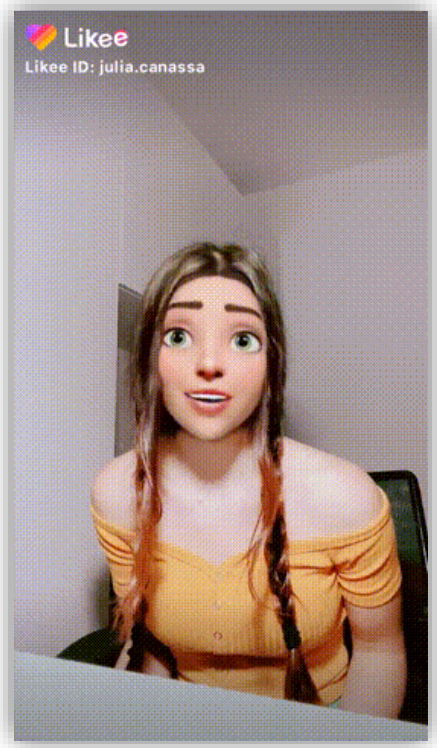
Towards High-fidelity Generative Modeling: From 2D Image Generation to 3D Character Rendering

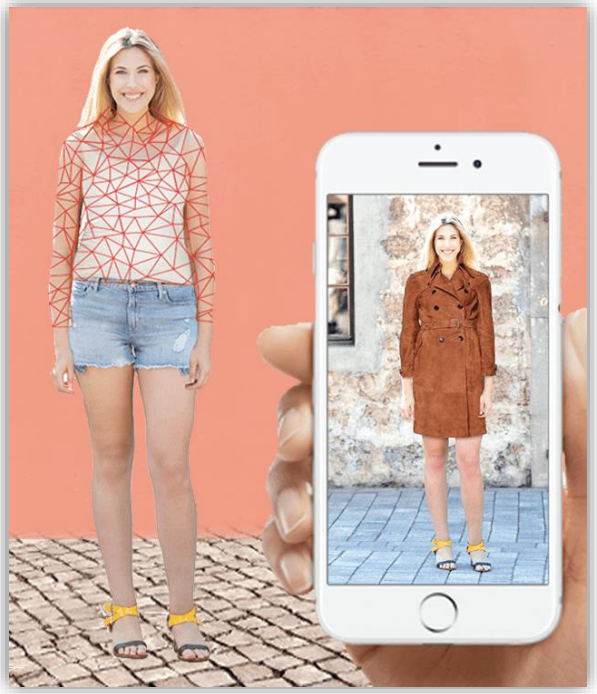
Bo Zhang
Microsoft Research Asia

Empower creativity with AI

Old photo restoration



Face cartoonization



Virtual try-on

Interactive creation

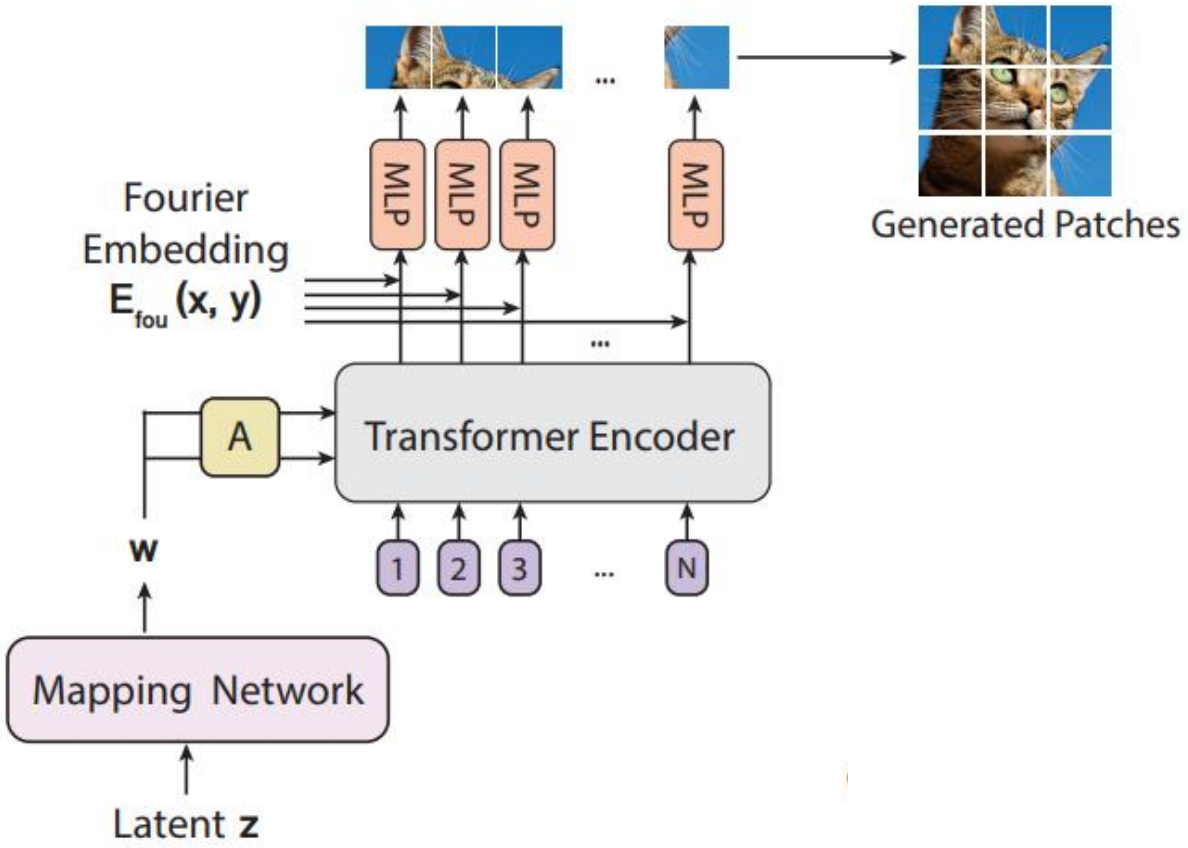
Agenda

- Transformer starts to dominate generative tasks
- The emergence of diffusion model
  - VQ-diffusion
  - Pretrained generative prior
- Towards realistic avatar
  - GAN-based 2D rendering
  - Controllable 3D character
- Summary

Long-range dependencies is crucial to image synthesis

Can we reproduce the success of vision transformer in discriminative tasks to generation?

