Diffusion Model Track

- Methods and Application in Image Synthesis

Sakura 2024/11/17

bili_sakura@zju.edu.cn

Timeline of novel works on diffusion (image) model

- **OpenAI** (Tim Brooks, Yang Song)

iGPT (Chen et al., 2020) → ADM (Dhariwal & Nichol, 2021) → DALL-E (Ramesh et al., 2021) → GLIDE (Nichol et al., 2022) → DALL-E 2 (Ramesh et al., 2022) → DALL-E 3 (Betker et al., 2023) → Consistency Models (Lu & Song, 2024)

- **Google** (Jonathan Ho)

Classifier-Free Guidance (Ho et al., 2021) → Prompt-to-Prompt (Hertz et al., 2022) →
Imagen (Saharia et al., 2022) → Null-Text Inversion (Mokady et al., 2023) → Imagen3 (Imagen 3 Team, 2024)

- **Stanford (Ermon Group)**

Score-based Generative Models (Song & Ermon, 2019) → Stochastic Differential Equations (Song et al., 2020) →
→ DDIM (Song et al., 2021) → SDEdit (Meng et al., 2021) → Distillation (Meng et al., 2023) →
→ DPO-Diffusion(Wallace et al., 2024)

- **NVIDIA** (Jiaming Song)

K-Diffusion (Karras et al., 2022) → VideoLDM (Blattmann et al., 2023) → DiffiT (Hatamizadeh et al., 2024)

- **Stability AI** (Robin Rombach)

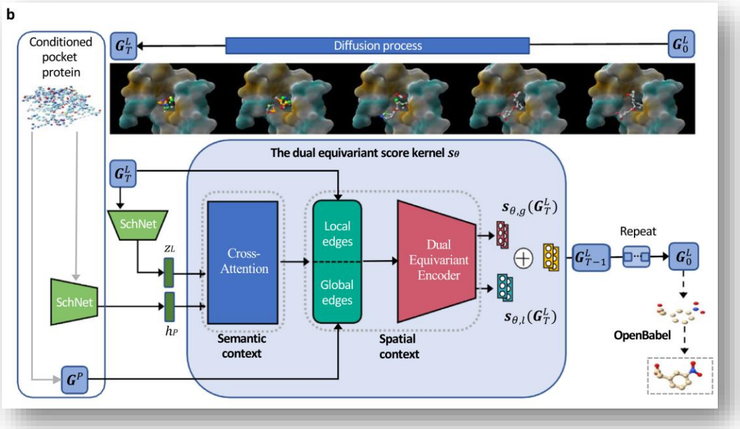
SD (Rombach et al., 2022) → SDXL (Podell et al.,2023) → SDXL Turbo (Nichol et al., 2024) → SD3(Esser et al., 2024)

- Unclassified

DDPM (Ho et al., 2020) ; ControlNet (Zhang et al., 2023) ; InstructPix2Pix (Brooks et al., 2023) ;

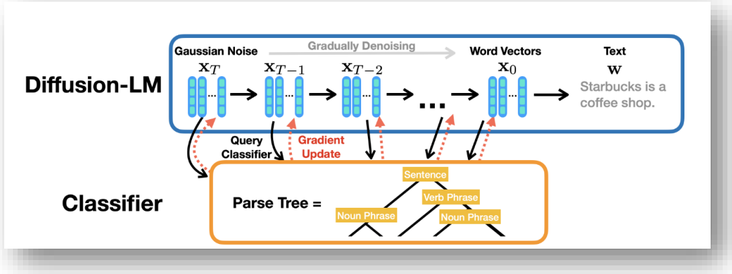
Note: Citations denote the time when work is accepted but ordered in preprint date.

