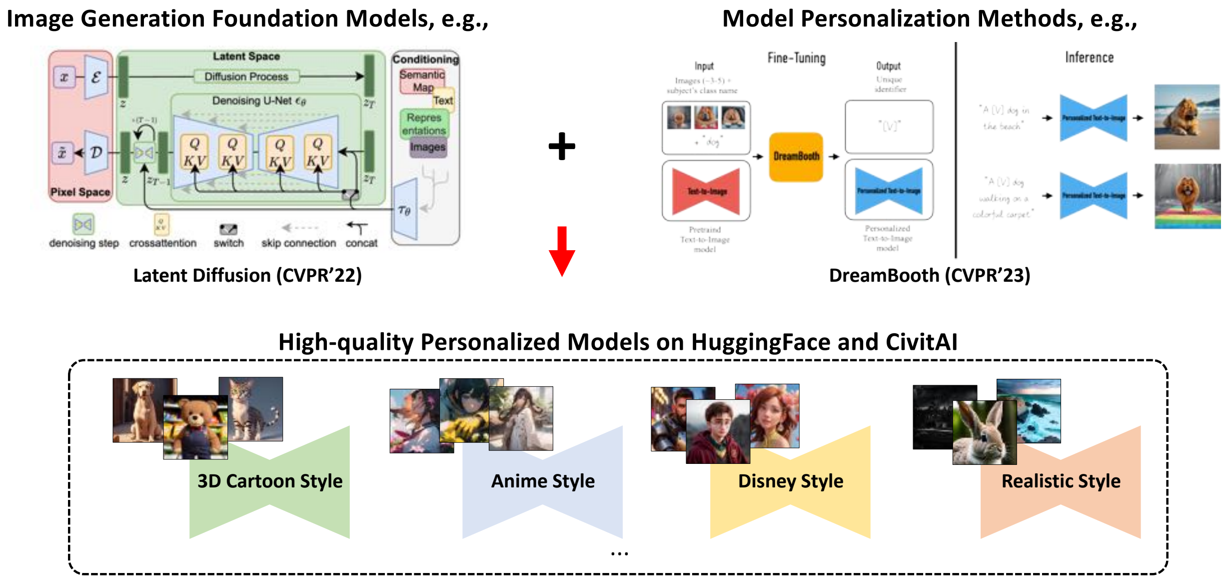
1. Repeat the image generator along the time axis (e.g., 16/24 frames)
2. **Enable cross-frame information exchange**



AnimateDiff:

Animate Your Personalized Text-to-Image Diffusion Model without Specific Tuning

AnimateDiff: Repurpose image diffusion model for video generation

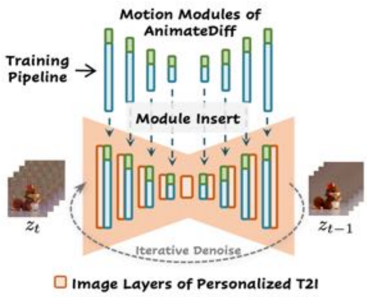


AnimateDiff: Method

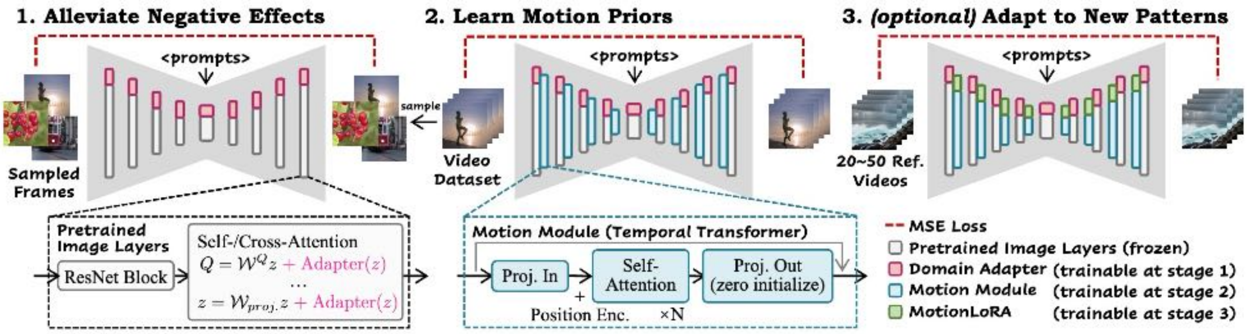
Our Goals: Animating personalized text-to-image diffusion models with a motion module, which

- Preserves original models' visual quality, ⟶ 1st stage, Domain Adapter
- Learns transferable motion priors from real-world videos, and ⟶ 2nd stage, Motion Module
- Efficiently adapts to specific motion patterns. ⟶ 3rd stage, MotionLoRA

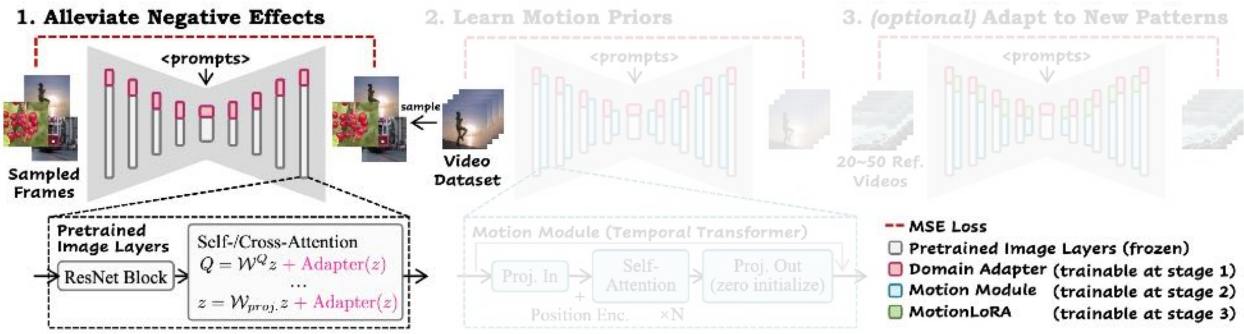
At inference, we directly insert the pre-trained motion module without needing specific tuning.



AnimateDiff: Method



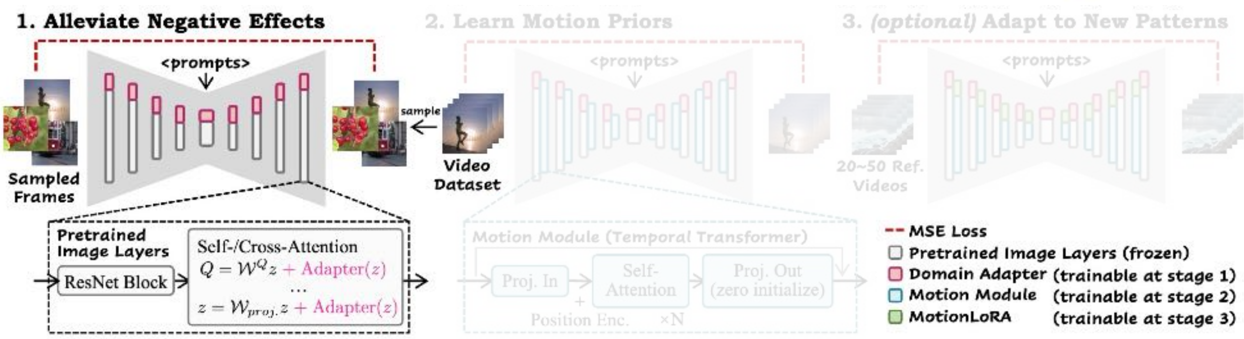
AnimateDiff: Method



Training Domain Adapter (1st stage): Alleviate Negative Effects from Training Data

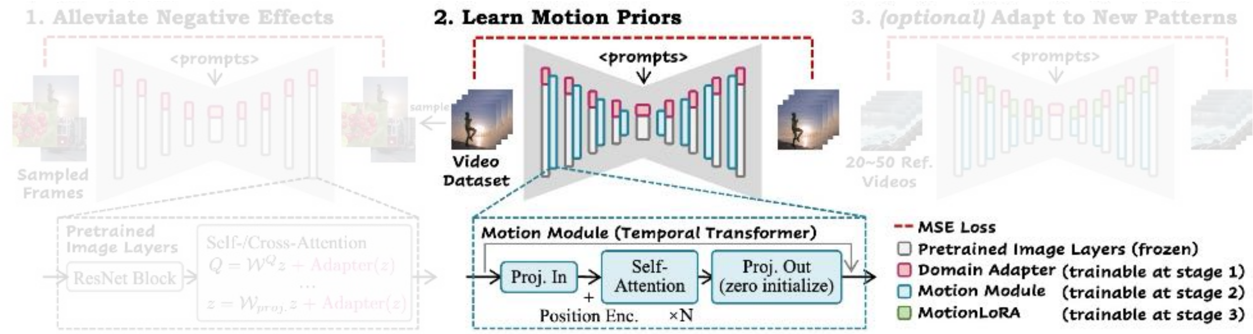
- Video datasets' lower quality: watermarks, motion blurs, and compression artifacts
- Solution: learning such visual patterns with domain adapter and removing it at inference

AnimateDiff: Method



Ablation Study: a lower domain adapter's effect helps preserve the original model's visual quality

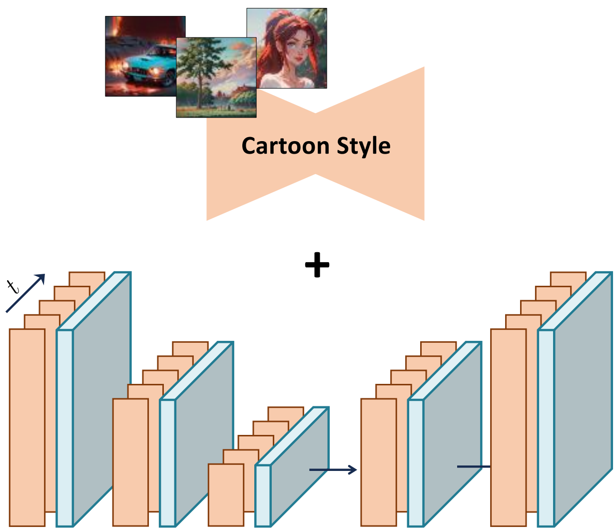
AnimateDiff: Method



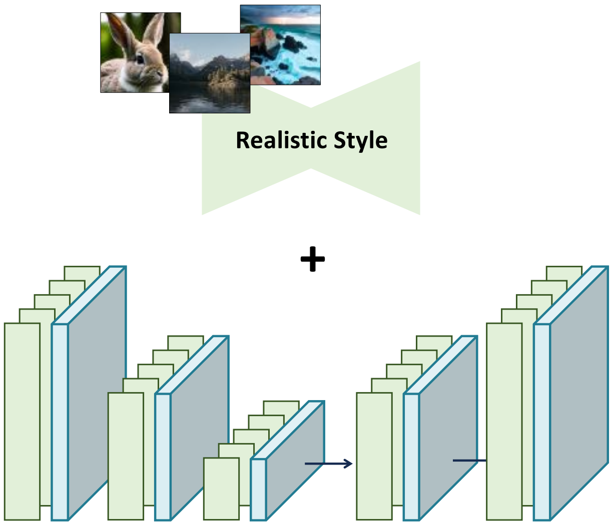
Training Motion Module (2nd stage): learning general motion priors from real-world videos

- **Model inflation:** from 2D image to 3D video
- **Temporal self-attention + position embeddings:** modeling cross-frame interactions
- Motion modules are inserted between frozen 2D image layers

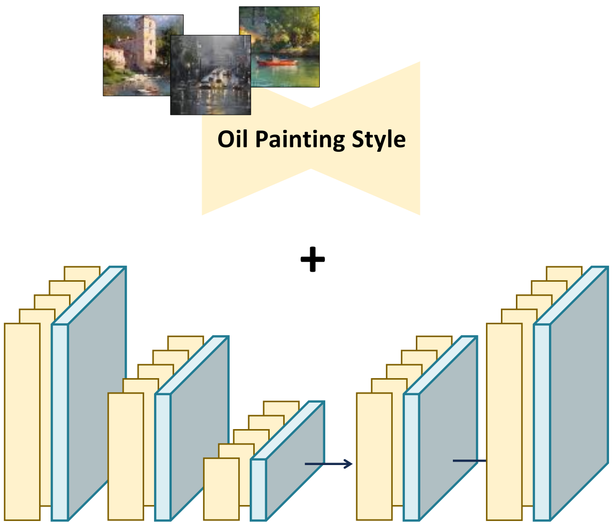
AnimateDiff: Method



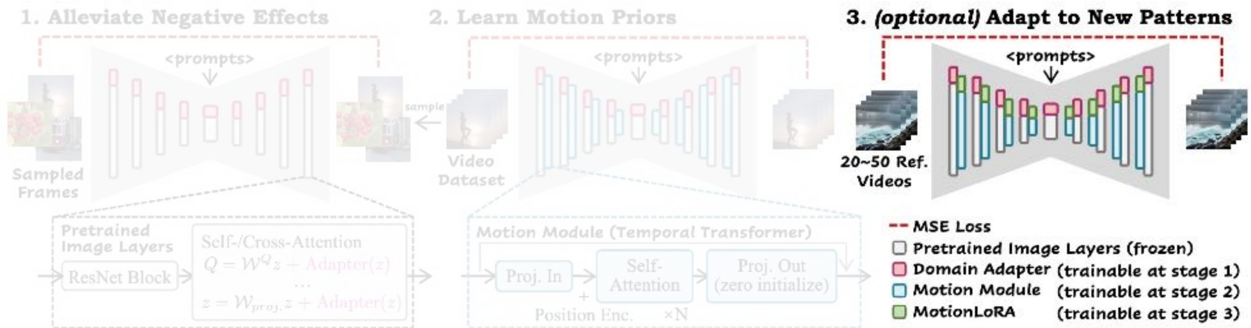
AnimateDiff: Method



AnimateDiff: Method



AnimateDiff: Method



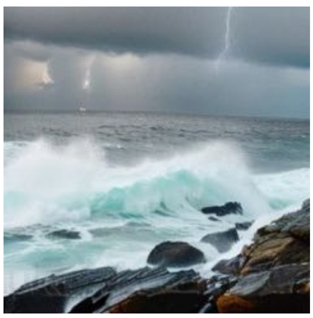
Training MotionLoRA (3rd stage, optional): adapting to specific motion patterns

- Motion patterns like zooming and rolling are common in productions
- **Solution:** training additional LoRA adapter upon motion module's pre-trained weights, with few numbers of reference videos

AnimateDiff: Method

zoom in

rolling



zoom out + rolling

right + up

AnimateDiff: Experiments

Training

- Dataset: WebVid-10M
- Pre-trained text-to-image model: Stable Diffusion V1.5

Evaluations

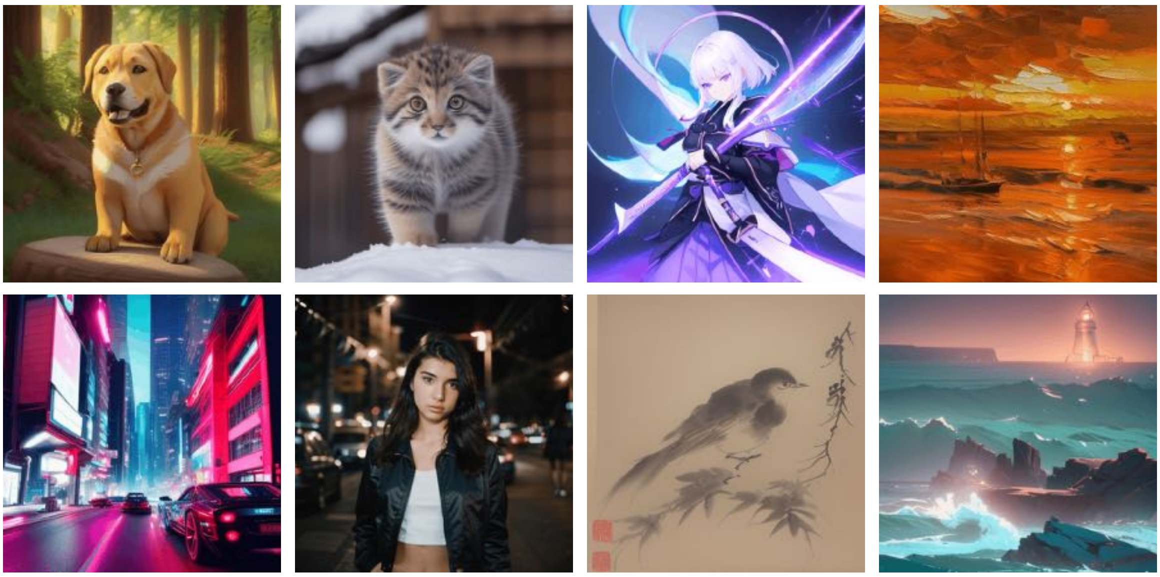
- Diverse model collected from the community

Model Name	Domain	Type
ToonYou	2D Cartoon	T2I Base Model
MeinaMix	2D Anime	T2I Base Model
Lyriel	Stylistic	T2I Base Model
RCNZ Cartoon 3d	3D Cartoon	T2I Base Model
epiC Realism	Realistic	T2I Base Model
Realistic Vision	Realistic	T2I Base Model
Oil painting	Stylistic	LoRA
MoXin	Stylistic	LoRA
TUSUN	Concept	LoRA

AnimateDiff: Experiments

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Qualitative Results: on eight different community model



AnimateDiff: Experiments

Quantitative Evaluation

- Our method is preferred by user study and CLIP metrics in text/domain fidelity and temporal smoothness

Method	User Study (↑)			CLIP Metric (↑)		
	Text.	Domain.	Smooth.	Text.	Domain.	Smooth.
Text2Video-Zero	1.620	**2.620**	1.560	32.04	84.84	96.57
Tune-a-Video	2.180	1.100	1.615	**35.98**	80.68	97.42
Ours	**2.210**	2.280	**2.825**	31.39	**87.29**	**98.00**

AnimateDiff: Experiments

Compatibility with Text-to-Image Models' Adapter

- AnimateDiff can be directly used with pre-trained T2I adapters, e.g., ControlNet, for controllable generations
- Depth-guided generation with ControlNet-depth

Generating Higher Spatial/Temporal Resolution

Cascaded pipeline

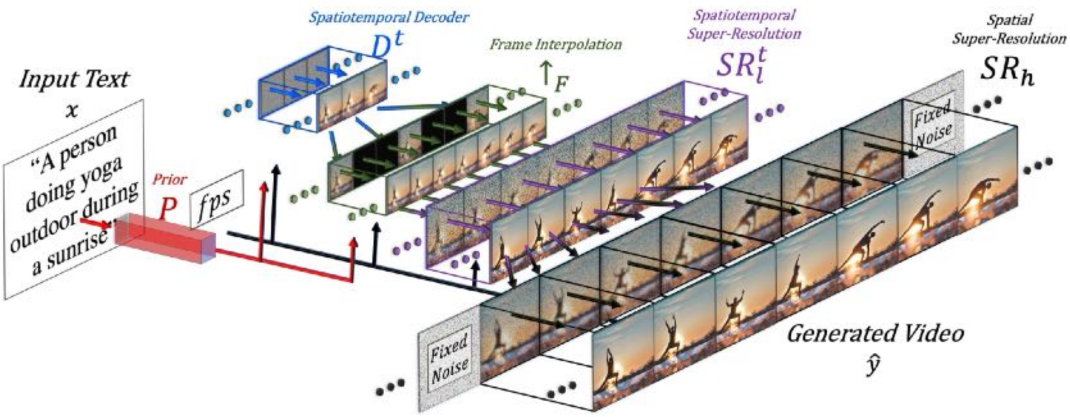
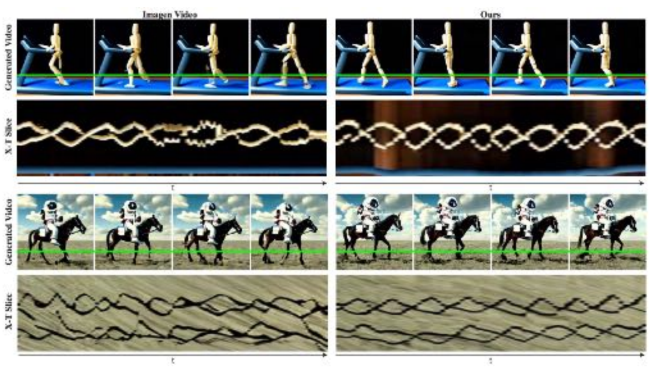


Image credit: Make-A-Video

Generating Higher Spatial/Temporal Resolution

Spatial-Temporal Architecture



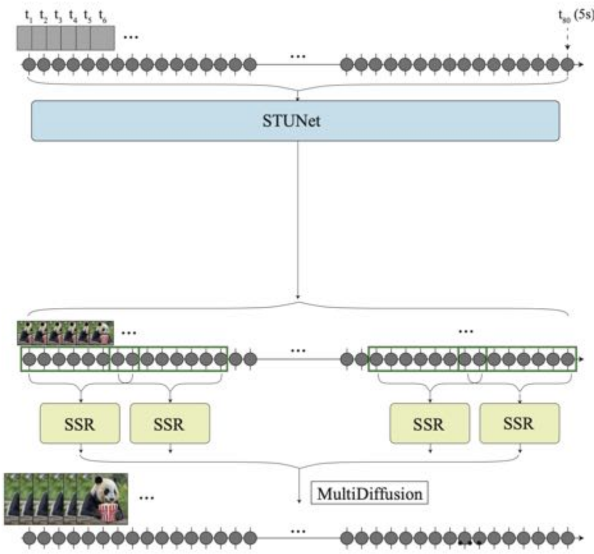


Image credit: Lumiere

Generating Higher Spatial/Temporal Resolution

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Spatial-Temporal Architecture

Transformer-based Approaches

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

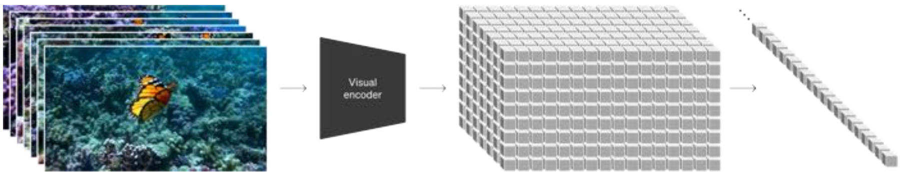


Image credit: Sora, OpenAI

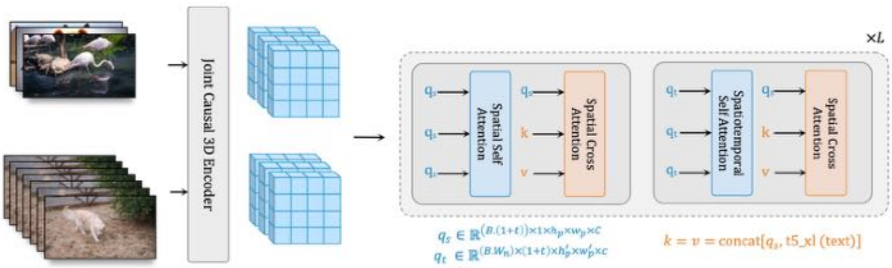
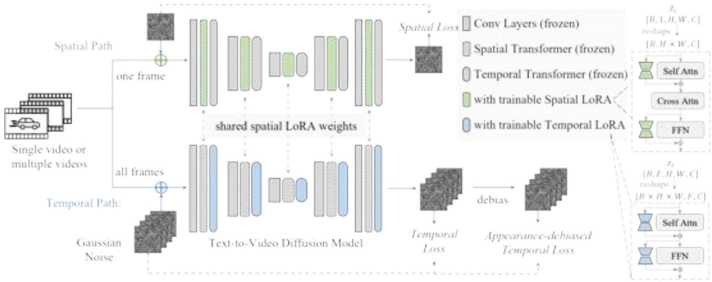
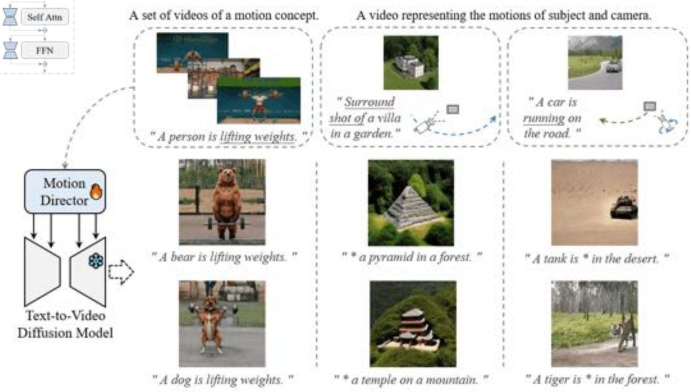