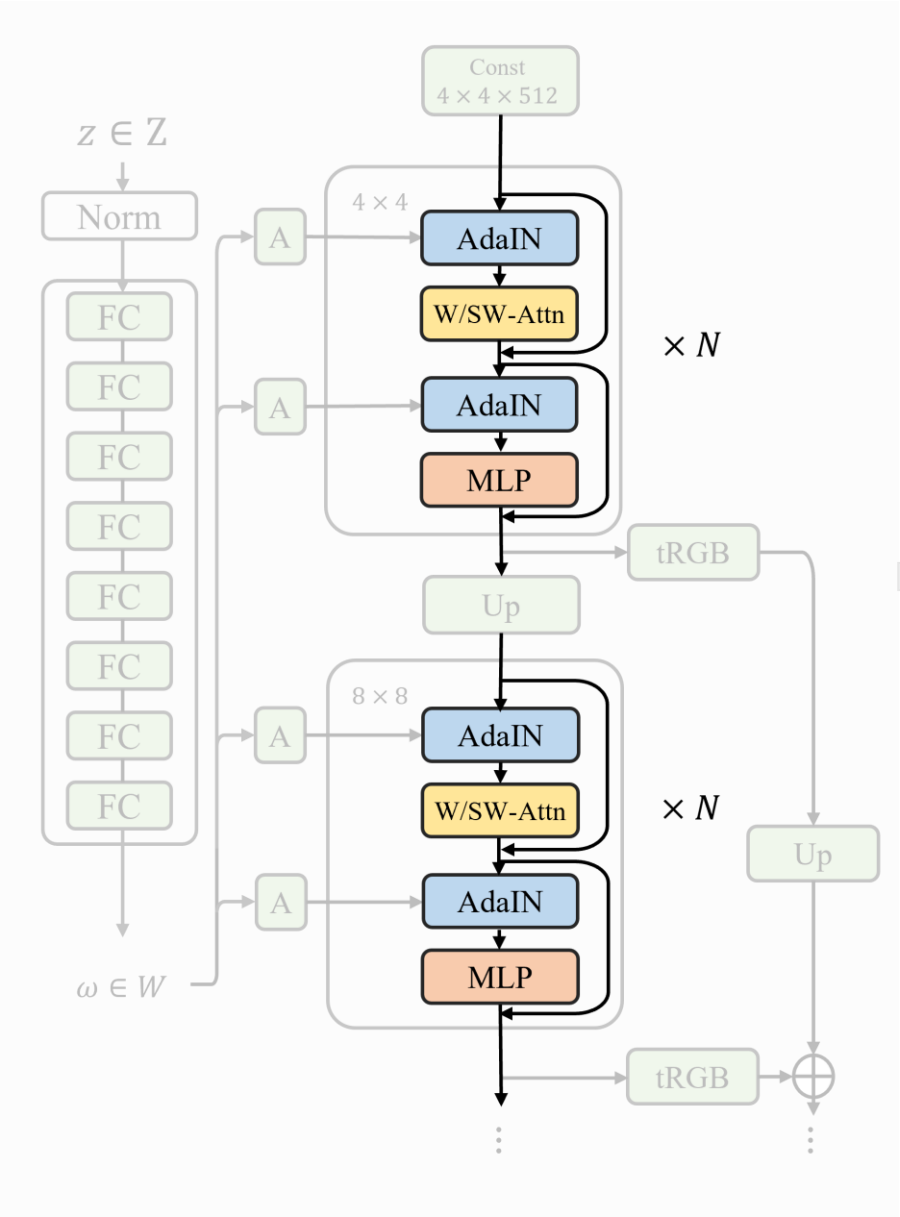
Concurrent work – ViTGAN by Google

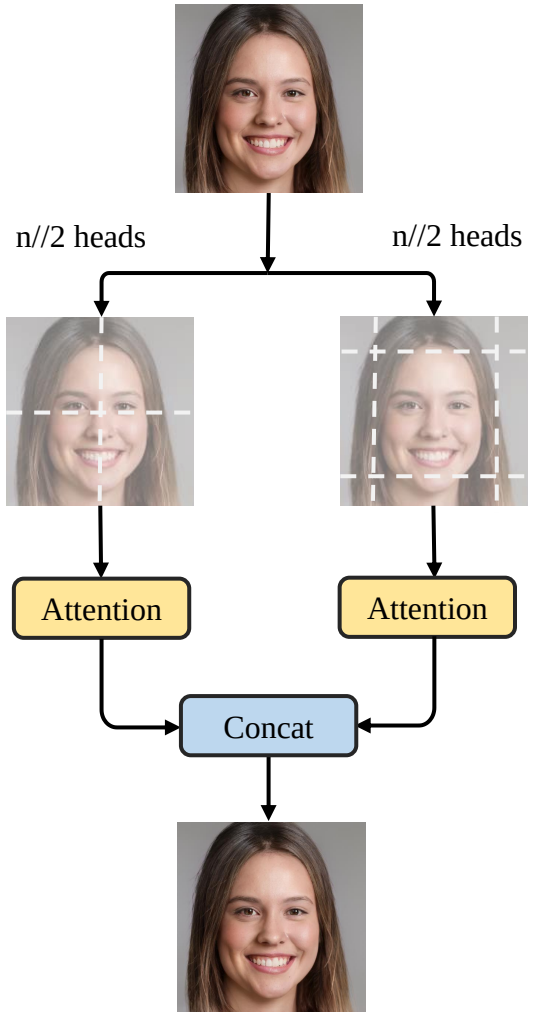


ViTGAN Generator

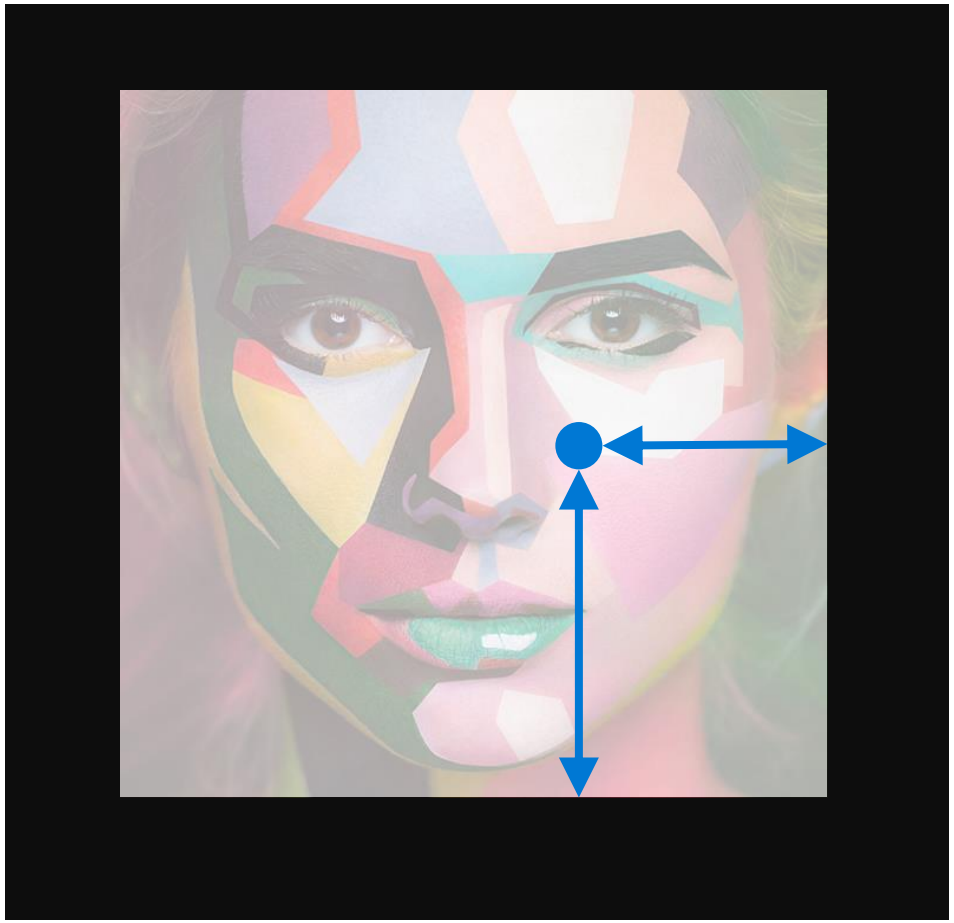
Instable adversarial training
Limited resolution: 64x64

StyleSwin (CVPR 2022)

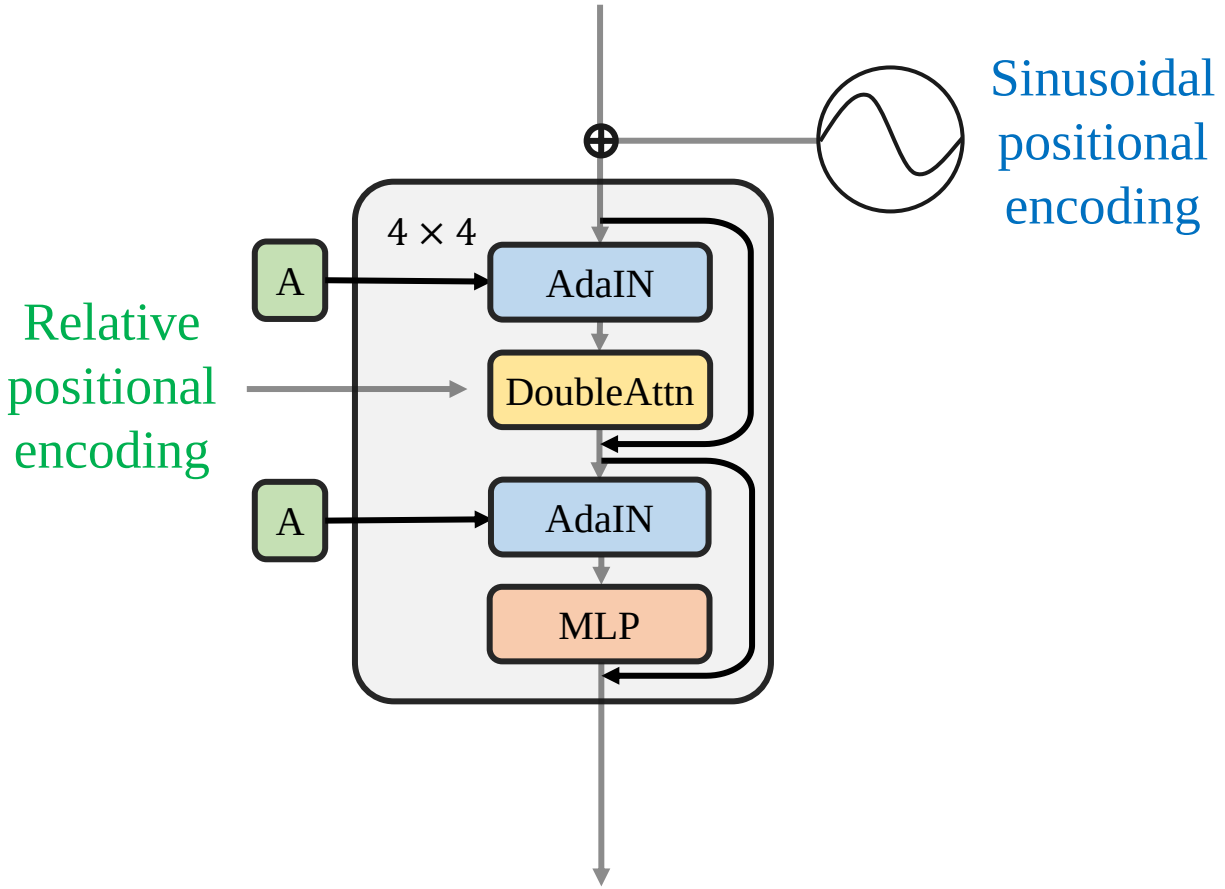




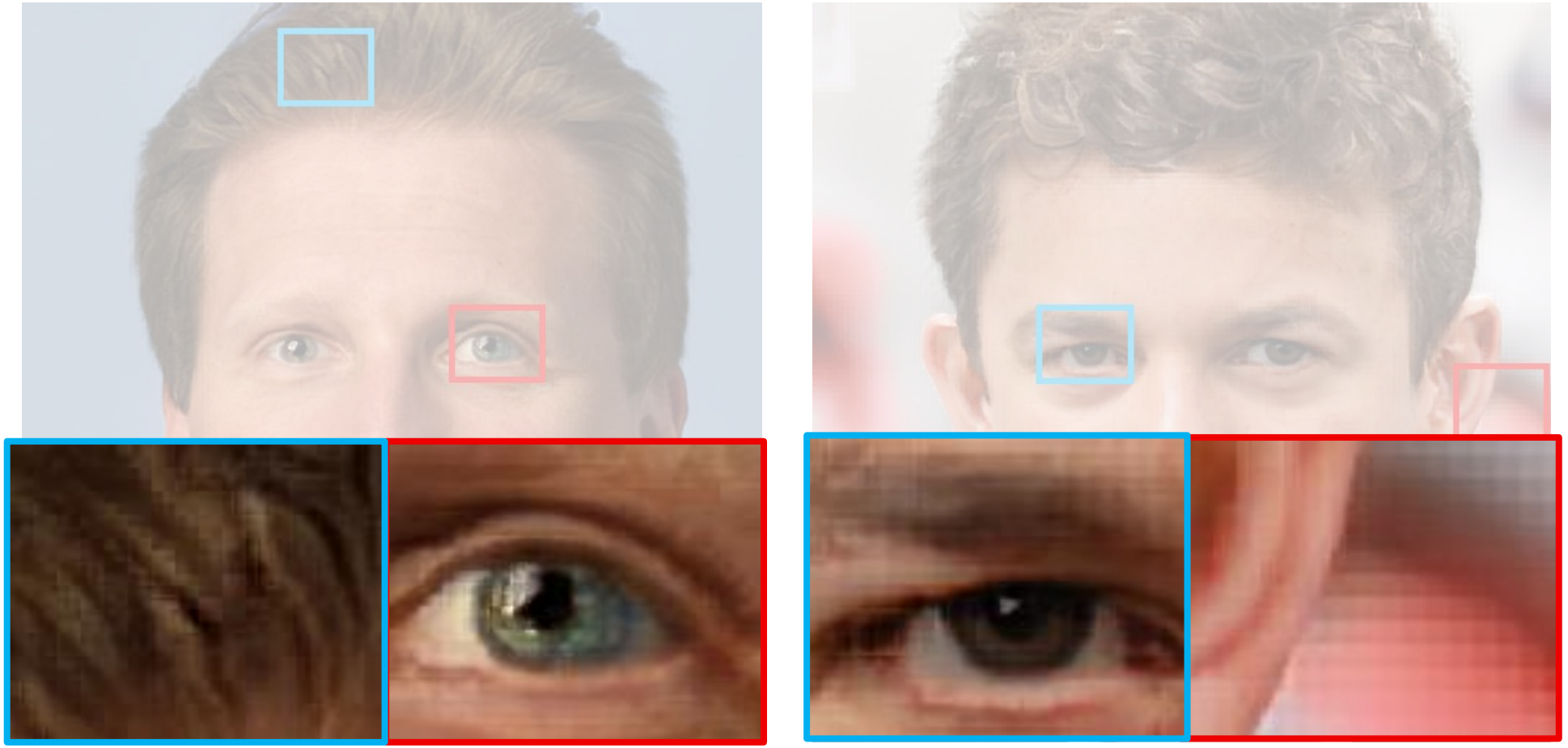
Position information is missing when replacing Convs



