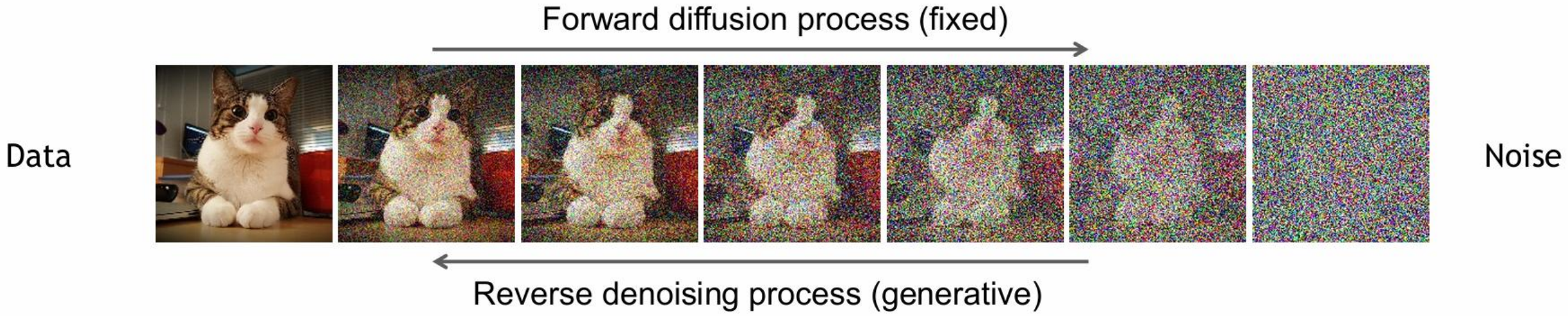
Unit 9
Generative AI for Multimedia

賴尚宏 教授
清華大學資工系

Denoising Diffusion Models

Denoising diffusion models consist of two processes:

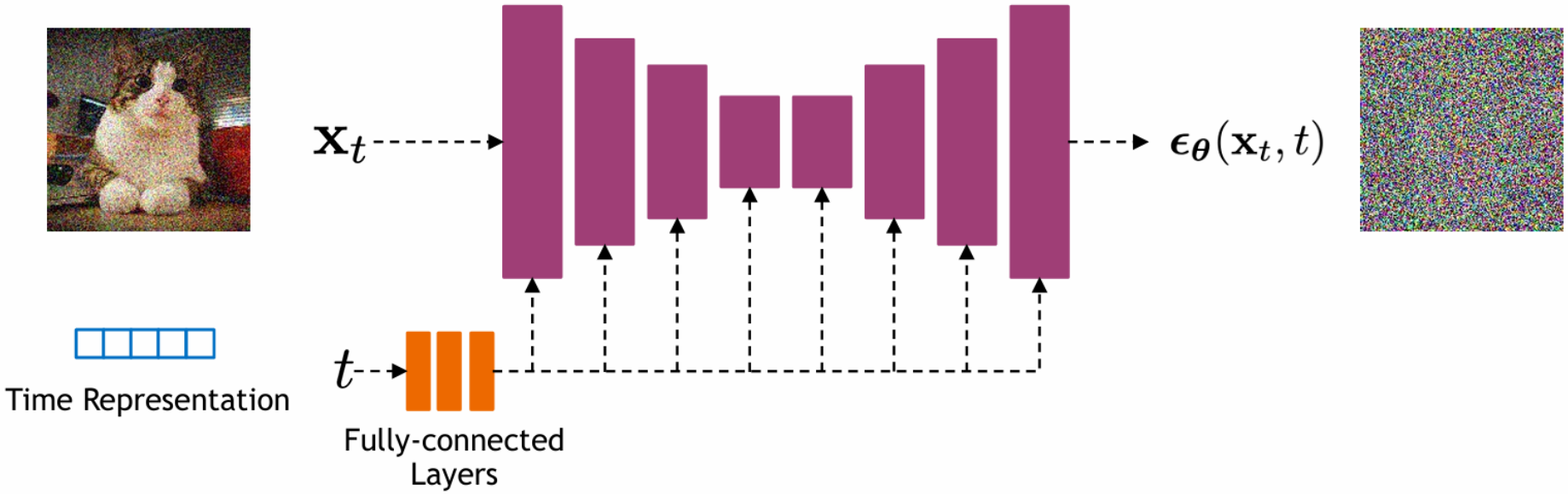
- Forward diffusion process that gradually adds noise to input
- Reverse denoising process that learns to generate data by denoising



Sohl-Dickstein et al., Deep Unsupervised Learning using Nonequilibrium Thermodynamics, ICML 2015
Ho et al., Denoising Diffusion Probabilistic Models, NeurIPS 2020
Song et al., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021

Network Architectures

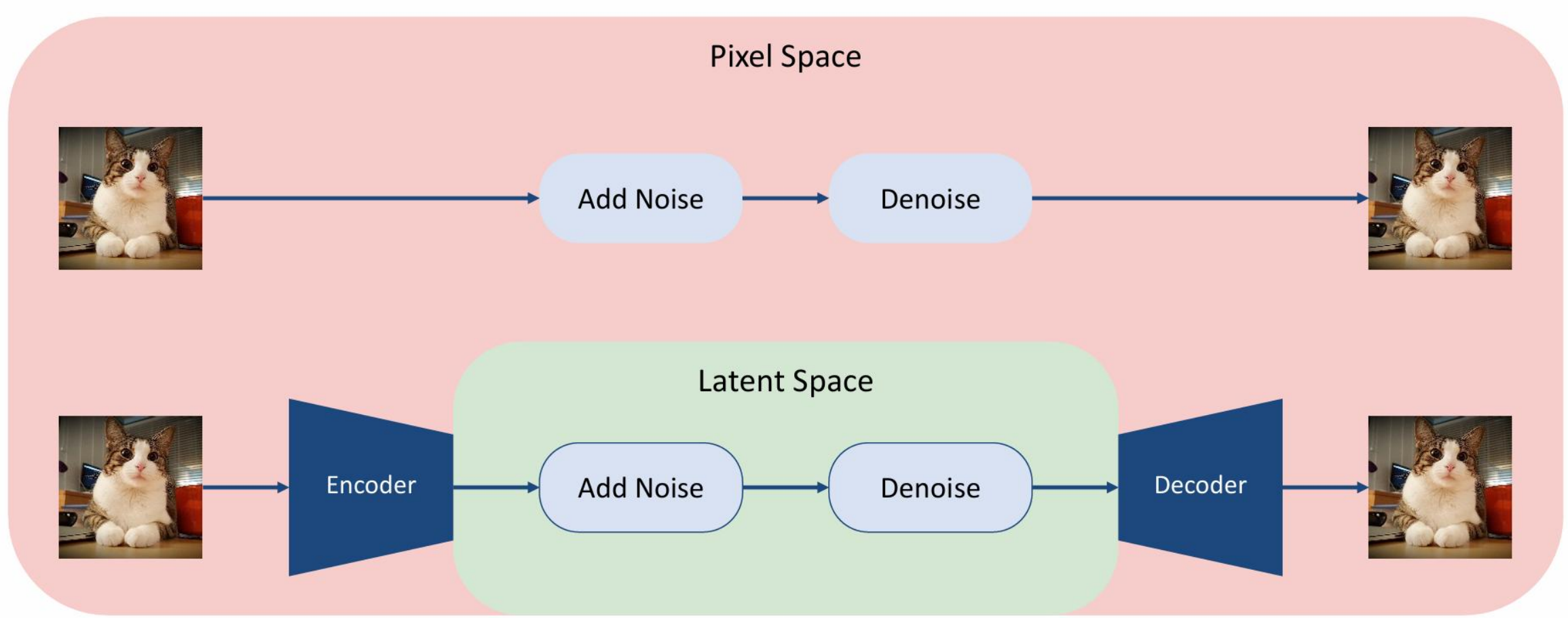
Diffusion models often use U-Net architectures with ResNet blocks and self-attention layers to represent $\epsilon_\theta(\mathbf{x}_t, t)$



Time representation: sinusoidal positional embeddings or random Fourier features.

Time features are fed to the residual blocks using either simple spatial addition or using adaptive group normalization layers. (see Dharivwal and Nichol NeurIPS 2021)

Latent Diffusion



Conditional Diffusion Models

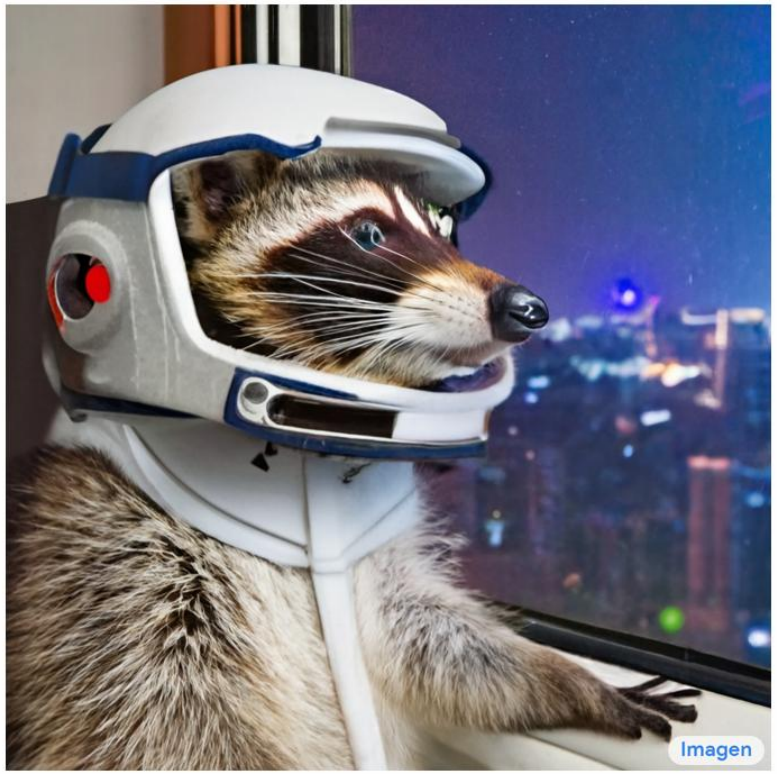
Text-to-image generation

DALL·E 2

"a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese"

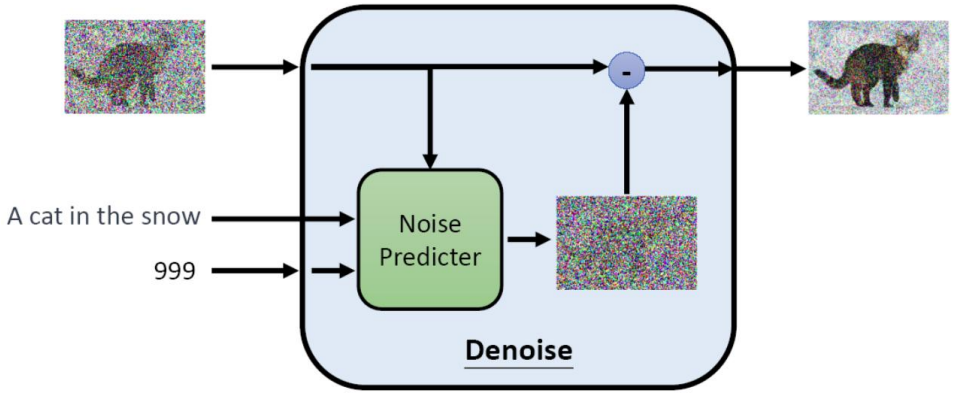
IMAGEN

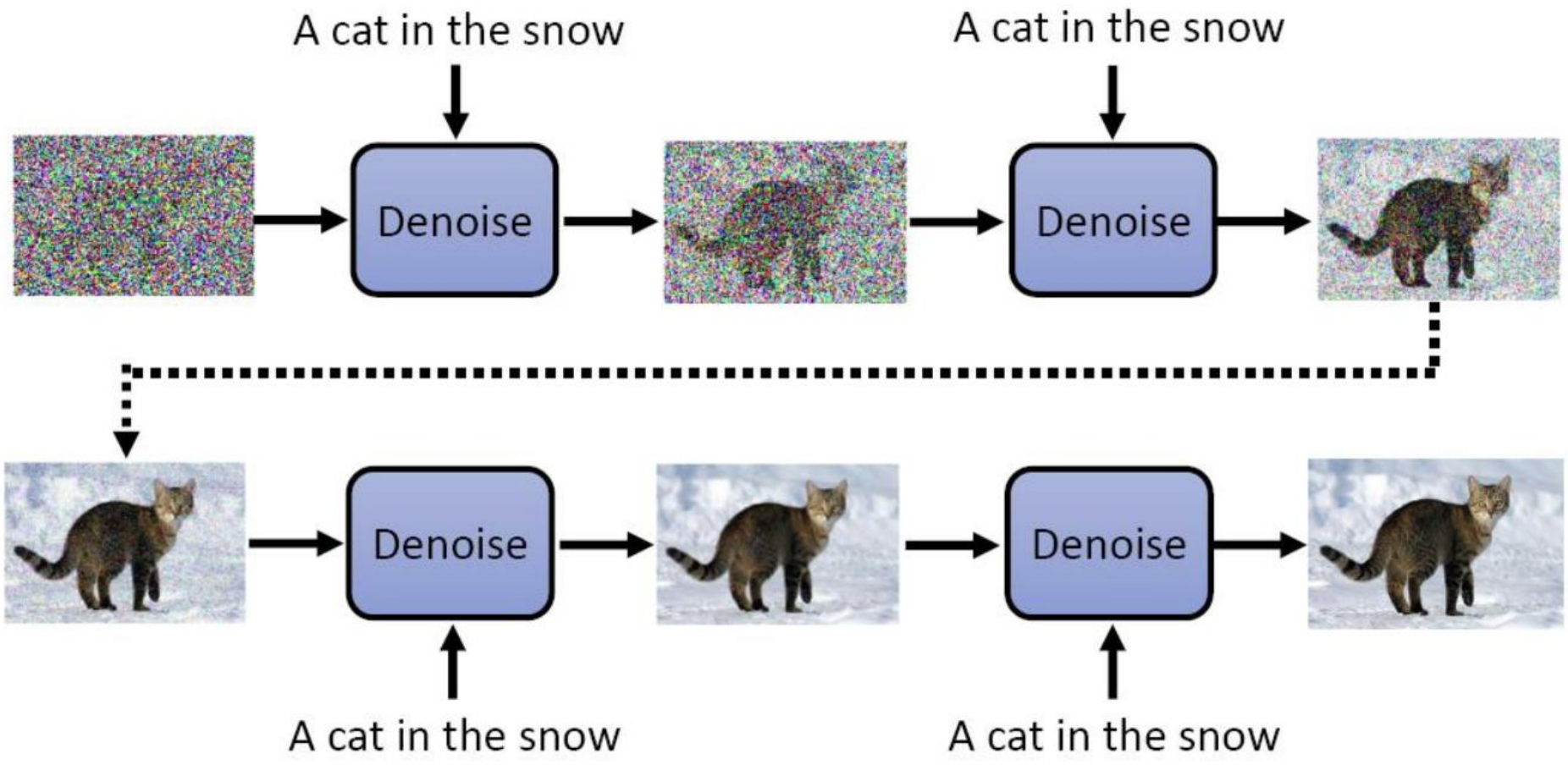
"A photo of a raccoon wearing an astronaut helmet, looking out of the window at night."



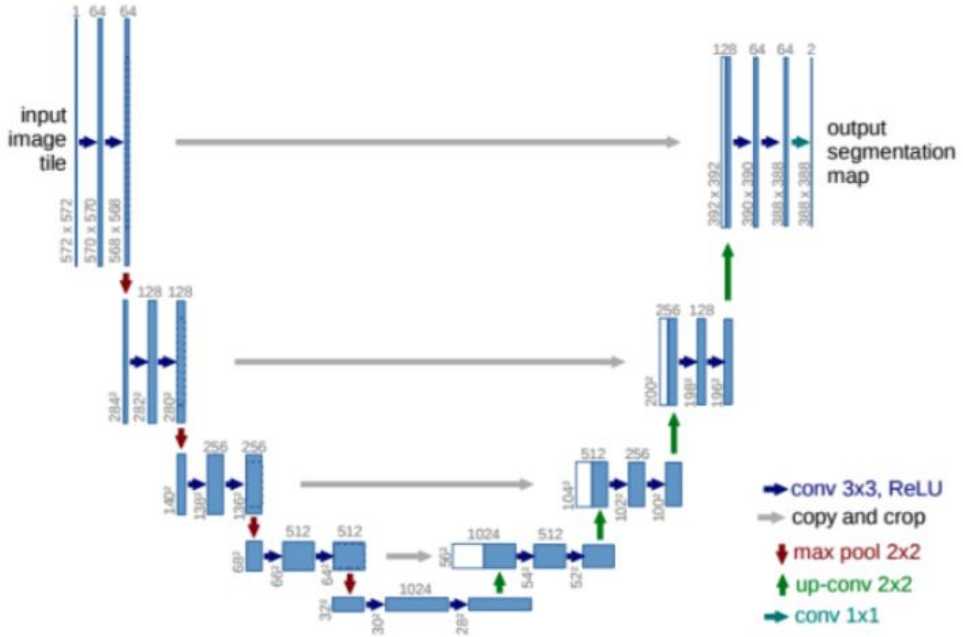
Ramesh et al., "Hierarchical Text-Conditional Image Generation with CLIP Latents", arXiv 2022.
Saharia et al., "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding", arXiv 2022.