Image credit: W.A.L.T

Controllable Generation

Motion: MotionDirector: Motion Customization of Text-to-Video Diffusion Models

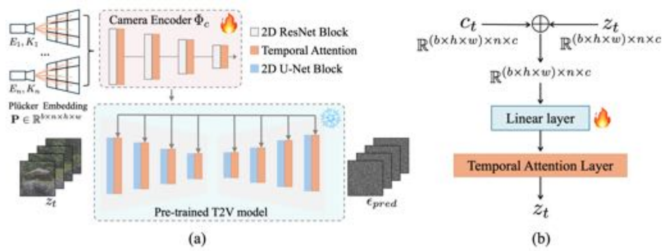




Controllable Generation

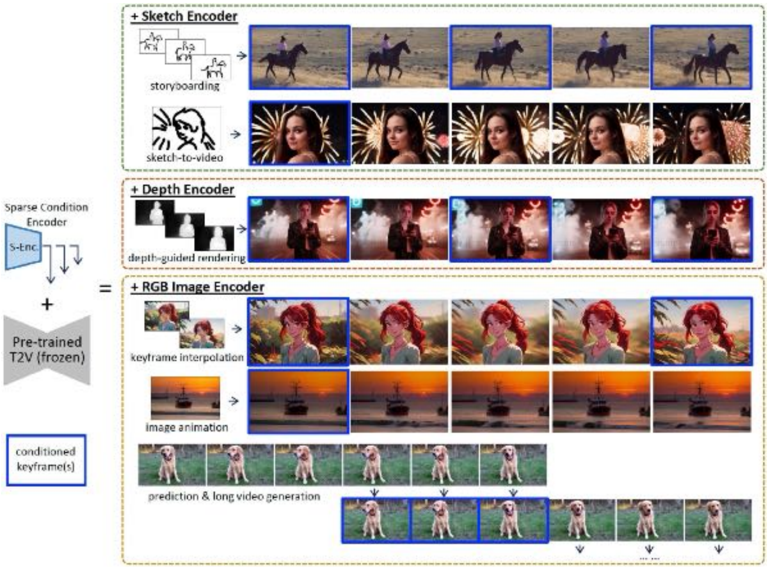
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Camera: CameraCtrl: Enabling Camera Control for Video Diffusion Models



Controllable Generation

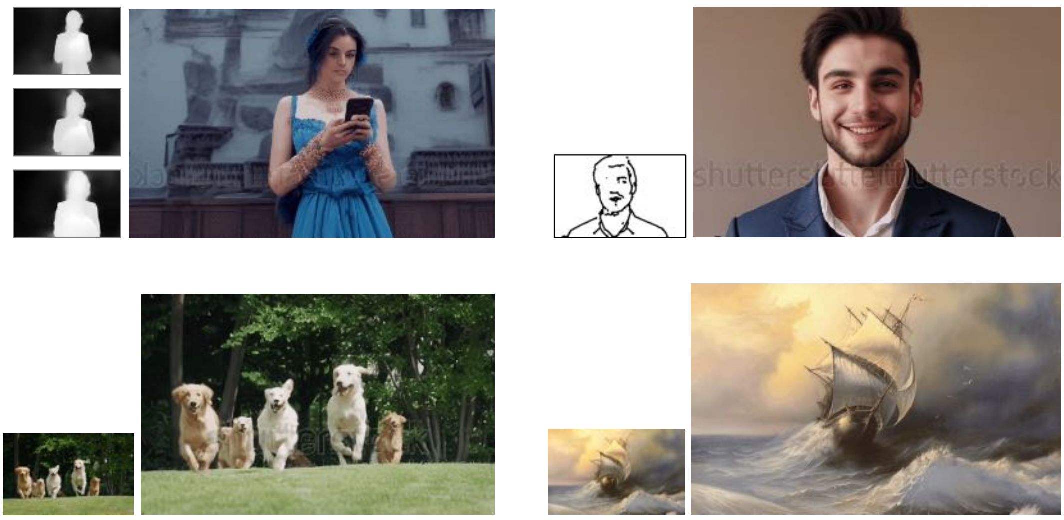
Layout & Pixel: SparseCtrl: Adding Sparse Controls to Text-to-Video Diffusion Models



SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Controllable Generation

Layout & Pixel: SparseCtrl: Adding Sparse Controls to Text-to-Video Diffusion Models



Diffusion Based Video Editing

Global Stylization: Structure and Content-Guided Video Synthesis with Diffusion Models (GEN-1)

Decouple the structure and appearance via depth maps

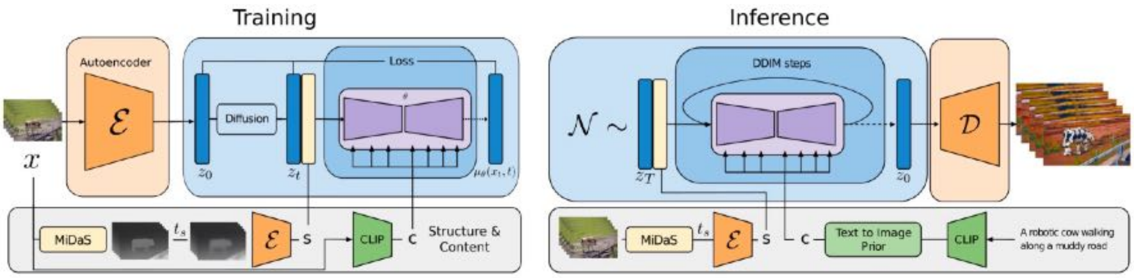


Image credit: GEN-1

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Diffusion Based Video Editing

Global Stylization: Structure and Content-Guided Video Synthesis with Diffusion Models (GEN-1)

Decouple the structure and appearance via depth maps

Diffusion Based Video Editing

Local: Fate/Zero: Fusing Attentions for Zero-shot Text-based Video Editing

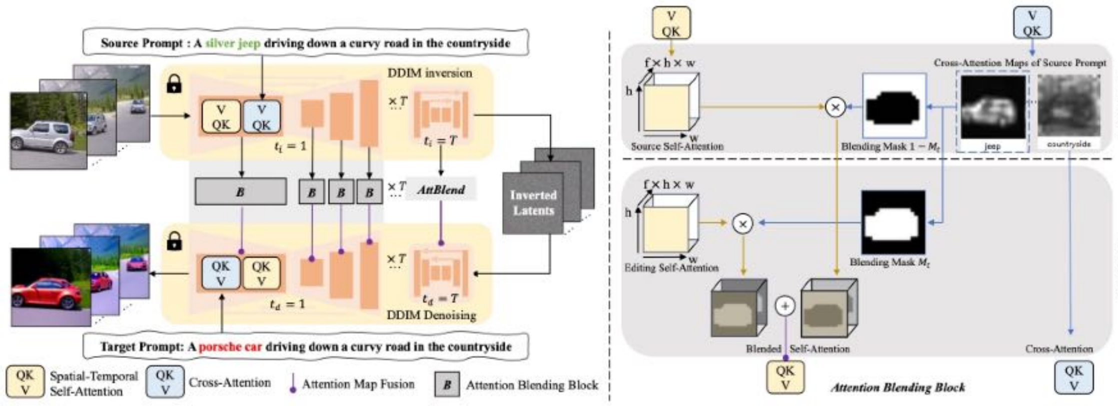


Image credit: FateZero

Diffusion Based Video Editing

Local: Fate/Zero: Fusing Attentions for Zero-shot Text-based Video Editing

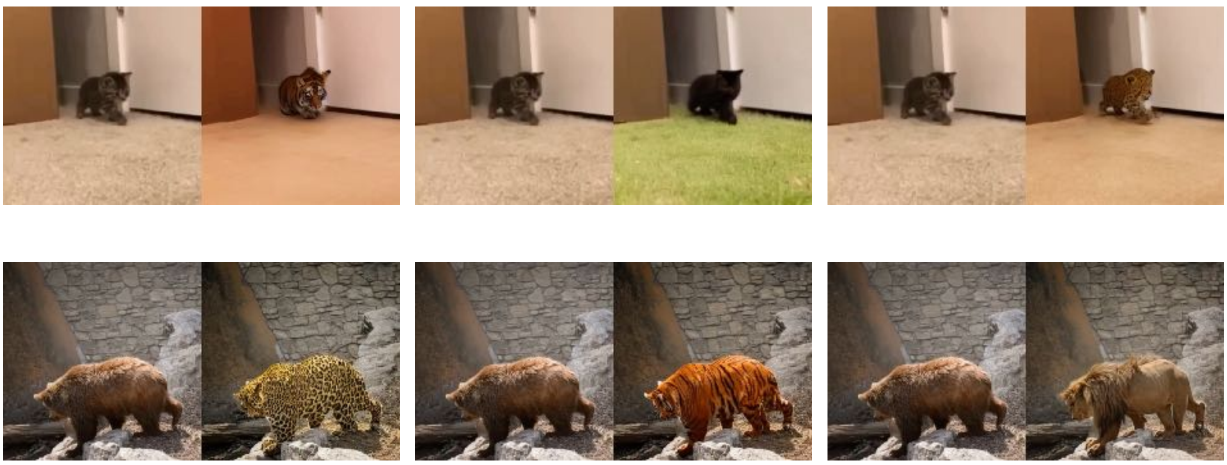
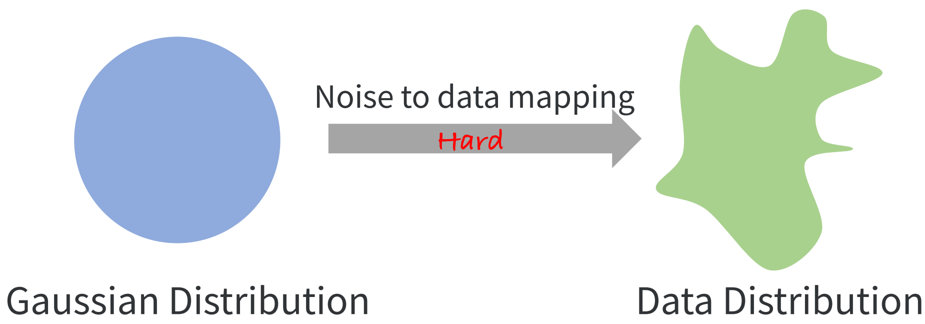


Image credit: FateZero

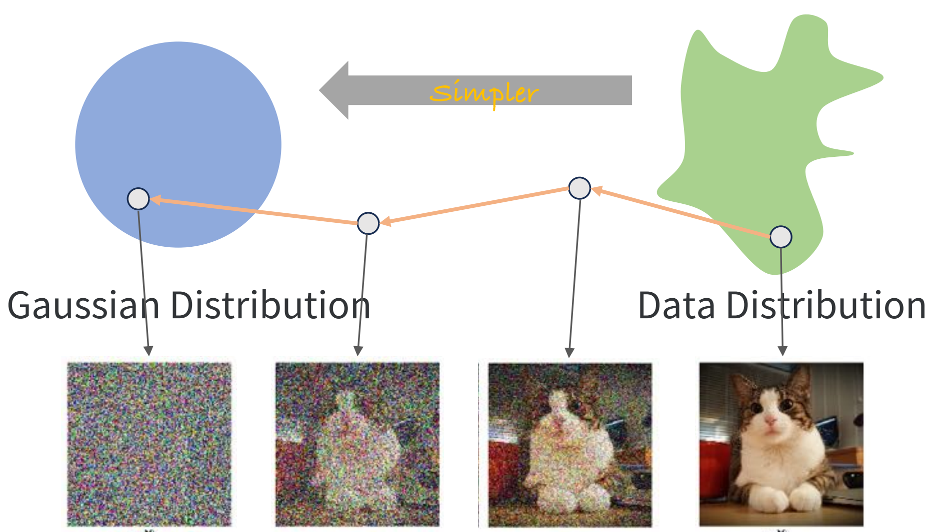
Recap Diffusion

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

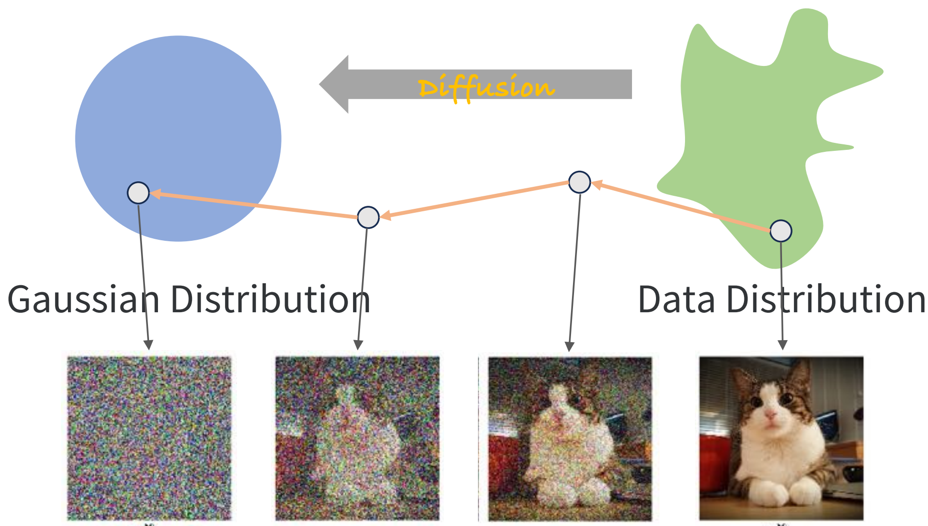


Recap Diffusion

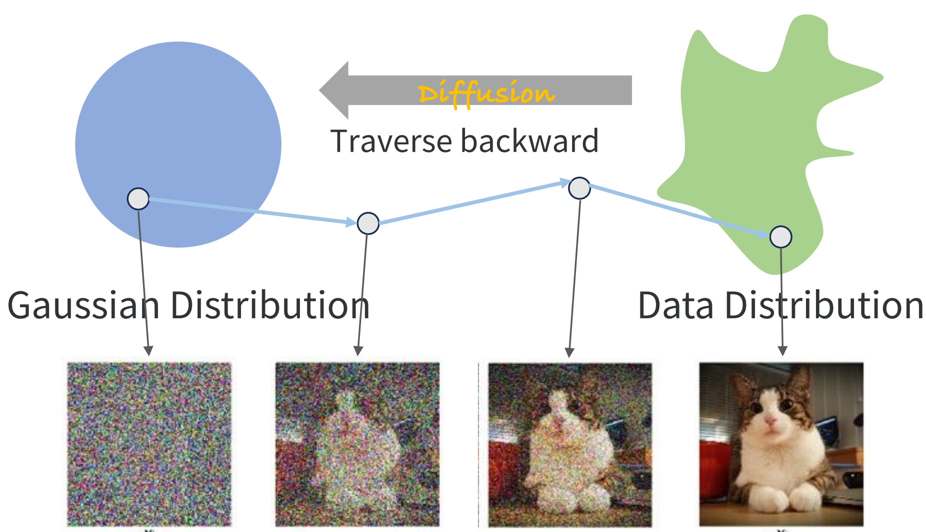
SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG



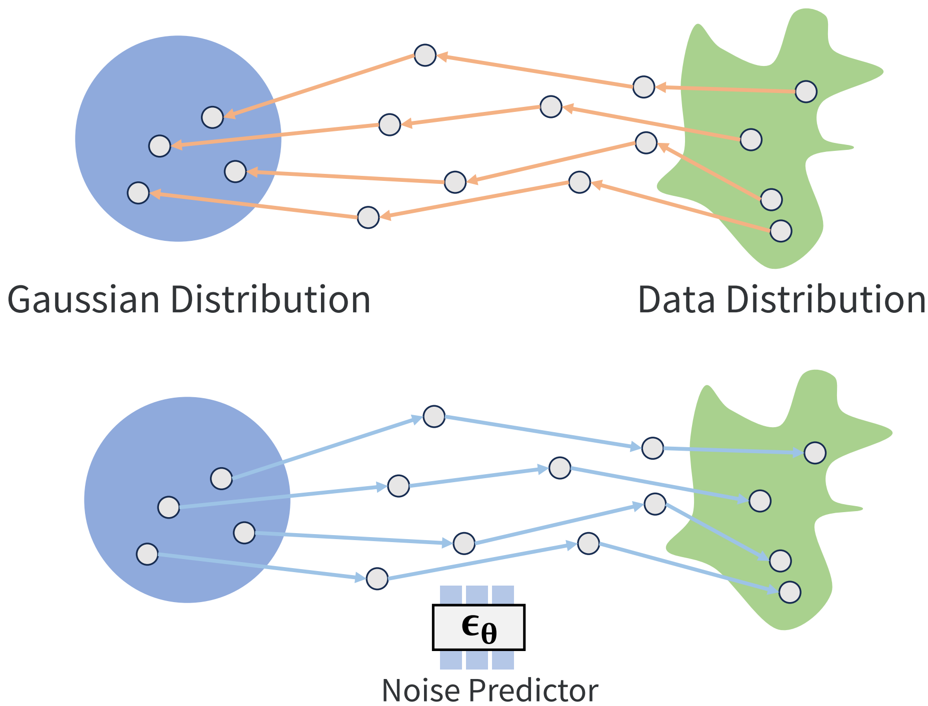
Recap Diffusion



Recap Diffusion

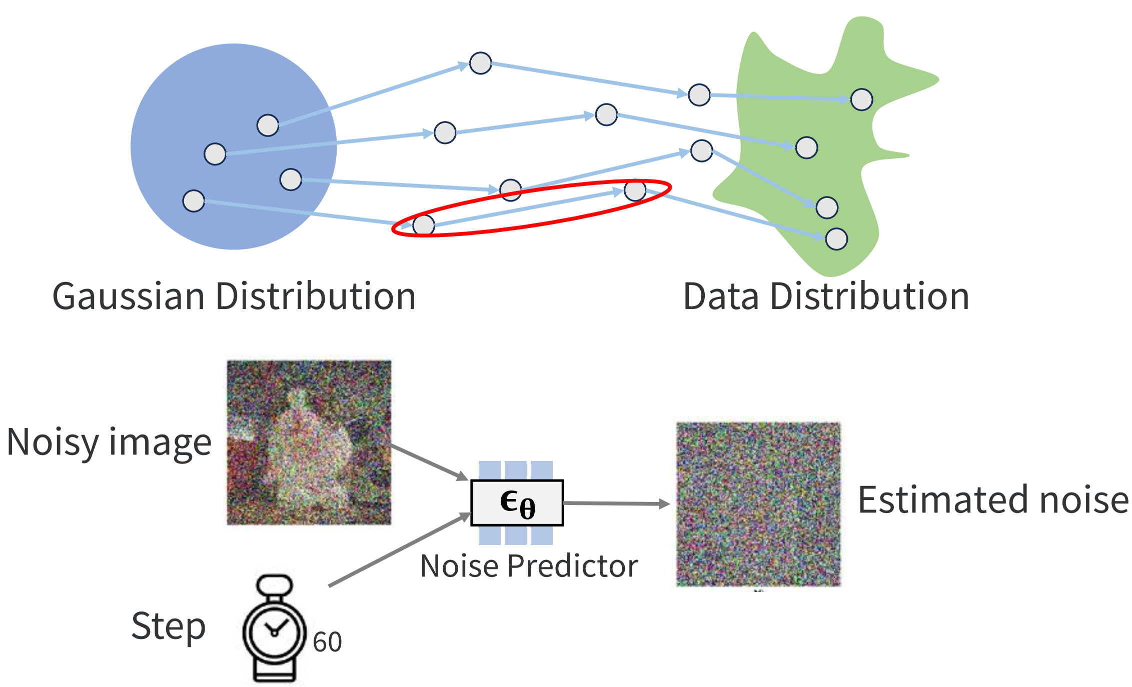


Recap Diffusion

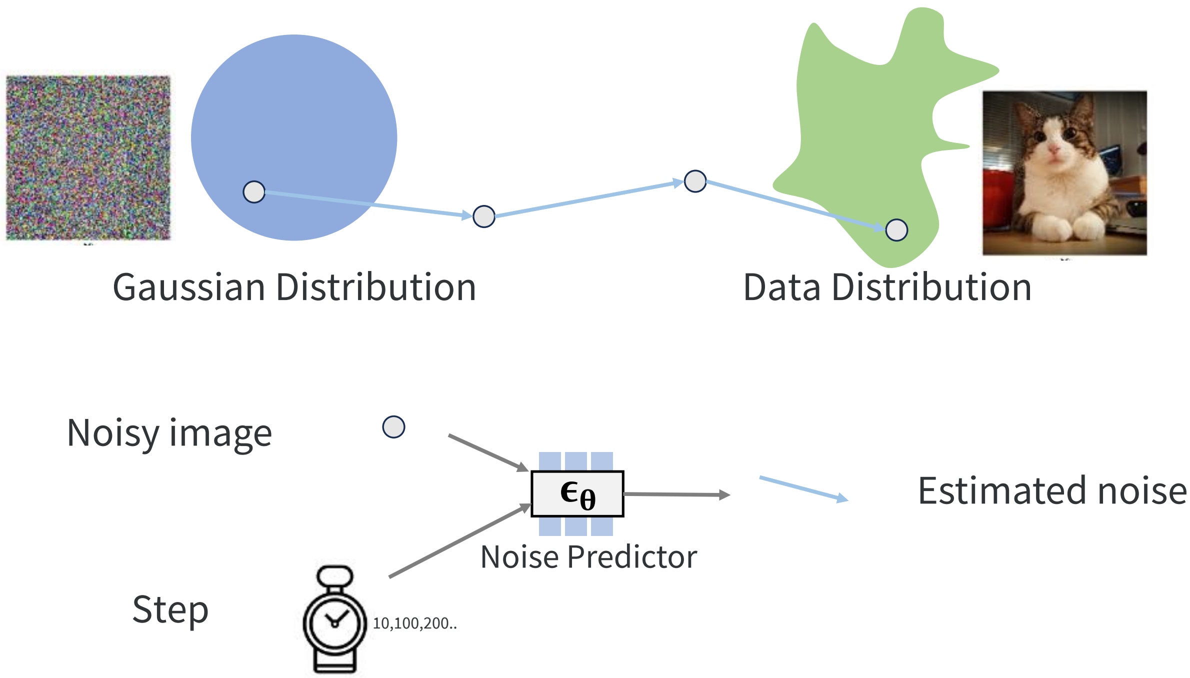


Recap Diffusion

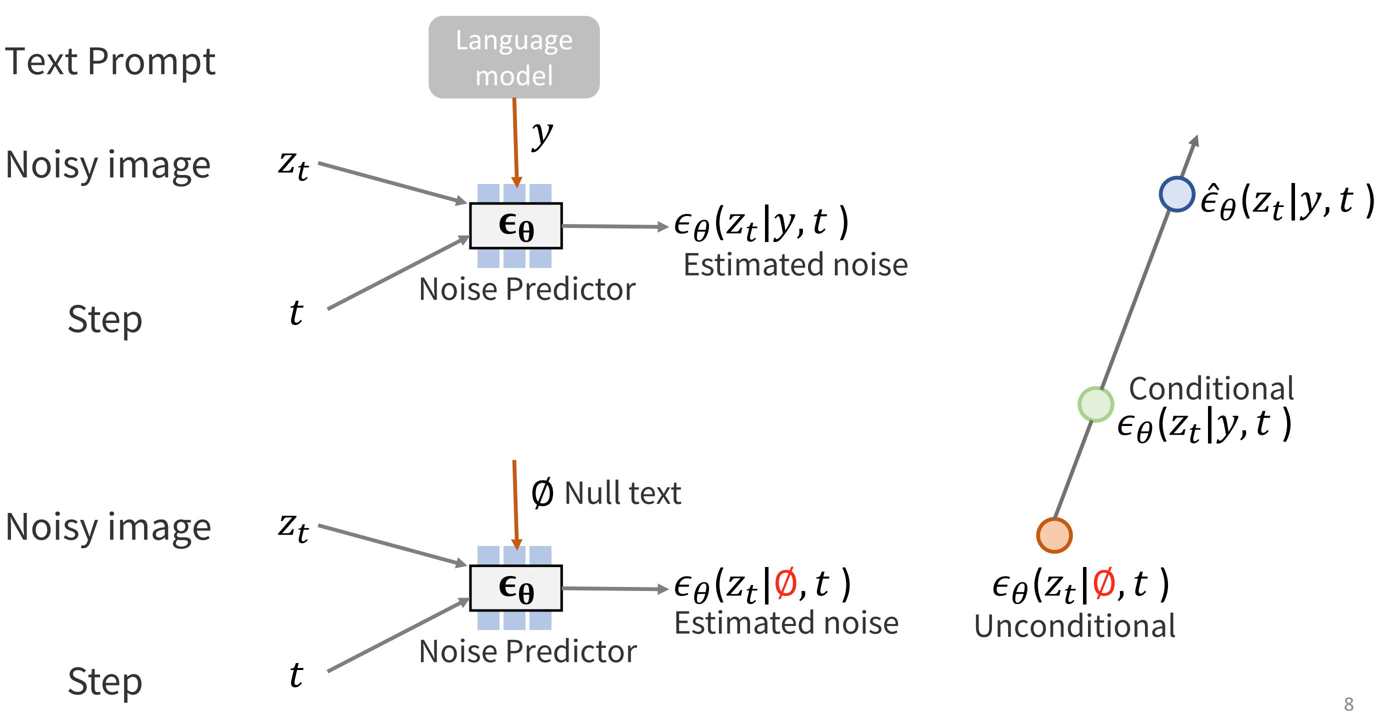
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG



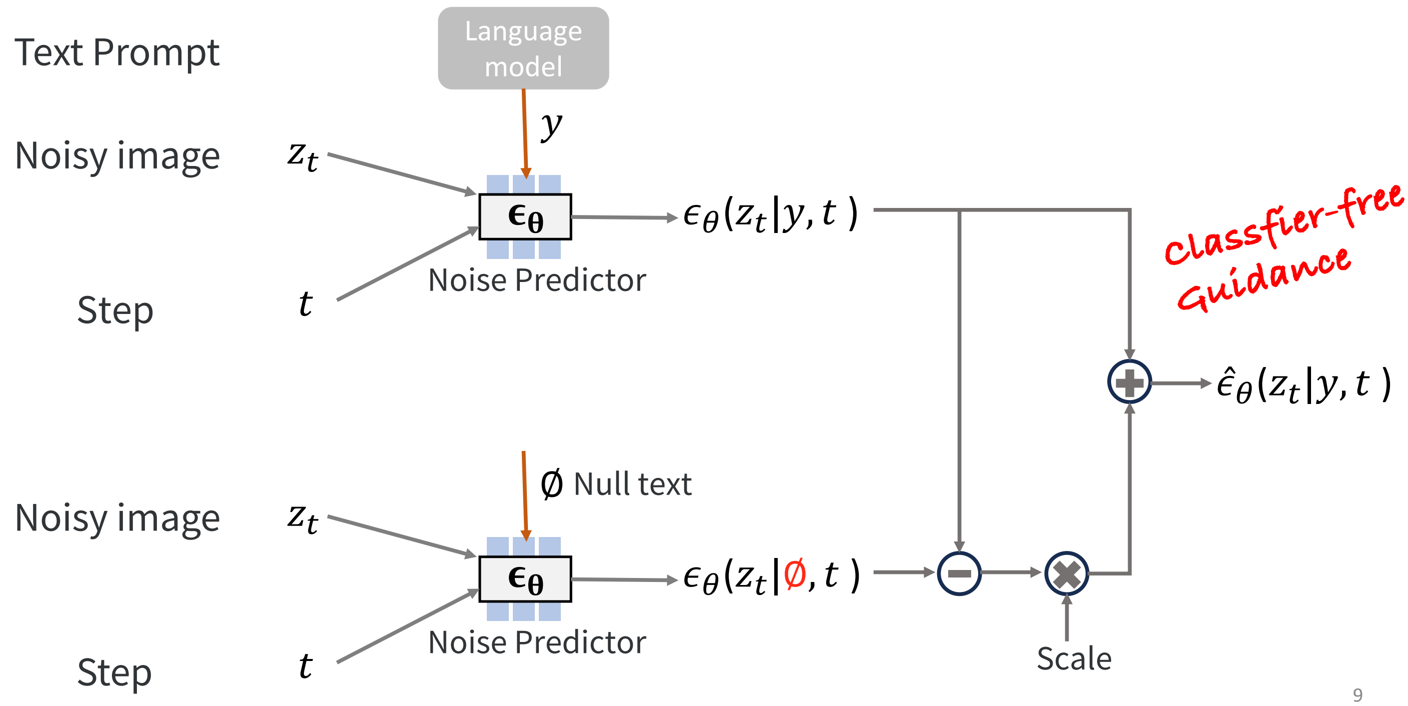
Recap Diffusion



Text Conditioned Diffusion

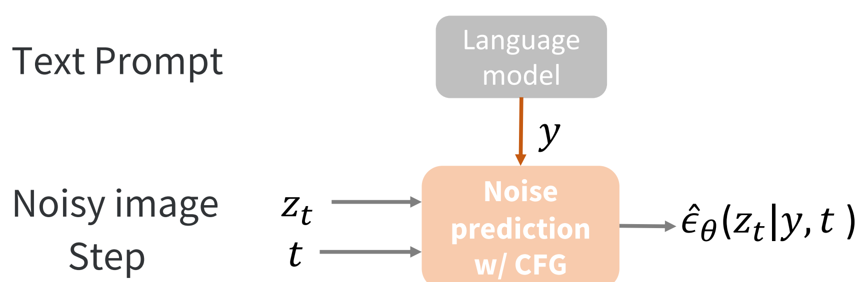


Text Conditioned Diffusion

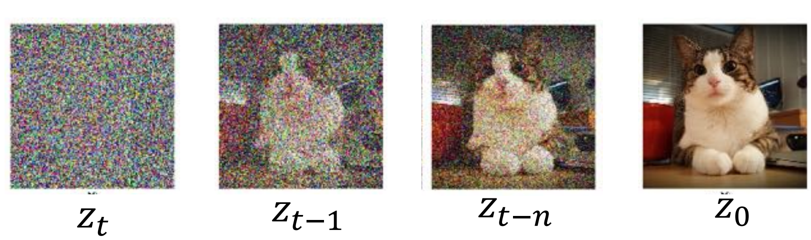


2D Diffusion to 3D

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG



update samples in pixel space ------- 2D images



2D Diffusion to 3D

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Text Prompt → Language model → y

Noisy image z_t →
Step t →

Noise prediction w/ CFG → $\hat{\epsilon}_\theta(z_t|y, t)$

Update samples in pixel space -------- 2D images

How about 3D?

Parameter

$$x = g(\theta)$$

Rendered Image

Differentiable Rendering

Update in parameter space?

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

$$L_{diff}(\phi, x) = \mathbb{E}_{t\sim\mathcal{U}(0,1),\epsilon\sim\mathcal{N}(0,I)}[w(t)\|\epsilon_\phi(z_t|y,t) - \epsilon\|_2^2]$$

$$z_t = a_t x + \sigma_t \epsilon$$

Training a diffusion model:
$\phi^* = argmin_\phi \, L_{diff}(\phi, x)$

With a trained diffusion model:
$x^* = argmin_x \, L_{diff}(\phi, x)$

Parameter

$$x = g(\theta)$$

Rendered Image

Differentiable Rendering

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024 DENVER+ 28 JUL — 1 AUG

$$L_{diff}(\phi, g(\theta)) = \mathbb{E}_{t\sim\mathcal{U}(0,1),\epsilon\sim\mathcal{N}(0,I)}[w(t)\| \epsilon_\phi(z_t|y,t) - \epsilon\|_2^2]$$

$$\nabla_\theta L_{diff}(\phi, g(\theta)) = \mathbb{E}_{t,\epsilon}[w(t)\underbrace{(\hat{\epsilon}_\phi(z_t|y,t) - \epsilon)}_{\text{Noise Residual}} \underbrace{\frac{\partial \hat{\epsilon}_\phi(z_t|y,t)}{\partial z_t}}_{\text{U-Net Jacobian}} \underbrace{\frac{\partial x}{\partial \theta}}_{\text{Generator Jacobian}}]$$

Update 3D representation w/ gradient descent

Text Prompt → Language model → y

Noisy image z_t →

Step t → Noise prediction w/ CFG → $\hat{\epsilon}_\theta(z_t|y,t)$

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024 DENVER+ 28 JUL — 1 AUG

$$L_{diff}(\phi, g(\theta)) = \mathbb{E}_{t\sim\mathcal{U}(0,1),\epsilon\sim\mathcal{N}(0,I)}[w(t)\|\epsilon_\phi(z_t|y,t) - \epsilon\|_2^2]$$

$$\nabla_\theta L_{diff}(\phi, g(\theta)) = \mathbb{E}_{t,\epsilon}[w(t)\underbrace{(\hat{\epsilon}_\phi(z_t|y,t) - \epsilon)}_{\text{Noise Residual}} \underbrace{\cancel{\frac{\partial \hat{\epsilon}_\phi(z_t|y,t)}{\partial z_t}}}_{\text{U-Net Jacobian}} \underbrace{\frac{\partial x}{\partial \theta}}_{\text{Generator Jacobian}}]$$

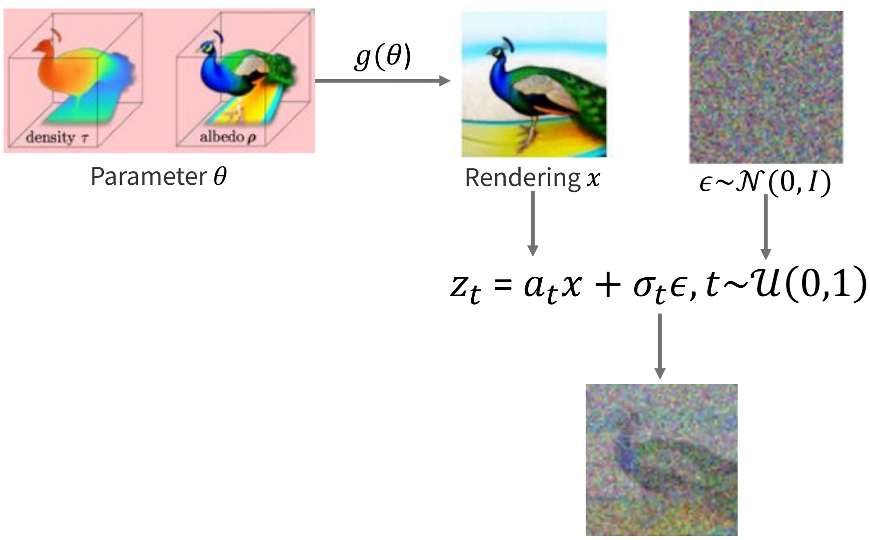
$$\nabla_\theta L_{SDS}(\phi, g(\theta)) = \mathbb{E}_{t,\epsilon}[w(t)(\hat{\epsilon}_\phi(z_t|y,t) - \epsilon) \frac{\partial x}{\partial \theta}]$$

Score Distillation Sampling

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

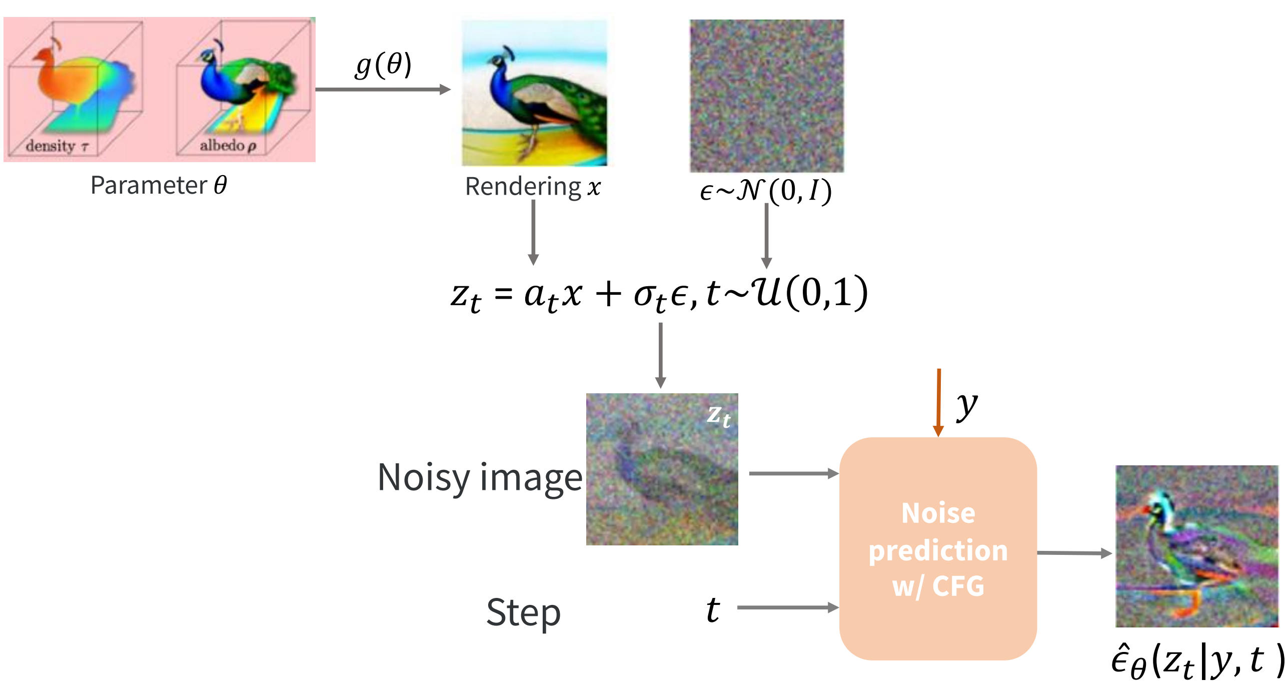
Score Distillation Sampling (SDS) Loss



3D Content Generation *Empowered by 2D Diffusion Priors*

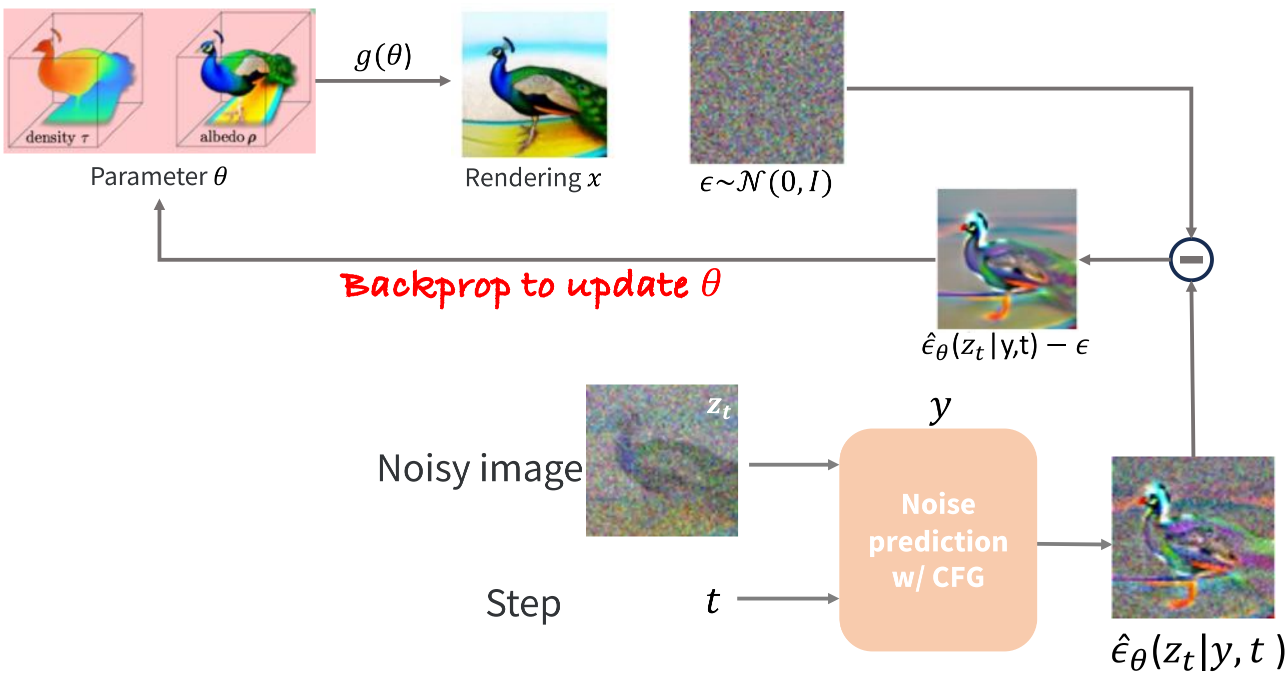
SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

Score Distillation Sampling (SDS) Loss



3D Content Generation *Empowered by 2D Diffusion Priors*

Score Distillation Sampling (SDS) Loss



3D Content Generation *Empowered by 2D Diffusion Priors*

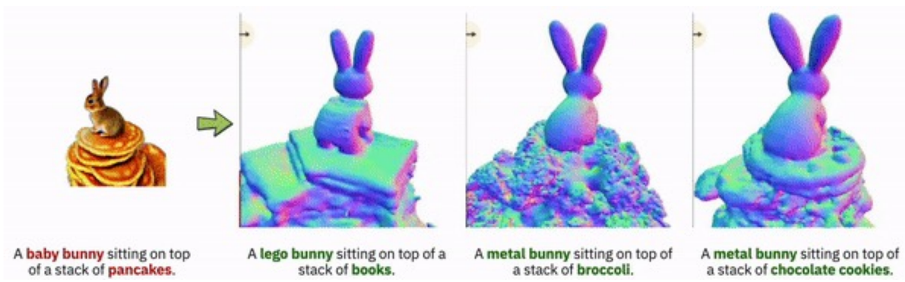
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

DreamFusion: Text-to-3D using 2D Diffusion

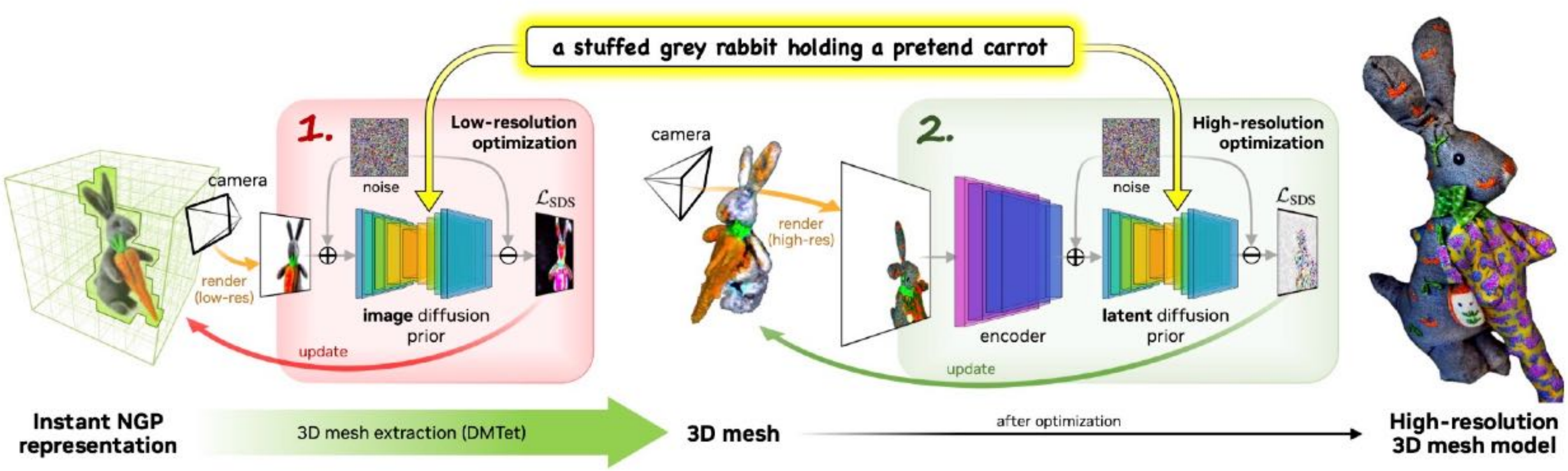
Magic3D: High-Resolution Text-to-3D Content Creation



3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Magic3D: High-Resolution Text-to-3D Content Creation



3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

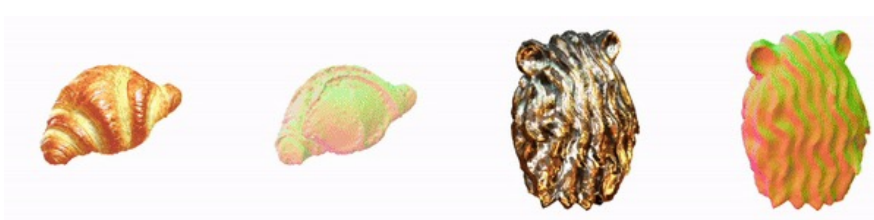
3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

DreamFusion: Text-to-3D using 2D Diffusion

Richer appearance

Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation

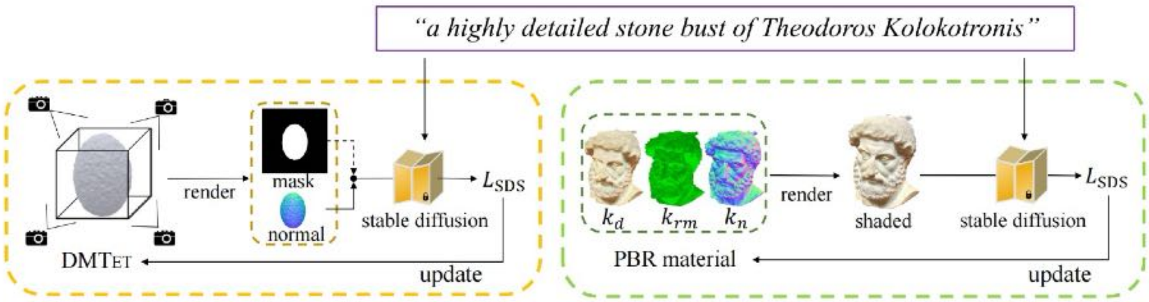


A delicious croissant   a metal sculpture of a lion's head, highly detailed

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation



(b) Geometry modeling (c) Appearance modeling

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

DreamFusion: Text-to-3D using 2D Diffusion

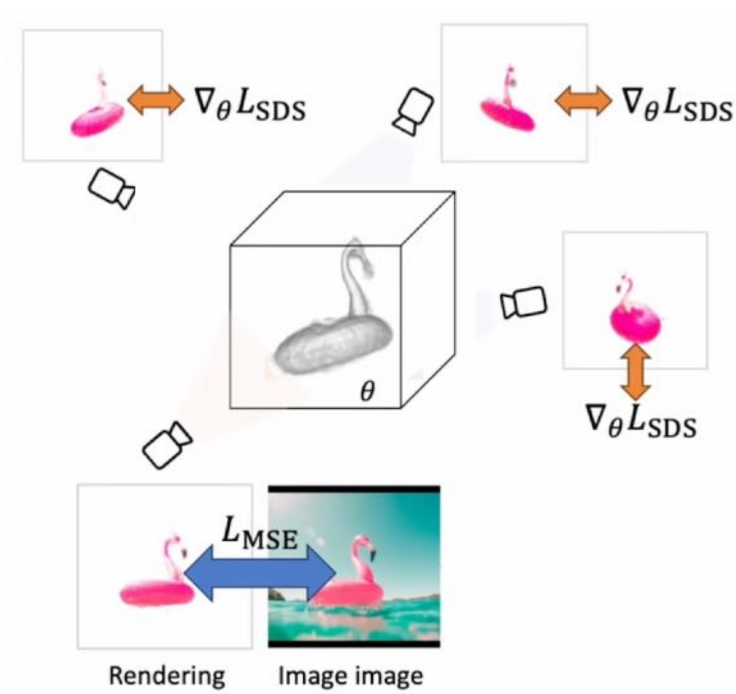
Single image to 3D

RealFusion: 360° Reconstruction of Any Object from a Single Image

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

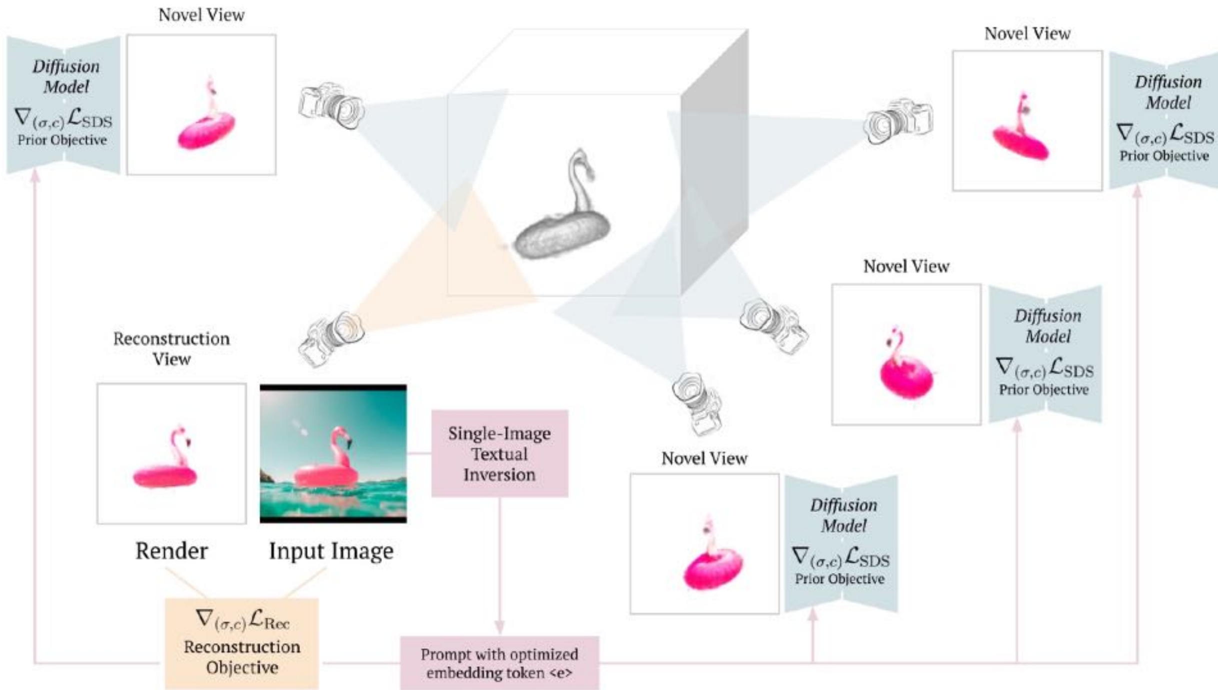
RealFusion: 360° Reconstruction of Any Object from a Single Image