Conv Nets use zero padding to locate pixels

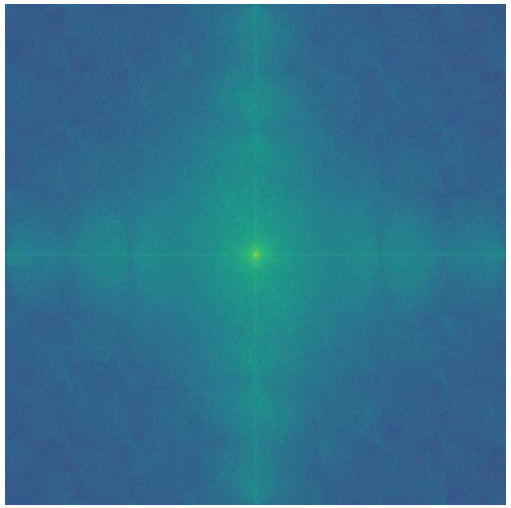


Blocking artifacts due to shifted window processing

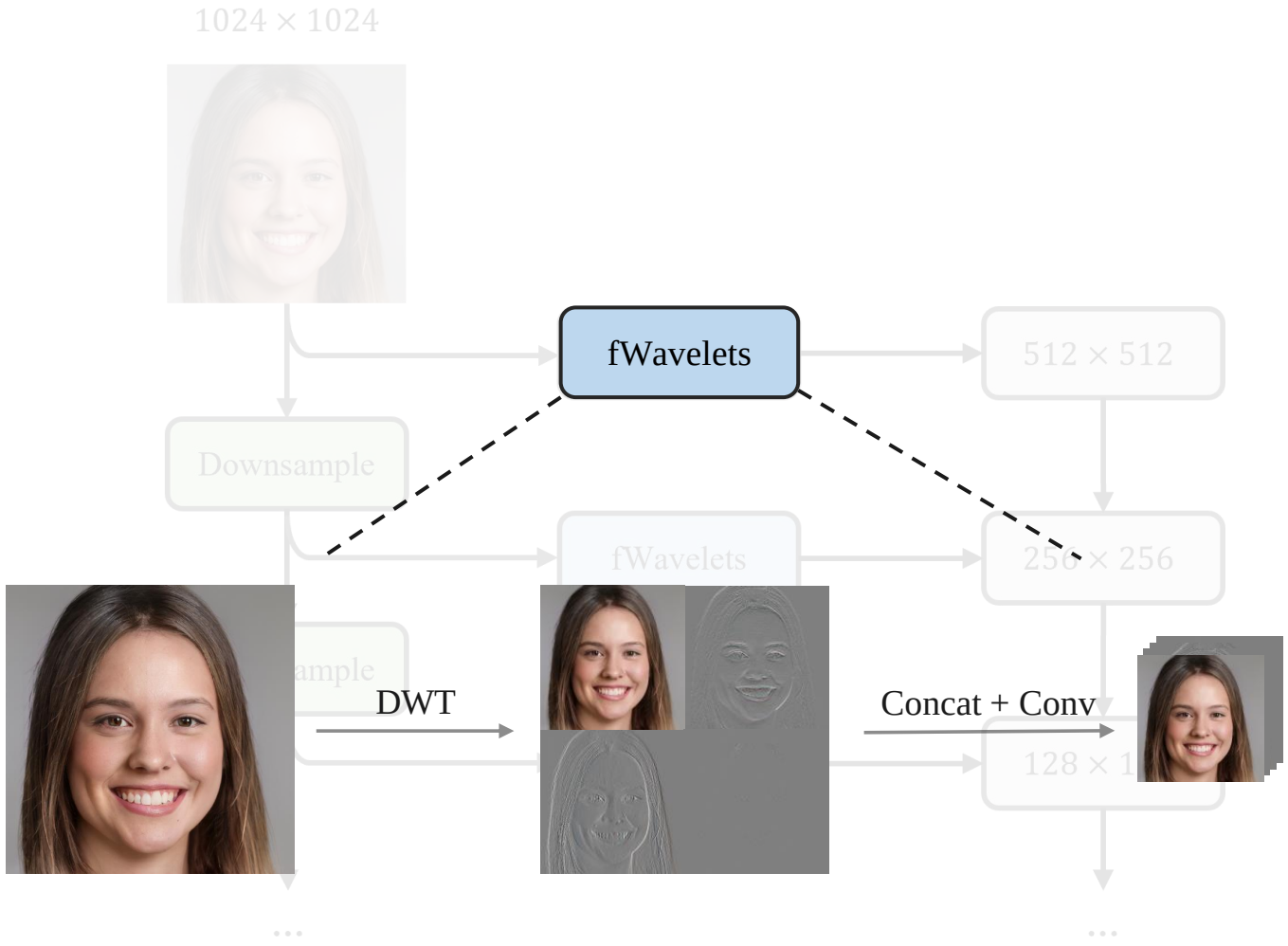


Wavelet discriminator for blocking artifact suppression

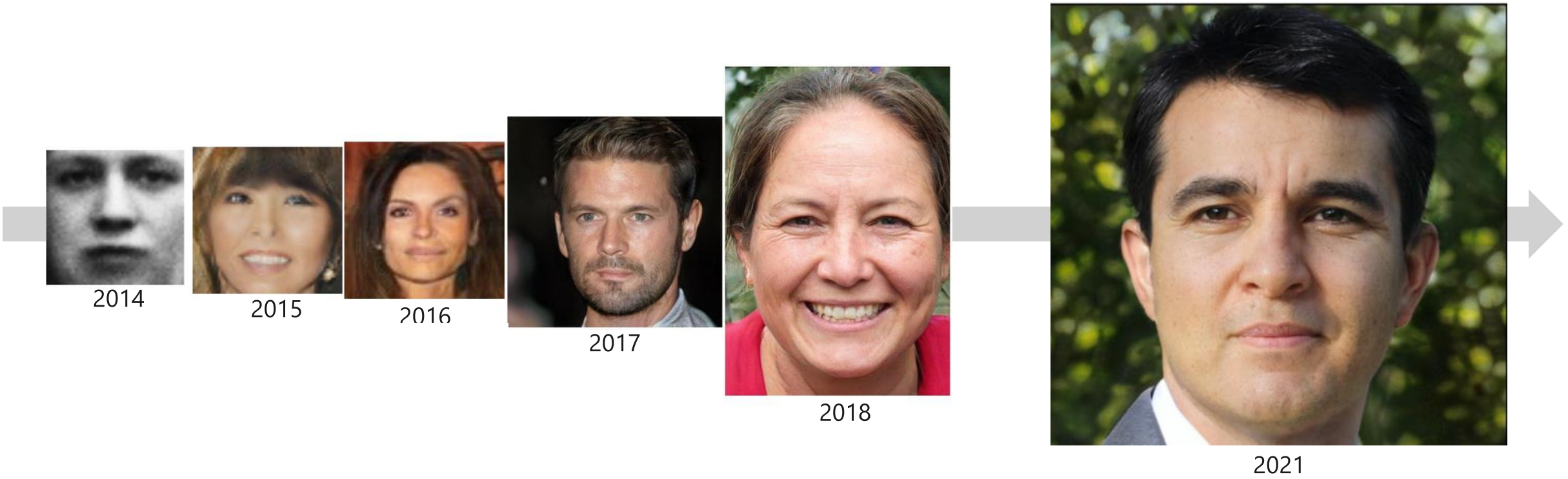
Image with blocking artifacts



Fourier spectrum



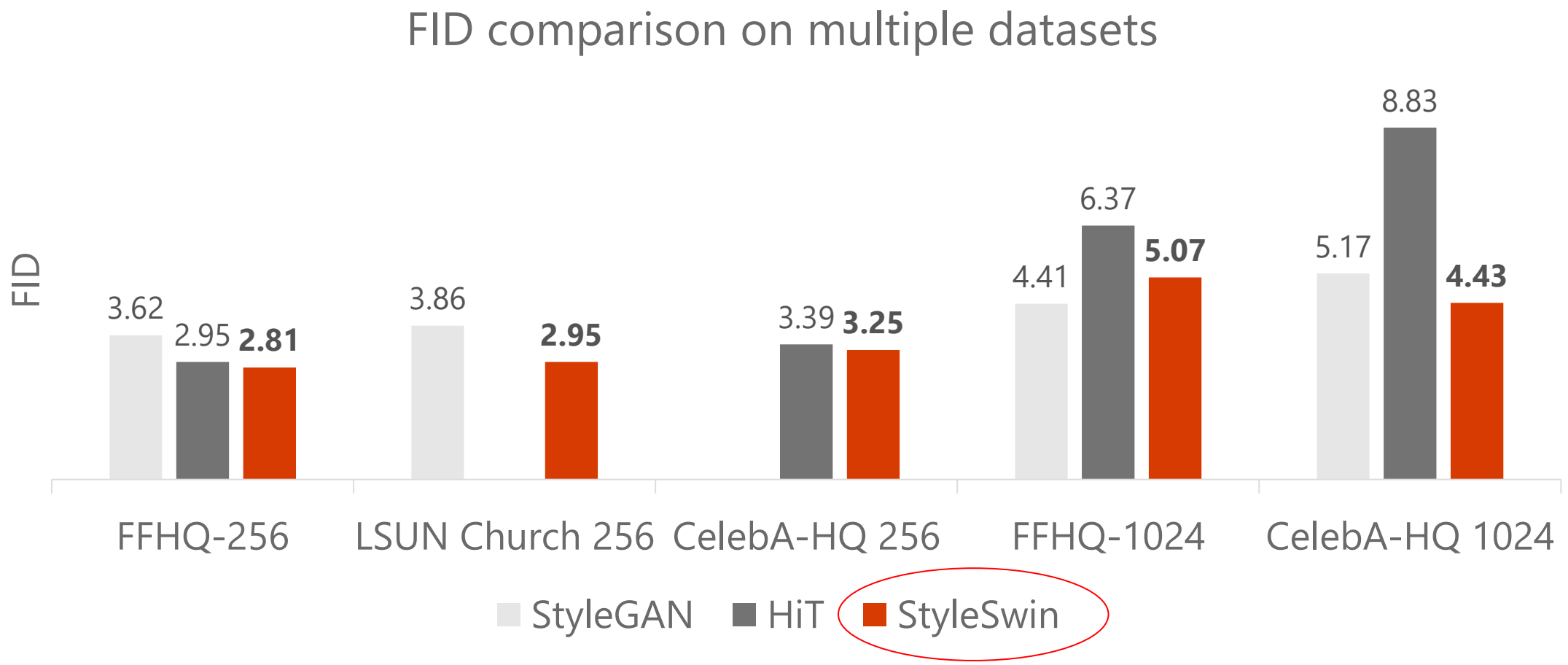
Progress of face generation



New state-of-the-art of year 2022: StyleSwin

FID score on FFHQ dataset: 3.62 (StyleGAN) → 2.81 (StyleSwin)