Text-to-Image Generation

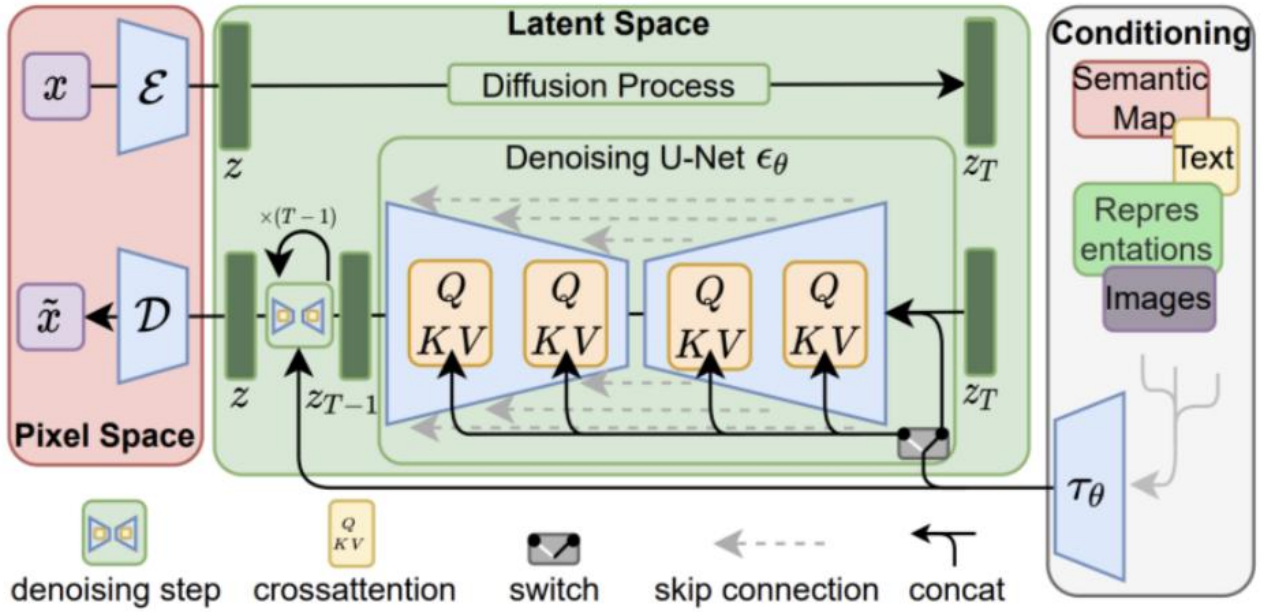




U-Net Architecture



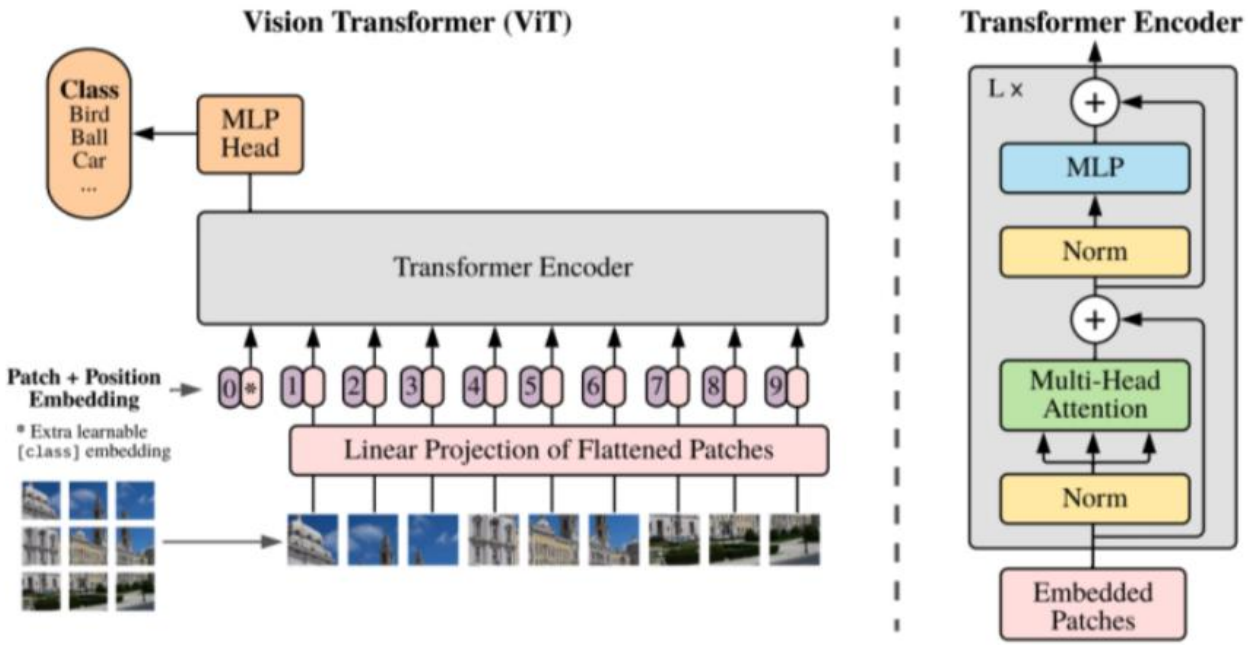
U-Net architecture
Image source: Ronneberger et al.



U-Net based diffusion architecture
Image source: Rombach et al.

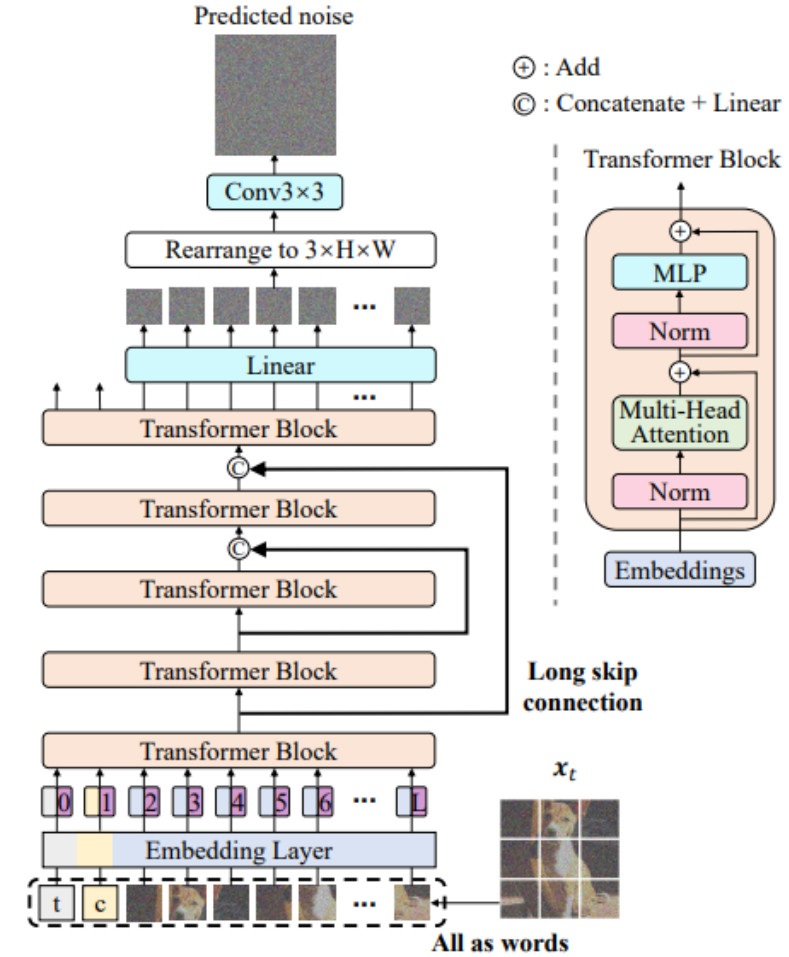
Ronneberger et al., “U-Net: Convolutional Networks for Biomedical Image Segmentation”, MICCAI 2015
Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models", CVPR 2022

Transformer Architecture



Vision transformer

Image source: Dosovitskiy et al.



Transformer based diffusion model

Image source: Bao et al.

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 Diffusion Models", CVPR 2023

DALL-E 3 (OpenAI)

Midjourney

Emu (Meta)

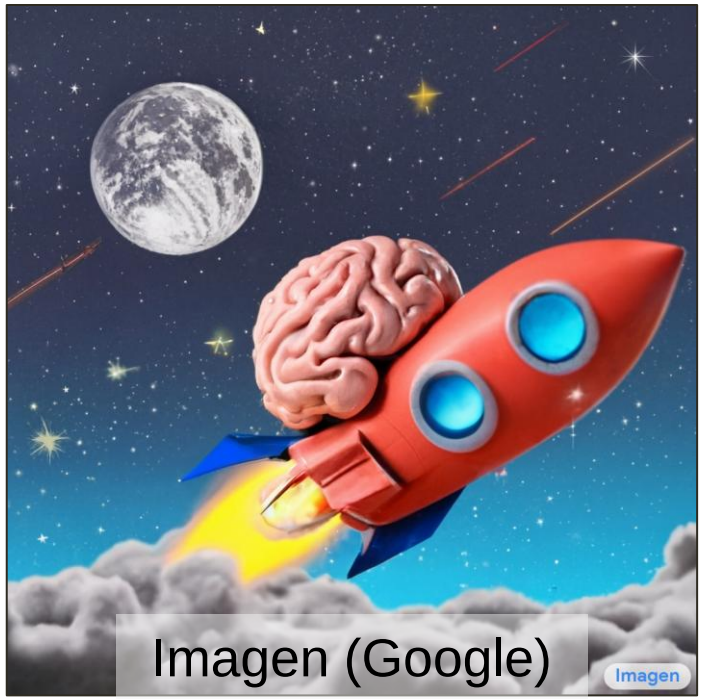
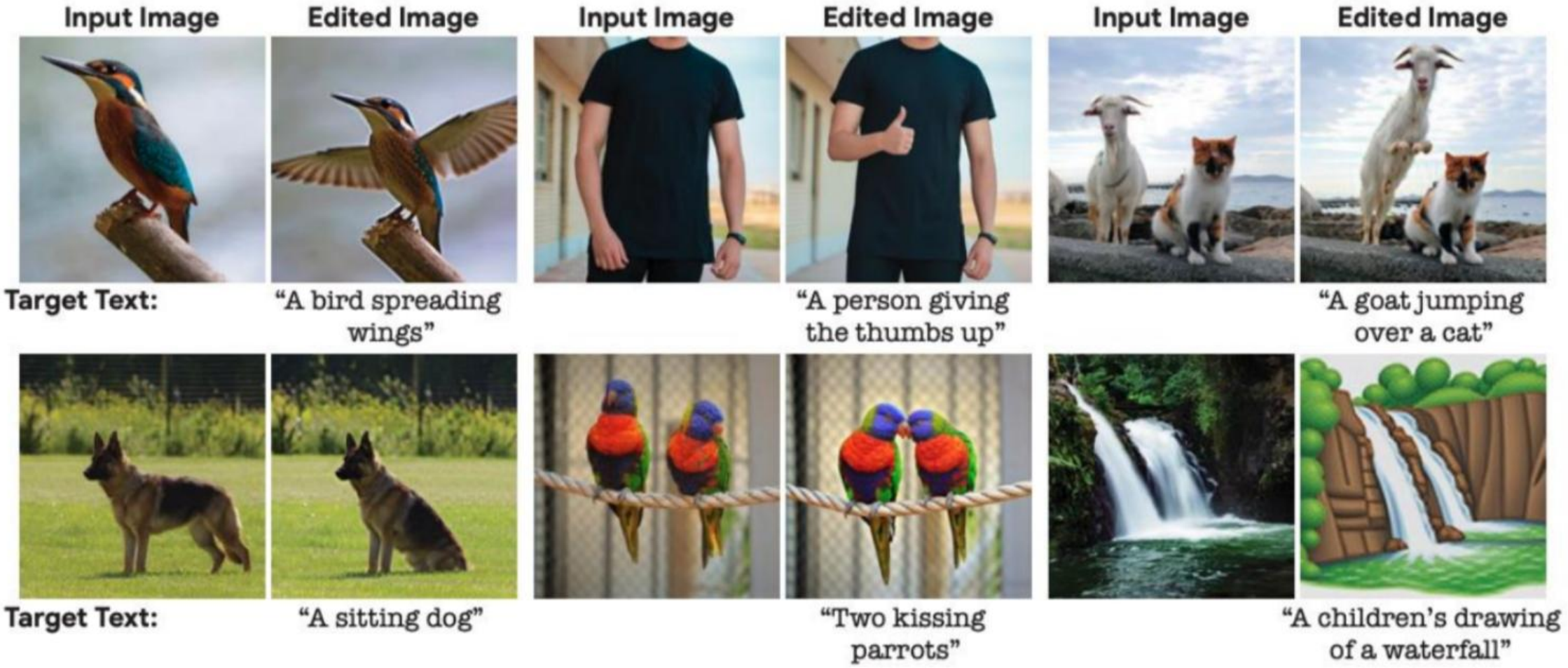


Imagen (Google)

SDXL (Stability.ai)

Firefly (Adobe)

Imagic: Text-Based Real Image Editing with Diffusion Models



Kawar et al., "Imagic: Text-Based Real Image Editing with Diffusion Models", CVPR 2023

Imagic: Text-Based Real Image Editing with Diffusion Models

Kawar et al., "Imagic: Text-Based Real Image Editing with Diffusion Models", CVPR 2023