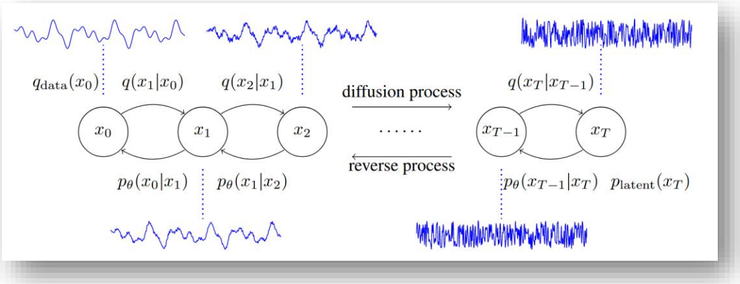
Broad Applications of Diffusion Models



Drug

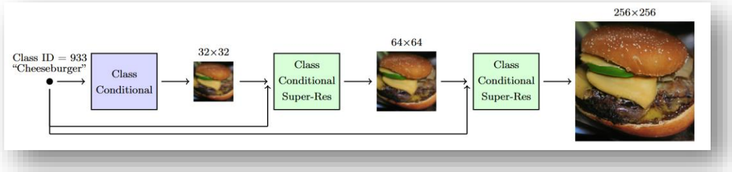
NLP





Audio

CV



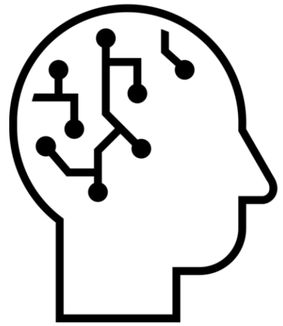
Huang et al. A Dual Diffusion Model Enables 3D Molecule Generation and Lead Optimization Based on Target Pockets. NC 2024
Ho et al. Cascaded Diffusion Models for High Fidelity Image Generation. JMLR 2022
Li et al. Diffusion-LM Improves Controllable Text Generation. NuerIPS 2022
Kong et al. DiffWave: A Versatile Diffusion Model for Audio Synthesis. ICLR 2020

Deep Generative Learning

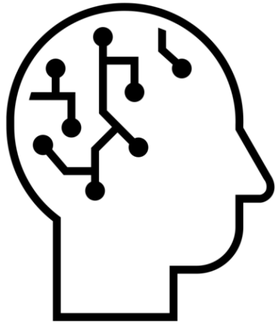
Learning to generate data

Samples from a Data Distribution

Train



Neural Network



Sample

Image Credit: CVPR 2023 Tutorial Denoising Diffusion-Based Generative Modeling: Foundations and Applications