Comparison with state of the arts

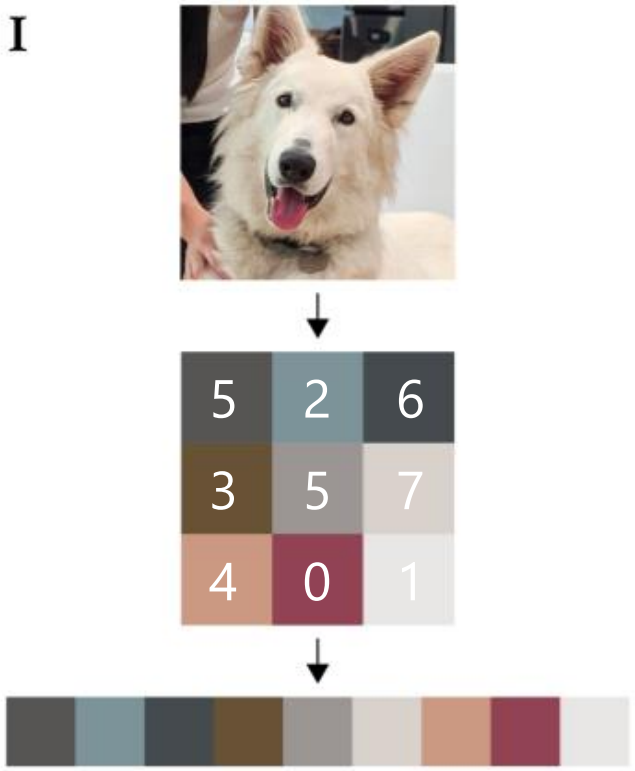


Part II: The emergence of diffusion model

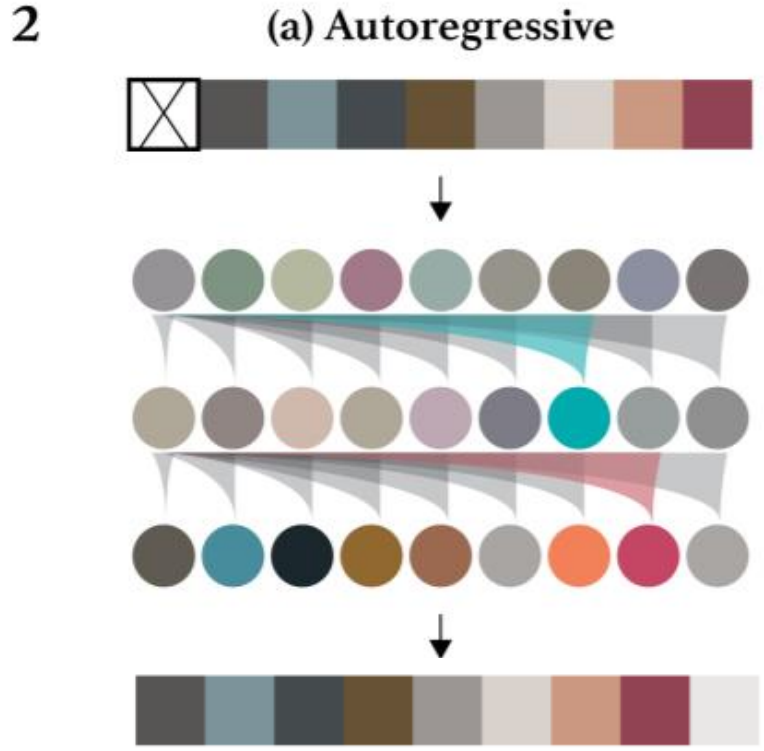
The autoregressive starts in 2021

I

The autoregressive starts in 2021



Discrete image tokens



Text-to-token synthesis

Language model inspired
Just like GPT 3

Taming Transformers for High-Resolution Image Synthesis, CVPR 2021
Zero-Shot Text-to-Image Generation, ICML 2021

OpenAI DALLE

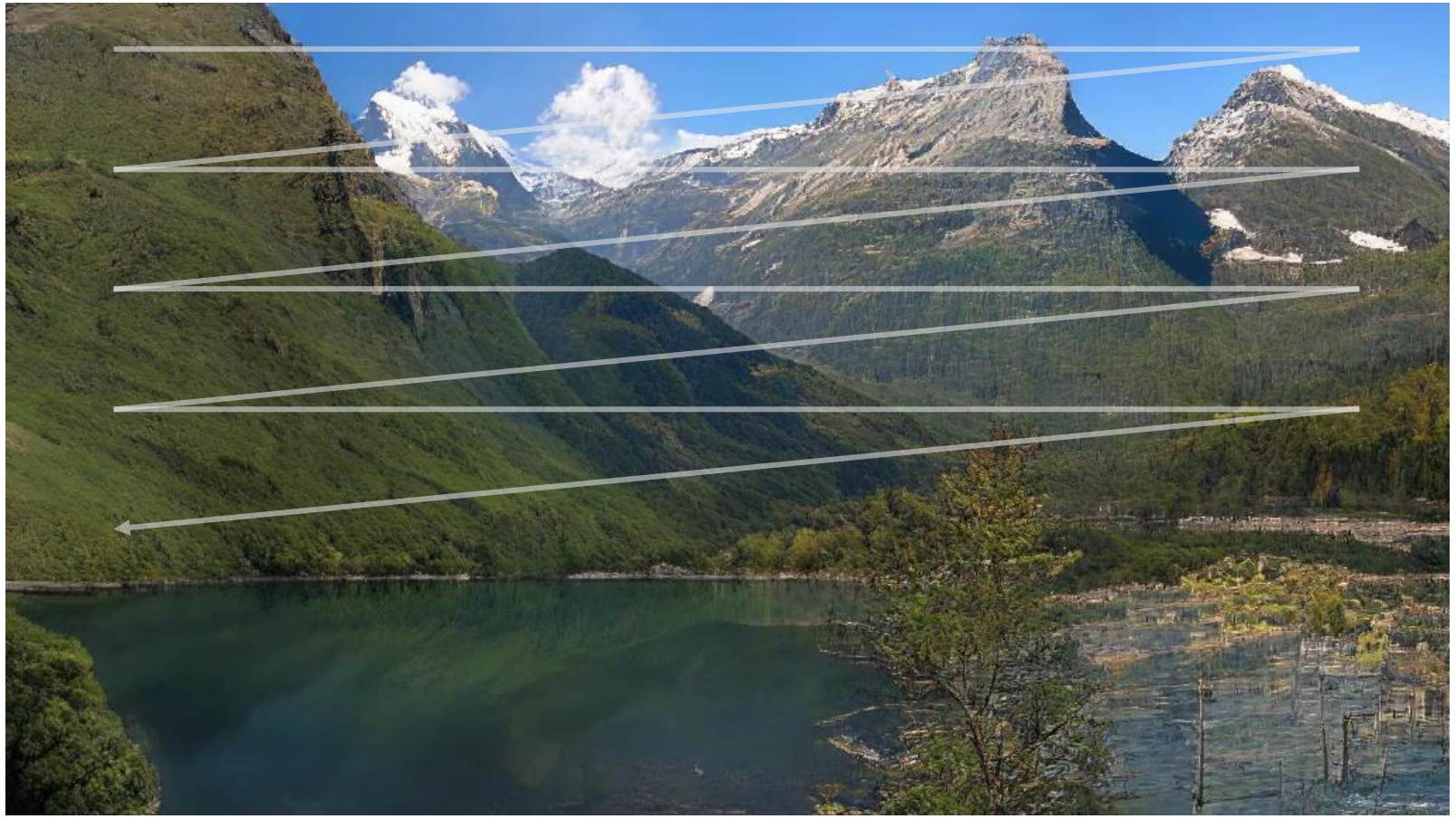
12-billion parameters

250-million training pairs

"An illustration of a baby hedgehog in a Christmas sweater walking a dog"

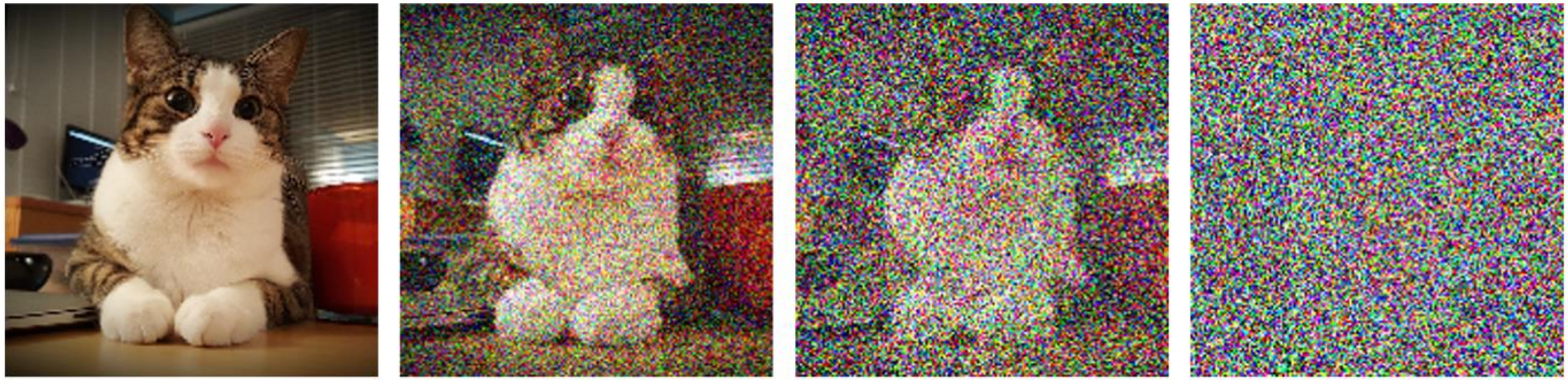
Issues of auto-regressive generation

- Slow to inference
- Directional bias
- Error accumulation



Motivated by “Diffusion model beats GAN”

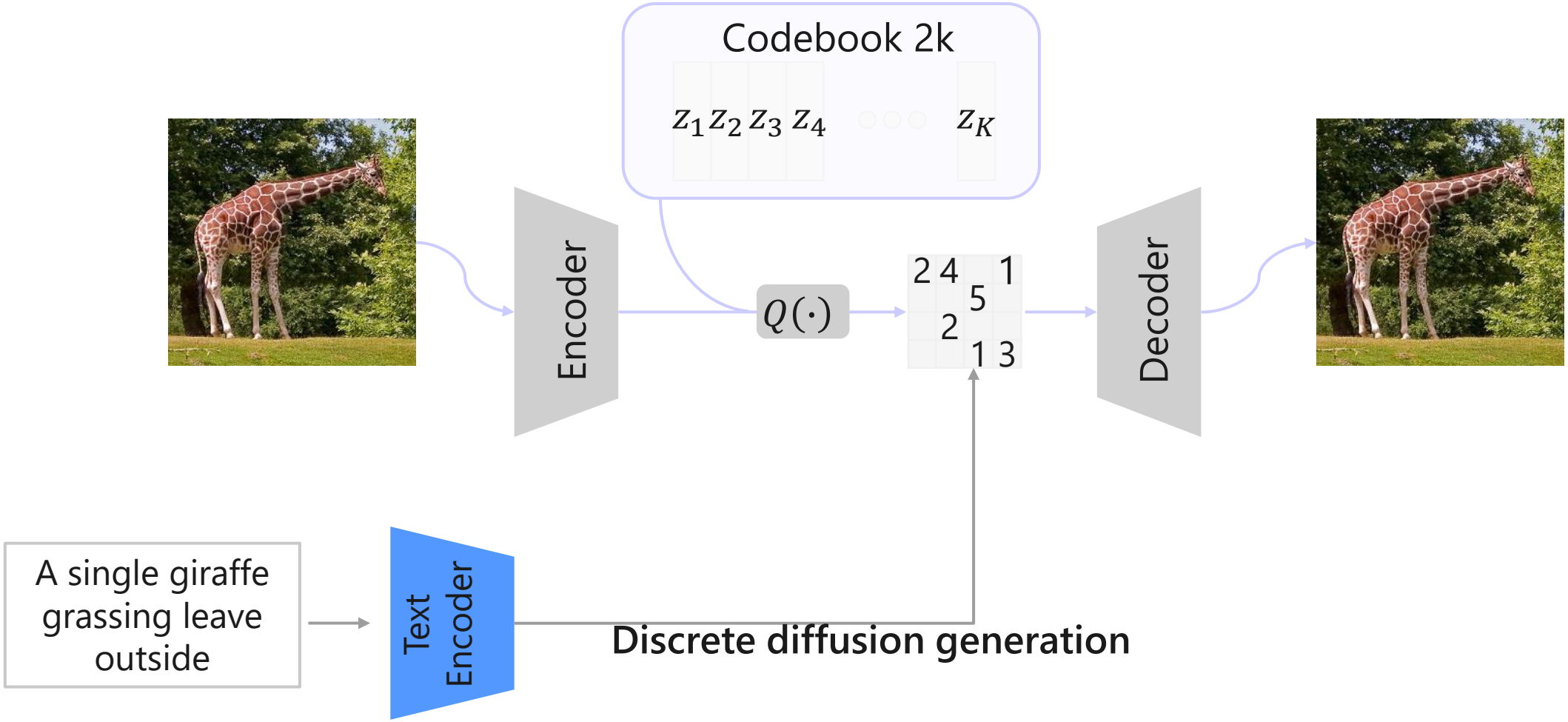
Feedforward diffusion process



Generative reverse process

- The denoising model takes the noisy input x_t and predicts the added noise

Our VQ-Diffusion (CVPR 2022, oral)



15x speedup comparing to DALLE!