InstructPix2Pix: Learning to Follow Image Editing Instructions

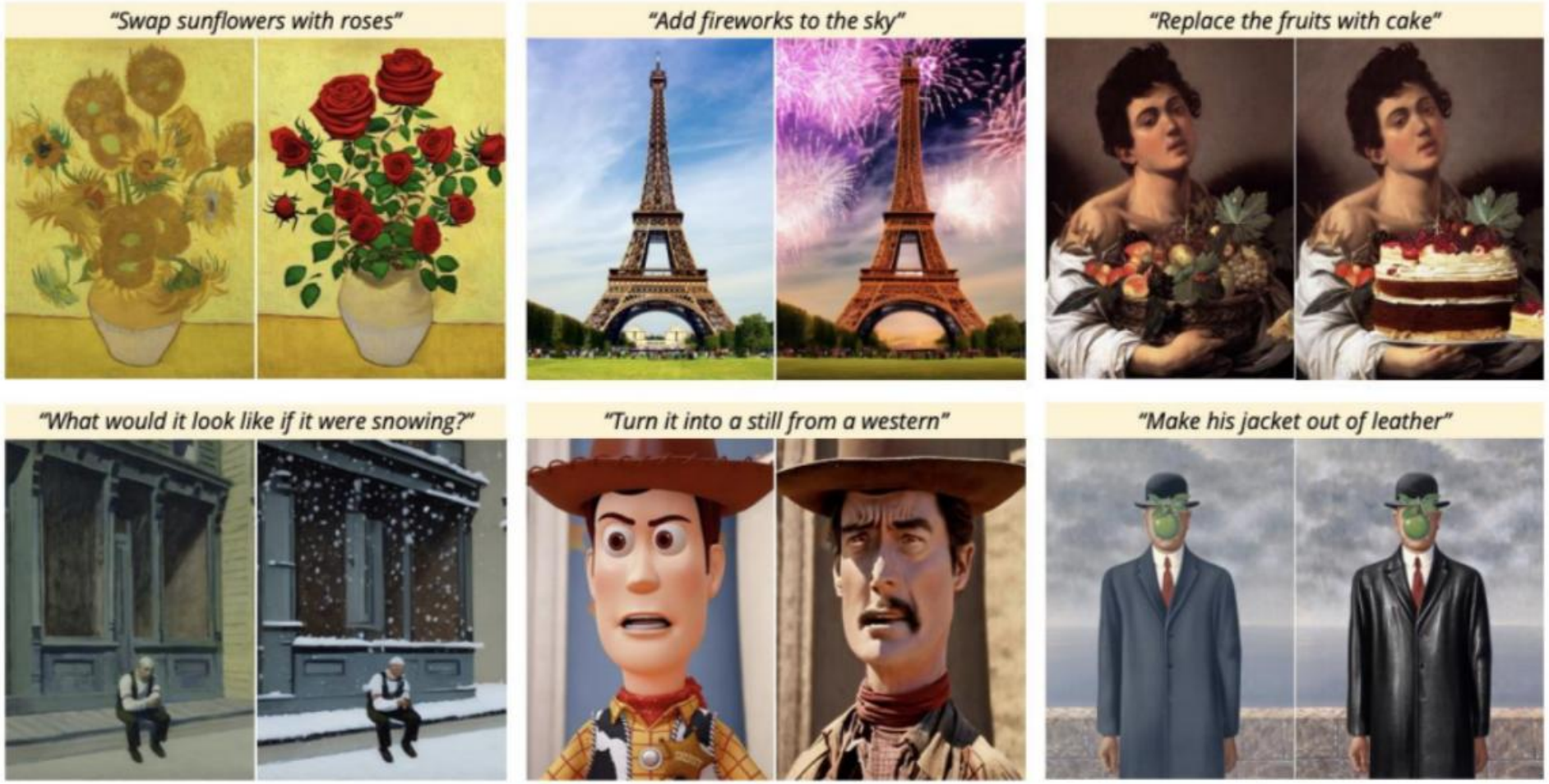
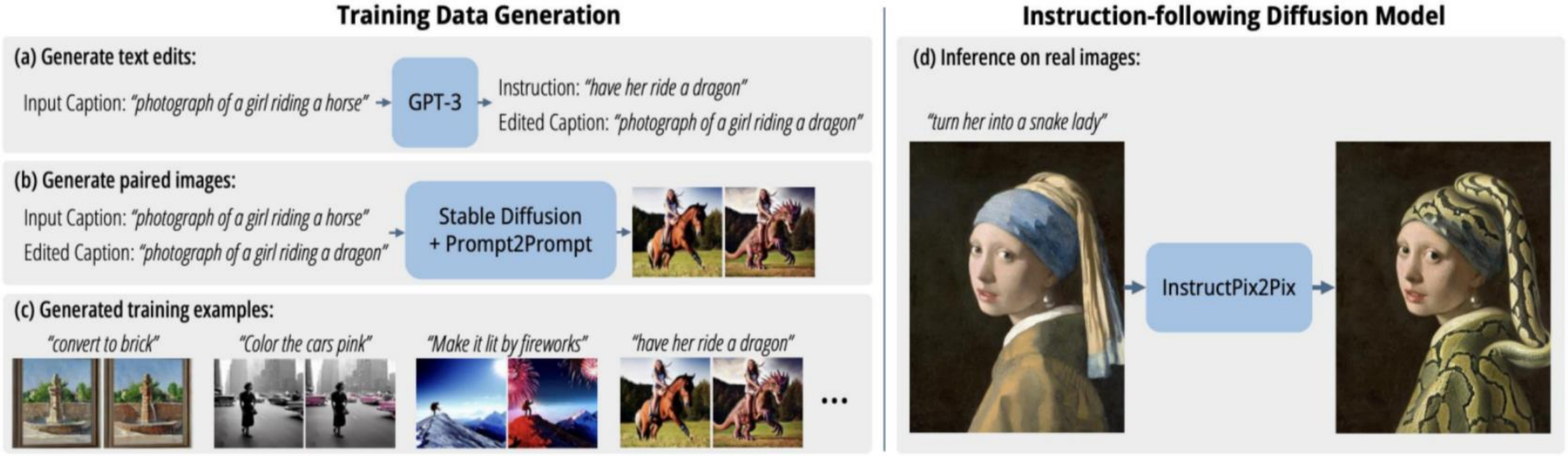


Figure 1. Given **an image** and **an instruction** for how to edit that image, our model performs the appropriate edit. Our model does not require full descriptions for the input or output image, and edits images in the forward pass without per-example inversion or fine-tuning.

Brooks et al., "Instructpix2pix: Learning to follow image editing instructions", CVPR 2023

InstructPix2Pix: Learning to Follow Image Editing Instructions



Brooks et al., "Instructpix2pix: Learning to follow image editing instructions", CVPR 2023

Personalization with diffusion models

Ruiz et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation", CVPR 2023

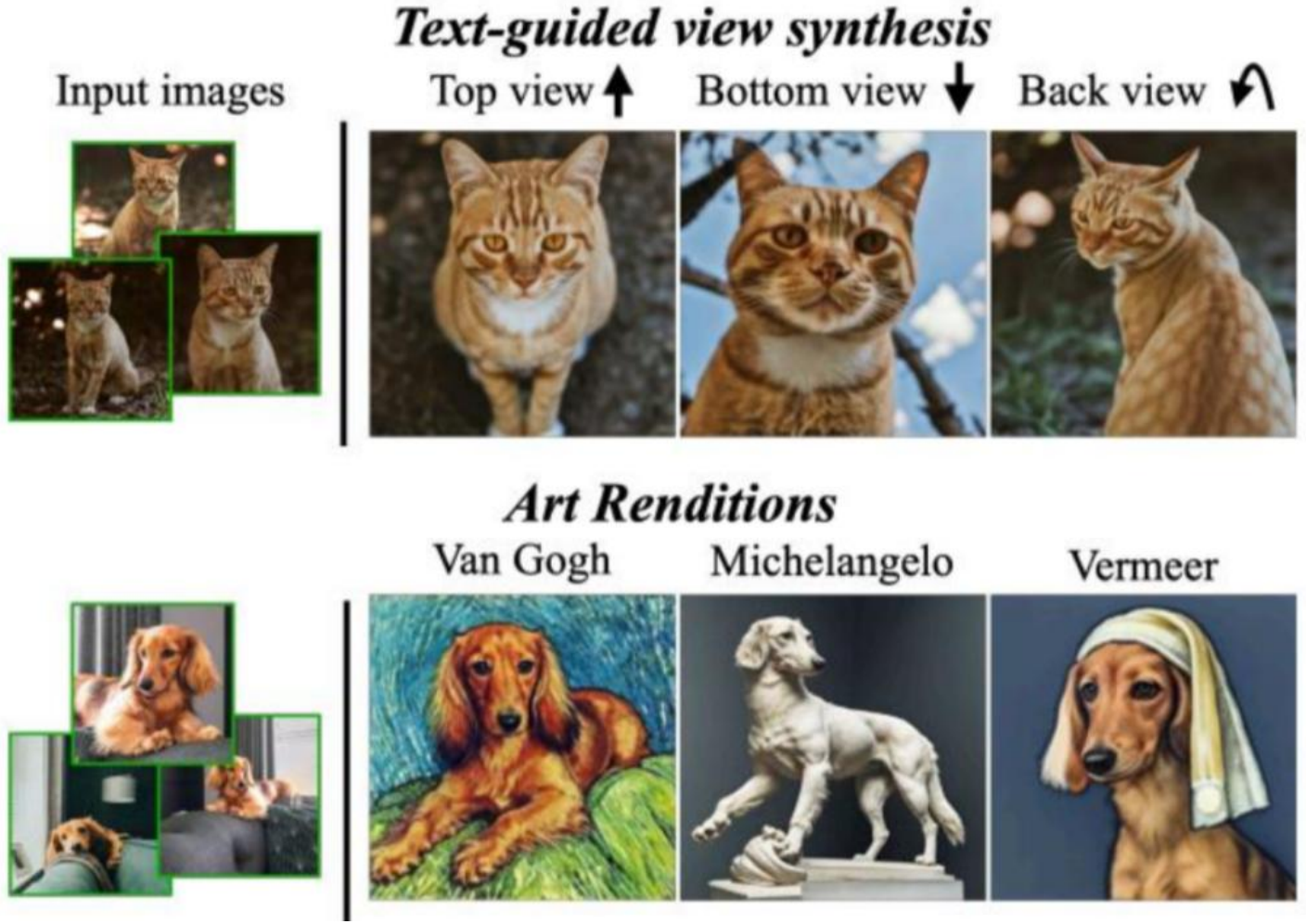
DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

Ruiz et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation", CVPR 2023

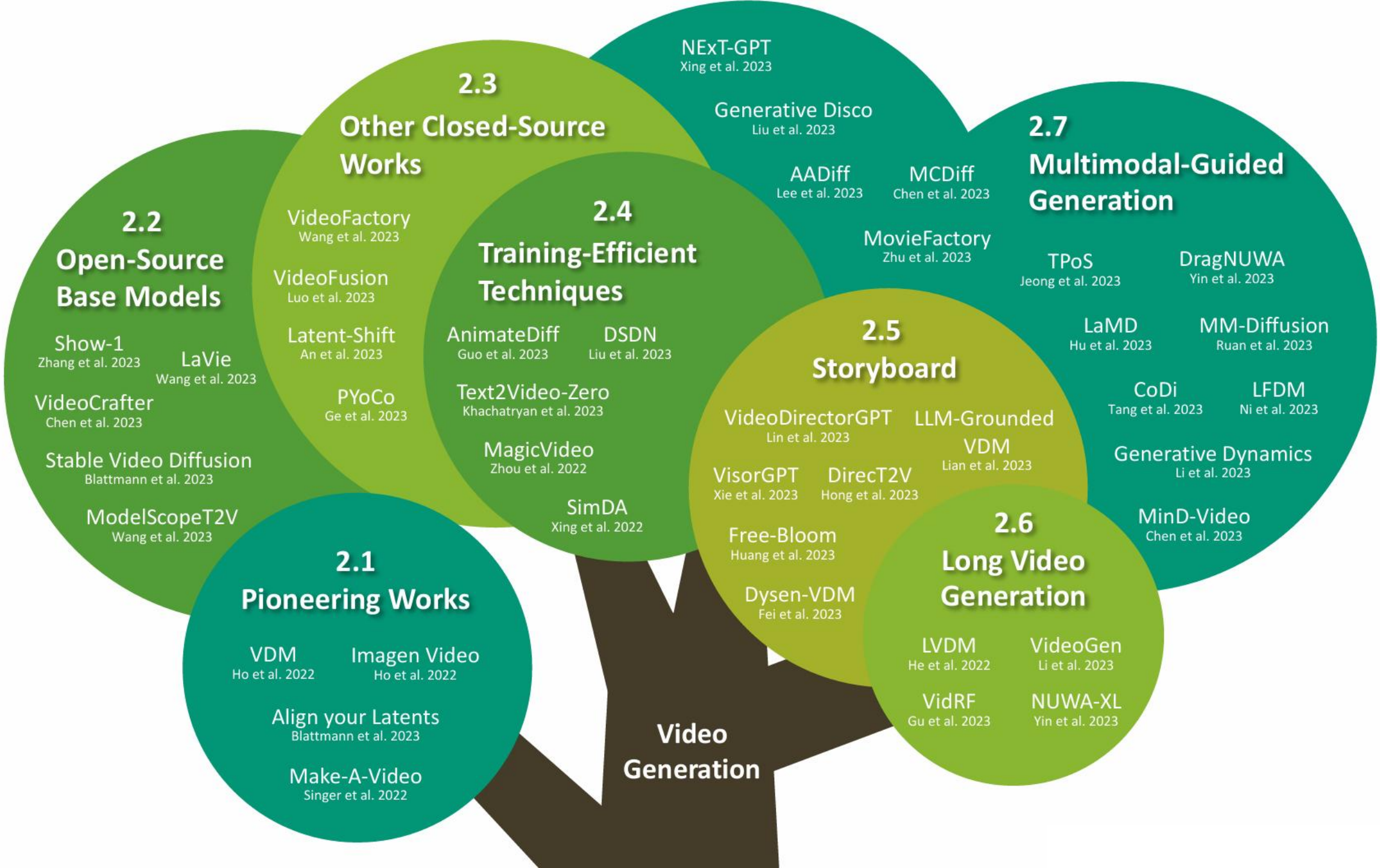
DreamBooth Results

Ruiz et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation", CVPR 2023

DreamBooth Applications



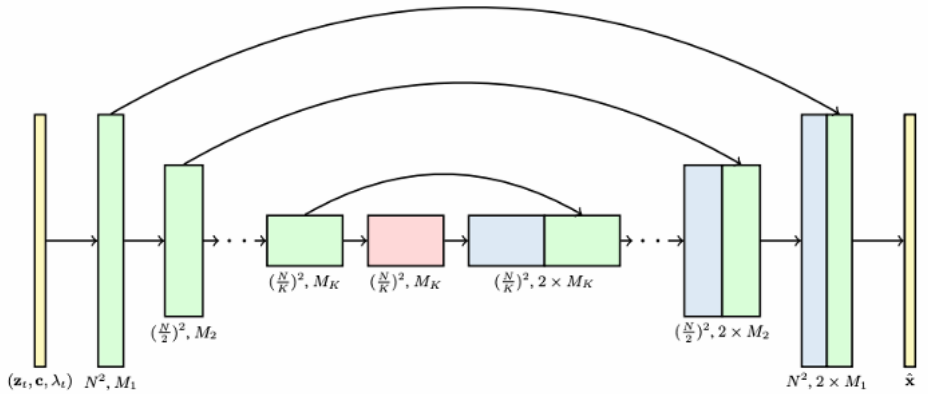
Ruiz et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation", CVPR 2023


Video Generation
2.1 Pioneering Works
VDM
Ho et al. 2022
Imagen Video
Ho et al. 2022
Align your Latents
Blattmann et al. 2023
Make-A-Video
Singer et al. 2022
2.2 Open-Source Base Models
Show-1
Zhang et al. 2023
LaVie
Wang et al. 2023
VideoCrafter
Chen et al. 2023
Stable Video Diffusion
Blattmann et al. 2023
ModelScopeT2V
Wang et al. 2023
2.3 Other Closed-Source Works
VideoFactory
Wang et al. 2023
VideoFusion
Luo et al. 2023
Latent-Shift
An et al. 2023
PYoCo
Ge et al. 2023
2.4 Training-Efficient Techniques
AnimateDiff
Guo et al. 2023
DSDN
Liu et al. 2023
Text2Video-Zero
Khachatryan et al. 2023
MagicVideo
Zhou et al. 2022
SimDA
Xing et al. 2022
2.5 Storyboard
VideoDirectorGPT
Lin et al. 2023
LLM-Grounded VDM
Lian et al. 2023
VisorGPT
Xie et al. 2023
DirecT2V
Hong et al. 2023
Free-Bloom
Huang et al. 2023
Dysen-VDM
Fei et al. 2023
2.6 Long Video Generation
LVDM
He et al. 2022
VideoGen
Li et al. 2023
VidRF
Gu et al. 2023
NUWA-XL
Yin et al. 2023
NExT-GPT
Xing et al. 2023
Generative Disco
Liu et al. 2023
AADiff
Lee et al. 2023
MCDiff
Chen et al. 2023
MovieFactory
Zhu et al. 2023
2.7 Multimodal-Guided Generation
TPoS
Jeong et al. 2023
DragNUWA
Yin et al. 2023
LaMD
Hu et al. 2023
MM-Diffusion
Ruan et al. 2023
CoDi
Tang et al. 2023
LFDM
Ni et al. 2023
Generative Dynamics
Li et al. 2023
MinD-Video
Chen et al. 2023

Video Diffusion Models

3D UNet from a 2D UNet.