The Landscape of Deep Generative Learning

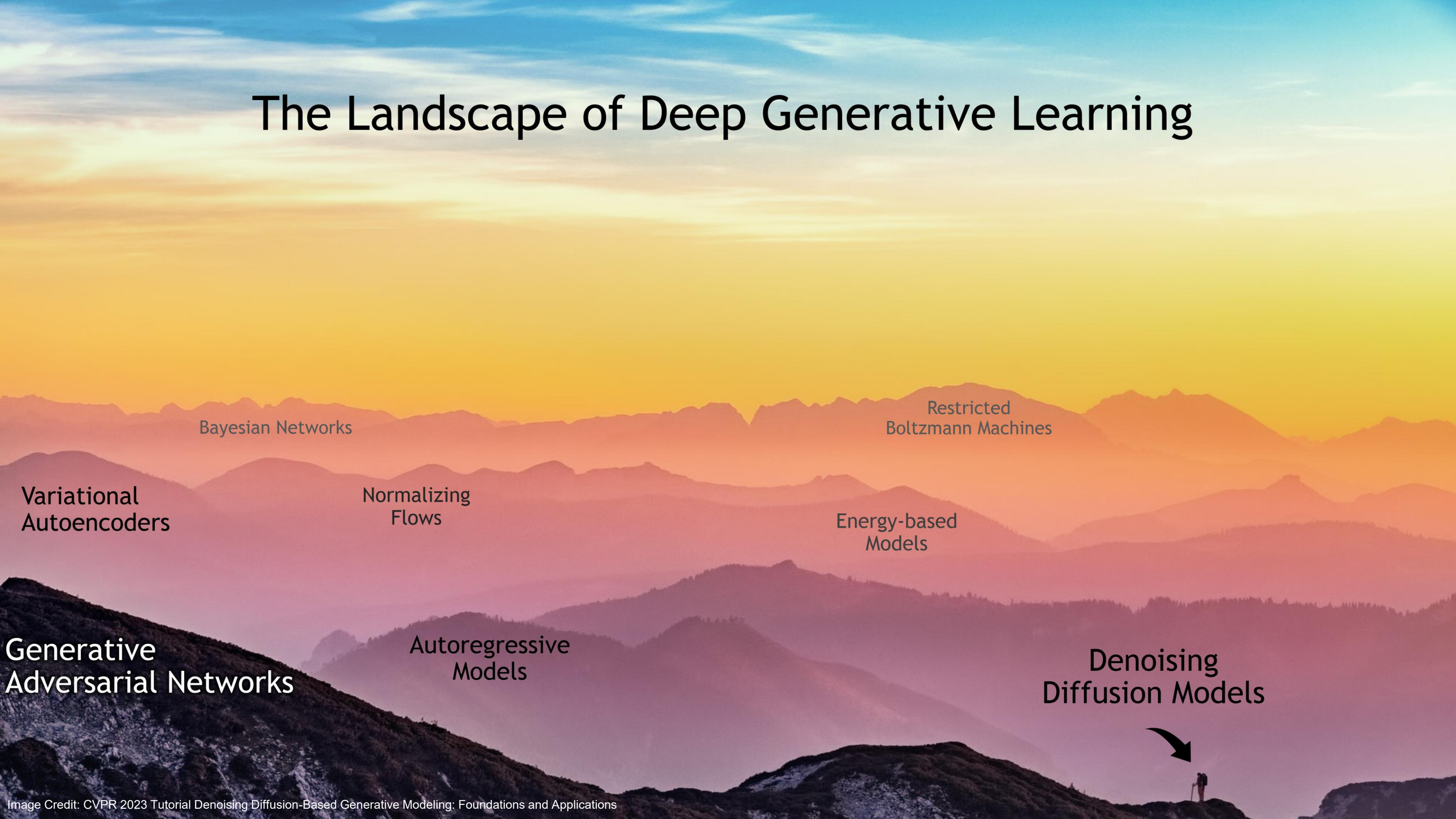
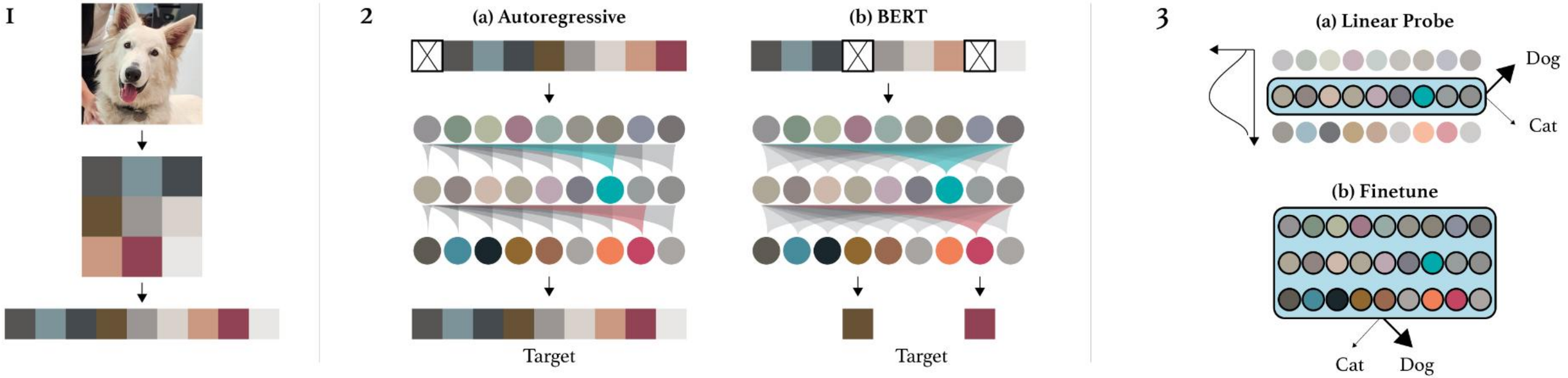


Image Credit: CVPR 2023 Tutorial Denoising Diffusion-Based Generative Modeling: Foundations and Applications

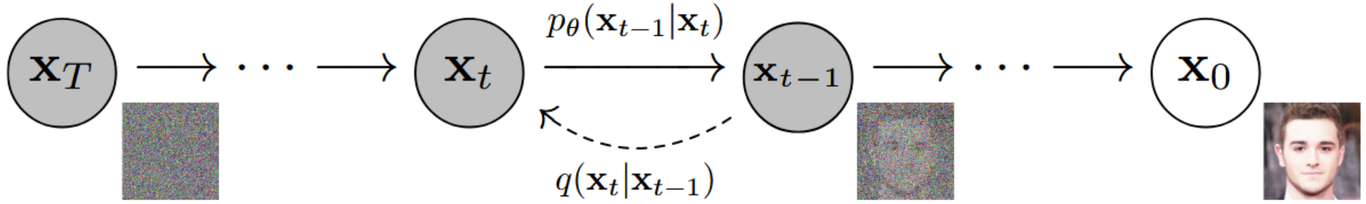
OpenAI - iGPT (Chen et al., 2020)