Comparison with state-of-the-arts

- Our method is trained using a subset of Conceptual Captions (~7M/15M) and LAION-400M datasets (~44M/400M)

	MSCOCO	CUB-200	Oxford-102
StackGAN [70]	74.05	51.89	55.28
StackGAN++ [71]	81.59	15.30	48.68
EFF-T2I [60]	-	11.17	16.47
SEGAN [61]	32.28	18.17	-
AttnGAN [67]	35.49	23.98	-
DM-GAN [73]	32.64	16.09	-
DF-GAN [63]	21.42	14.81	-
DAE-GAN [51]	28.12	15.19	-
12B parameters DALLE [48]	27.50	56.10	-
Cogview [13]	27.10	-	-
34M parameters VQ-Diffusion-S	30.17	12.97	14.95
VQ-Diffusion-B	19.75	11.94	14.88
370M parameters VQ-Diffusion-F	**13.86**	**10.32**	**14.10**

Text-to-image synthesis result

The sunset on the beach is wonderful

A picture of a very tall stop sign

Two girls in cartoon style

A man with beard in 1920s

A mountain near the lake

The face of Bill Gates

A picture of some food in the plate

Two smiling beautiful ladies are standing together

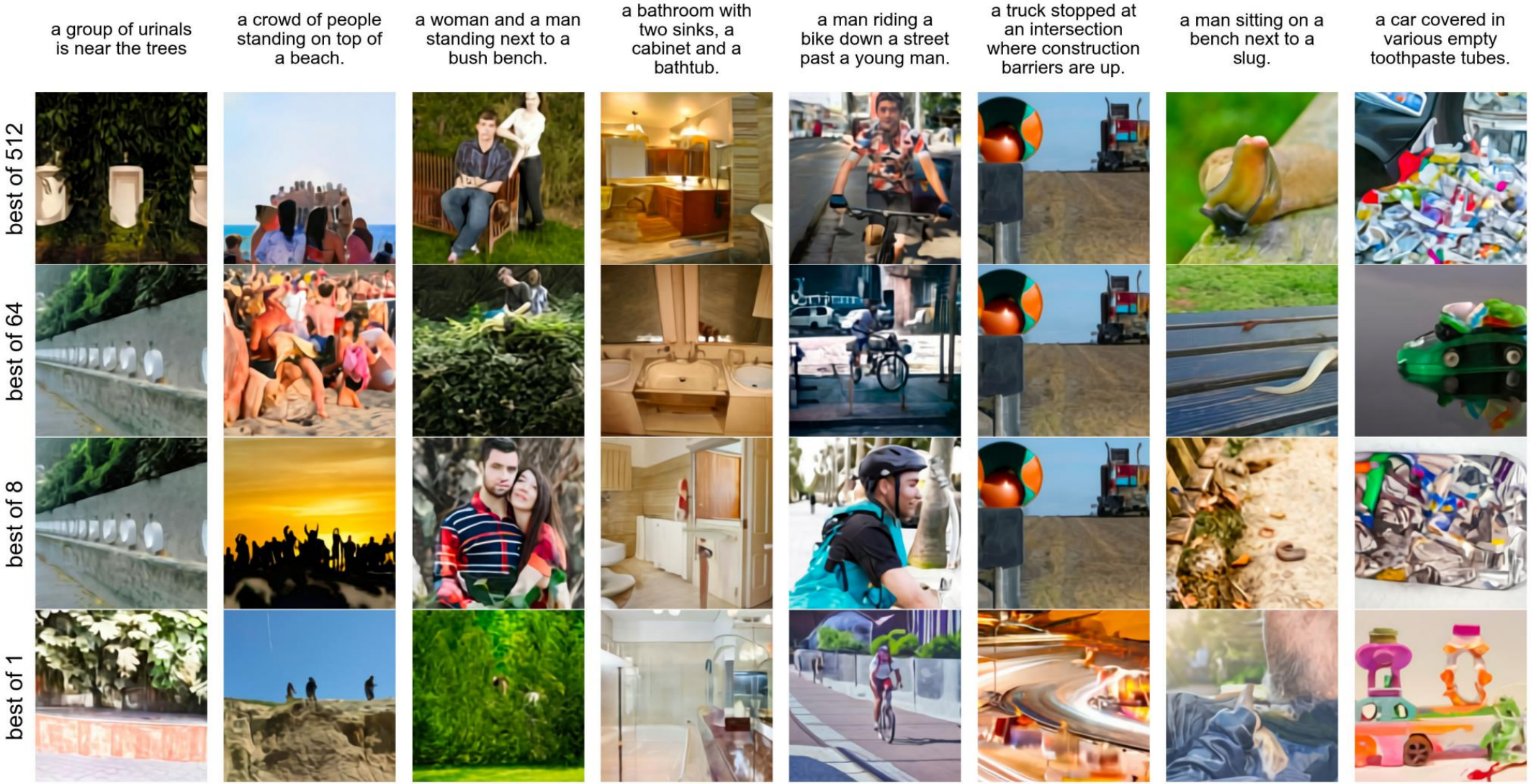
A red bus is driving on the road