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



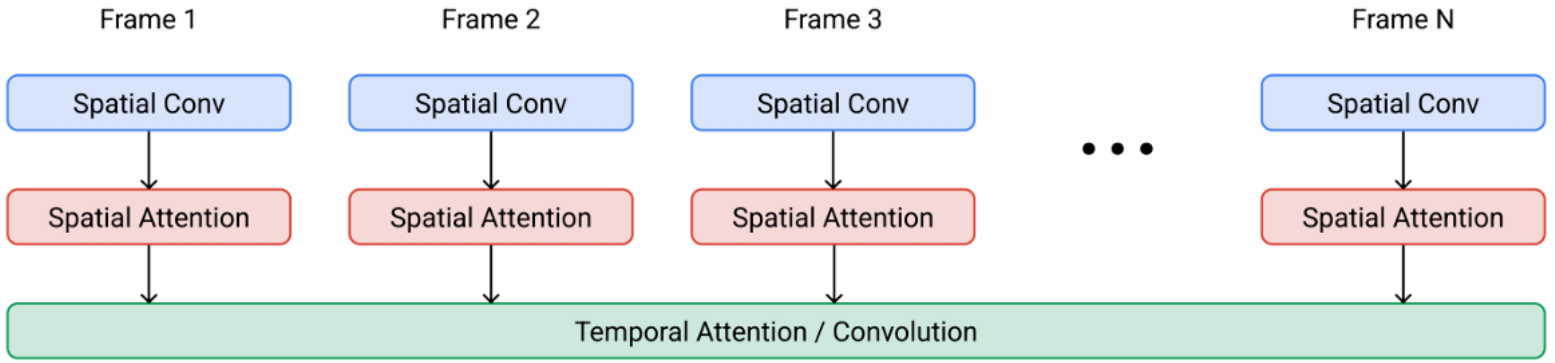
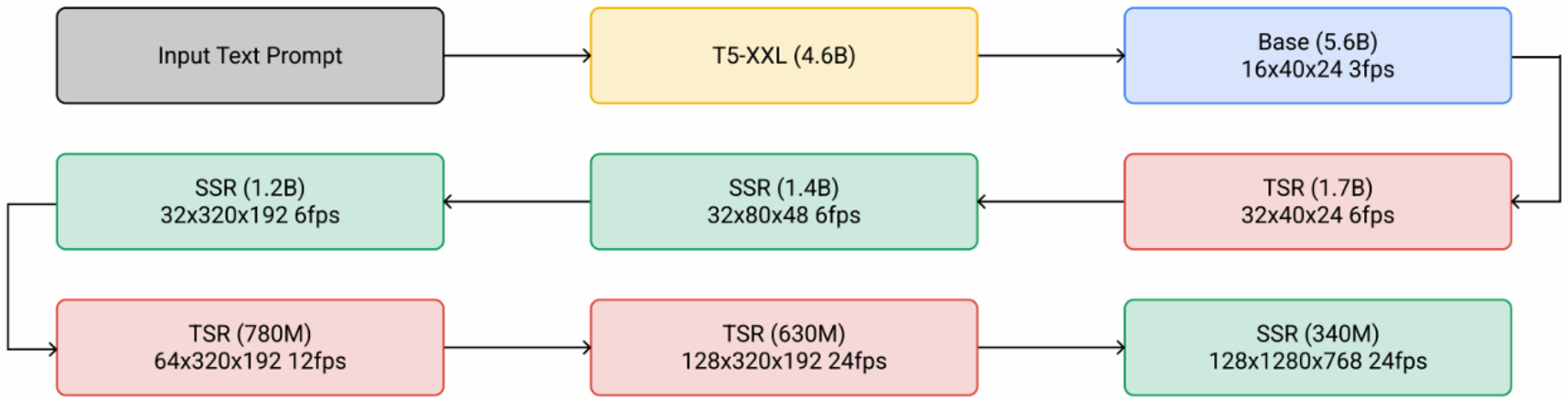


Illustration on how the 3d attention is factorized (from Imagen video)

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

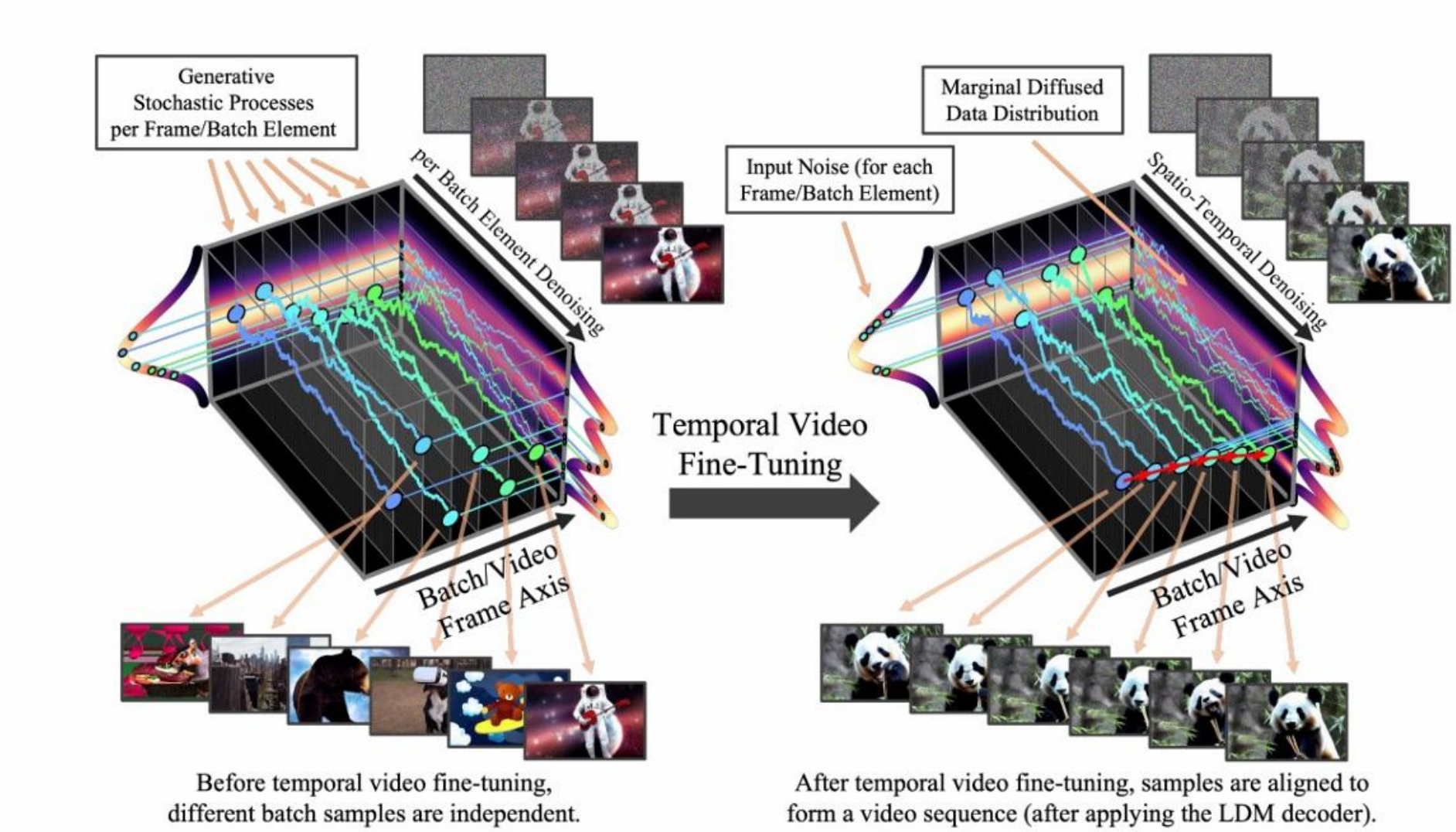
Imagen Video: : Large Scale Text-to-Video

- 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

Video LDM



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

Gen-1

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

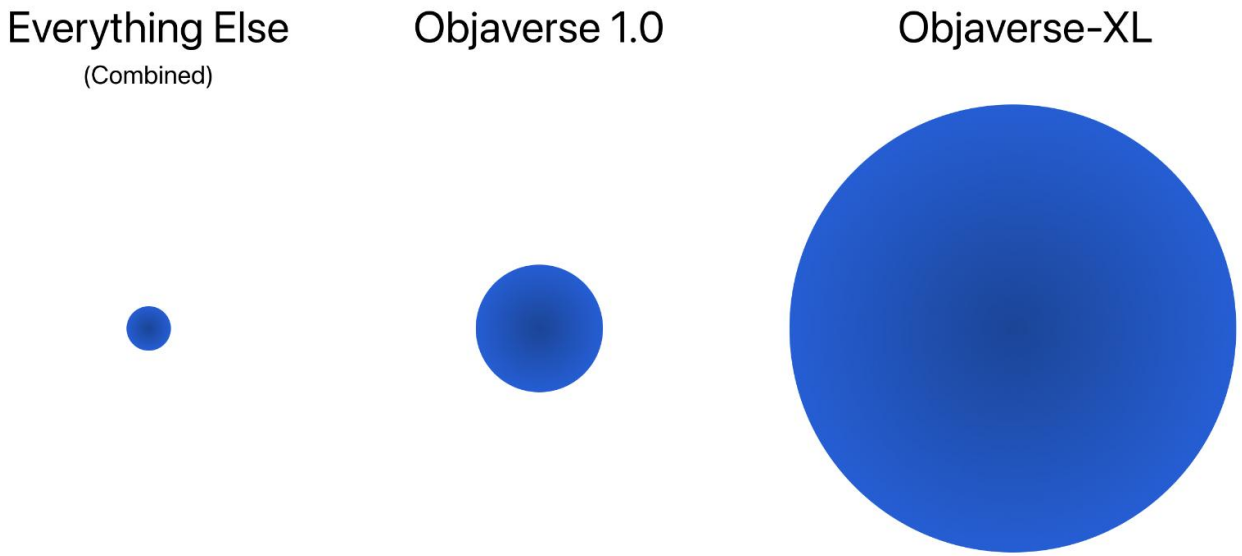
Prompt	Driving Video (top) and Result (bottom)
a man using a laptop inside a train, anime style	
a woman and man take selfies while walking down the street, claymation	

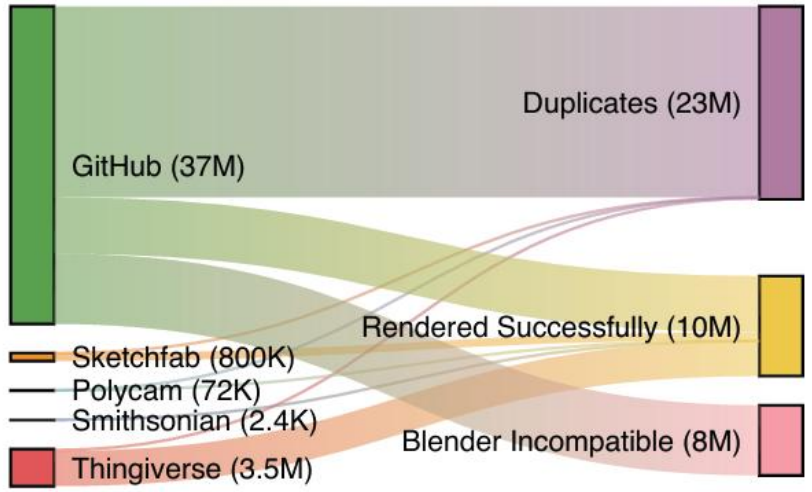
Esser et al., "Structure and Content-Guided Video Synthesis with Diffusion Models", arXiv 2023

3D Datasets

Objaverse-XL extends Objaverse 1.0 to an even larger 3D dataset of 10.2M unique objects from a diverse set of sources, object shapes, and categories.

Source	# Objects
IKEA [32]	219
GSO [17]	1K
EGAD [41]	2K
OmniObject3D [63]	6K
PhotoShape [46]	5K
ABO [13]	8K
Thingi10K [67]	10K
3d-Future [19]	10K
ShapeNet [9]	51K
Objaverse 1.0 [14]	800K
Objaverse-XL	**10.2M**





NeRF

Representing Scenes as Neural Radiance Fields for View Synthesis

ECCV 2020 Oral - Best Paper Honorable Mention

Ben Mildenhall* UC Berkeley

Pratul P. Srinivasan* UC Berkeley

Matthew Tancik* UC Berkeley

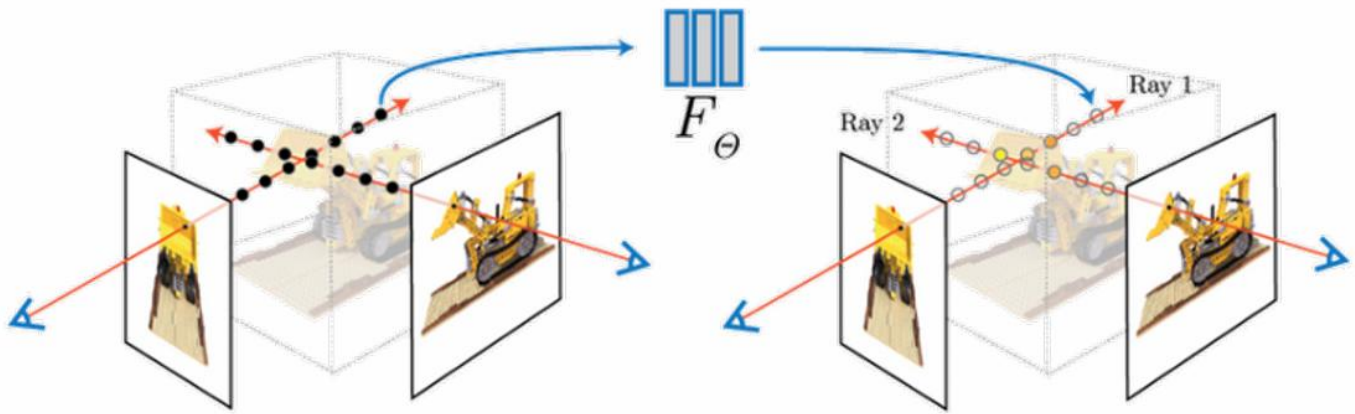
Jonathan T. Barron Google Research

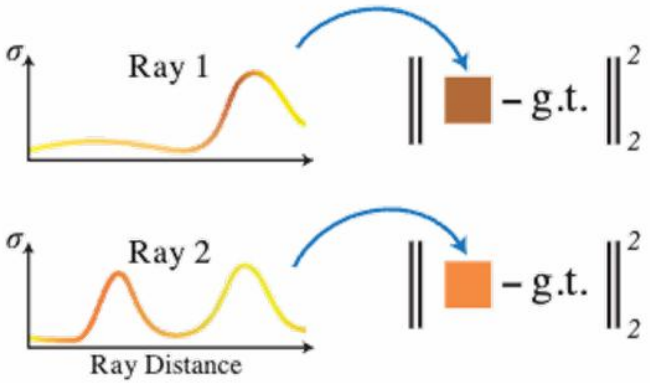
Ravi Ramamoorthi UC San Diego

Ren Ng UC Berkeley

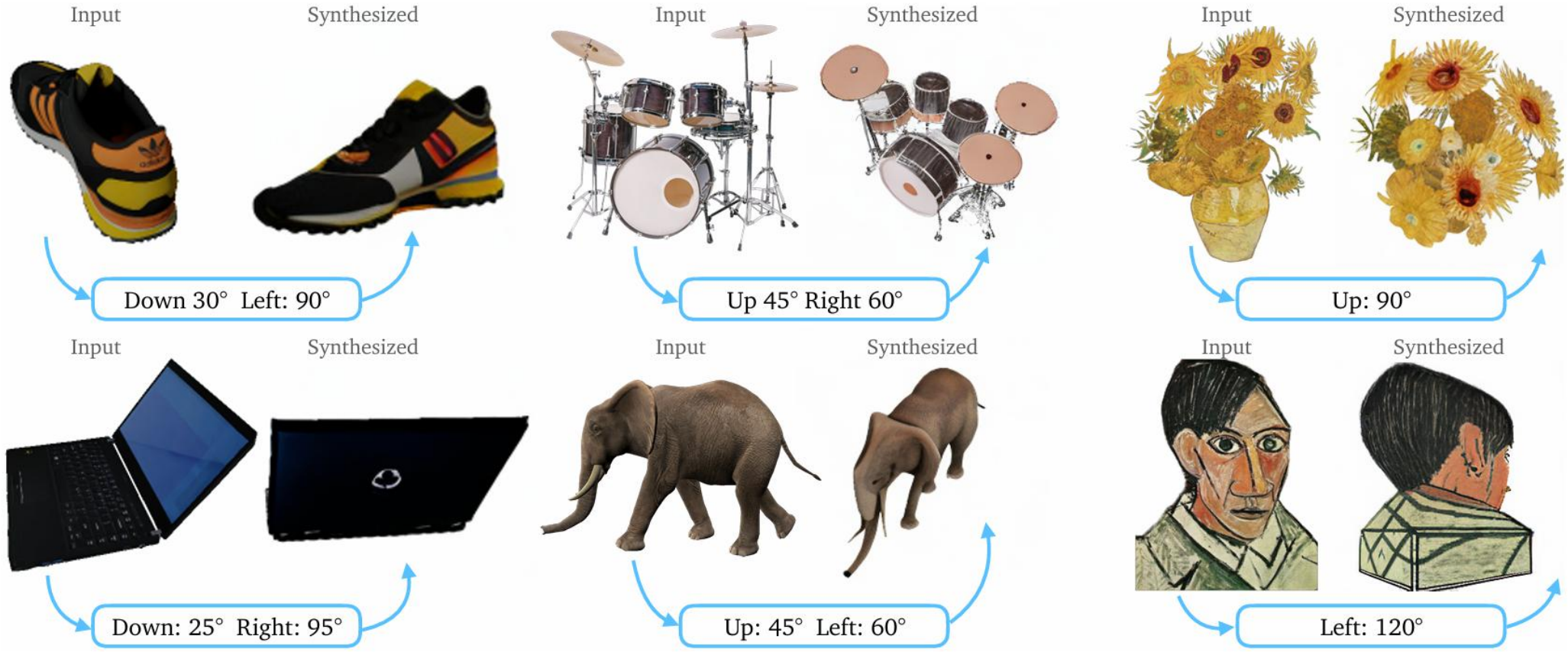
$$(x,y,z,\theta,\phi) \rightarrow \boxed{\quad} \rightarrow (RGB\sigma)$$