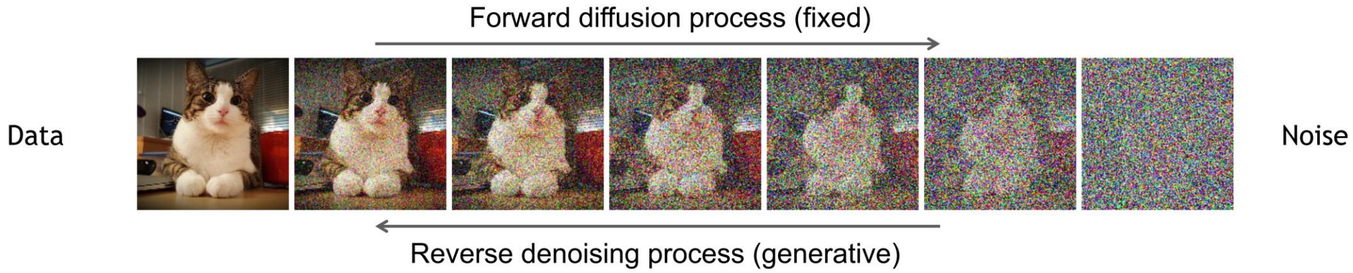
Class-**unconditional** samples from iGPT-L trained on input images of resolution 96×96 (Chen et al., 2020).

Chen et al. “Generative Pretraining From Pixels.” ICML 2020

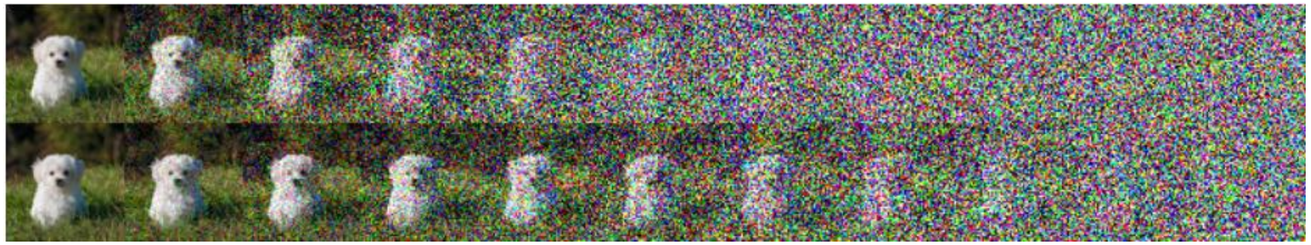
OpenAI – Improved DDPM (Nichol & Dhariwal, 2021)



DDPM Illustration (Ho et al., 2020)



Forward/Reverse Process Illustration (Song et al., 2021)



Instead of linear schedule (top) in DDPM, Improved DDPM uses cosine (bottom) schedules. The cosine schedule adds noise more slowly.

Iters	T	Schedule	Objective	NLL	FID
200K	1K	linear	L_{simple}	3.99	32.5
200K	4K	linear	L_{simple}	3.77	31.3
200K	4K	linear	L_{hybrid}	3.66	32.2
200K	4K	cosine	L_{simple}	3.68	**27.0**
200K	4K	cosine	L_{hybrid}	3.62	28.0
200K	4K	cosine	L_{vlb}	**3.57**	56.7
1.5M	4K	cosine	L_{hybrid}	3.57	**19.2**
1.5M	4K	cosine	L_{vlb}	**3.53**	40.1