A very cute giraffe making a funny face

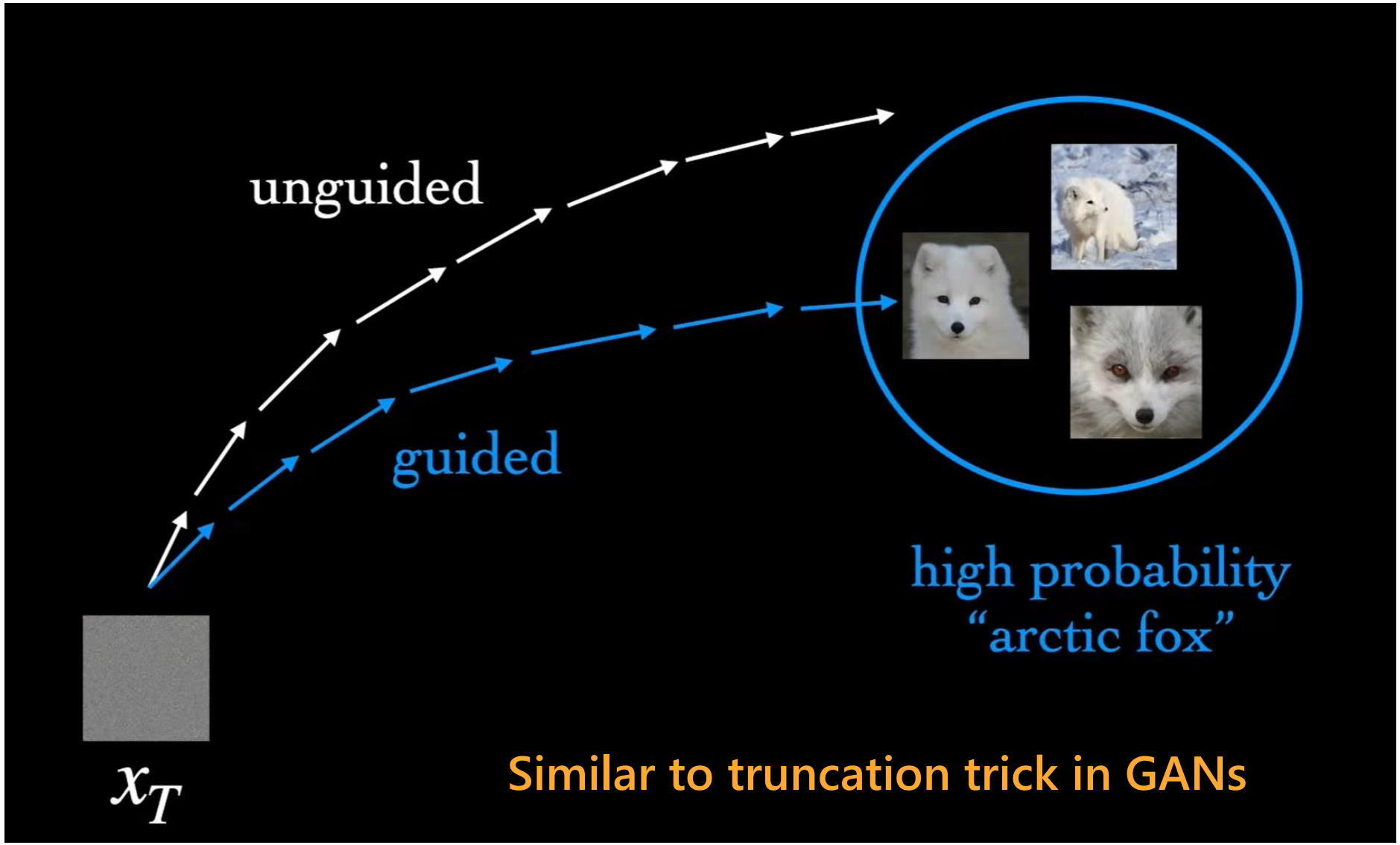
Black and white icon of man and woman

A woman with curly hairs and brown skin

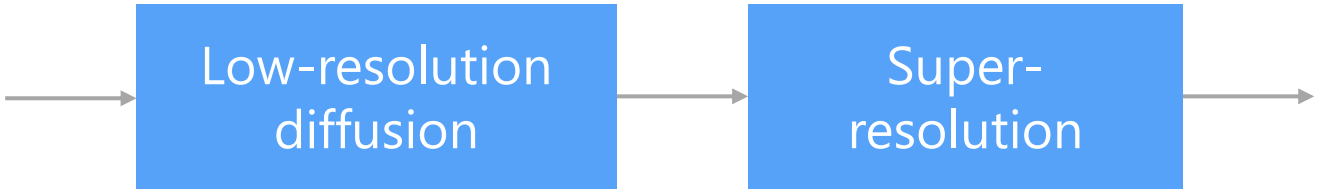
DALLE result recap



Improved VQ-diffusion (1): classifier-free guidance



Improved VQ-diffusion (2): hierarchical generation

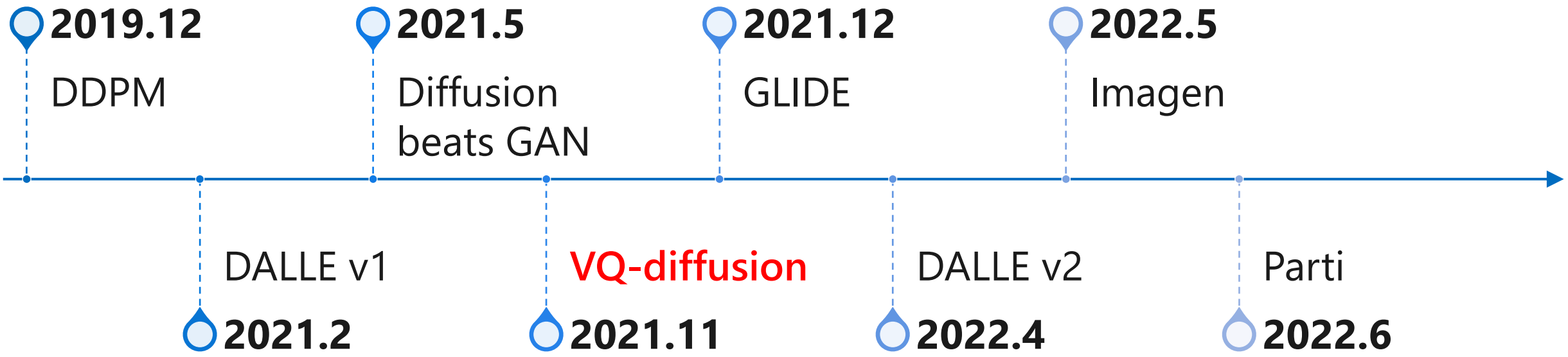


Teddy bear playing in the pool

Cactus with a Mexican hat on top of it

Fruits in plate, strawberry and blueberry

Milestone

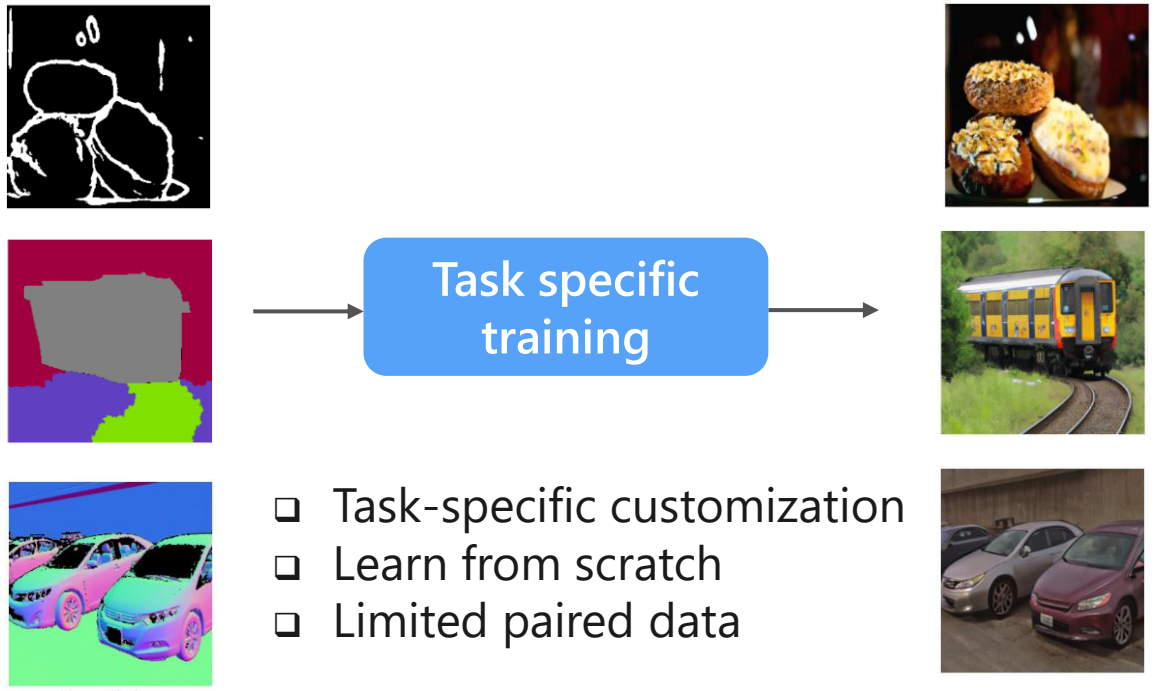


Discrete space diffusion enables more possibilities for multimodality generation

Can we derive a universal generative prior?

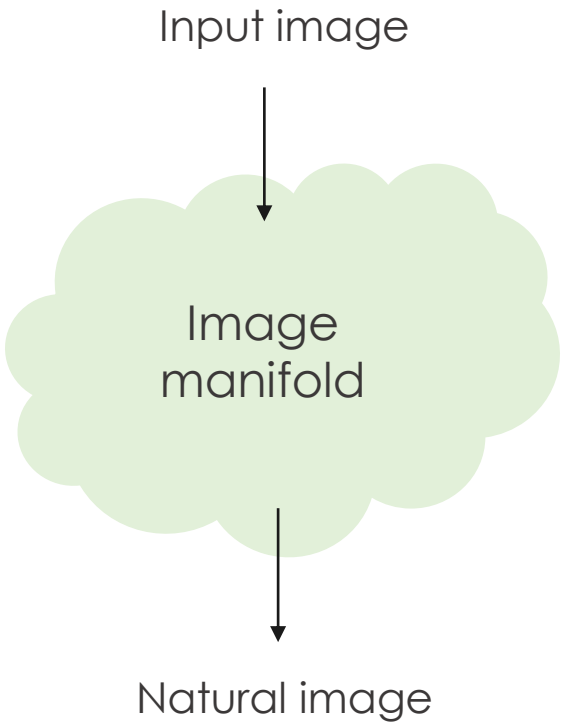
- Directly learn the domain mapping

$$\min f_\theta : D_A \rightarrow D_B$$



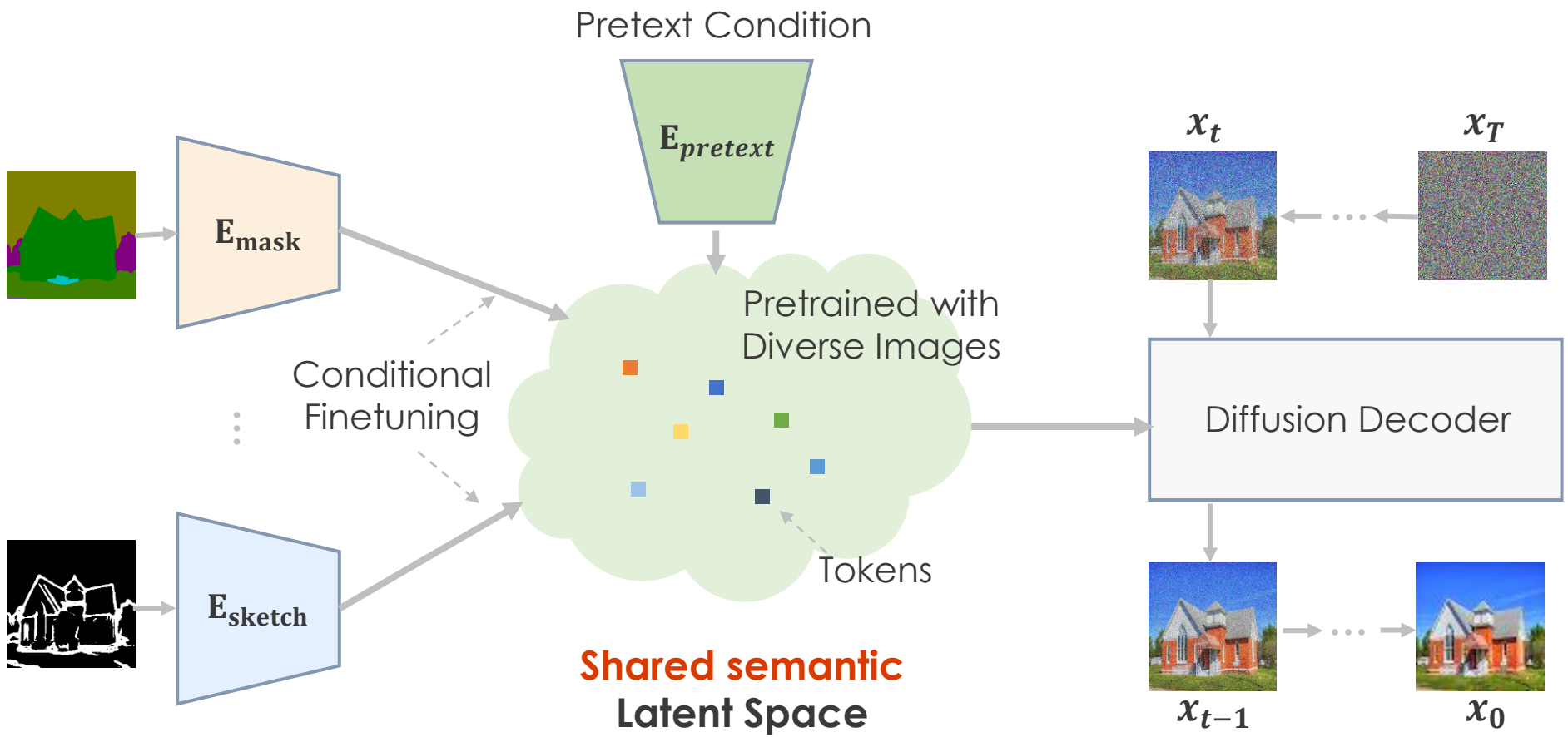
- Task-specific customization
- Learn from scratch
- Limited paired data

- Sample from the image manifold and choose the sample that mostly conforms to the input semantics