$$F_\Theta$$





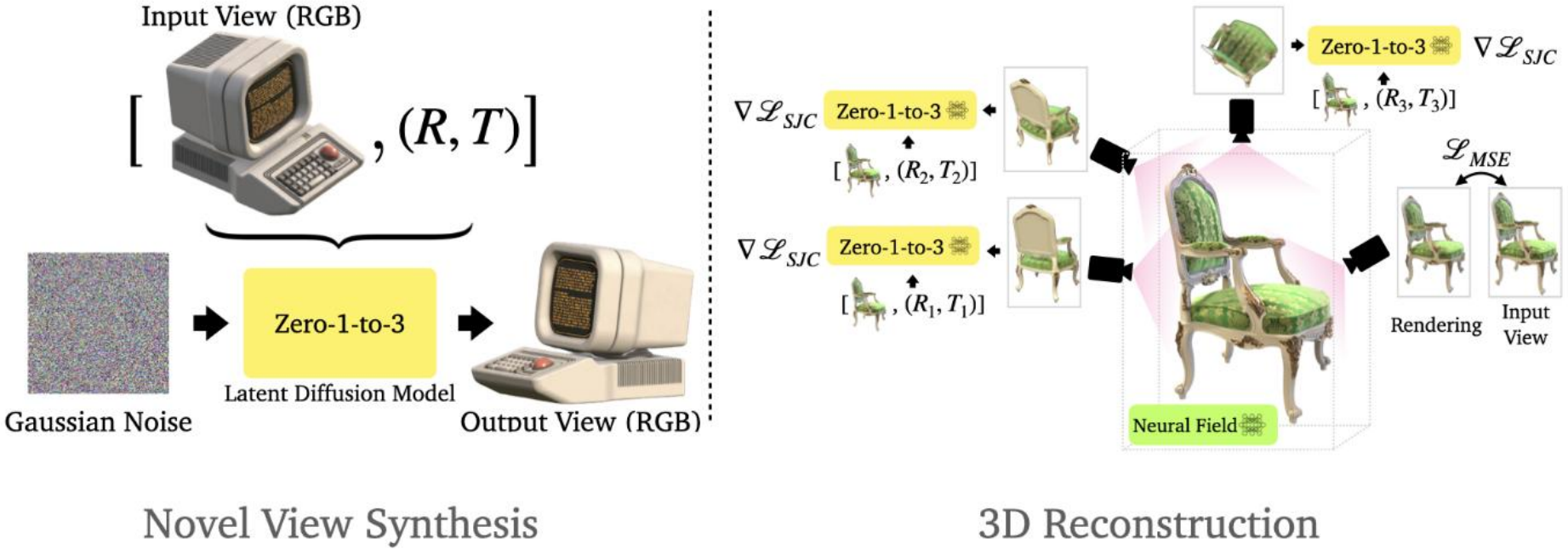
Zero-1-to-3: Zero-shot One Image to 3D Object



Zero-1-to-3: Zero-shot One Image to 3D Object, Liu et al., ICCV 2023

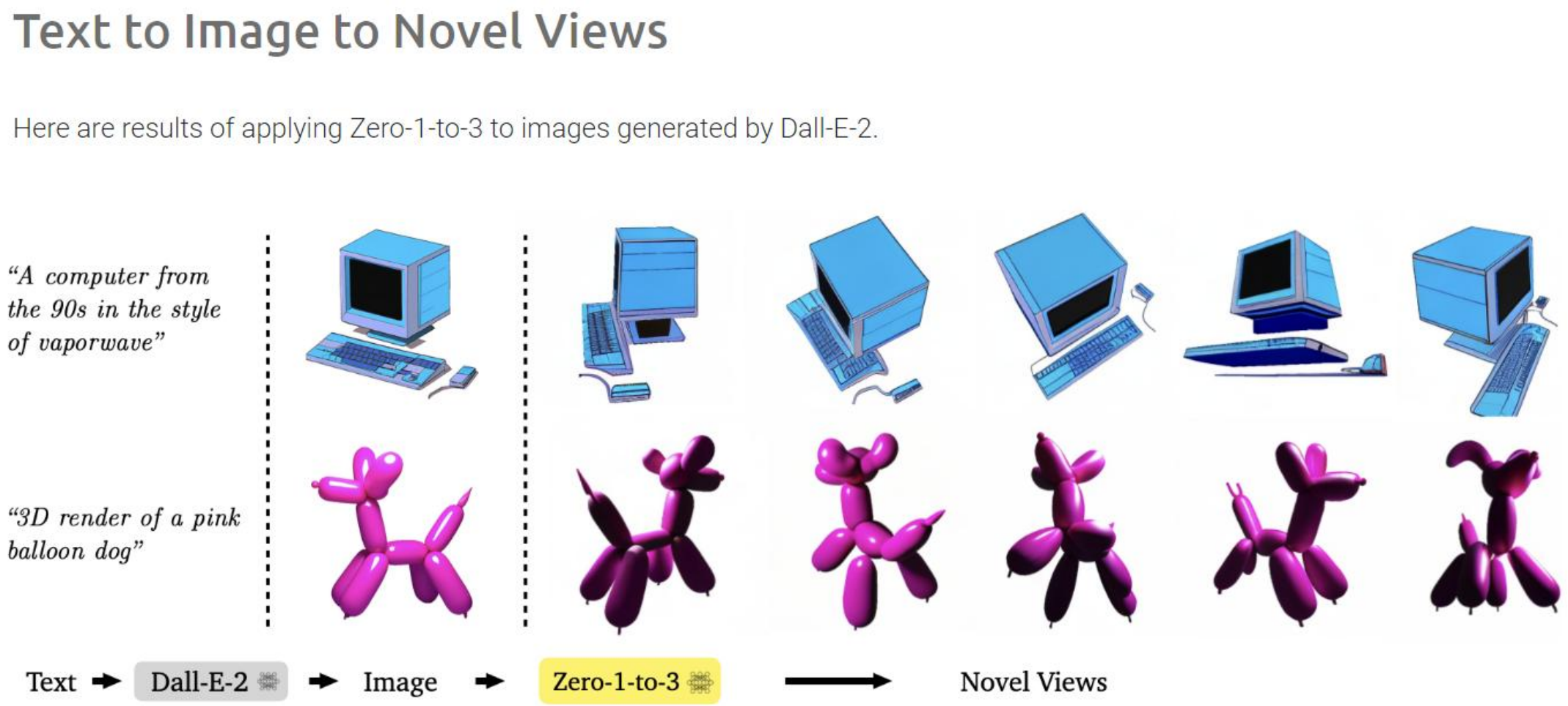
Zero-1-to-3: Zero-shot One Image to 3D Object

A view-conditioned diffusion model is learned to control the viewpoint of an image containing a novel object (left). Such diffusion model can also be used to train a NeRF for 3D reconstruction (right).



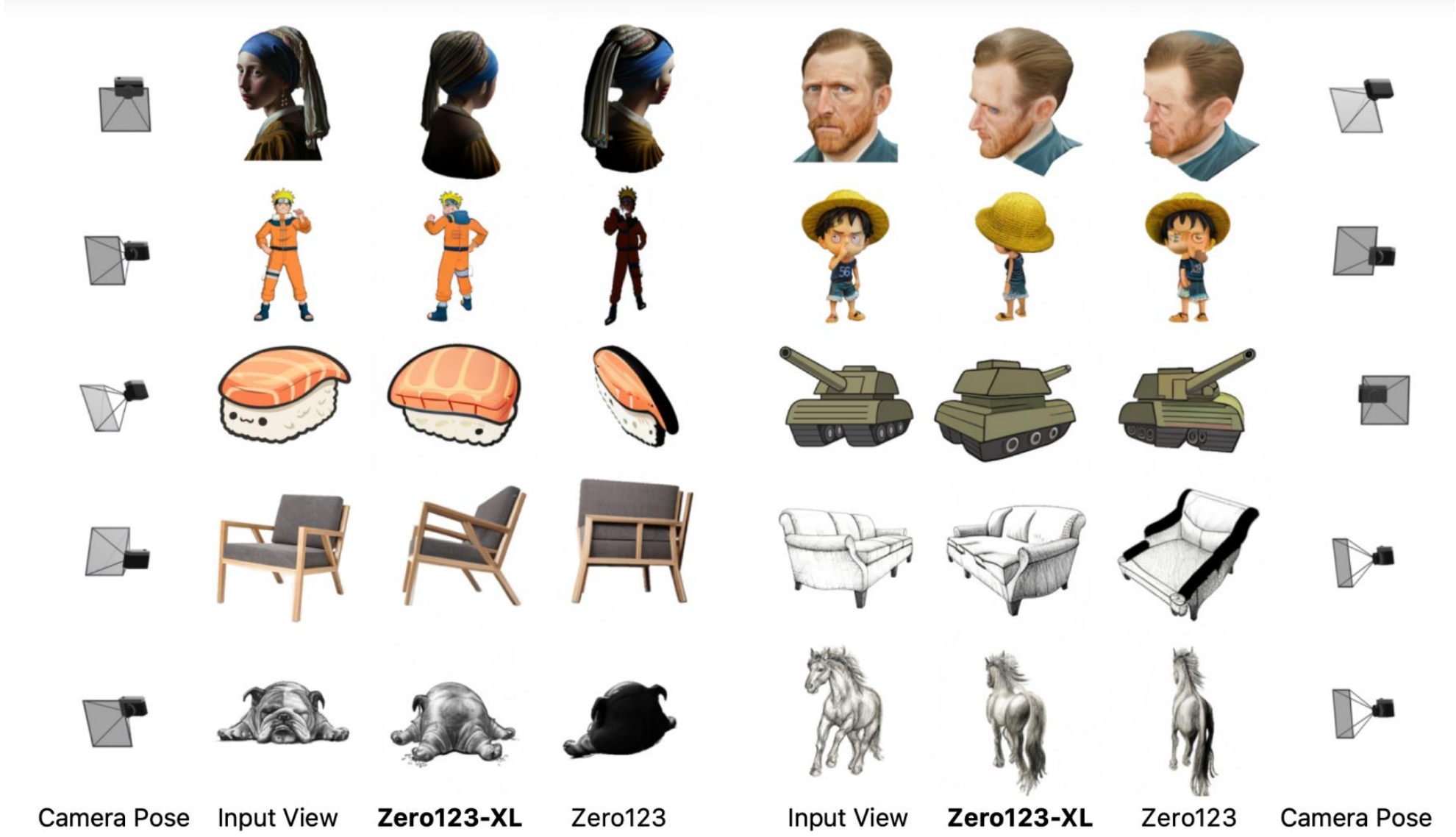
Zero-1-to-3: Zero-shot One Image to 3D Object, Liu et al., ICCV 2023

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



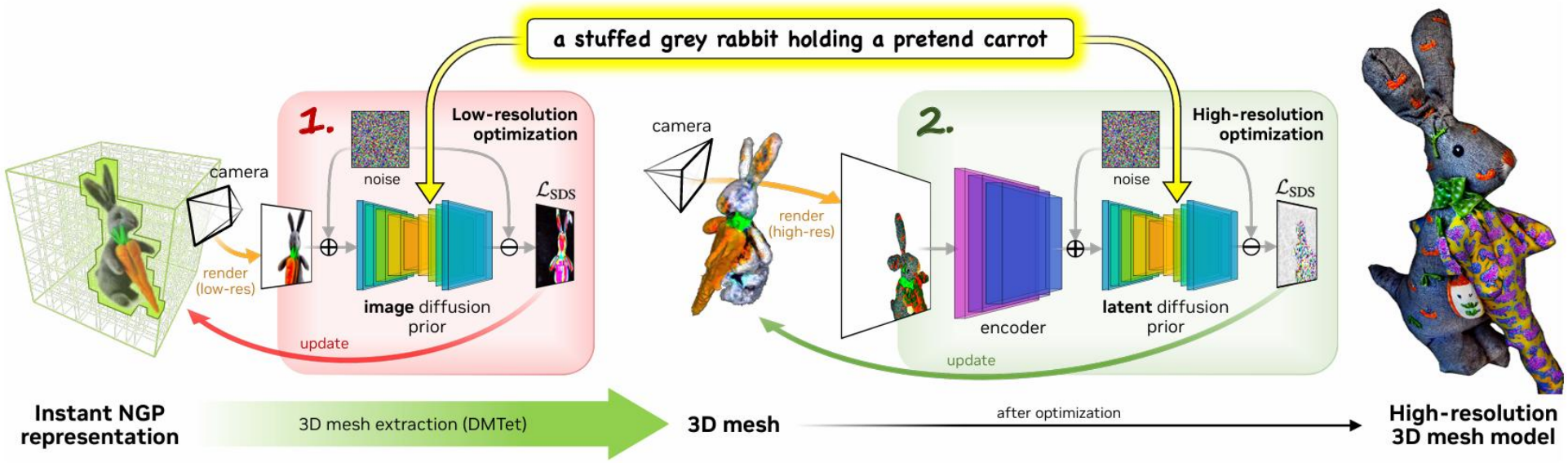
Zero-1-to-3: Zero-shot One Image to 3D Object, Liu et al., ICCV 2023

Novel View Generation for Objects



Magic3D

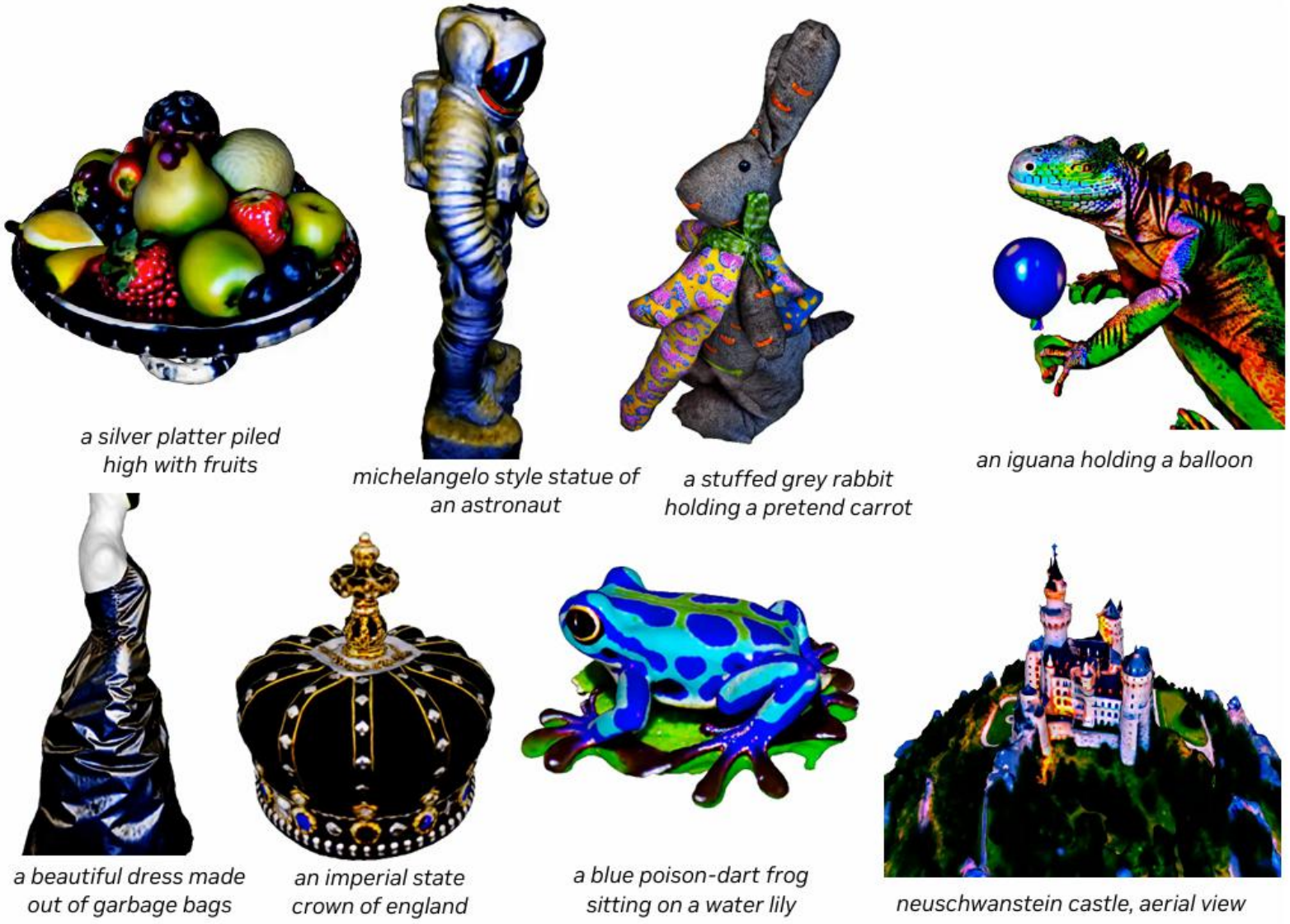
- Generating high-resolution 3D content from an input text prompt in a coarse-to-fine manner
- In the first stage, a low-resolution diffusion prior is used to optimize neural field representations (color, density, and normal fields) for a coarse model. Then, a textured 3D mesh is extracted from the density and color fields of the coarse model. Then, it is fine-tuned using a high-resolution latent diffusion model. After optimization, the model generates high-quality 3D meshes with detailed textures.