3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

RealFusion: 360° Reconstruction of Any Object from a Single Image

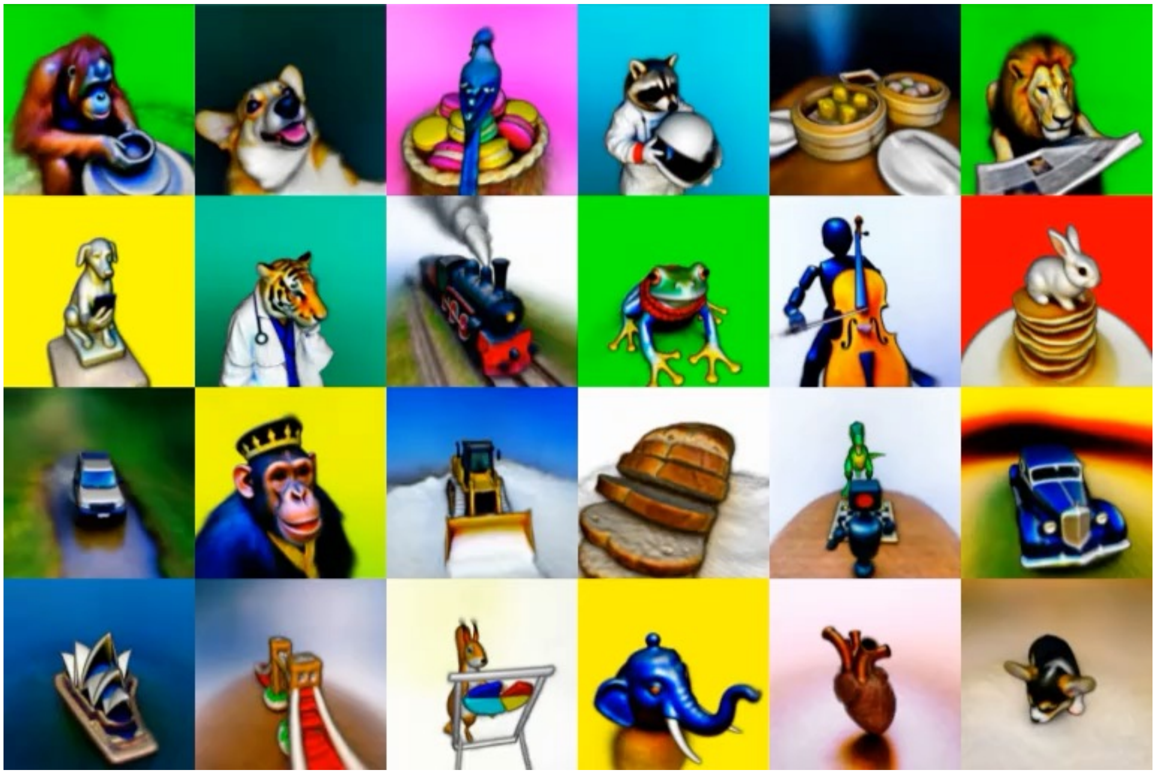


3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG



3D Content Generation *Empowered by 2D Diffusion Priors*

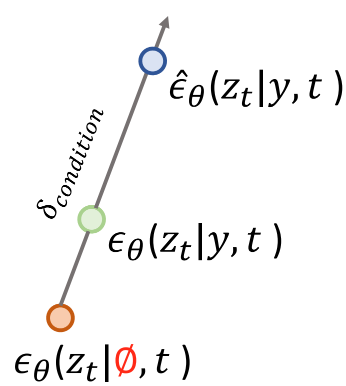
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

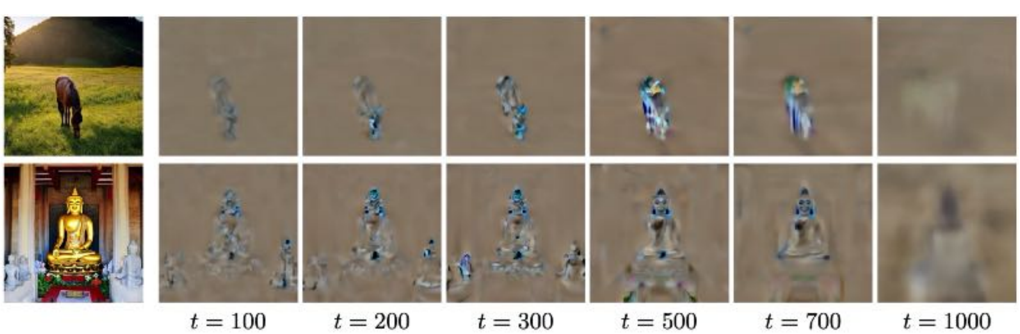
Photorealistic appearance?

$$\nabla_\theta L_{SDS}(\phi, g(\theta)) = \mathbb{E}_{t,\epsilon}[w(t)(\boxed{\hat{\epsilon}_\theta(z_t|y,t)} - \epsilon)\frac{\partial x}{\partial \theta}]$$

Estimated noise
with classifier free guidance

$$\hat{\epsilon}_\theta(z_t|y,t) = \epsilon_\phi(z_t|\emptyset,t) + s\underbrace{(\epsilon_\phi(z_t|y,t) - \epsilon_\phi(z_t|\emptyset,t))}_{\delta_{condition}}$$





aligned with the condition;
uncorrelated with the added noise ϵ

3D Content Generation *Empowered by 2D Diffusion Priors*

Photorealistic appearance?

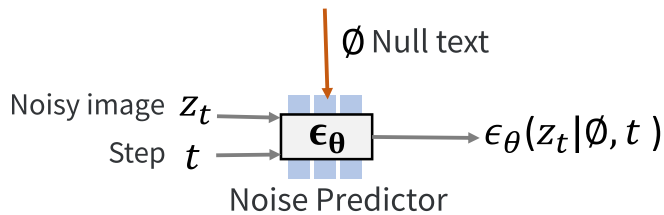
$$\nabla_\theta L_{SDS}(\phi, g(\theta)) = \mathbb{E}_{t,\epsilon}[w(t)(\boxed{\hat{\epsilon}_\theta(z_t|y, t)} - \epsilon)\frac{\partial x}{\partial \theta}]$$

Estimated noise
with classifier free guidance

$$\hat{\epsilon}_\theta(z_t|y, t) = \boxed{\epsilon_\phi(z_t|\emptyset, t)} + s\delta_{condition}$$

$z_t = a_t x + \sigma_t \epsilon$

Training: Real image x
SDS: Rendered image $x = g(\theta)$
Domain difference



3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Photorealistic appearance?

$$\nabla_\theta L_{SDS}(\phi, g(\theta)) = \mathbb{E}_{t,\epsilon}[w(t)(\boxed{\hat{\epsilon}_\theta(z_t|y, t)} - \epsilon)\frac{\partial x}{\partial \theta}]$$

Estimated noise
with classifier free guidance

$$\hat{\epsilon}_\theta(z_t|y, t) = \boxed{\epsilon_\phi(z_t|\emptyset, t)} + s\delta_{condition}$$

$$\epsilon_\phi(z_t|\emptyset, t) = \delta_{domain} + \delta_{denoising}$$

x_{ID} | $\delta_{denoising}$ | x_{OOD} | δ_{domain} | $x_{OOD} + \delta_{domain}$

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Photorealistic appearance?

$$\nabla_\theta L_{SDS}(\phi, g(\theta)) = w(t)(\delta_{domain} + s\delta_{condition} + \delta_{denoising} - \epsilon)\,\frac{\partial x}{\partial \theta}$$

Domain-correction

Align w/ text

Diff b/w predicted noise
& added noise

Training: Real image x
SDS: Rendered image $x = g(\theta)$

Domain difference

x_{ID}

x_{OOD}

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Photorealistic appearance?

$$\nabla_\theta L_{SDS}(\phi, g(\theta)) = w(t)(\delta_{domain} + s\delta_{condition} + \delta_{denoising} - \epsilon)\,\frac{\partial x}{\partial \theta}$$

Domain-correction

Align w/ text

Diff b/w predicted noise
& added noise

Disgard this one!

$$s\delta_{condition} = s(\underbrace{\epsilon_\phi(z_t|y,t)}_{\text{Conditional}} - \underbrace{\epsilon_\phi(z_t|\emptyset,t)}_{\text{Unconditional}})$$

$$\epsilon_\phi(z_t|\emptyset,t) = \delta_{domain} + \delta_{denoising}$$

🤔 hard to separate

Assumption: $\delta_{condition=p_{neg}} = -\delta_{domain}$

" unrealistic, blurry, low quality, out of focus, ugly, low contrast, dull, dark, low-resolution, gloomy"

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Photorealistic appearance?

$$\nabla_\theta L_{SDS}(\phi, g(\theta)) = w(t)(\delta_{domain} + s\delta_{condition} + \delta_{denoising} - \epsilon)\,\frac{\partial x}{\partial \theta}$$

Domain-correction

Align w/ text

Diff b/w predicted noise
& added noise

Disgard this one!

$$s\delta_{condition} = \underbrace{s(\epsilon_\phi(z_t|y,t)}_{\text{Conditional}} - \underbrace{\epsilon_\phi(z_t|\emptyset,t))}_{\text{Unconditional}}$$

$$\epsilon_\phi(z_t|\emptyset,t) = \delta_{domain} + \delta_{denoising}$$

🤔 **hard to separate**

$$\epsilon_\phi(z_t|\emptyset,t) - \epsilon_\phi(z_t|y = p_{neg},t) =$$
$$\delta_{domain} + \cancel{\delta_{denoising}} - (\cancel{\delta_{domain}} + \cancel{\delta_{denoising}} + \cancel{\delta_{condition=p_{neg}}})$$

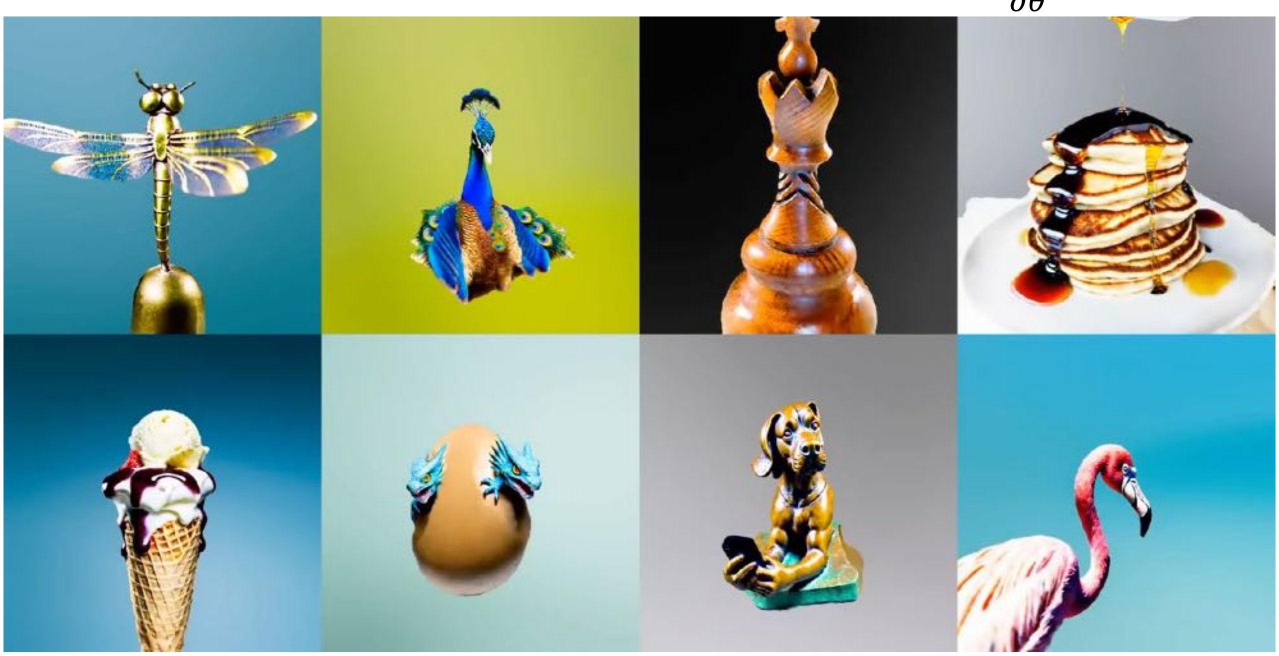
Assumption: $\delta_{condition=p_{neg}} = -\delta_{domain}$

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

NFSD: Noise Free Score Distillation

$$\nabla_\theta L_{NFSD}(\phi, g(\theta)) = w(t)(\delta_{domain} + s\delta_{condition})\,\frac{\partial x}{\partial \theta}$$



3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

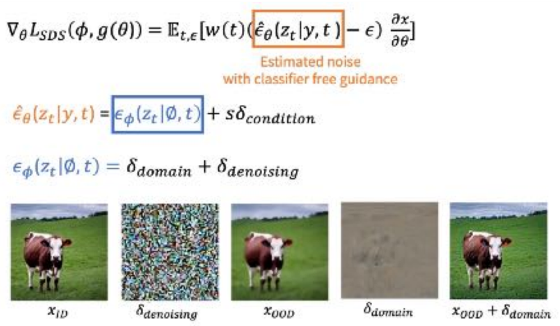
ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation

$$\nabla_\theta L_{VSD}(\phi, g(\theta)) = w(t)(\delta_{domain} + s\delta_{condition})\ \frac{\partial x}{\partial \theta}$$

$$\nabla_\theta L_{\text{VSD}}(\phi, g(\theta)) = w(t)(\underbrace{\hat{\epsilon}_\theta(z_t|y,t)}_{\text{Conditional model}} - \underbrace{\epsilon_{\text{LoRA}}(z_t|y,t,c)}_{\text{LoRA model, trained on the rendered images}})\ \frac{\partial x}{\partial \theta}]$$

$$\epsilon_{\text{LoRA}}(z_t|y,t,c) = \delta_{denoising}$$

No domain difference this time!



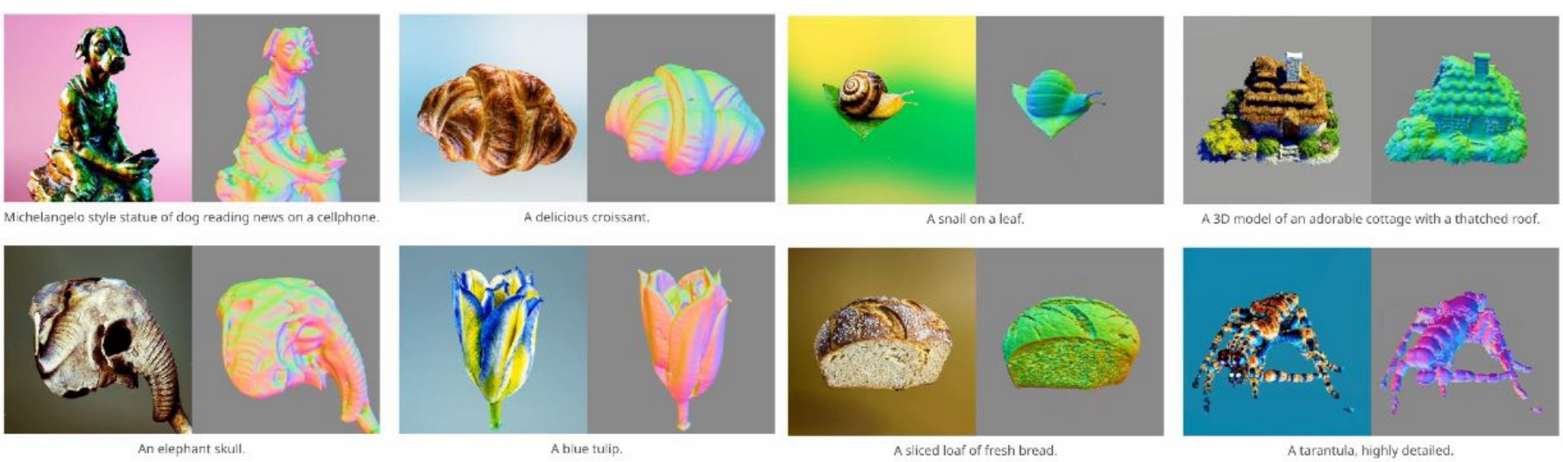
$$\hat{\epsilon}_\theta(z_t|y,t) - \epsilon_{\text{LoRA}}(z_t|y,t,c) = \delta_{domain} + \cancel{\delta_{denoising}} + s\delta_{condition} - \cancel{\delta_{denoising}}$$

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation

$$\nabla_\theta L_{VSD}(\phi, g(\theta)) = w(t)(\delta_{domain} + s\delta_{condition})\ \frac{\partial x}{\partial \theta}$$



3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Janus (multi-face) problem

3D Content Generation *Empowered by 2D Diffusion Priors*

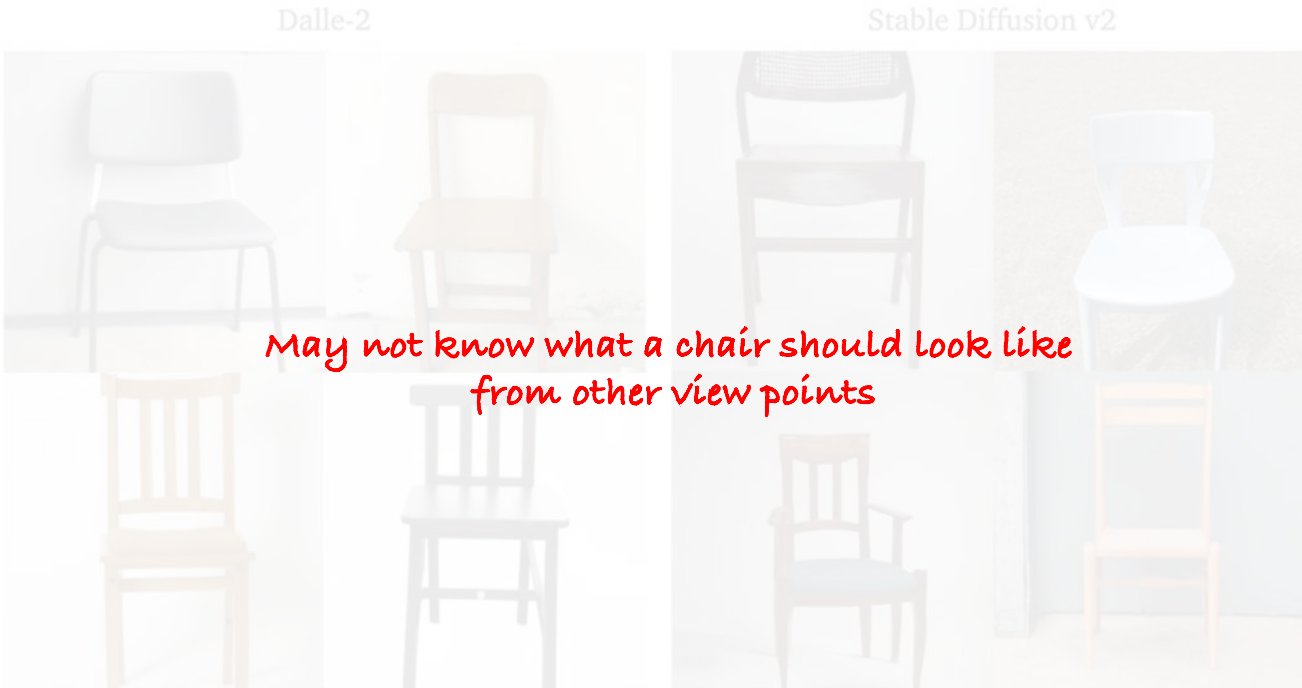
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Janus (multi-face) problem

3D Content Generation *Empowered by 2D Diffusion Priors*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Janus (multi-face) problem



May not know what a chair should look like
from other view points

3D Content Generation *Learn from 3D data*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Janus (multi-face) problem

How to use 3D data?

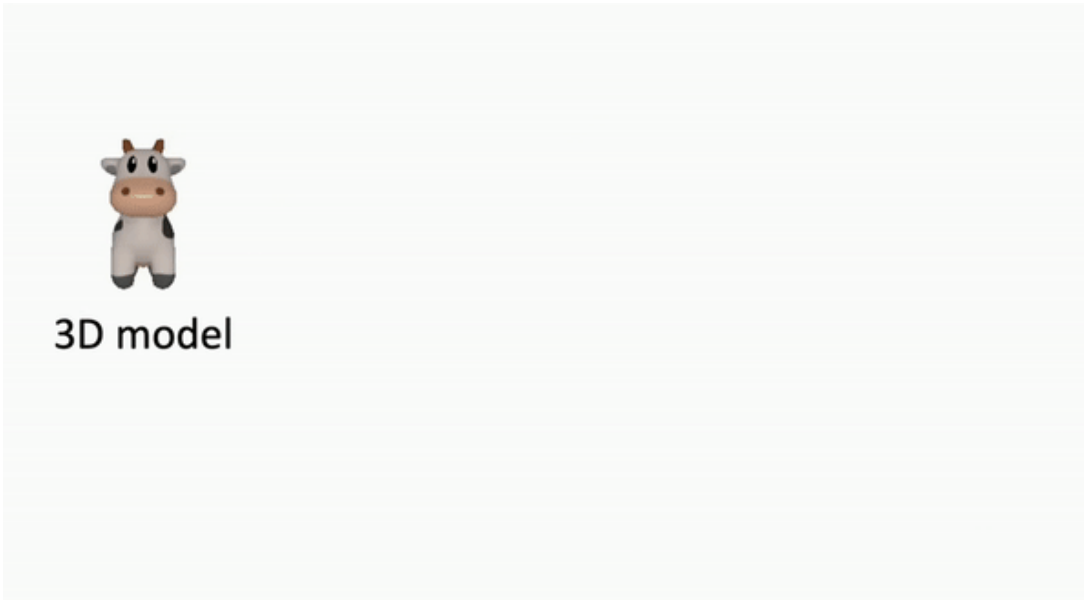
A Universe of
Annotated 3D Objects
Objaverse 1.0 is a Massive Dataset with 800K+ Annotated 3D Objects

Objaverse-XL
A Universe of 10M+ 3D Objects

📝 Image dataset: 5B

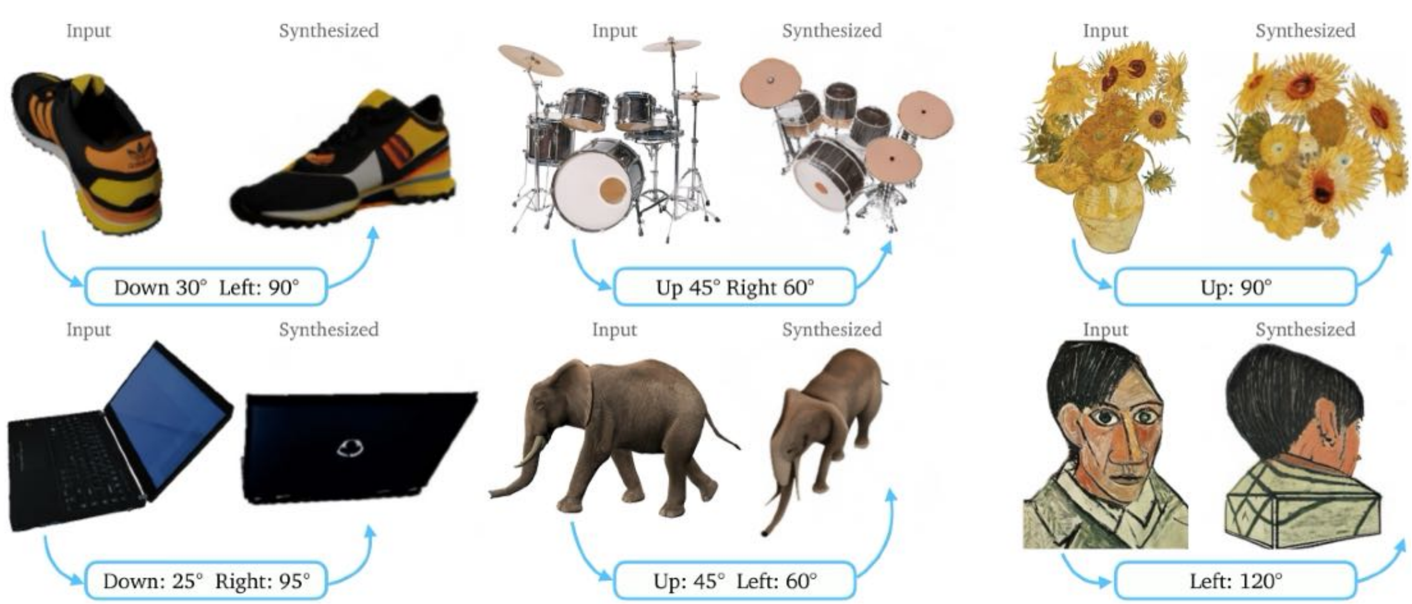
3D Content Generation *Learn from 3D data*