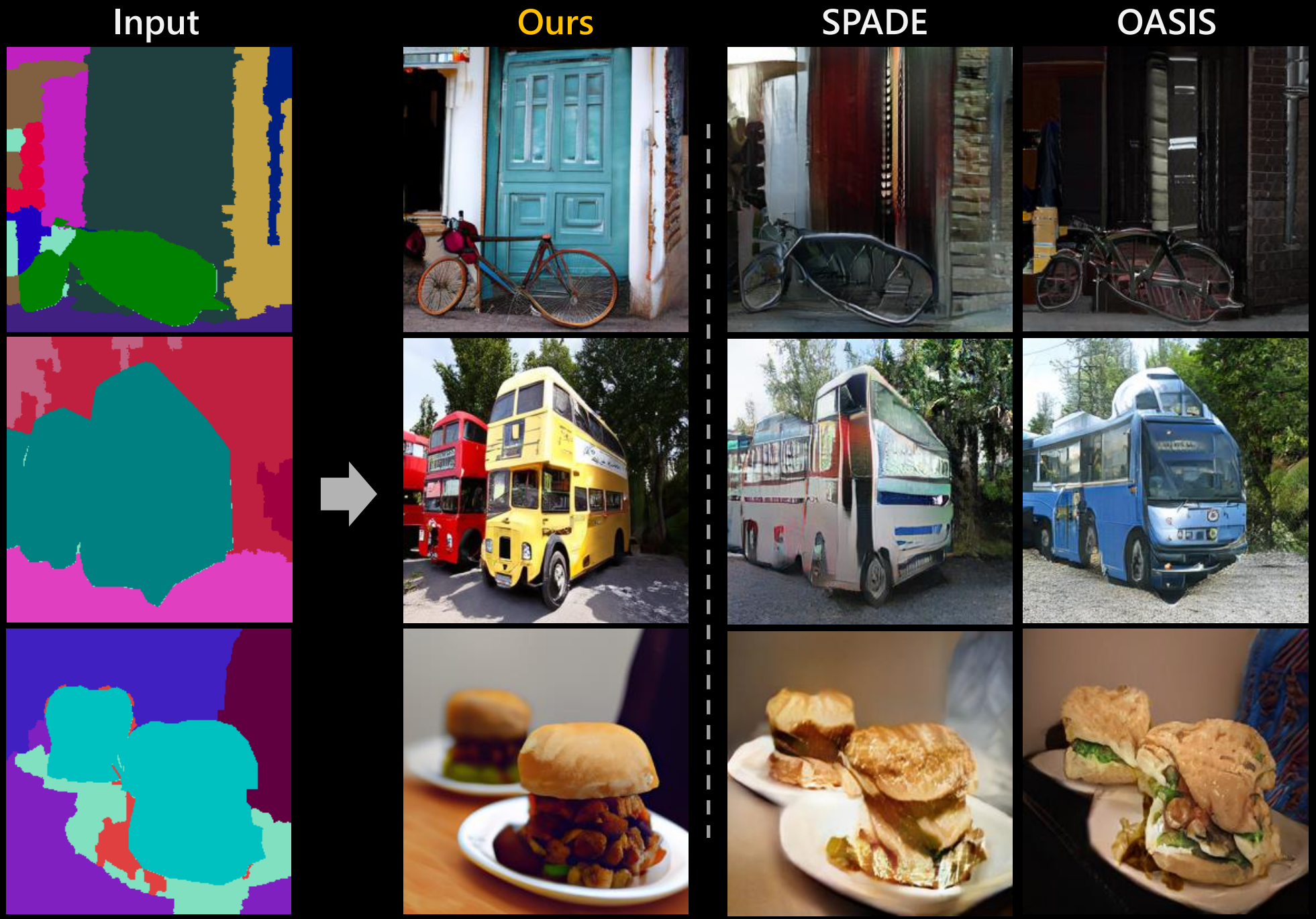


Pretraining is All You Need for Image-to-Image Translation, arXiv 2022

Pretraining-based image-to-image translation (PITI)



Sparse coding, low-rank, generative prior, …, -> Universal generative prior

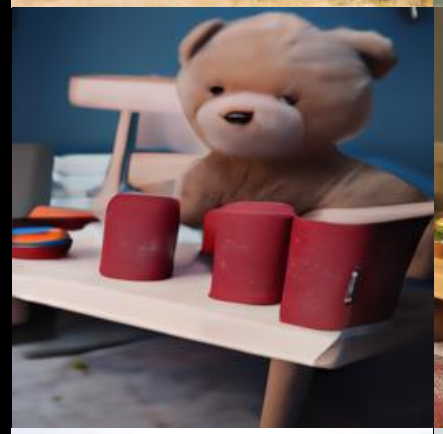

Input
Ours
SPADE
OASIS

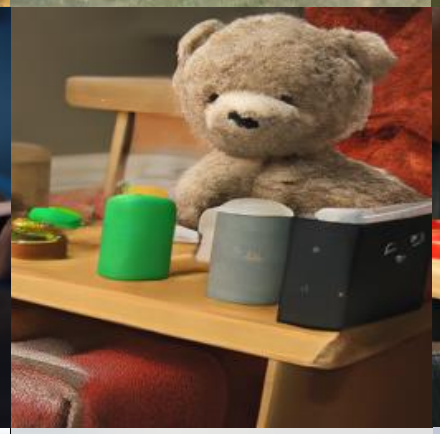
Input

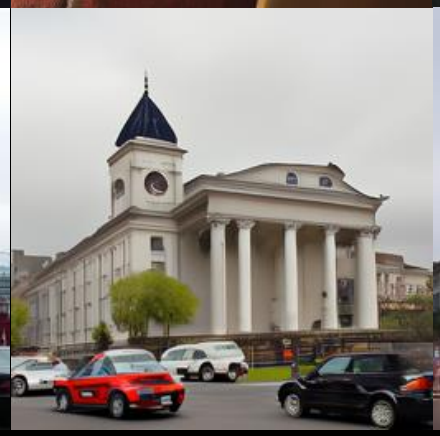
Sample1

Sample2

Sample3







Pretraining is all you need for I2I translation!

Realistic 3D avatar generation

Traditional avatar generation is inadequate

Rendered by Unity, unreal engine, …

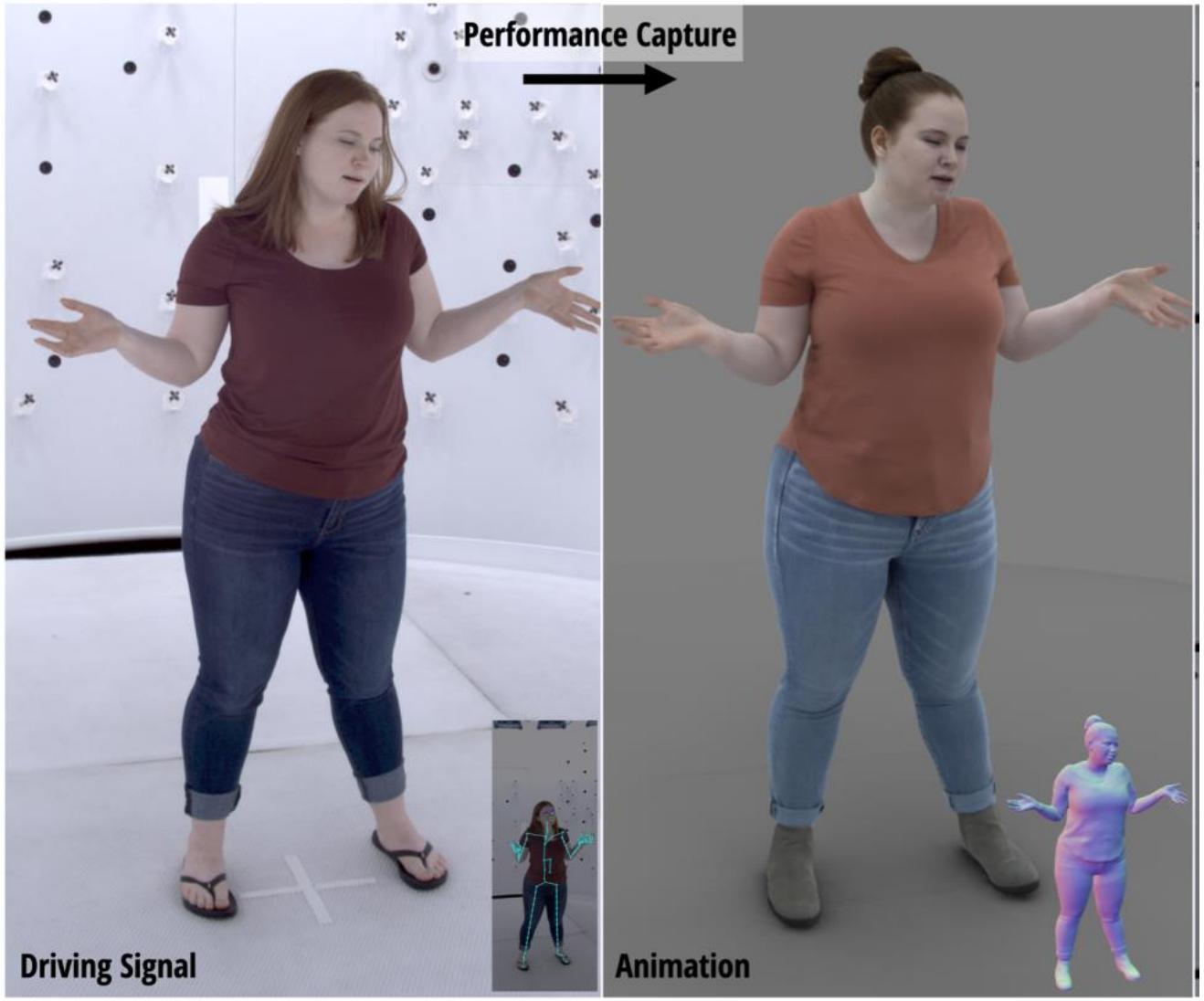
user

avatar

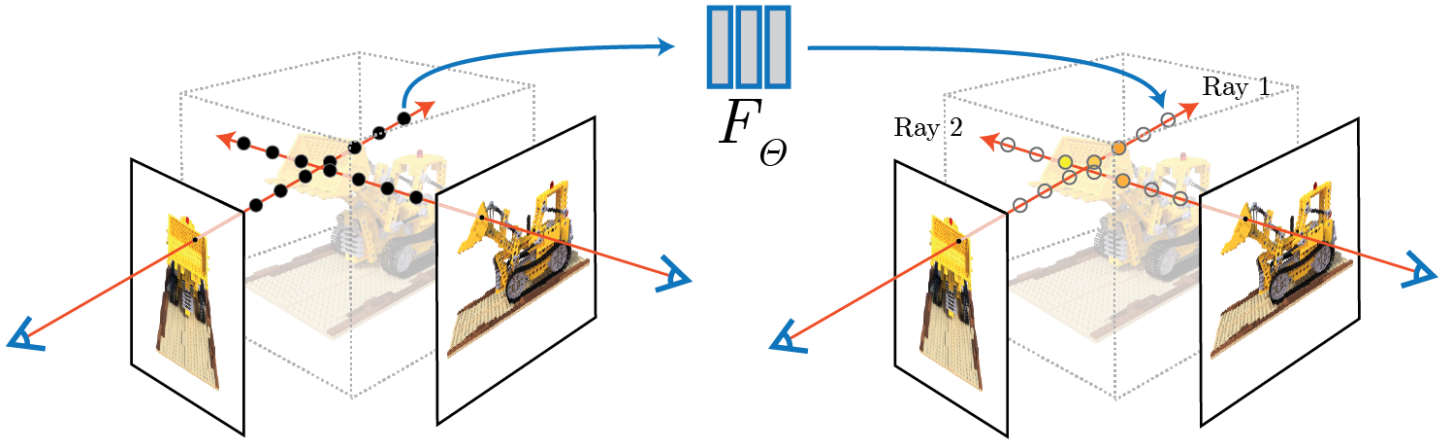
We need personalized avatar!