Magic3D: High-Resolution Text-to-3D Content Creation, CVPR 2023

Magic3D: High-resolution text-to-3D generation

Magic3D can generate high-quality and high-resolution 3D models from text prompts.



Magic3D: High-Resolution Text-to-3D Content Creation, CVPR 2023

Migic3D: High-resolution Prompt-based Editing

Magic3D can edit 3D models by fine-tuning with the diffusion prior using a different prompt. Taking the low-resolution 3D model as the input, Magic3D can modify different parts of the 3D model corresponding to different input text prompts. Together with various creative controls on the generated 3D models, Magic3D is a convenient tool for augmenting 3D content creation.

Magic3D: High-Resolution Text-to-3D Content Creation, CVPR 2023

Instruct-NeRF2NeRF

Edit a 3D scene with text instructions

Original NeRF

"Turn him into the Tolkien Elf"

"Make it look like a Fauvism painting"

"Make it look like an Edward Munch Painting"

"Turn him into Lord Voldemort"

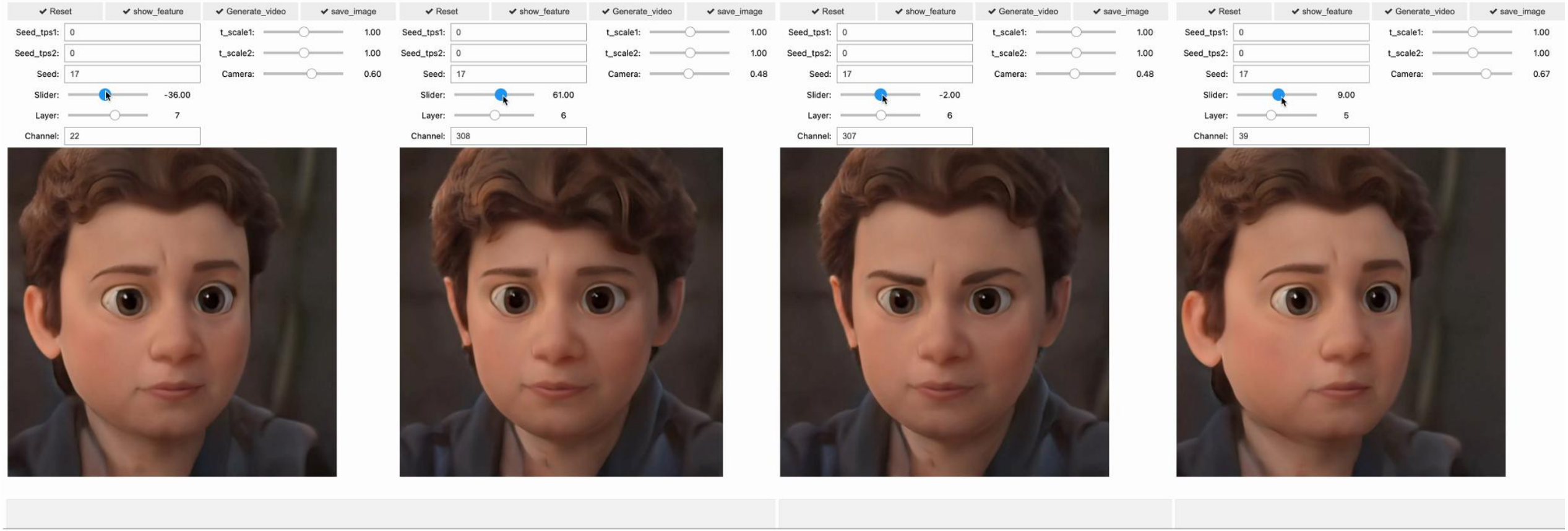
"Make him look like Vincent Van Gogh"

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

3DAvatarGAN

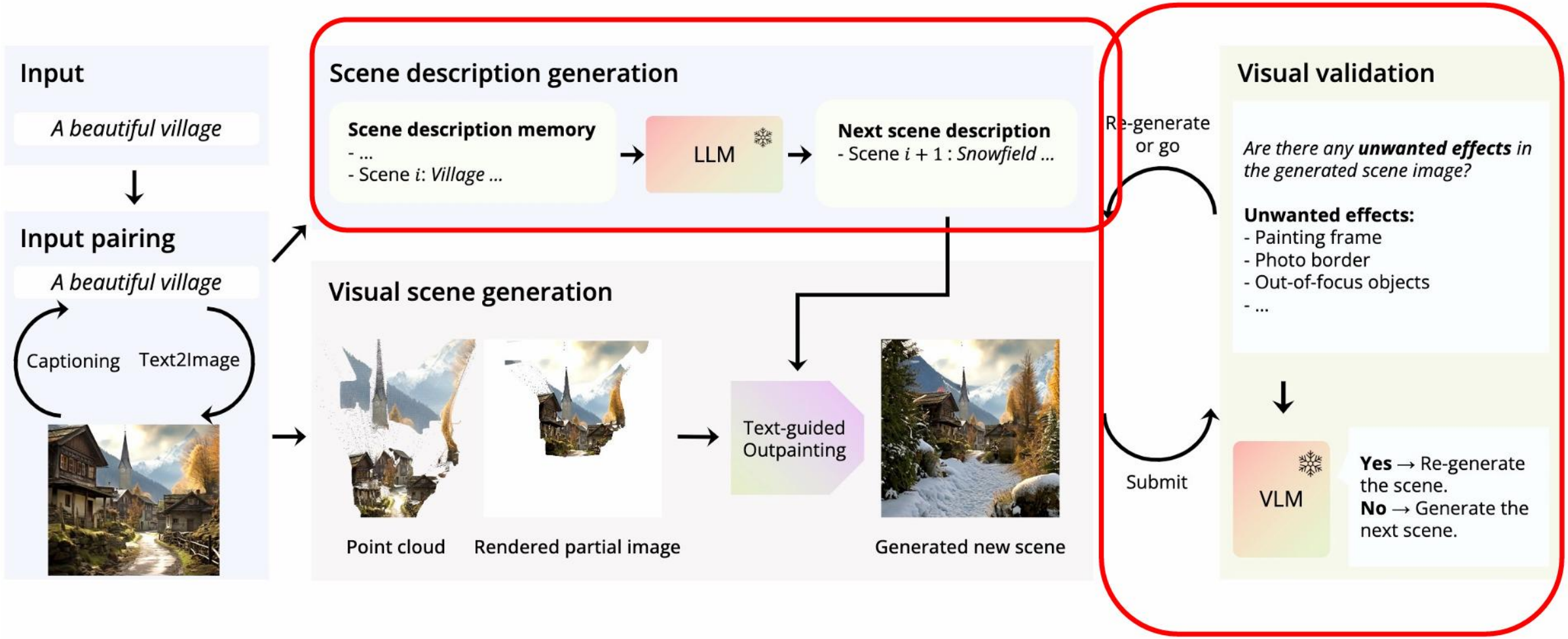
Rameen et al., **3DAvatarGAN: Bridging Domains for Personalized Editable Avatars,** CVPR'2023

3DAvatarGAN



Rameen et al., **3DAvatarGAN: Bridging Domains for Personalized Editable Avatars,** CVPR'2023

WonderJourney

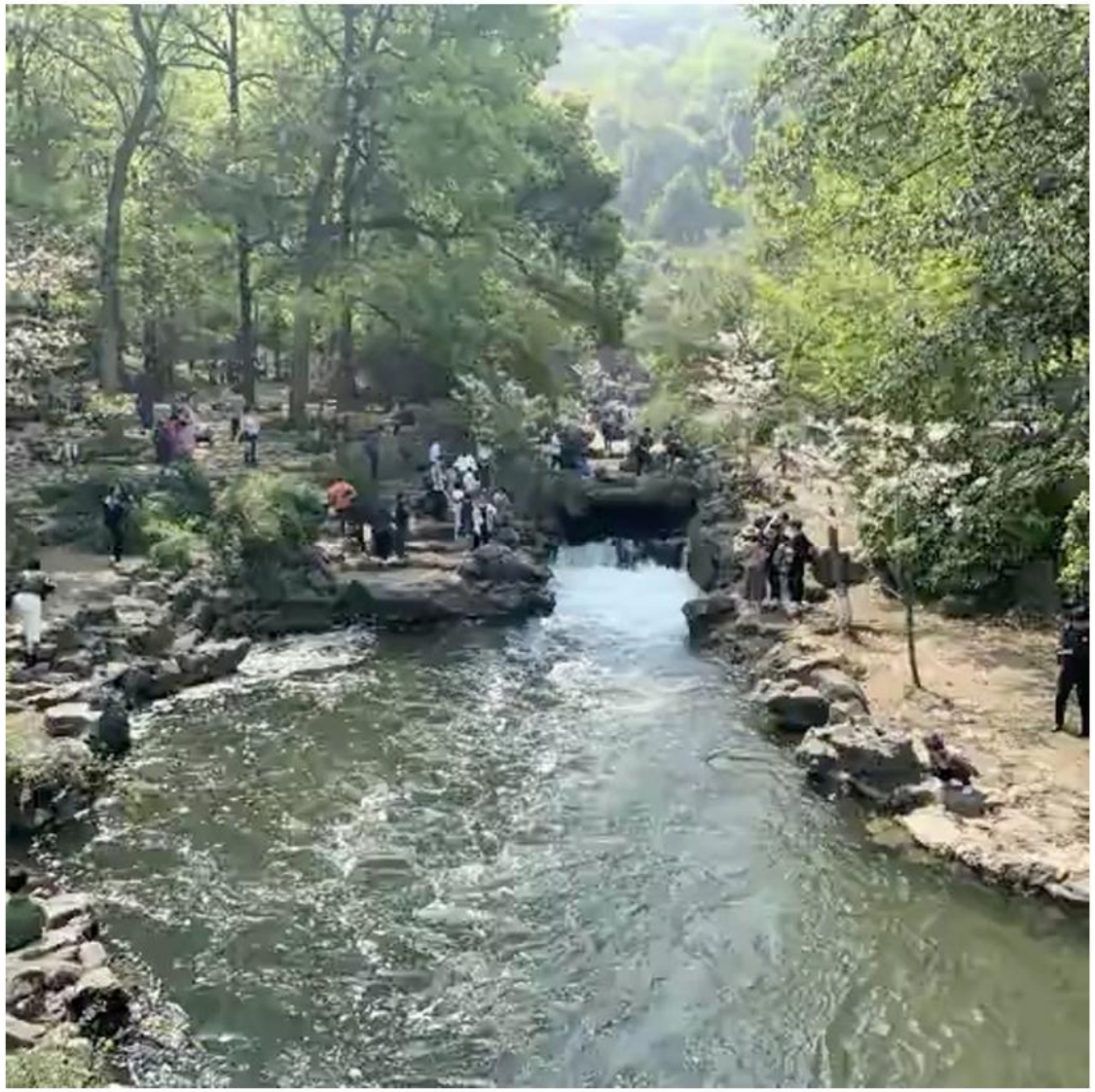


WonderJourney: Going from Anywhere to Everywhere, CVPR'24

WonderJourney

WonderJourney: Going from Anywhere to Everywhere, CVPR'24

WonderJourney



WonderJourney: Going from Anywhere to Everywhere, CVPR'24

Applications of Generative AI in CV

1. Image Synthesis and Enhancement
2. Style Transfer and Artistic Creation
3. Data Augmentation for Machine Learning
4. Medical Imaging
5. Virtual Try-On and Augmented Reality (AR)
6. Gaming and Virtual Worlds
7. Face Recognition and Privacy Preservation
8. 3D Model Generation and Reconstruction
9. Video Generation and Manipulation
10. Autonomous Vehicles

Image Synthesis and Enhancement

- **Content Creation:** Generative AI can create realistic images from scratch, which is useful in industries like advertising, entertainment, and media for generating high-quality visuals.
- **Image Inpainting:** Filling in missing parts of an image or removing unwanted objects, which is useful in photo editing and restoration.

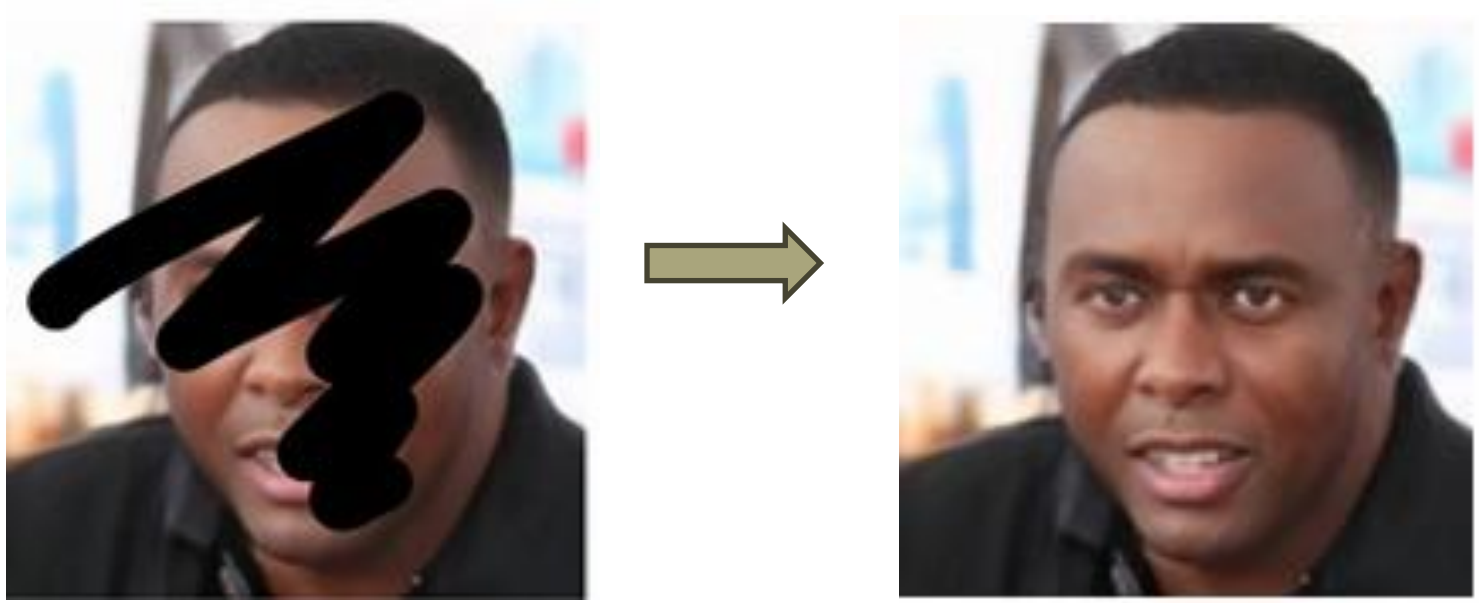


Image Synthesis and Enhancement

Image Super-resolution: Improving the resolution of low-quality images for applications in surveillance, medical imaging, and photography.

$P(High\ resolution\ image \mid Low\ resolution\ image)$

$P(High\ quality\ image \mid Low\ quality\ image)$

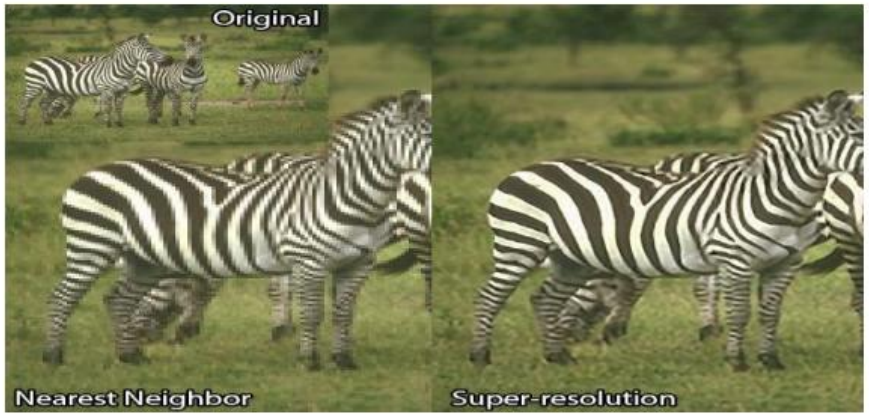


Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. C. Ledig, L. Theis et al. CVPR 2017

Style Transfer and Artistic Creation

- **Art Generation:** Transforming photos into artworks by applying the style of famous painters or generating entirely new artistic pieces.
- **Fashion Design:** Assisting designers in creating new clothing designs by mixing styles or generating novel patterns.