Use 3D models to get multi-view images



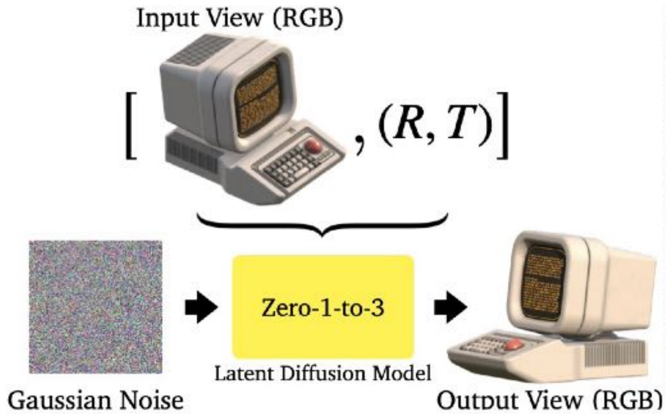
3D Content Generation *Learn from 3D data*

Zero-1-to-3: Zero-shot One Image to 3D Object



3D Content Generation *Learn from 3D data*

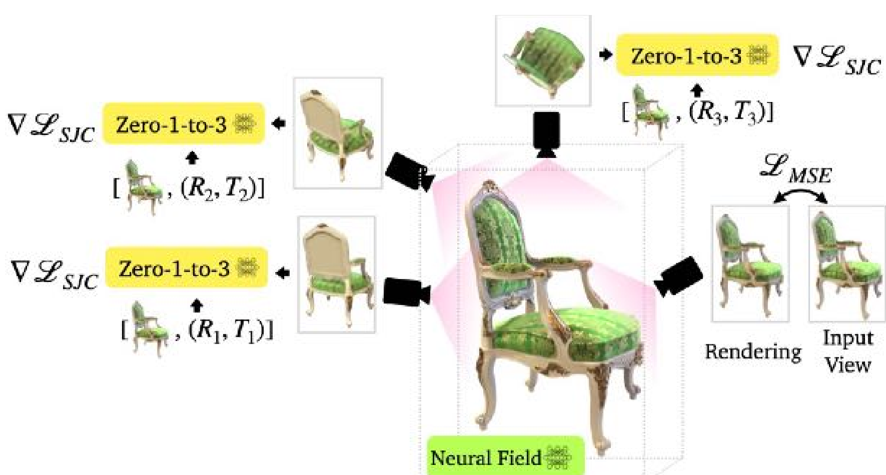
Zero-1-to-3: Zero-shot One Image to 3D Object



Idea: Finetune a 2D diffusion model to generate novel views (i.e. condition on camera pose)

3D Content Generation *Learn from 3D data*

Zero-1-to-3: Zero-shot One Image to 3D Object

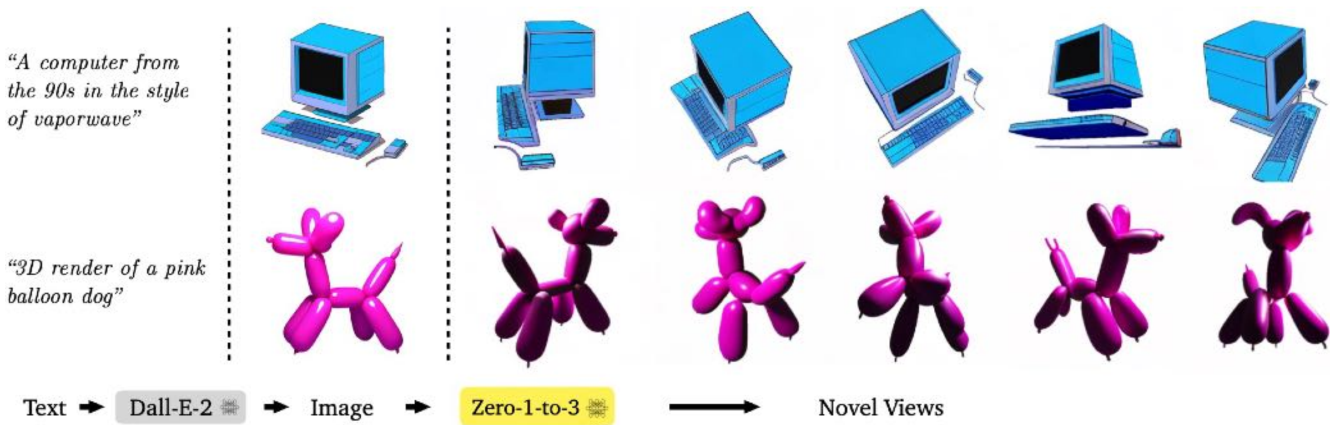


The trained model can be used for (single-view) 3D reconstruction

3D Content Generation *Learn from 3D data*

Zero-1-to-3: Zero-shot One Image to 3D Object

Text2img2NVS



3D Content Generation *Learn from 3D data*

Zero-1-to-3: Zero-shot One Image to 3D Object

Single view 3D reconstruction

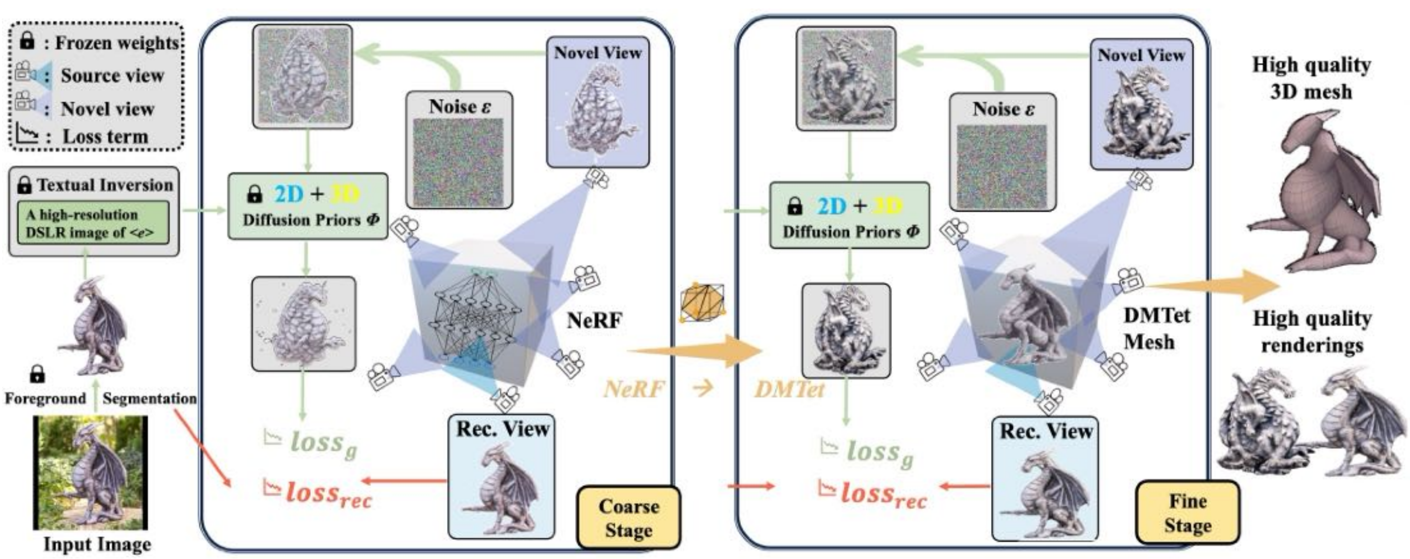
3D Content Generation *Learn from 3D data*

But 3D prior only is blurry, 2D prior only lacks geometry...

3D Content Generation *Learn from 3D data*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

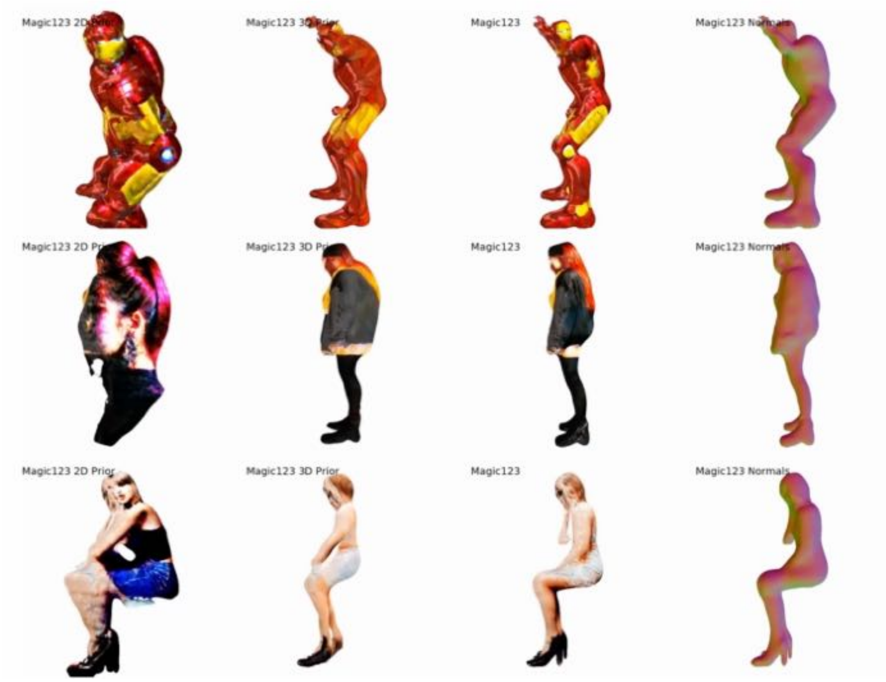
Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors



Idea: Combine both 2D and 3D diffusion priors

3D Content Generation *Learn from 3D data*

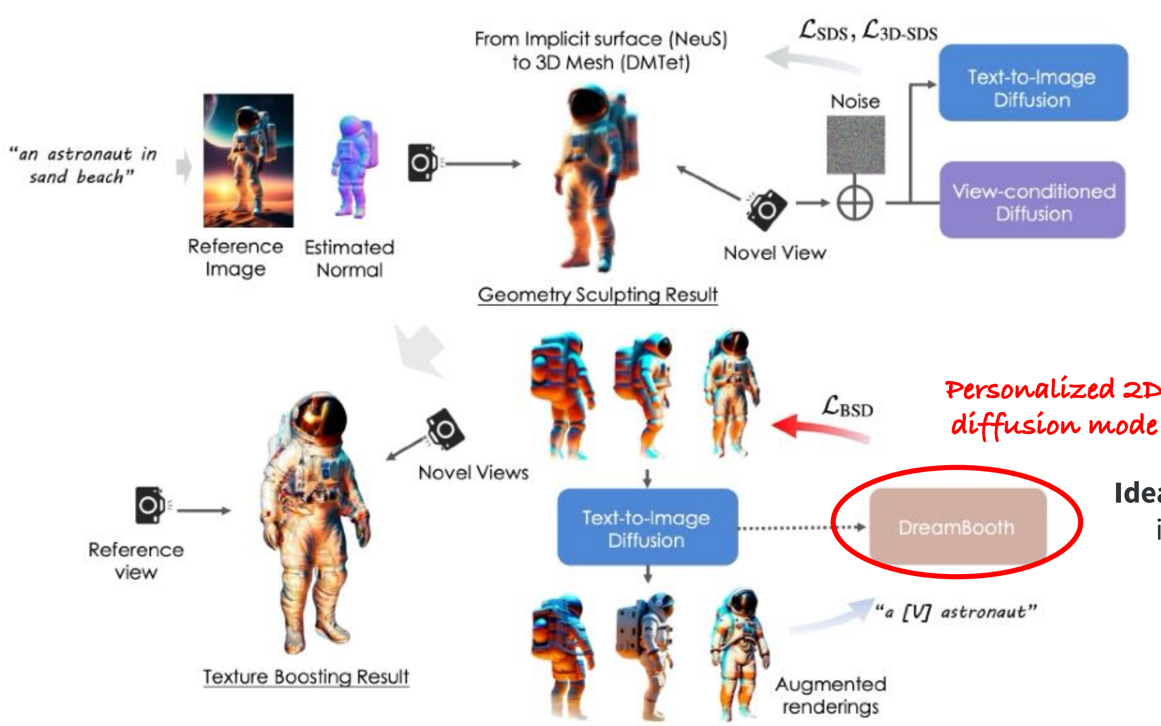
Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors



3D Content Generation *Learn from 3D data*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

DreamCraft3D: Hierarchical 3D Generation with Bootstrapped Diffusion Prior



Idea: Enhanced VSD loss to improve appearance

3D Content Generation *Learn from 3D data*

DreamCraft3D: Hierarchical 3D Generation with Bootstrapped Diffusion Prior

Personalized 2D diffusion mode

Idea: Enhanced VSD loss to improve appearance

3D Content Generation *Learn from 3D data*

DreamCraft3D: Hierarchical 3D Generation with Bootstrapped Diffusion Prior

Recall: $\nabla_\theta L_{\text{VSD}}(\phi, g(\theta)) = w(t)(\hat{\epsilon}_\theta(z_t|y,t,z) - \epsilon_{\text{LoRA}}(z_t|y,t,c,z))\ \frac{\partial x}{\partial \theta}]$

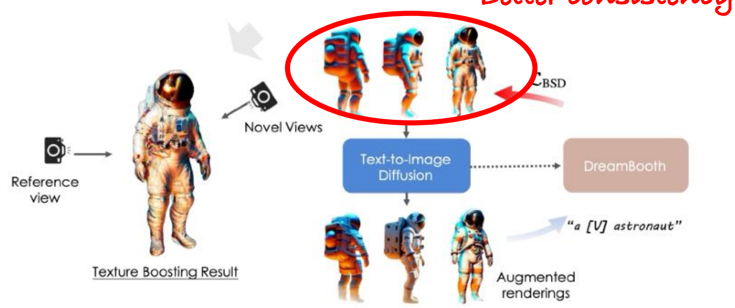
$$\nabla_\theta L_{BSD}(\phi, g(\theta)) = w(t)(\epsilon_{DreamBooth}(z_t|y,t,r_{t'}(z),v) - \epsilon_{\text{LoRA}}(z_t|y,t,z,v))\ \frac{\partial x}{\partial \theta}]$$

$r_{t'}(z)$: "Augmented" image renderings

Restore a heavily noised image

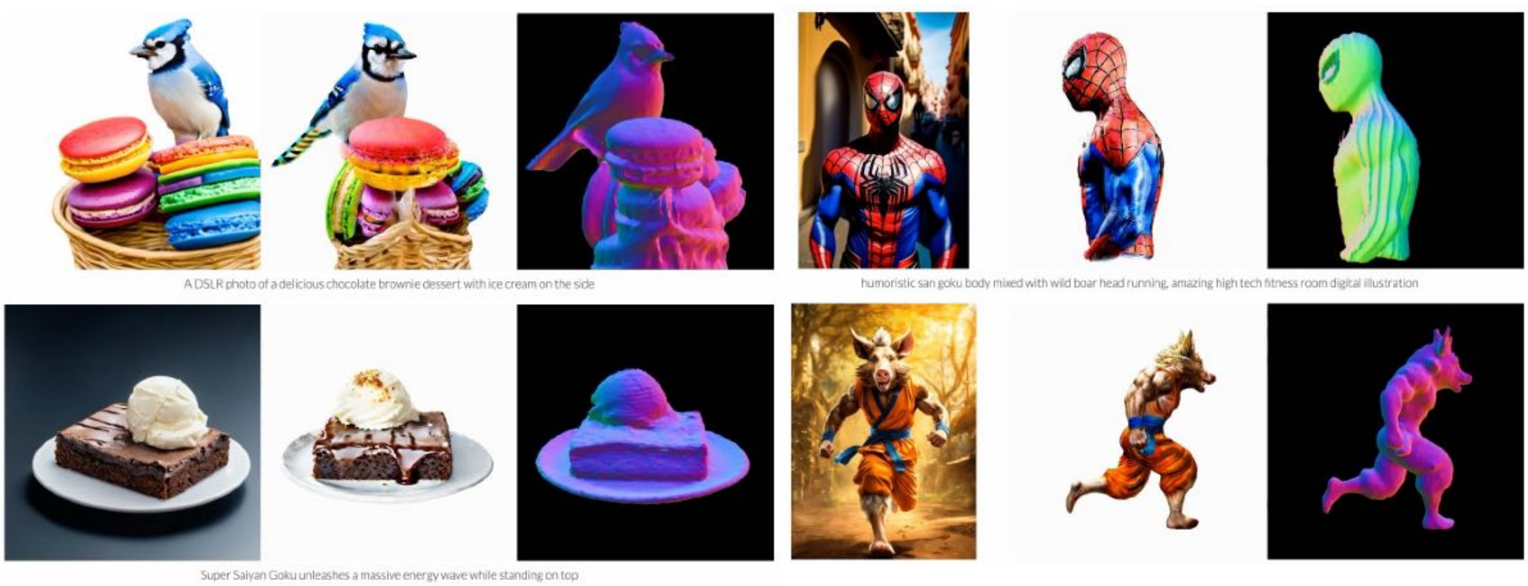
Reveal high-frequency details but sacrifice fidelity

Better quality
Identity preserved
Better consistency



3D Content Generation *Learn from 3D data*

DreamCraft3D: Hierarchical 3D Generation with Bootstrapped Diffusion Prior



3D Content Generation *Learn from 3D data*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

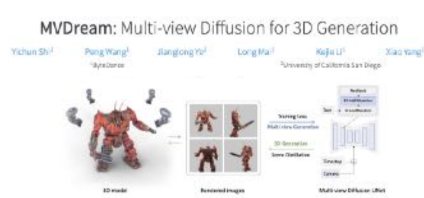
Single viewpoint prediction to multi viewpoint prediction

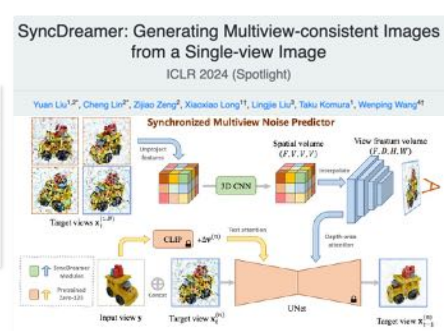
Ref image viewpoint
$v = (R, T)$

z_t, t → **Noise prediction w/ CFG** → $\hat{\epsilon}_\theta(z_t | y, t, v)$

Ref image viewpoint
$v = (R, T)$

$\{z_1 \dots z_t\}$, t → **Noise prediction w/ CFG** → $\hat{\epsilon}_\theta(z_t^1 | y, t, v)$ ⋮ $\hat{\epsilon}_\theta(z_t^N | y, t, v)$



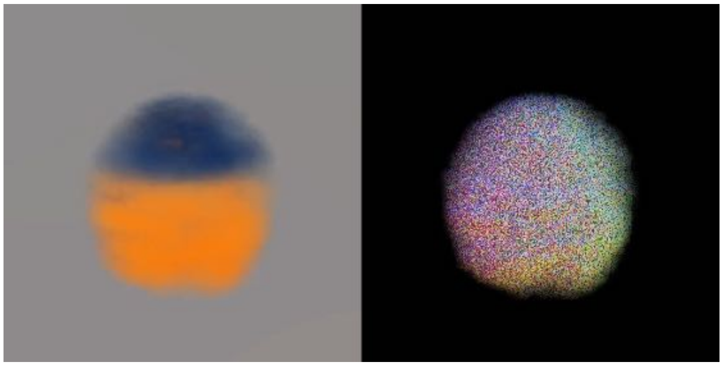


3D Content Generation *Learn from 3D data*

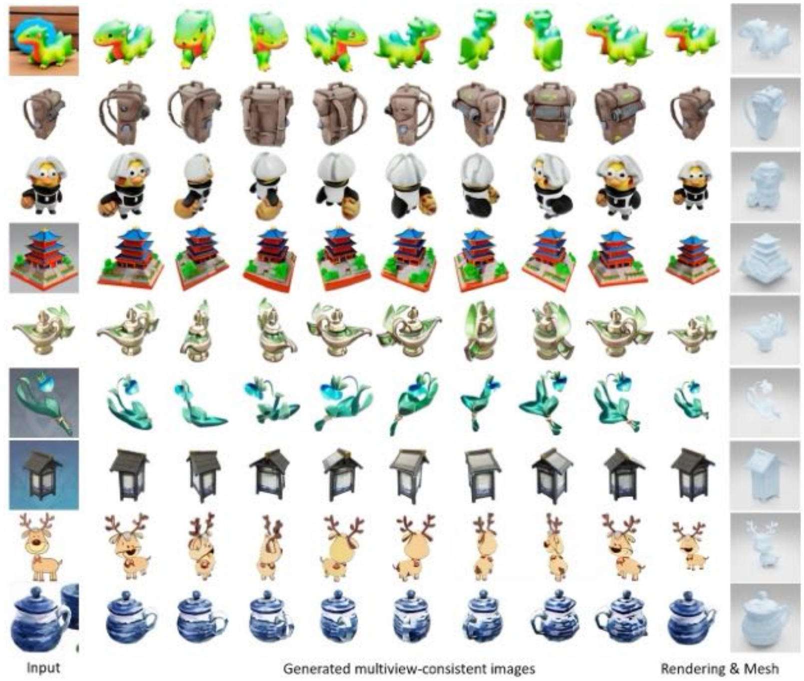
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Single viewpoint prediction to multi viewpoint prediction

MVDream



SyncDreamer



3D Content Generation *Learn from 3D data*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

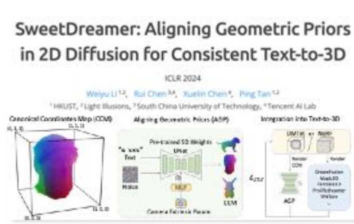
RGB + Geometry

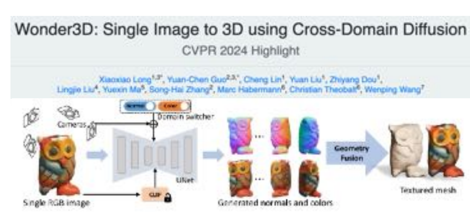
Ref image viewpoint
$v = (R, T)$

z_t, t → Noise prediction w/ CFG → $\hat{\epsilon}_\theta(z_t | y, t, v)$

Ref image viewpoint
$v = (R, T)$

$\{z_t^c, z_t^g\}$, t → Noise prediction w/ CFG → $\hat{\epsilon}_\theta(z_t^c | y, t, v)$ ⋮ $\hat{\epsilon}_\theta(z_t^g | y, t, v)$



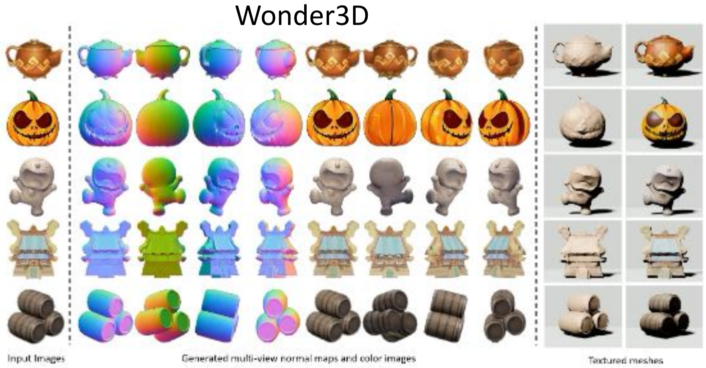


3D Content Generation *Learn from 3D data*

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

RGB + Geometry

Wonder3D



MVDreamer

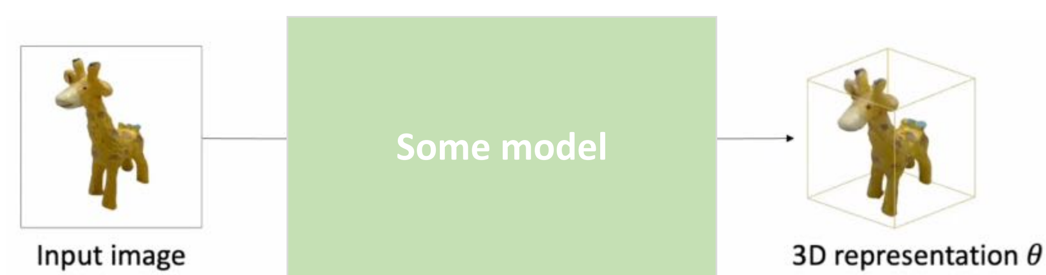
3D Content Generation *Learn from 3D data*

SIGGRAPH 2024
DENVER+ 28 JUL – 1 AUG

More direct approaches?