Audio-driven avatar

Towards 3D animatable avatar



Free-viewpoint rendering - NeRF

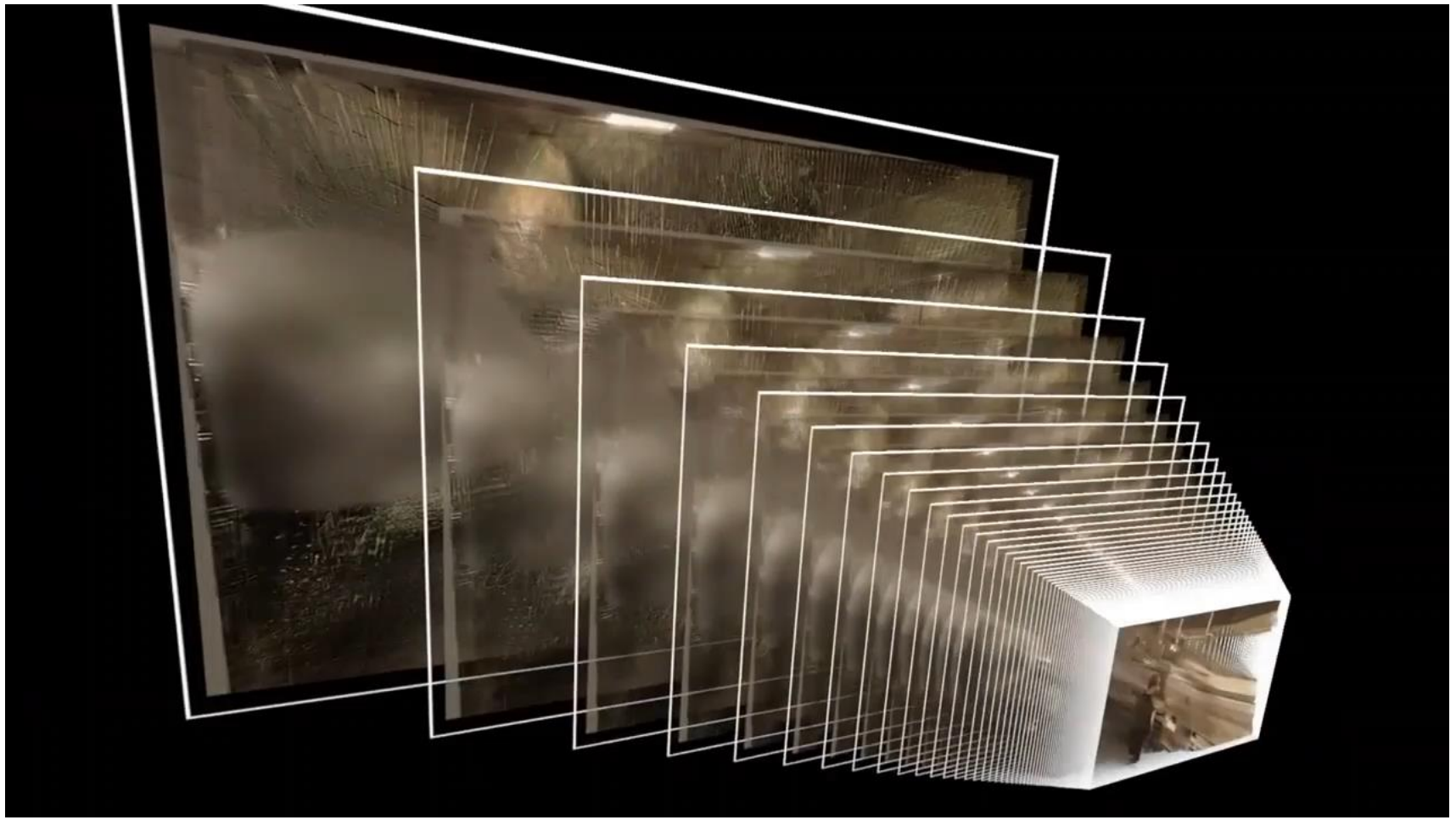


- Simply memorize the scene
- Static scene only
- Not controllable
- Slow to render ~20s

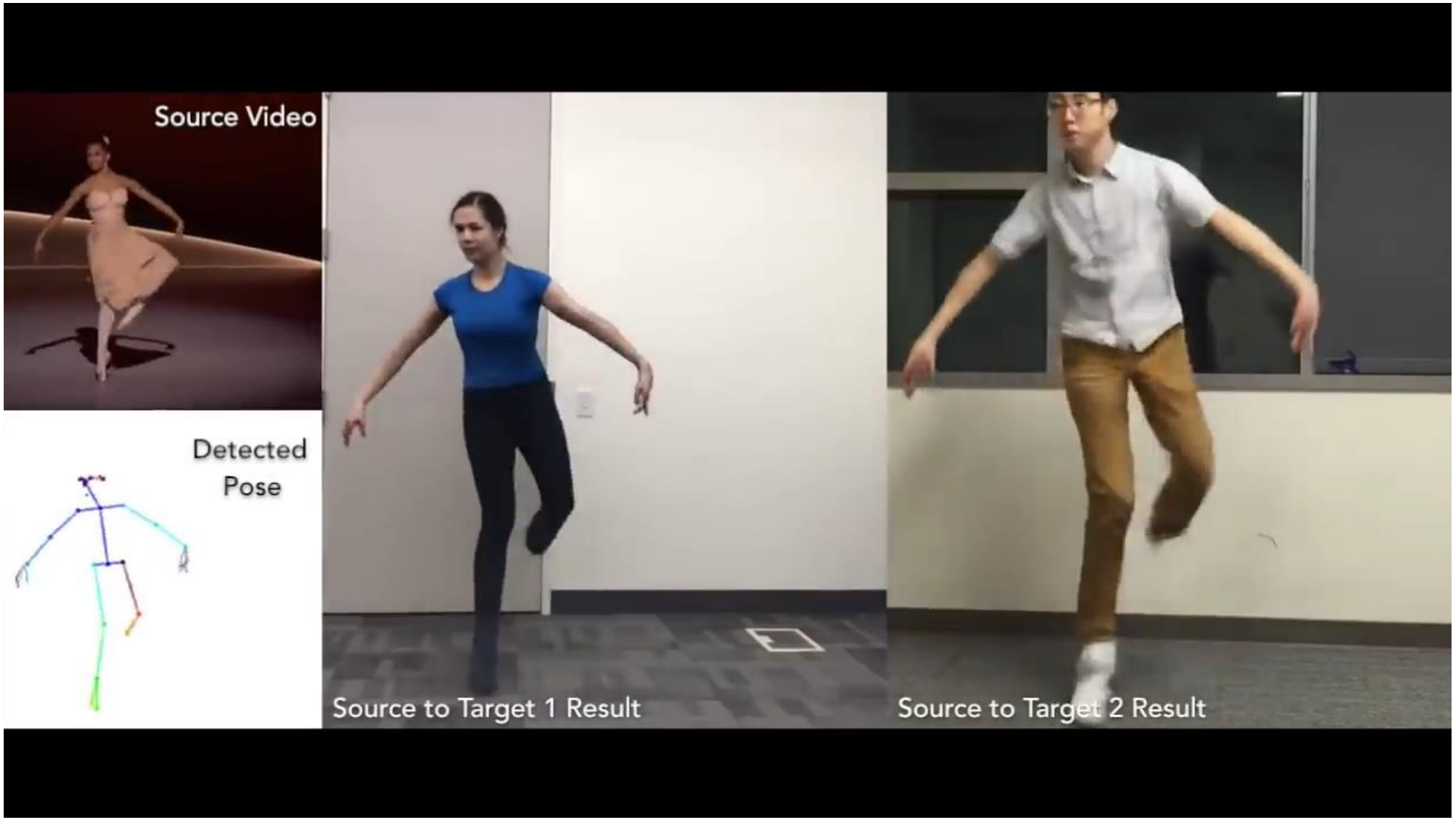
Can we synthesize a photorealistic character with controllable viewpoints and body poses in real-time?

Multiplane image representation (MPI)

- Good enough for front-facing scenes



Inspiration from pose-to-human synthesis



Video from "*Everybody Dance Now*"

Pose-guided Multiplane image synthesis (ECCV 2022)

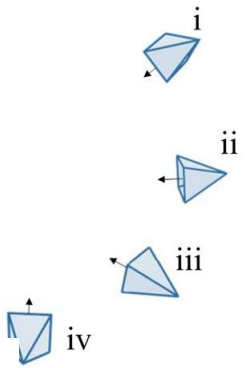
- Formulate the problem as **pose-to-MPI translation**
- Conv is fast, and generalizable

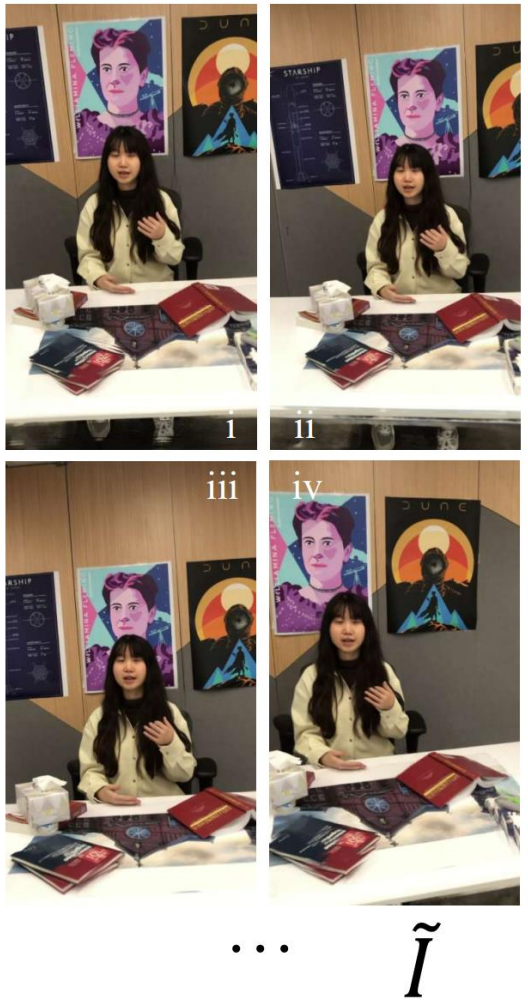
Human keypoint

2D CNN

$$\mathcal{G}_\theta : P_t \rightarrow M_t$$

MPI





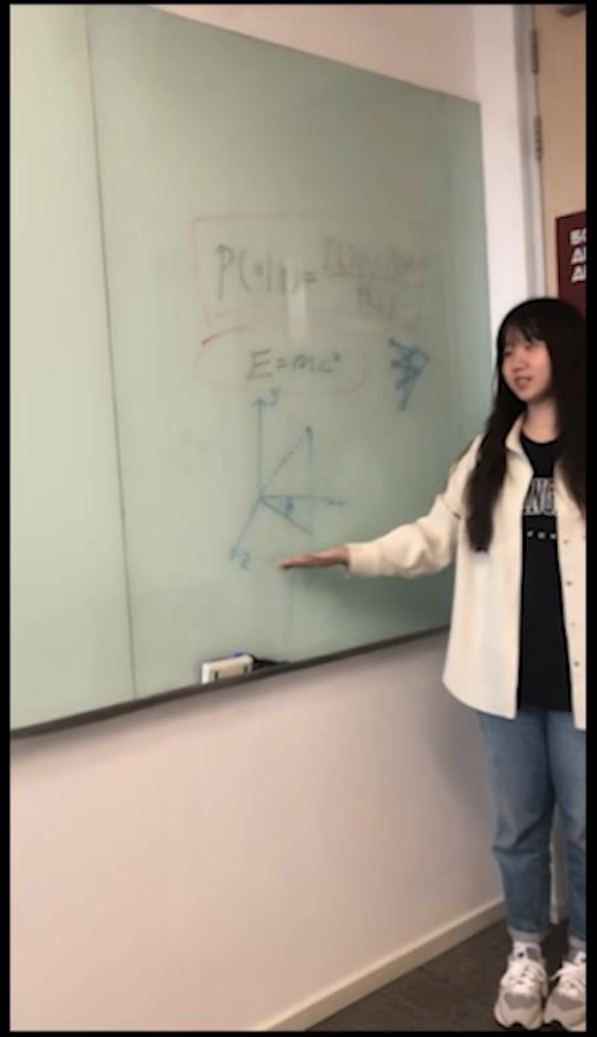
$\cdots$ $\tilde{I}$

Ground truth supervisions

Low-cost capturing setup

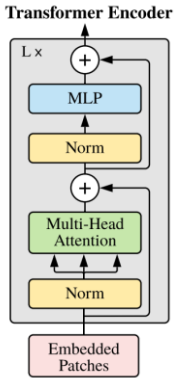

78


STARSHIP



Summary of this talk

z



Styleswin

A very cute giraffe making a funny face

PITi

VQ-diffusion

Realistic 3D avatar

Thank you!