Human-AI collaborative fashion designs made by Agnes Cameron from How To Generate Almost Anything

source- Adidas Artificial Intelligence Design | Webhead Studios

Data Augmentation for Machine Learning

- **Synthetic Data Generation:** Creating synthetic images to augment training datasets for improving the performance of machine learning models, especially when real data is scarce or expensive to obtain.
- **Domain Adaptation:** Generating images that mimic different environments or conditions to make models more robust and adaptable.

(a) DCFace [16].

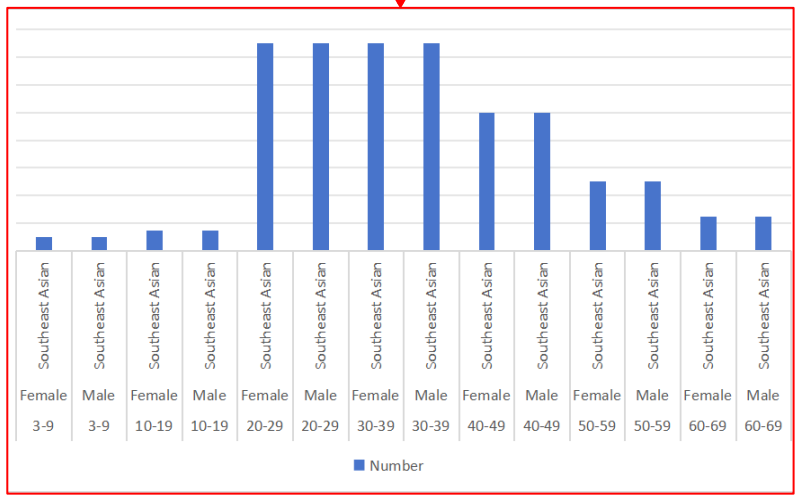
(b) GANDiffFace [28].

Leveraging Synthetic Data for Improving Face Recognition Fairness

MR-All	Mask	IJBC5	IJBC4	LFW	CFPFP	AgeDB	African	Caucasian	South Asian	East Asian	Children
85.271	81.368	94.636	96.768	99.75	98.886	97.7	83.139	91.153	85.88	65.406	64.67
85.358↑	81.275↓	93.854↓	96.4↓	99.75	98.857↓	97.917↑	83.223↑	91.137↓	86.514↑	65.954↑	67.683↑

- **The first column: using the wf19m + 10% mask face**
- **The second column: using the wf19m + 10% mask face + about 40,000 ids of synthetic face images**

Examples of synthetic faces of female 20-29 "Indian"

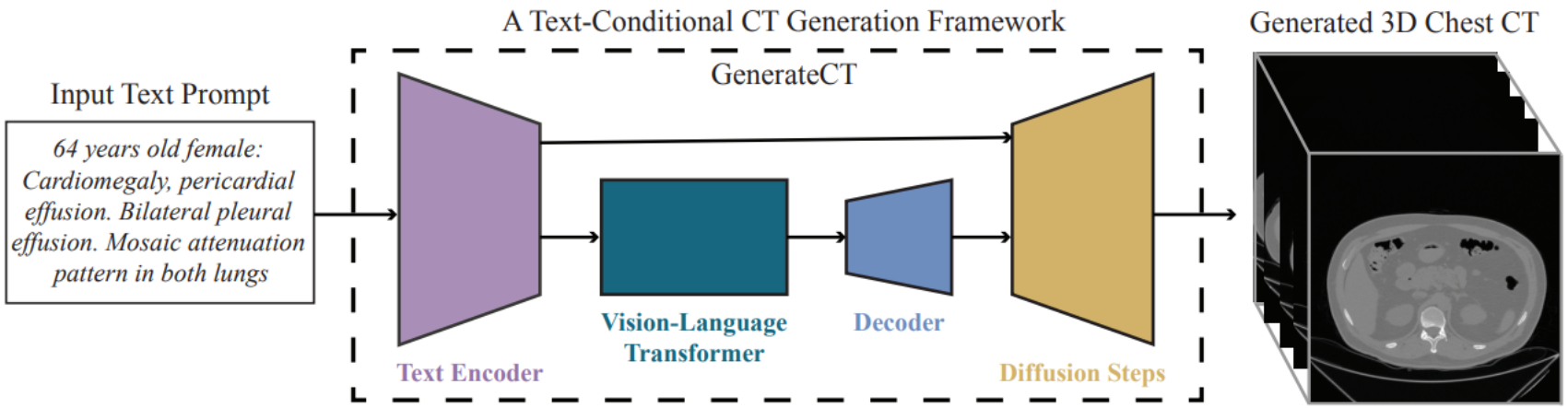


The distribution of the synthetic datasets [Southeast Asian]

We found that in face recognition with a small model, generating a dataset for a specific ethnicity that aligns with the original dataset's age distribution, and combining it with the original dataset for training, improves face recognition performance for that particular ethnicity. Additionally, by increasing the representation of younger ages for that ethnicity, there is also an improvement in face recognition for children.

Medical Imaging

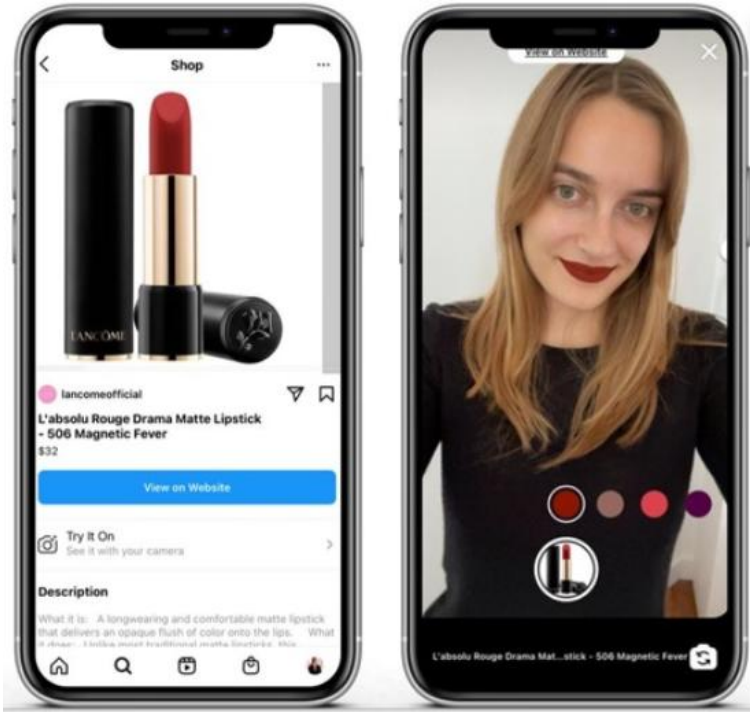
- **Medical Image Synthesis:** Generating realistic medical images, such as MRIs or CT scans, to aid in training medical professionals or for improving diagnostic tools.
- **Anomaly Detection:** Generating synthetic images of anomalies (like tumors) for training AI models to detect rare conditions.



GenerateCT: Text-Conditional Generation of 3D Chest CT Volumes, ECCV 2024

Virtual Try-On and Augmented Reality (AR)

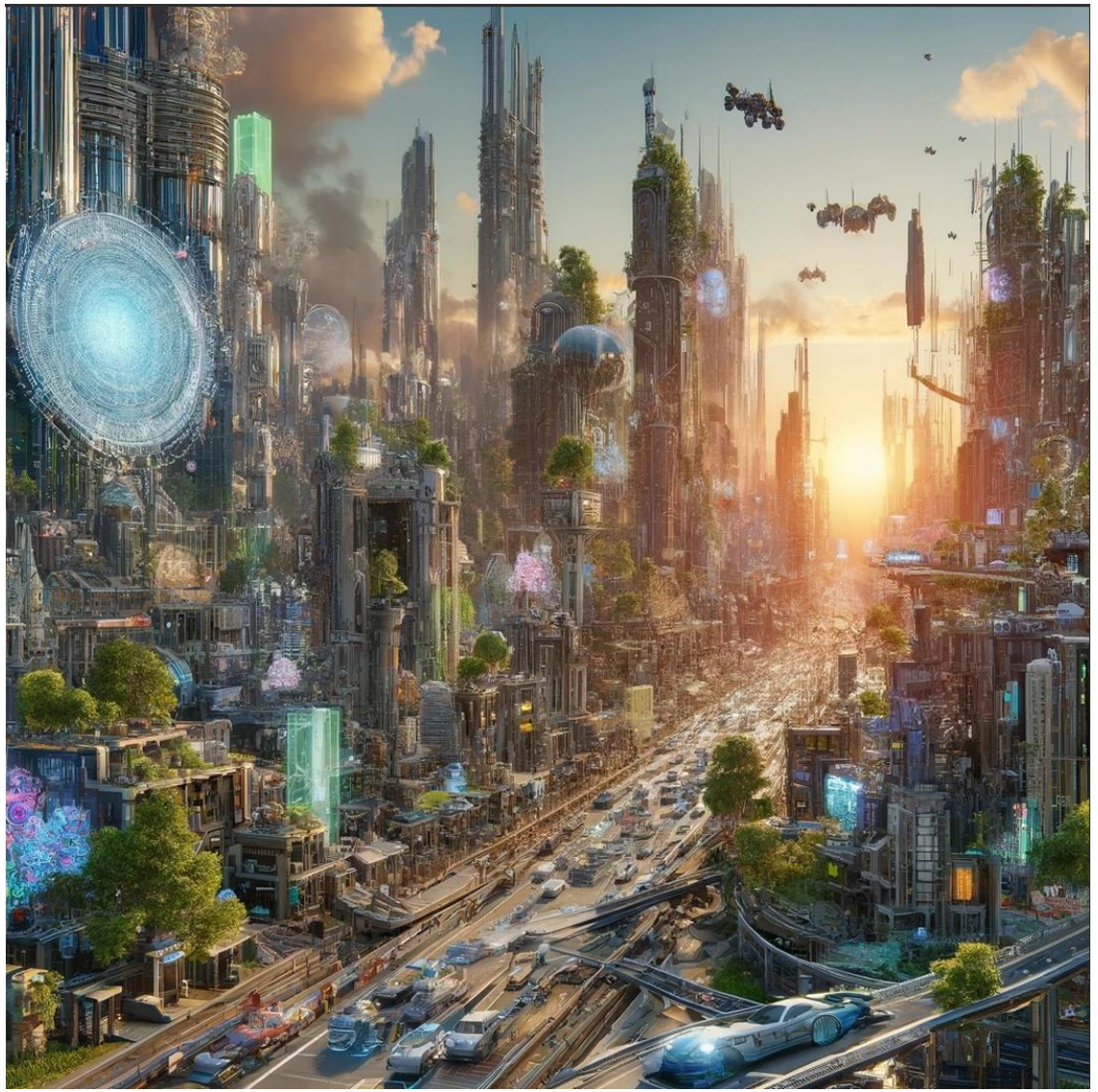
- **Virtual Fitting Rooms:** Allowing customers to virtually try on clothes, accessories, or makeup using generative models that map products onto their images or videos.
- **AR Applications:** Enhancing AR experiences by generating realistic virtual objects or environments that can be superimposed onto the real world.



https://www.drip.com/blog/virtual-try-on-examples

Gaming and Virtual Worlds

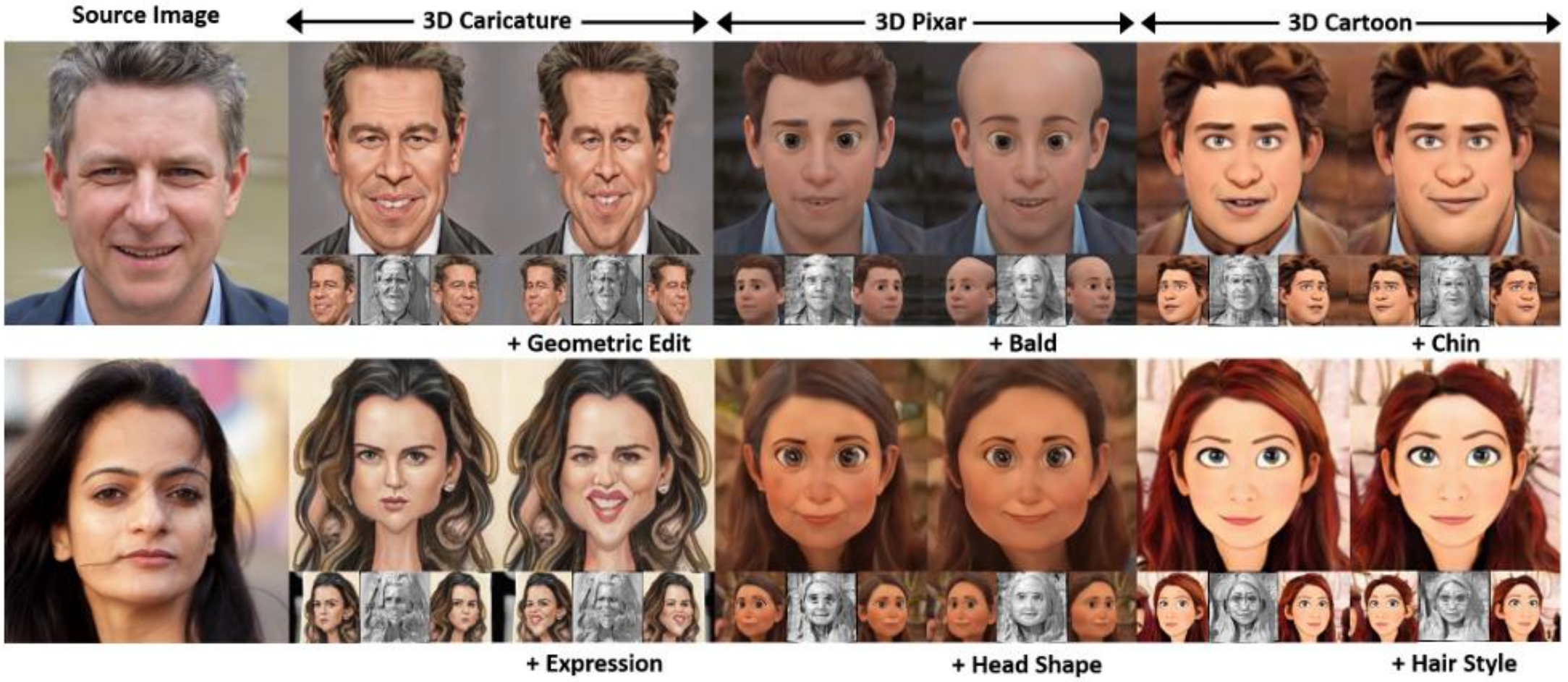
- **Procedural Content Generation:** Automatically creating landscapes, characters, or objects in video games to save time and resources for game developers.



https://snap-research.github.io/3D_4D_modeling_tutorial/

Gaming and Virtual Worlds

- **Avatar Creation:** Generating personalized avatars based on user's appearance or preferences.