Optimize 3D representations
1) ref view match the input image
2) Novel views are photorealistic and view-consistent

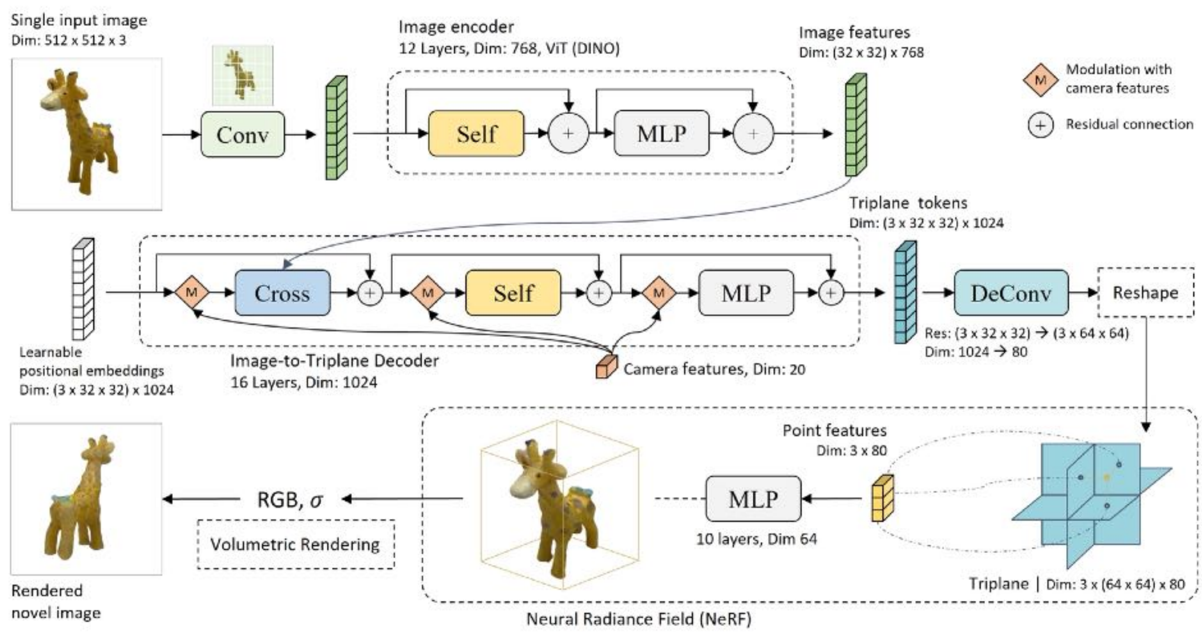
But is time consuming
A direct inference approach?



3D Content Generation *Learn from 3D data*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

LRM: Large Reconstruction Model for Single Image to 3D

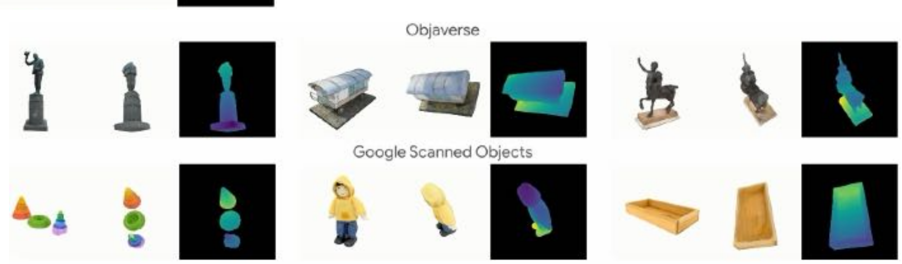


3D Content Generation *Learn from 3D data*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

LRM: Large Reconstruction Model for Single Image to 3D

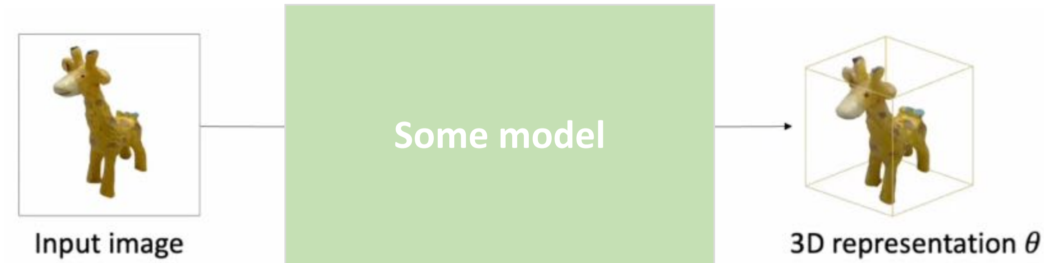
5 seconds image to 3D
A bit blurry 🤔



3D Content Generation *Learn from 3D data*

SIGGRAPH 2024 DENVER+ 28 JUL — 1 AUG

LRM: Large Reconstruction Model for Single Image to 3D

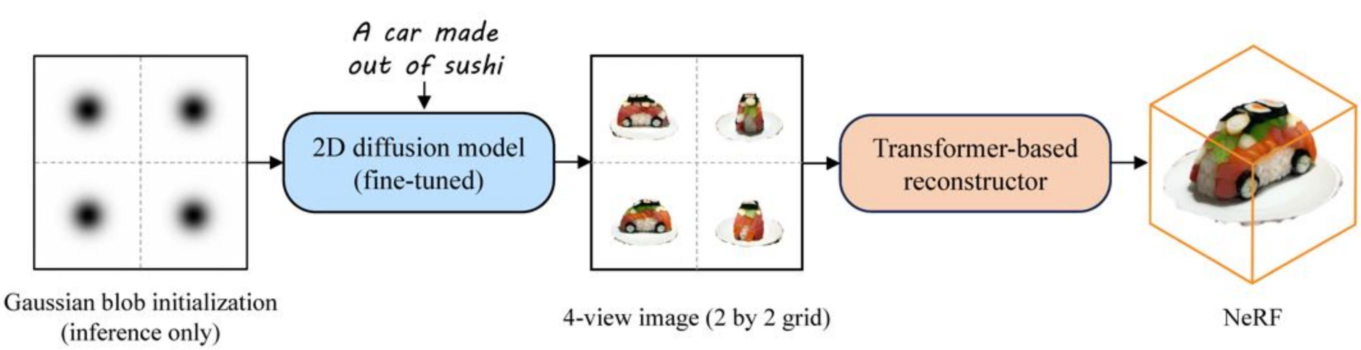


Instant3D: Fast Text-to-3D with Sparse-View Generation and Large Reconstruction Model

3D Content Generation *Learn from 3D data*

SIGGRAPH 2024 DENVER+ 28 JUL — 1 AUG

Instant3D: Fast Text-to-3D with Sparse-View Generation and Large Reconstruction Model

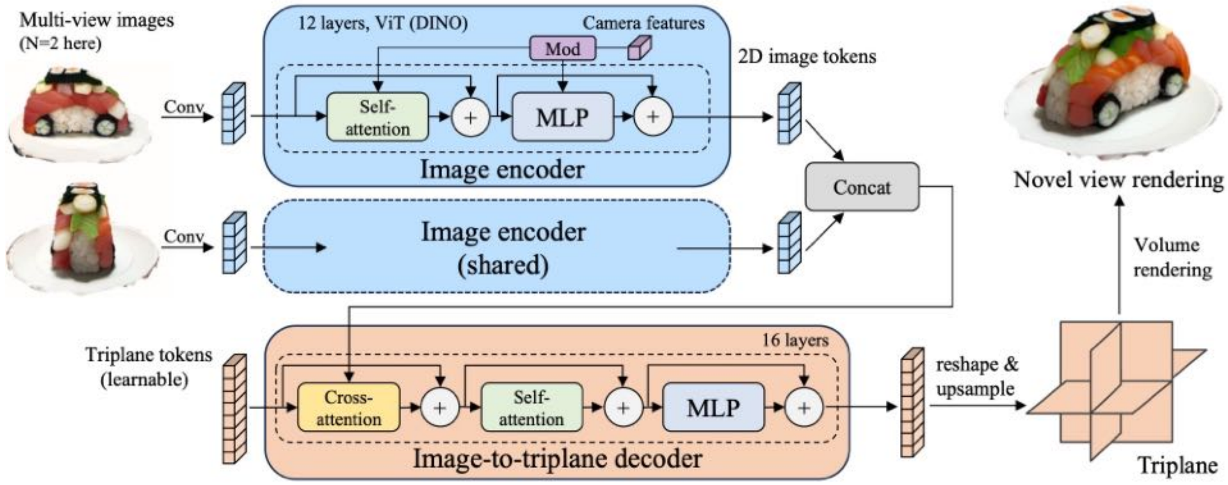


Idea: multi-view 2D diffusion + sparse view reconstruction

3D Content Generation *Learn from 3D data*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

INSTANT3D: FAST TEXT-TO-3D WITH SPARSE-VIEW GENERATION AND LARGE RECONSTRUCTION MODEL



3D Content Generation *Learn from 3D data*

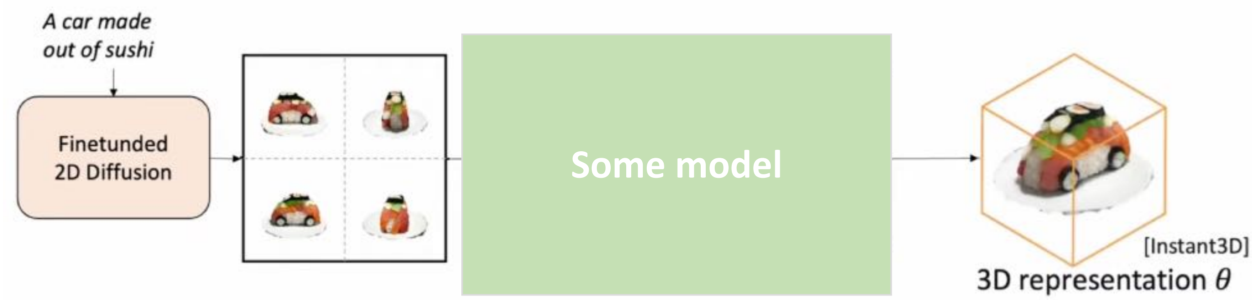
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

INSTANT3D: FAST TEXT-TO-3D WITH SPARSE-VIEW GENERATION AND LARGE RECONSTRUCTION MODEL

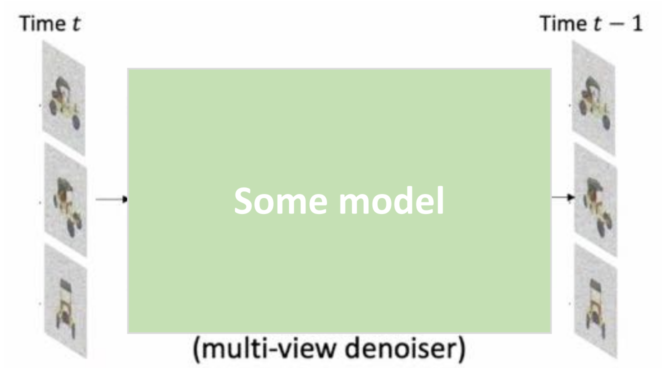
3D Content Generation *Learn from 3D data*

SIGGRAPH 2024 DENVER+ 28 JUL – 1 AUG

Instant3D: Fast Text-to-3D with Sparse-View Generation and Large Reconstruction Model



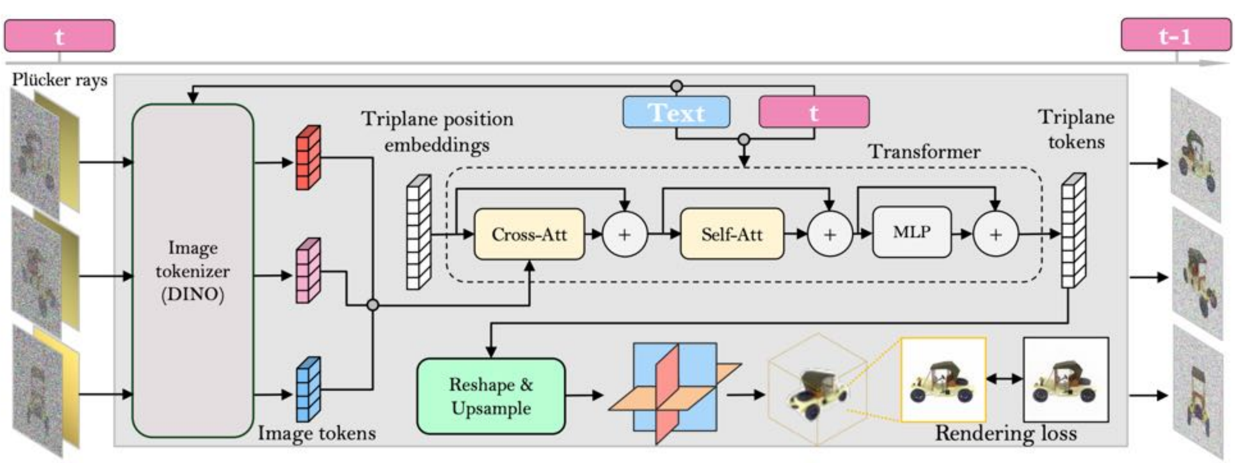
DMV3D: Denoising Multi-View Diffusion using 3D Large Reconstruction Model



3D Content Generation *Learn from 3D data*

SIGGRAPH 2024 DENVER+ 28 JUL – 1 AUG

DMV3D: Denoising Multi-View Diffusion using 3D Large Reconstruction Model



Input: noisy images

Predict: clean triplane

Add slighter noise on rendered images

3D Content Generation *Learn from 3D data*

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

DMV3D: Denoising Multi-View Diffusion using 3D Large Reconstruction Model

Conclusion

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

2D priors with Score Distillation Sampling
- Higher resolution
- Richer appearance
- Single-view to 3D
- Photorealistic appearance

3D priors
- View-conditioned diffusion
- Multi-view diffusion
- View-conditioned gemetry + appearance diffusion

Feed-forward models (empowered by data + transformer)
- Single-image to 3D
- Multi-view to 3D
- Multi-view diffusion

Sometimes we want to manipulate existing scenes…

Scene Manipulation/Editing/Generation

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Manipulate/Edit

SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

Interpretable?

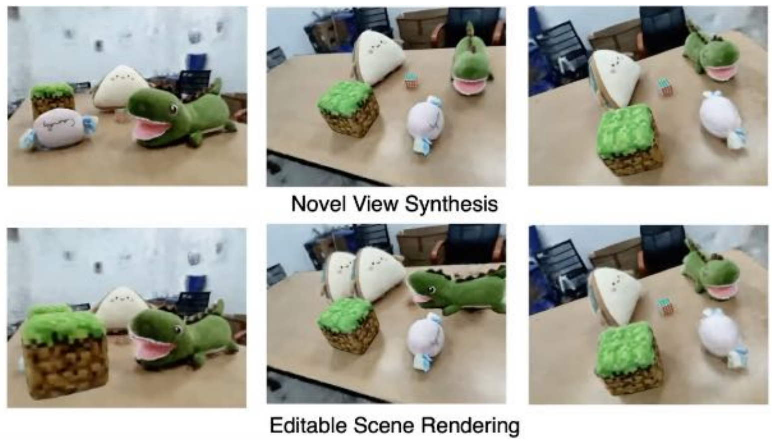
Kobayashi, S., Matsumoto, E., & Sitzmann, V. (2022). Decomposing NeRF for Editing via Feature Field Distillation.
Cen, J., Zhou, Z., Fang, J., Shen, W., Xie, L., Jiang, D., Zhang, X., & Tian, Q. (2023). Segment Anything in 3D with NeRFs.

Manipulate/Edit

Interpretable?

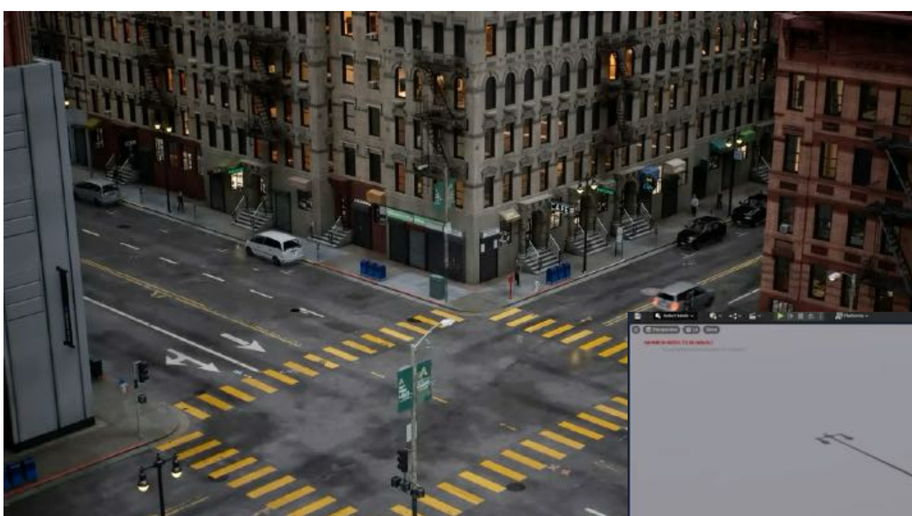
Flexible?

Interactive editing NSVF



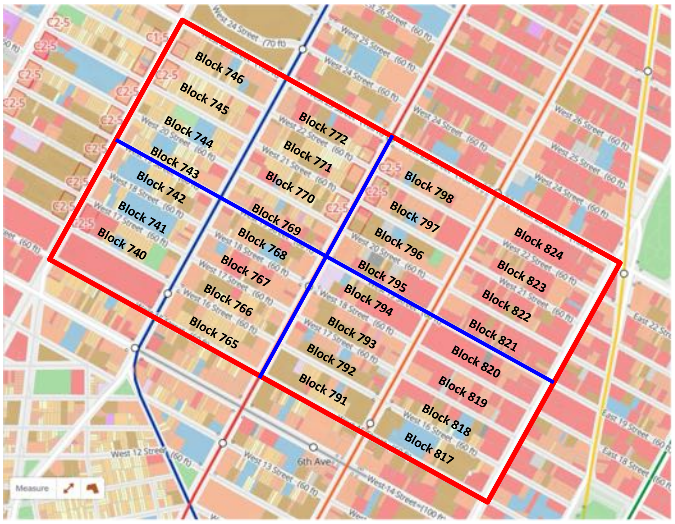
Liu, L., Gu, J., Lin, K.Z., Chua, T., & Theobalt, C. (2020). Neural Sparse Voxel Fields
Yang, B., Zhang, Y., Xu, Y., Li, Y., Zhou, H., Bao, H., Zhang, G., & Cui, Z. (2021). Learning Object-Compositional Neural Radiance Field for Editable Scene Rendering.

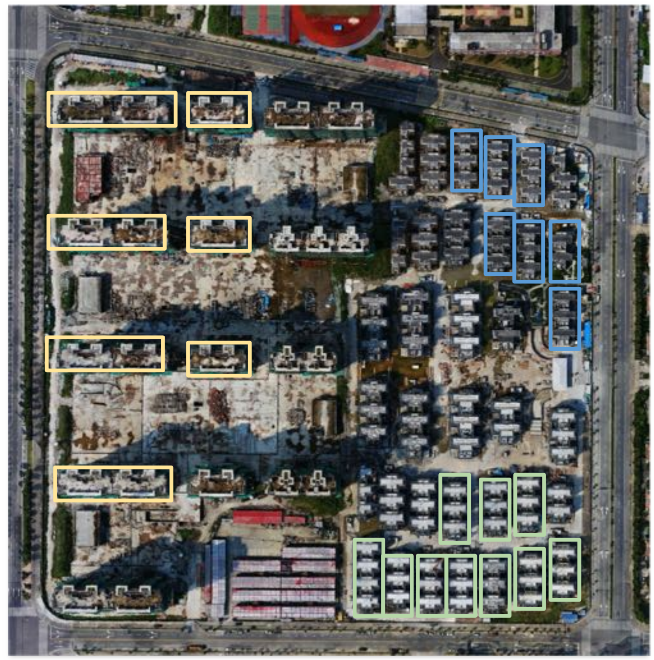
Scalable?



Urban Fabric

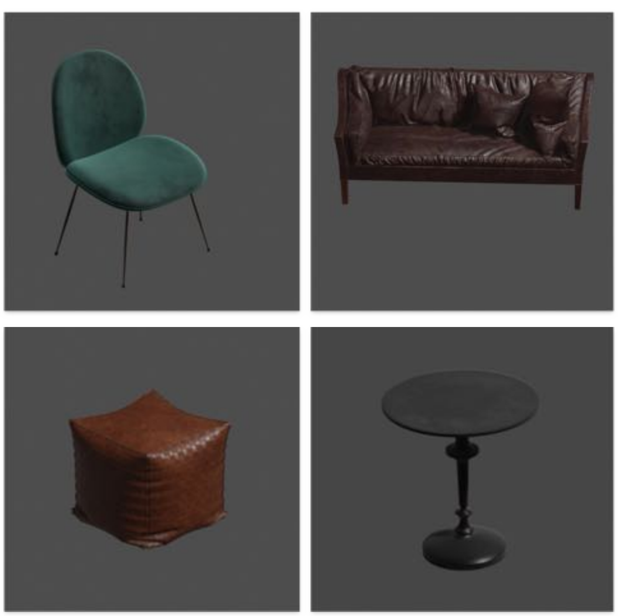
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

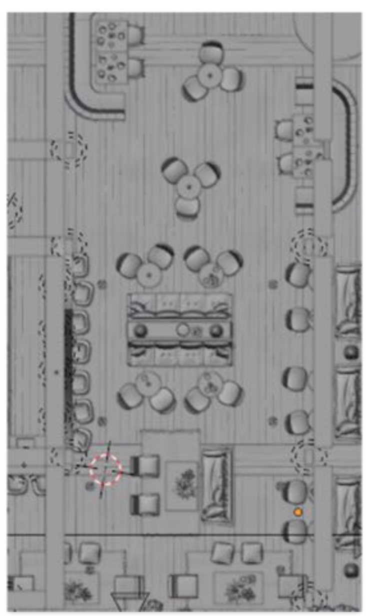




Interior Design

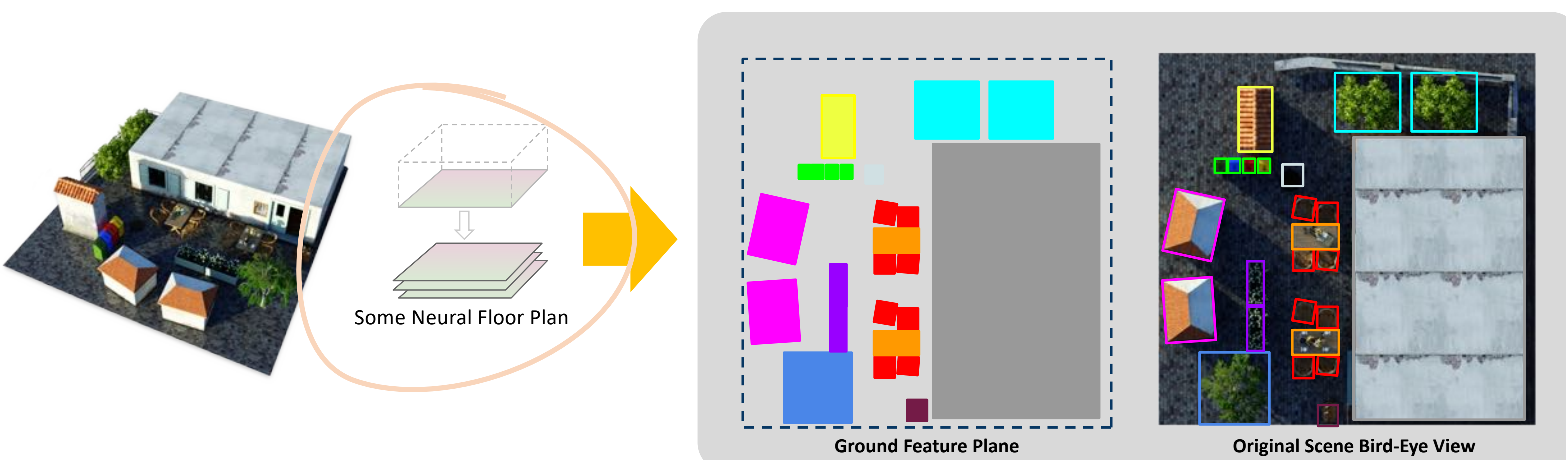
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG





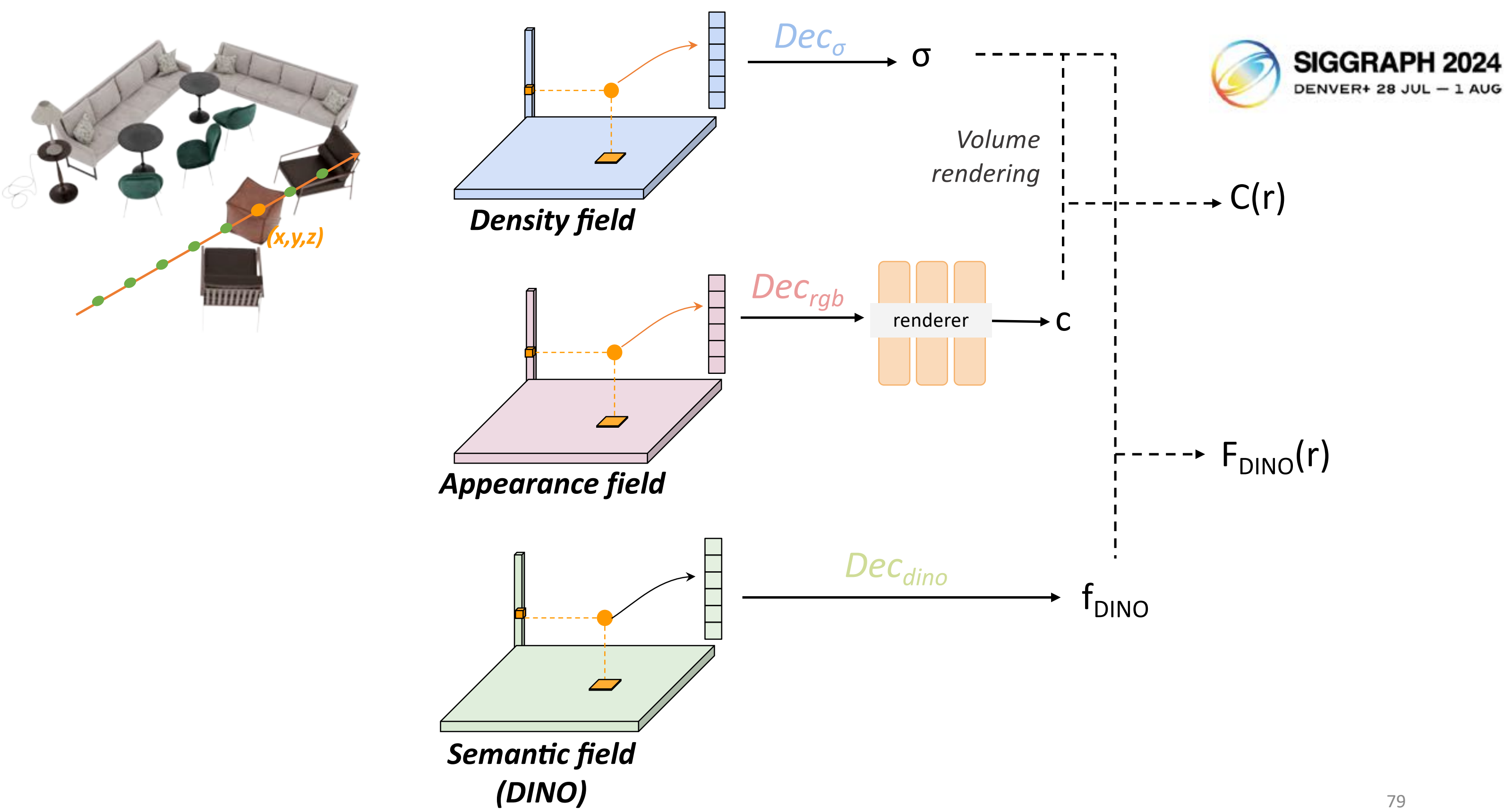
Extract Assets and Layout

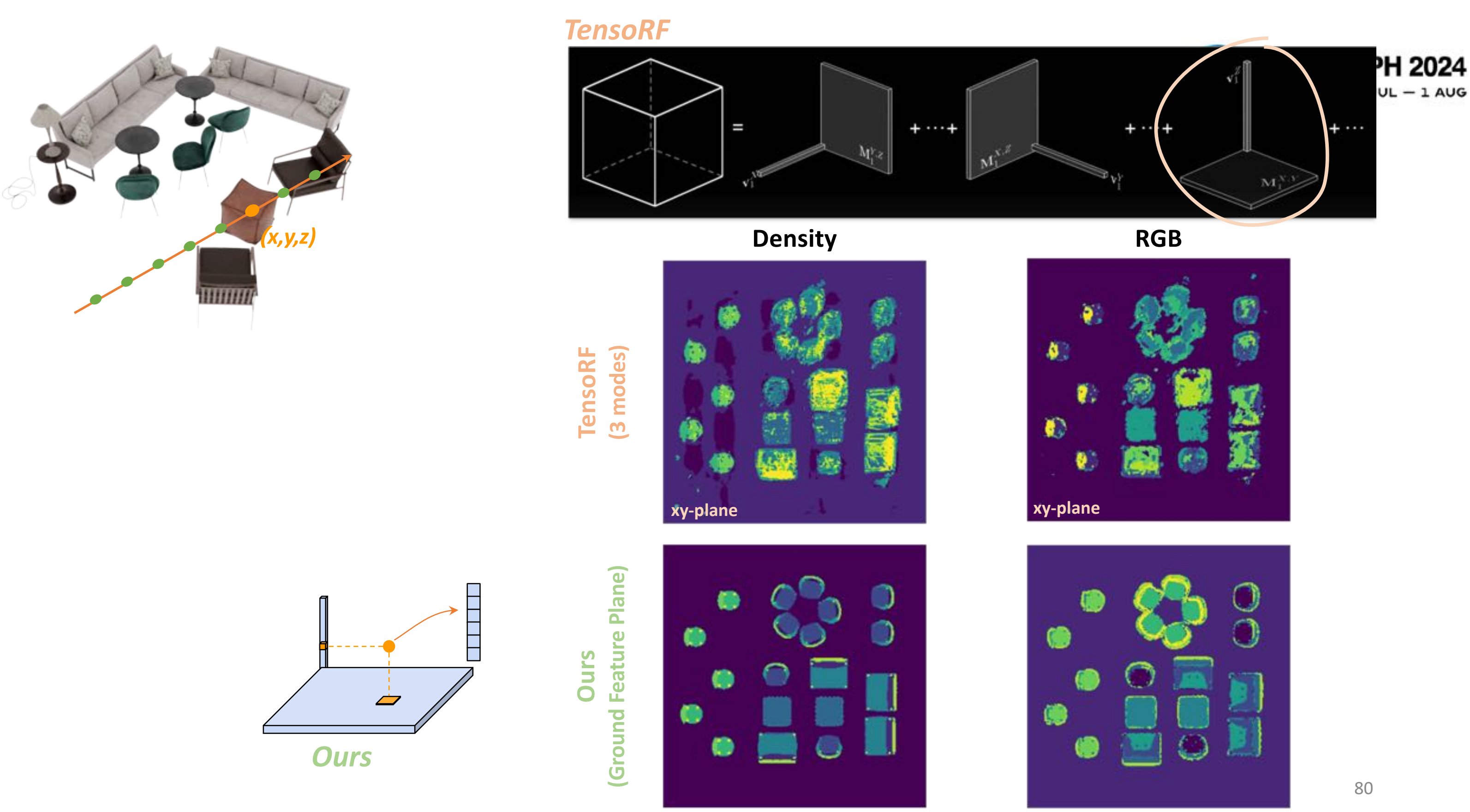
Can we find and categorize objects on this floor plan?

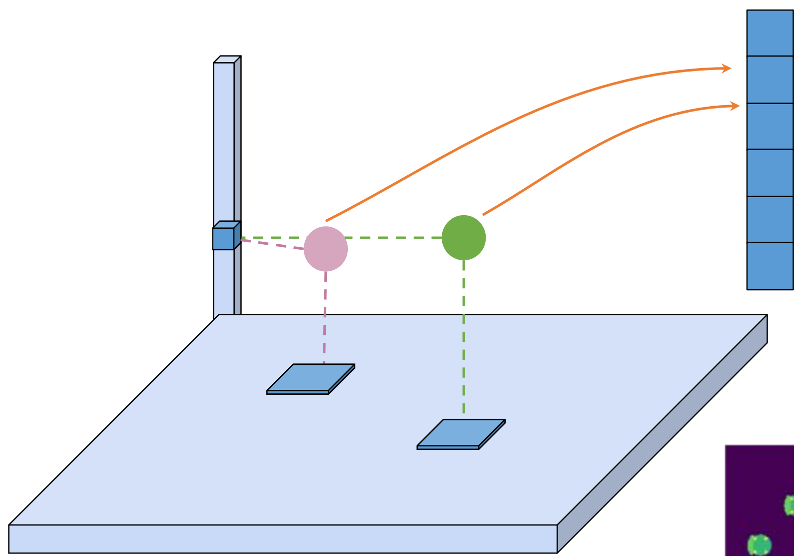


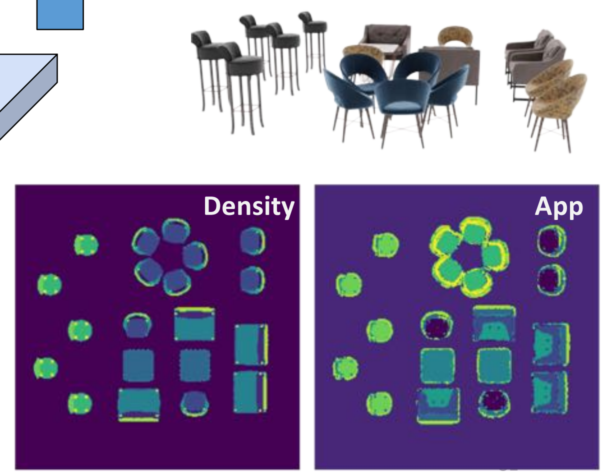
Clean *Sharp* *Object clues*

Xiangli, Y., Xu, L., Pan, X., Zhao, N., Dai, B., & Lin, D. (2023). AssetField: Assets Mining and Reconfiguration in Ground Feature Plane Representation.

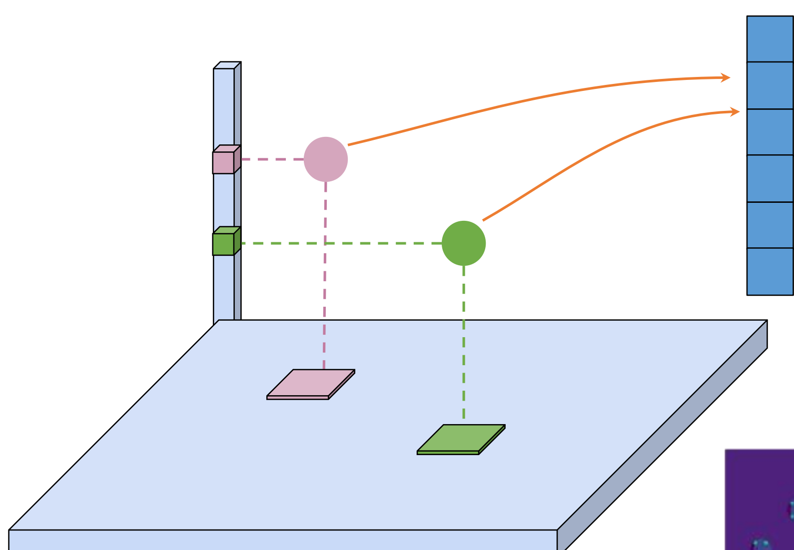

(x,y,z)
Density field
Appearance field
Semantic field (DINO)
Dec_σ
σ
Dec_rgb
renderer
c
Dec_dino
f_DINO
Volume rendering
C(r)
F_DINO(r)
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

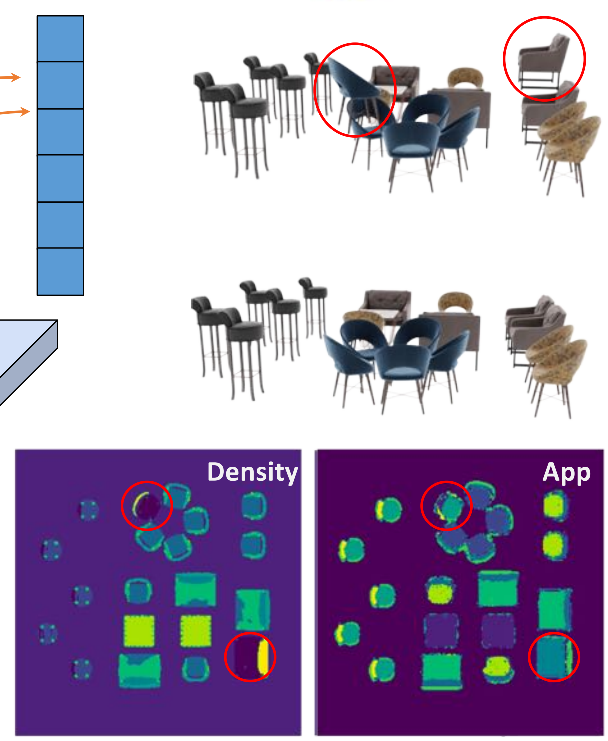

(x,y,z)
Ours
TensoRF
Density
RGB
TensoRF (3 modes)
xy-plane
xy-plane
Ours (Ground Feature Plane)



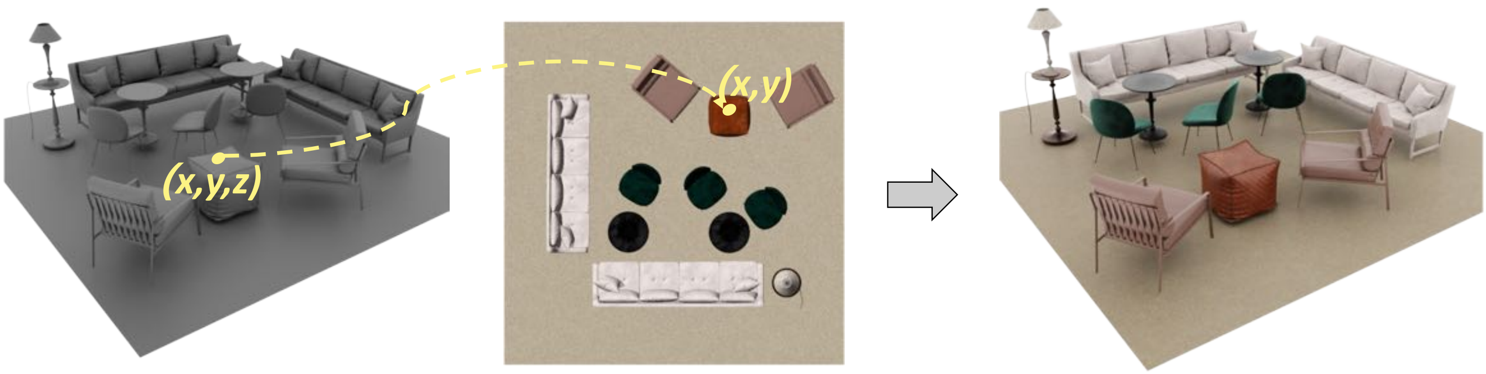


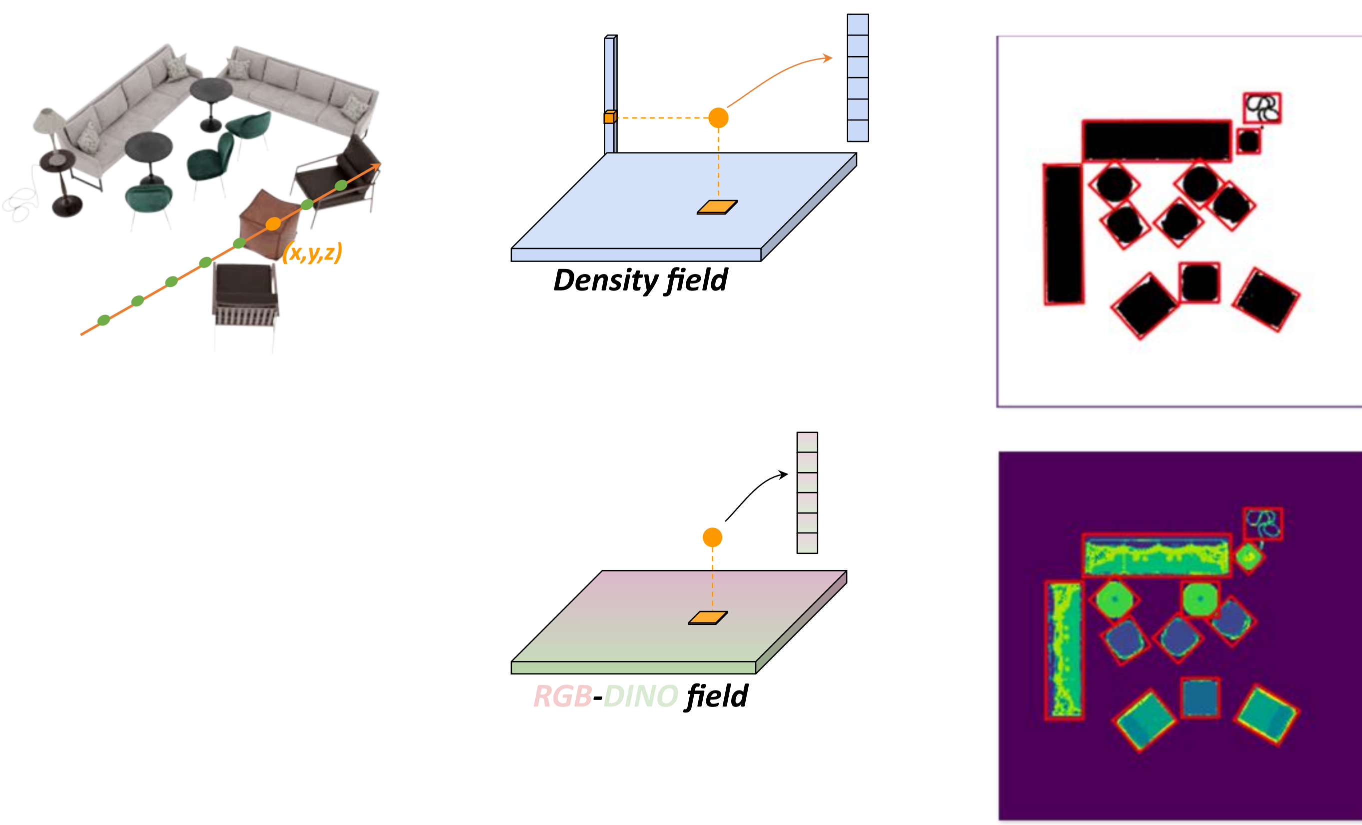
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG



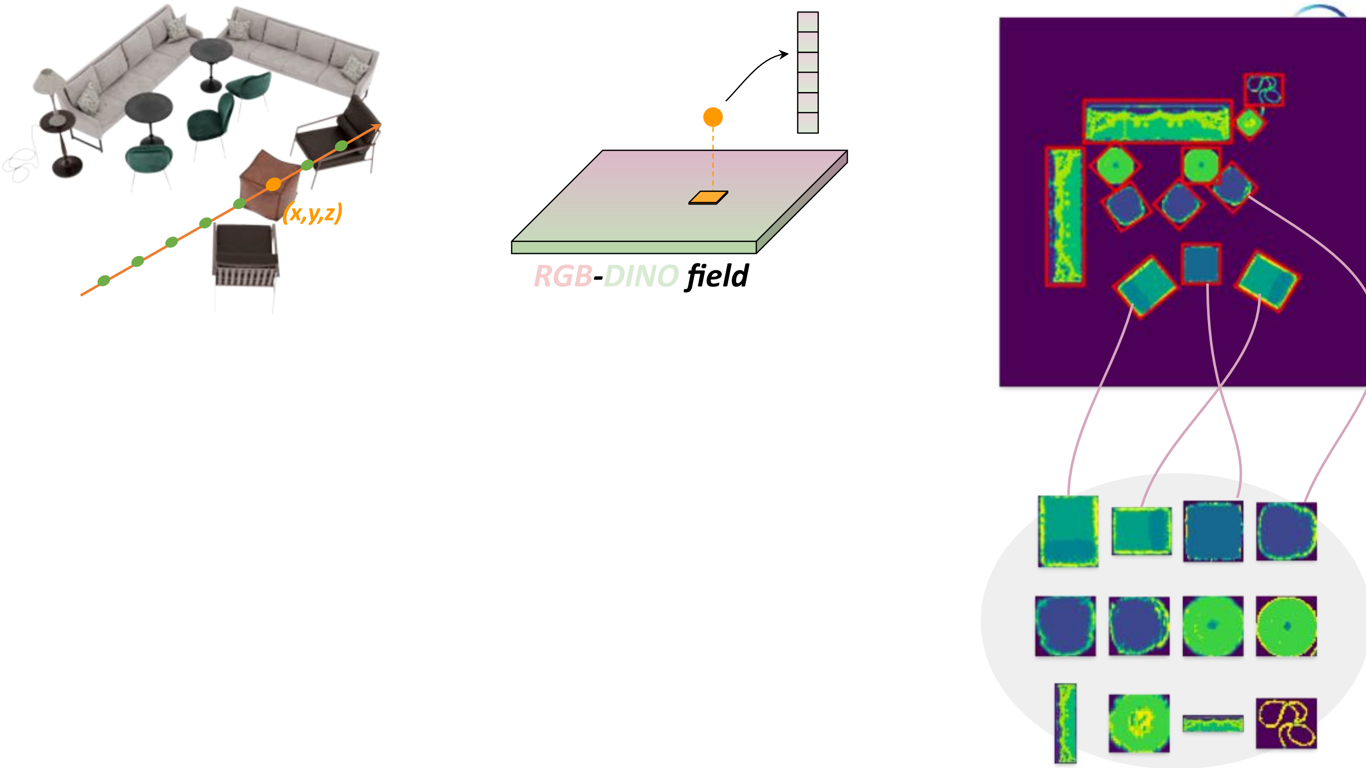


3D density feature **2D** RGB-DINO plane feature **3D** RGB-DINO feature

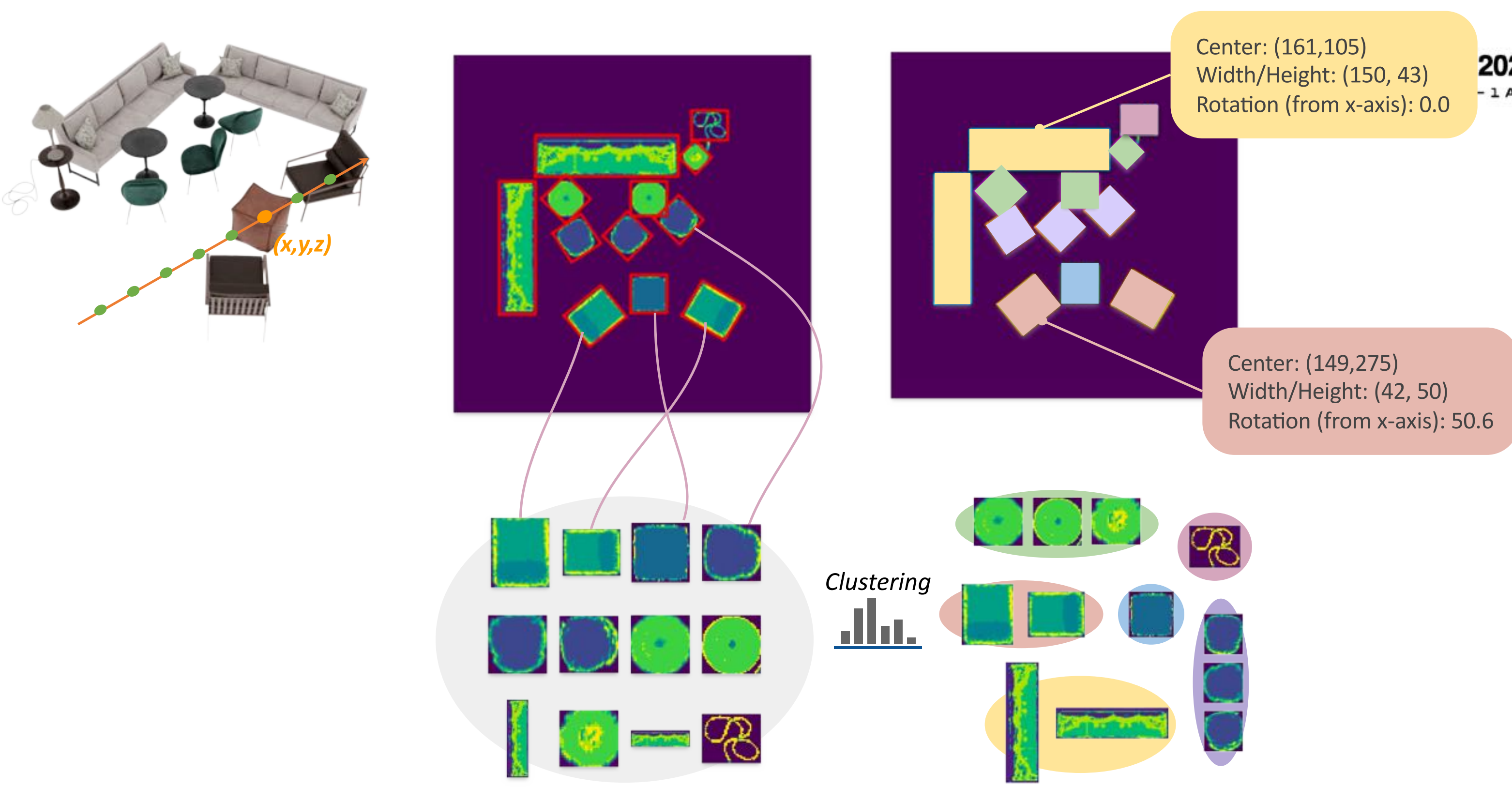


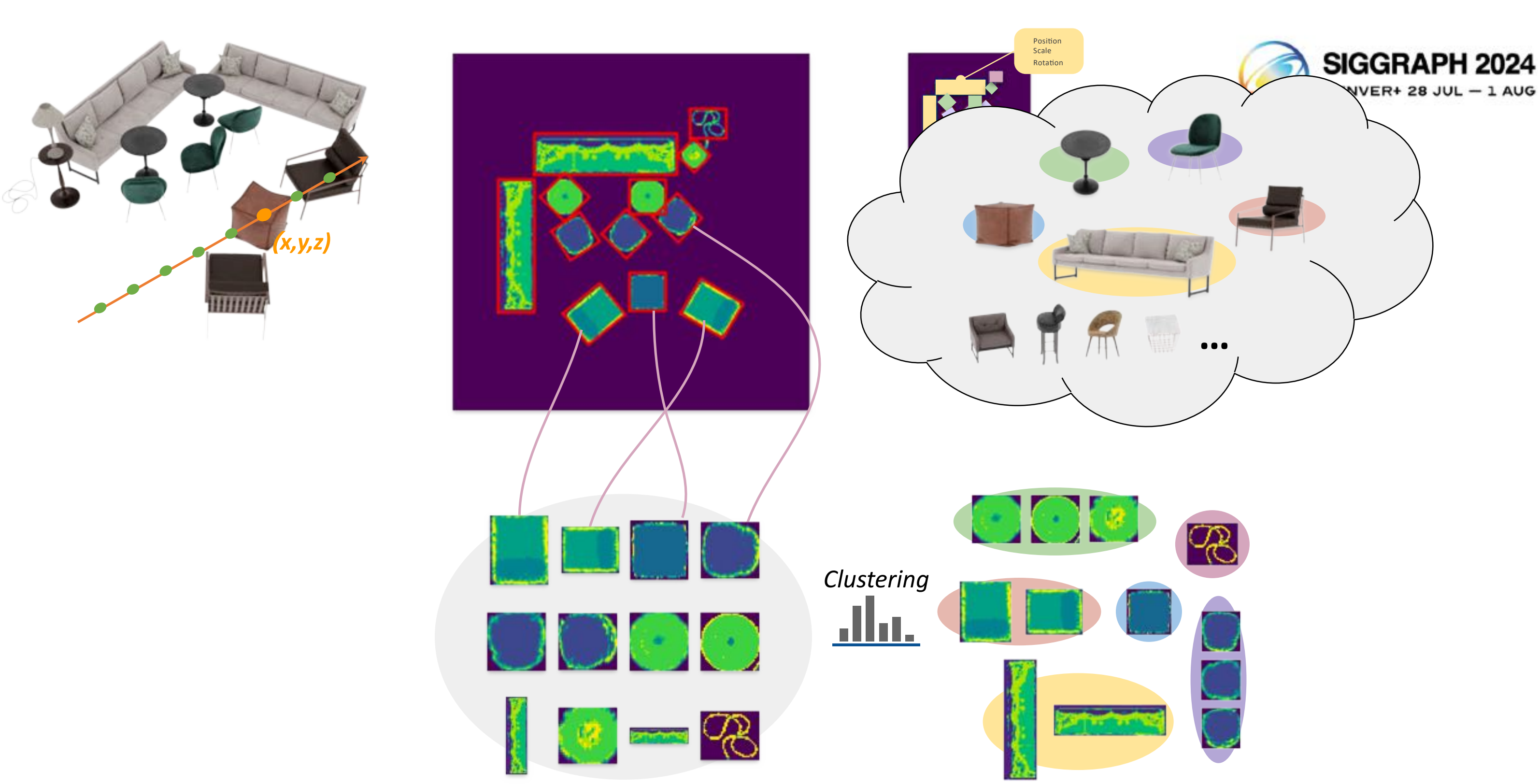


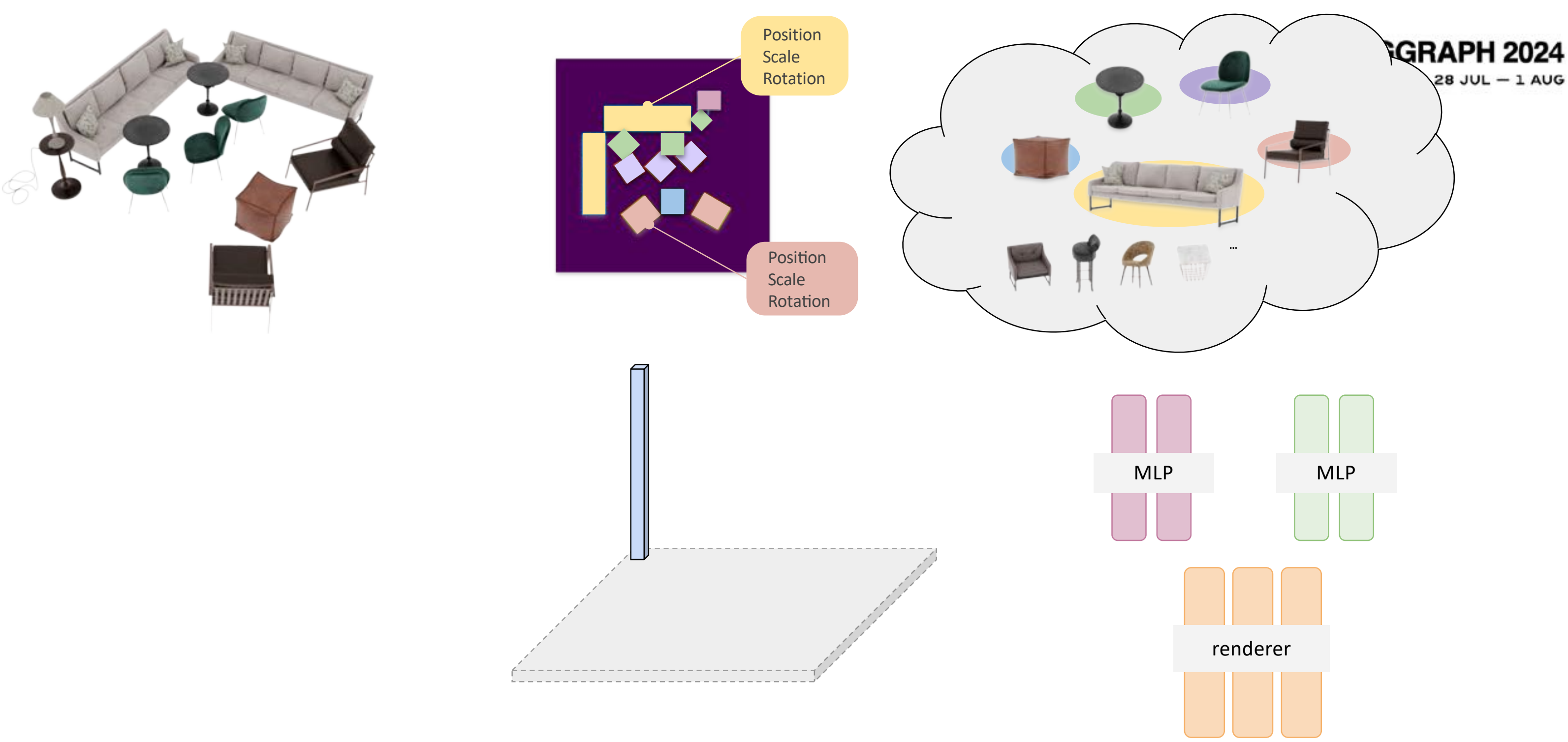
SIGGRAPH 2024
DENVER+ 28 JUL — 1 AUG

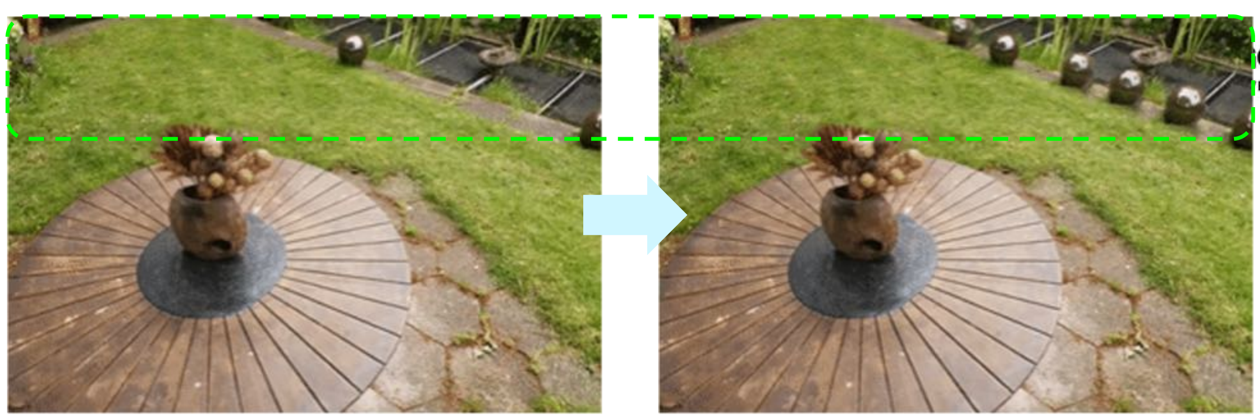


2024
– 1 AUG

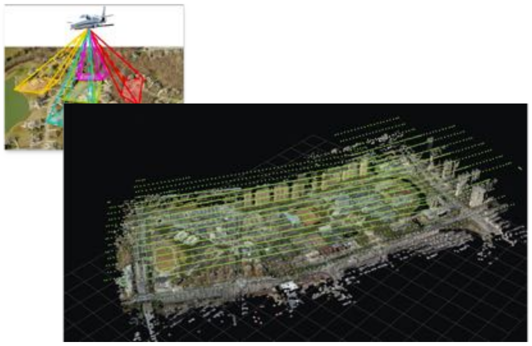


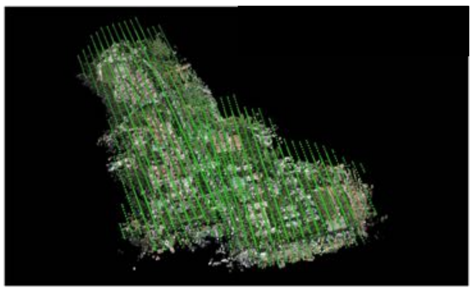

SIGGRAPH 2024
(x,y,z)
Position
Scale
Rotation
Clustering
...


Position
Scale
Rotation
Position
Scale
Rotation
MLP
MLP
renderer

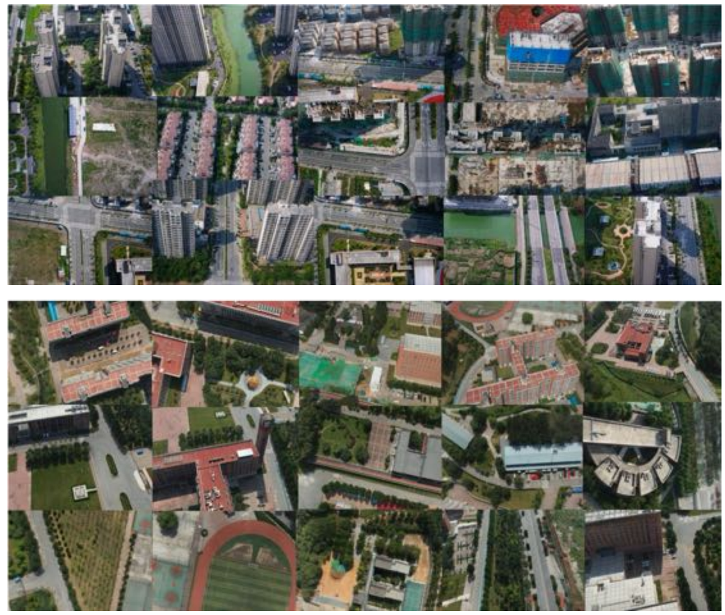


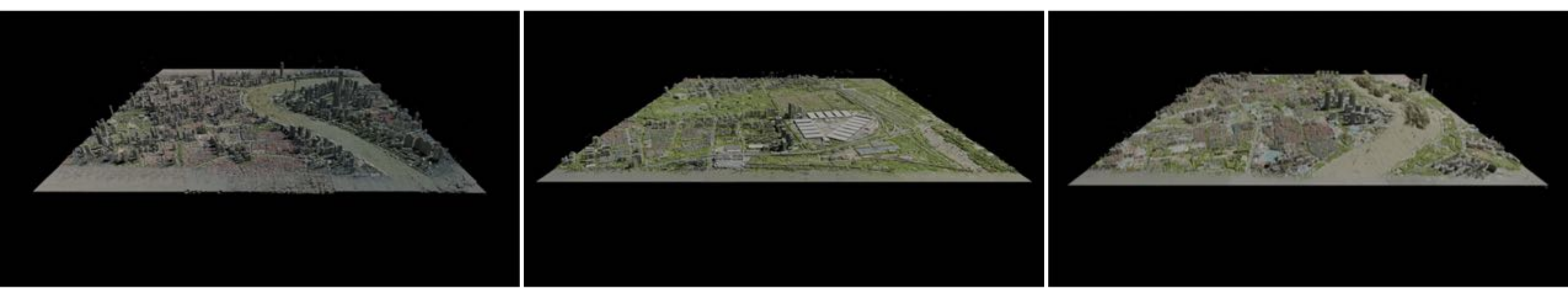
Oblique Photography



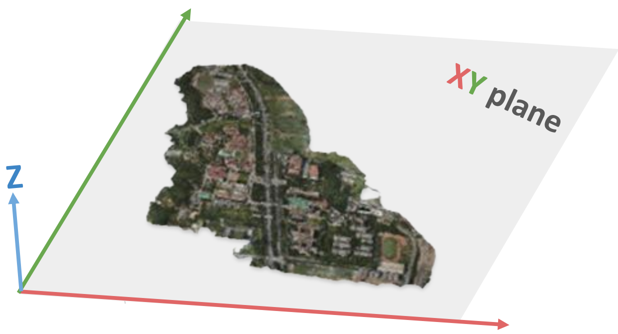


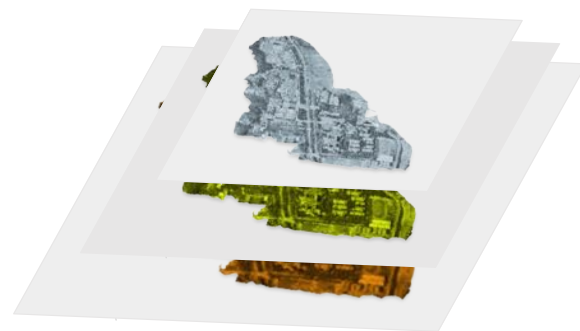
1km^2~10k images
1 image~50 megapixels

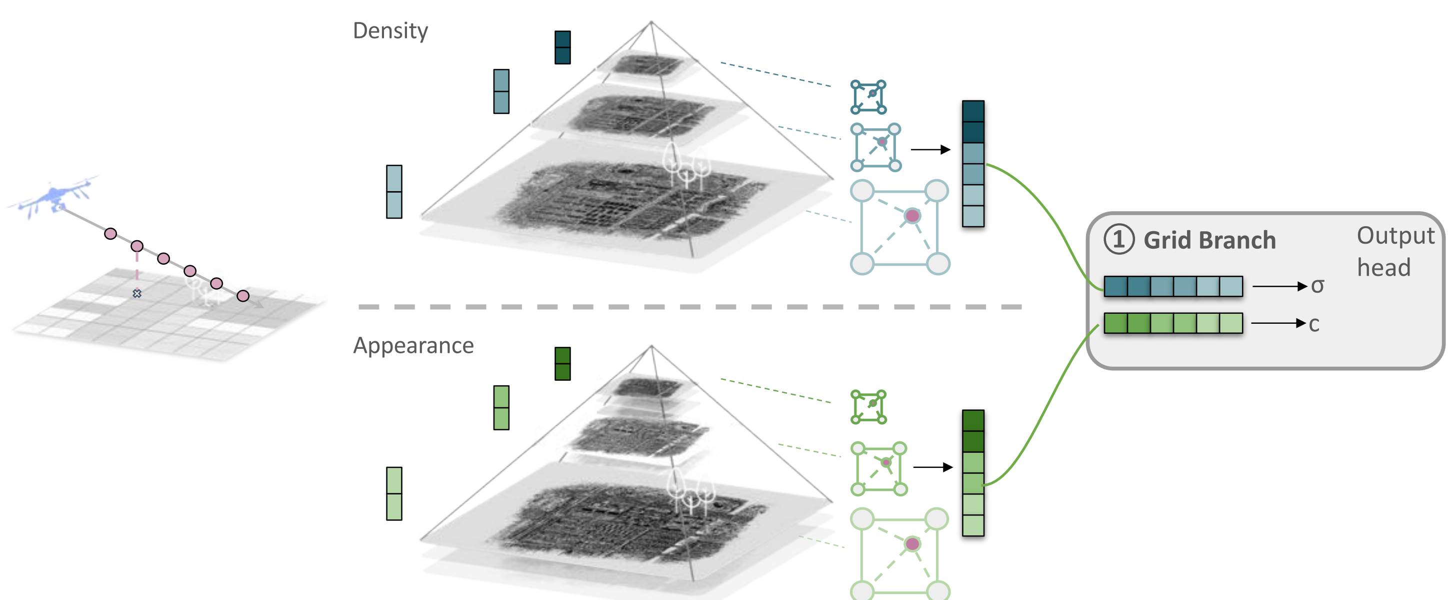


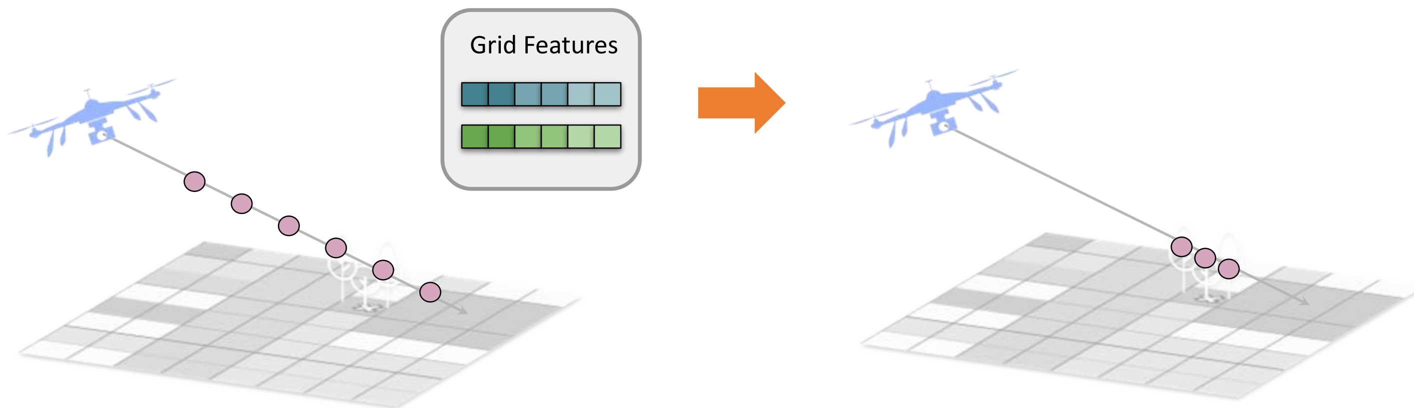


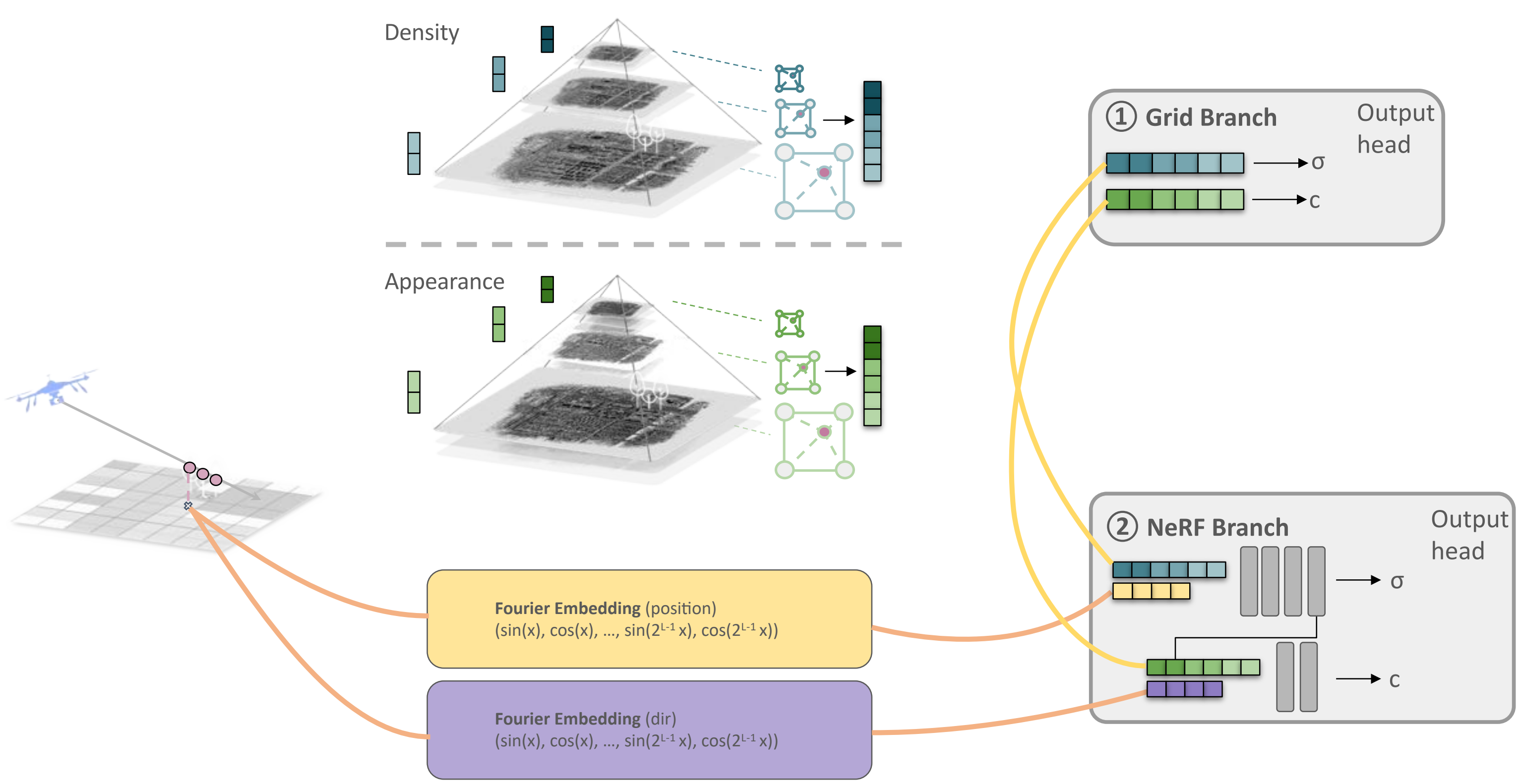
Plane-Axis Factorization

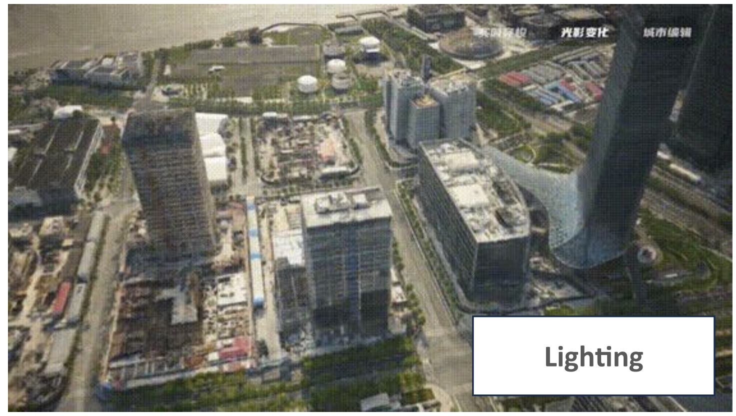


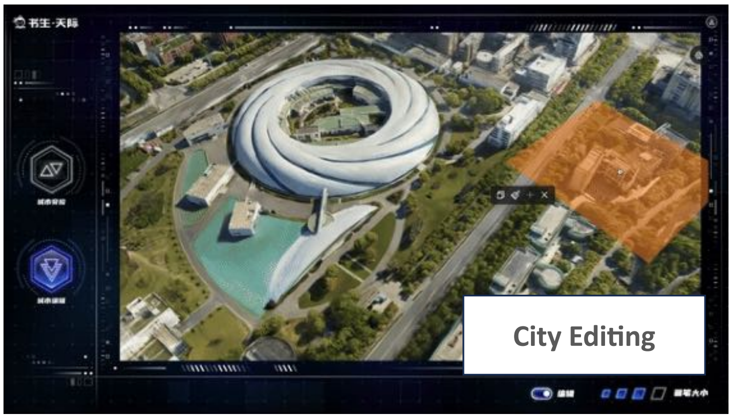





Grid Features


Density
Appearance
1 Grid Branch
Output head
σ
c
2 NeRF Branch
Output head
σ
c
Fourier Embedding (position)
(sin(x), cos(x), ..., sin(2^{L-1} x), cos(2^{L-1} x))
Fourier Embedding (dir)
(sin(x), cos(x), ..., sin(2^{L-1} x), cos(2^{L-1} x))





Creative Scene Editing