3DAvatarGAN: Bridging Domains for Personalized Editable Avatars, ArXiv 2023

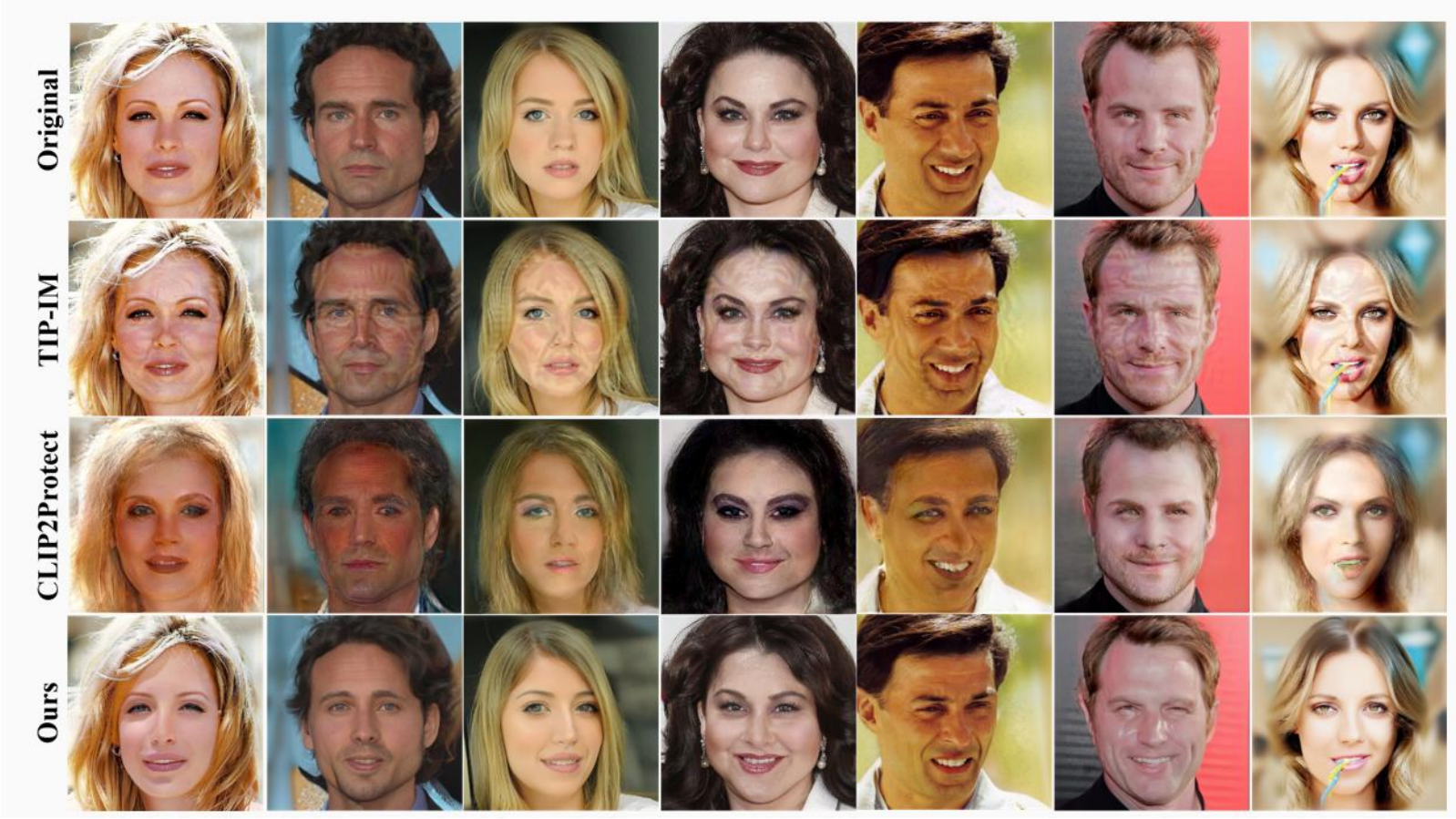
Face Recognition and Privacy Preservation

- **Deepfake Detection:** Generating deepfakes to train systems to detect them, which is crucial in combating misinformation and maintaining digital security.

Deepfake video detection: challenges and opportunities, Artificial Intelligence Review 2024

Face Recognition and Privacy Preservation

- **Anonymization:** Creating generative models that can obscure identities in images while retaining other key information, useful in privacy-preserving technologies.



HIDE: Hierarchical Identity Diffusers for Enhanced Face Privacy via Diffusion Models, T.-Y. Lu et al., under review

3D Model Generation and Reconstruction

- **3D Object Reconstruction:** Converting 2D images into 3D models for applications in virtual reality, robotics, and CAD.

Input images

*a [V] cat riding a bike**

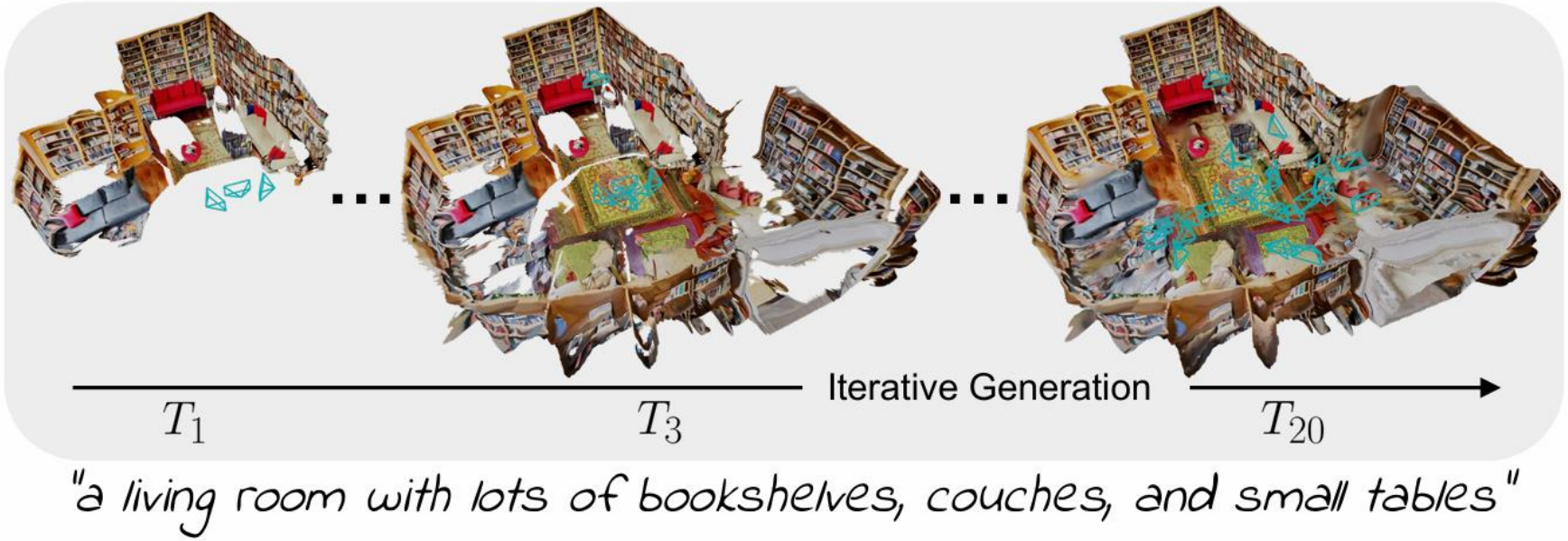
Input images

*a [V] dog running down the track**

Magic3D: High-Resolution Text-to-3D Content Creation, arXiv, 2023

3D Model Generation and Reconstruction

- **Interior Design:** Generating 3D models of furniture and room layouts based on user inputs for interior design planning.



(a) 3D Mesh Generation from Text

(b) Rendered Image + Mesh

Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models, ICCV'23

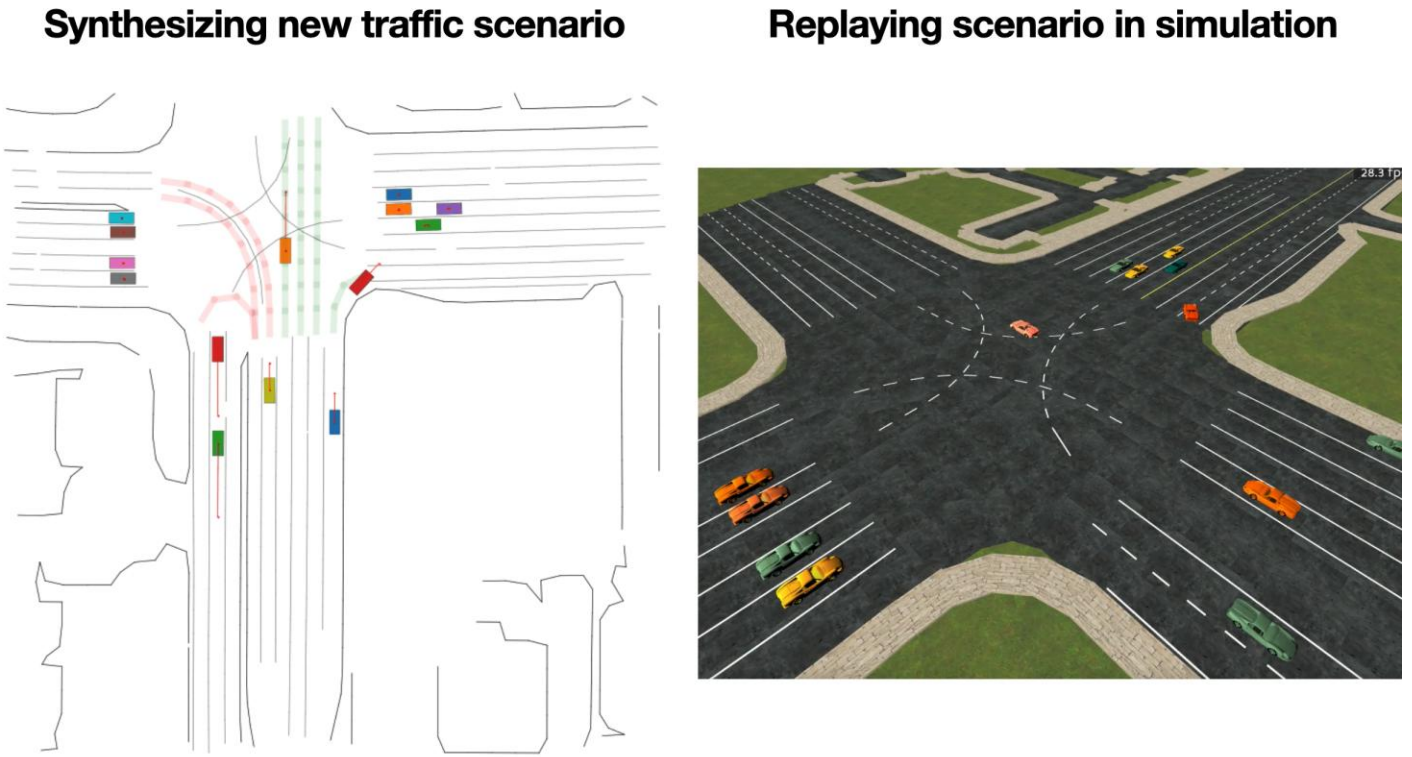
Video Generation and Manipulation

- **Video Editing:** Automating aspects of video creation, such as generating new frames to create slow-motion effects or removing unwanted objects.
- **Animated Content Creation:** Generating animated sequences for use in films, advertisements, and other media.

SORA demo video: https://youtu.be/HK6y8DAPN_0

Autonomous Vehicles

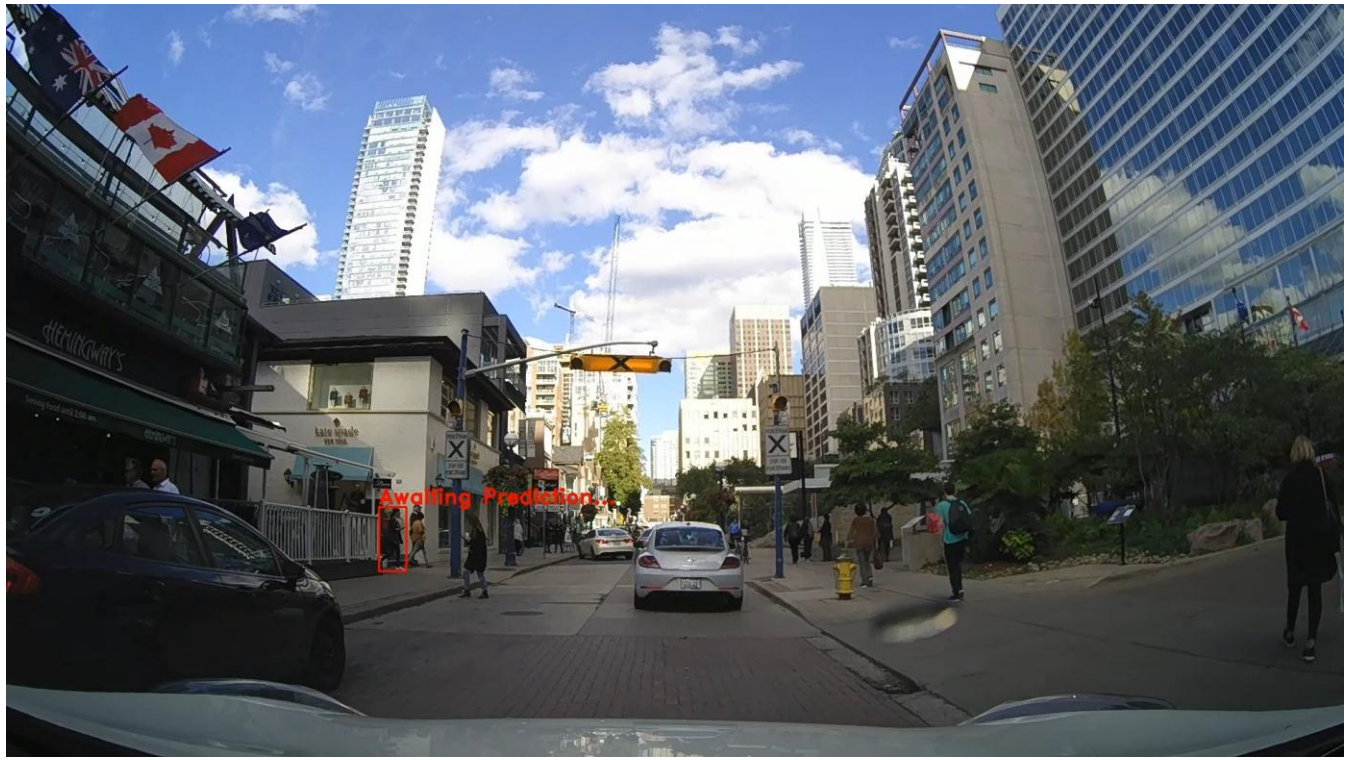
- **Simulated Training Environments:** Generating realistic driving scenarios to train autonomous vehicles in varied conditions without real-world risks.

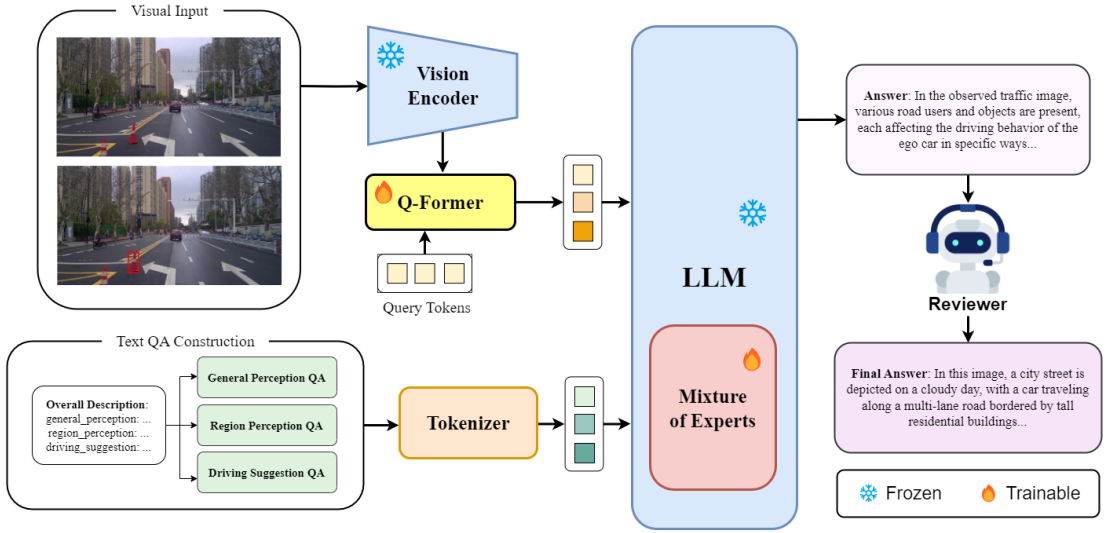


TrafficGen: Learning to Generate Diverse and Realistic Traffic Scenarios, ICRA 2023

Autonomous Vehicles

- **Scene Understanding:** Enhancing the vehicle's understanding of its environment through generated scenes that improve perception models.





A Lightweight Vision-Language Model Pipeline for Corner-Case Scene Understanding in Autonomous Driving, Y. Cheng, M.-H. Chen, S.-H. Lai, ECCV Workshop, 2024

Socially Responsible AI (SRA)

- **Components of the SRA**
  1. **Fairness:** AI should treat all individuals equitably, considering social, racial, and gender justice.
  2. **Transparency:** AI systems must explain how they make decisions to foster trust and accountability.
  3. **Safety:** Ensuring AI systems are robust and secure, minimizing risks of failure or exploitation.
- **User Feedback**
  Continuous feedback loops are essential for improving SRAs. User inputs help refine AI systems to meet ethical and societal standards.

Conclusion

- Applications of Generative AI technologies to multimedia content generation are rapidly increasing.
- Appropriate evaluation and comparison of generative AI models is required.
- Very high-quality image/video content can be generated with SOTA techniques
- Mis-information and deepfake detection is a major challenge today
- How to ensure proper use of generative AI is an urgent topic in our society.