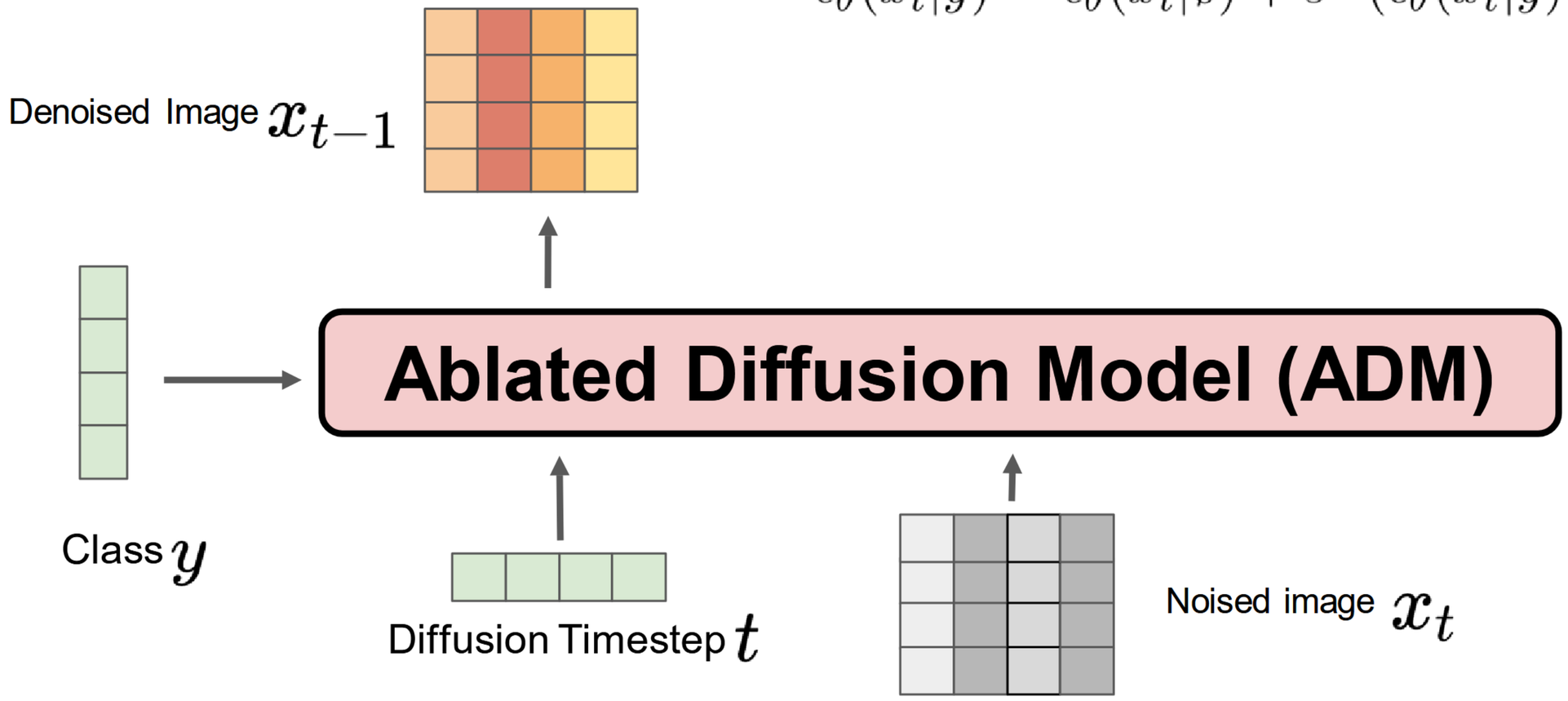
Ablating schedule and objective on ImageNet 64 × 64.

Class-conditional ImageNet 64 × 64 samples generated using 250 sampling steps from Lhybrid model (FID 2.92) which shows high diversity

Ho et al. “Denoising Diffusion Probabilistic Models.” NeurIPS 2020
Song et al., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021
Nichol & Dhariwal. “Improved Denoising Diffusion Probabilistic Models.” ICML 2021

Conditioned Diffusion Model

$$\hat{\epsilon}_\theta(x_t|y) = \epsilon_\theta(x_t|\emptyset) + s \cdot (\epsilon_\theta(x_t|y) - \epsilon_\theta(x_t|\emptyset))$$



Dhariwal & Nichol . “Diffusion Models Beat GANs on Image Synthesis.” NeurIPS 2021

*This slide is derived from https://www.crcv.ucf.edu/wp-content/uploads/2018/11/Group-6-Paper-2-Dalle-2.pdf

Recall that when sampling from a conditional model $p(\mathbf{x}|\mathbf{y})$ we basically need an estimation of $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{y})$

Using Bayes' rule, we have:

Classifier gradient

Score model

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

Introduce a guidance scale (w):

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{y}) \approx w\nabla_{\mathbf{x}_t} \log p(\mathbf{y}|\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$$

Samples from an unconditional diffusion model with classifier guidance to condition on the class "Pembroke Welsh corgi". Using classifier scale 1.0 (left; FID: 33.0) does not produce convincing samples in this class, whereas classifier scale 10.0 (right; FID: 12.0) produces much more class-consistent images.

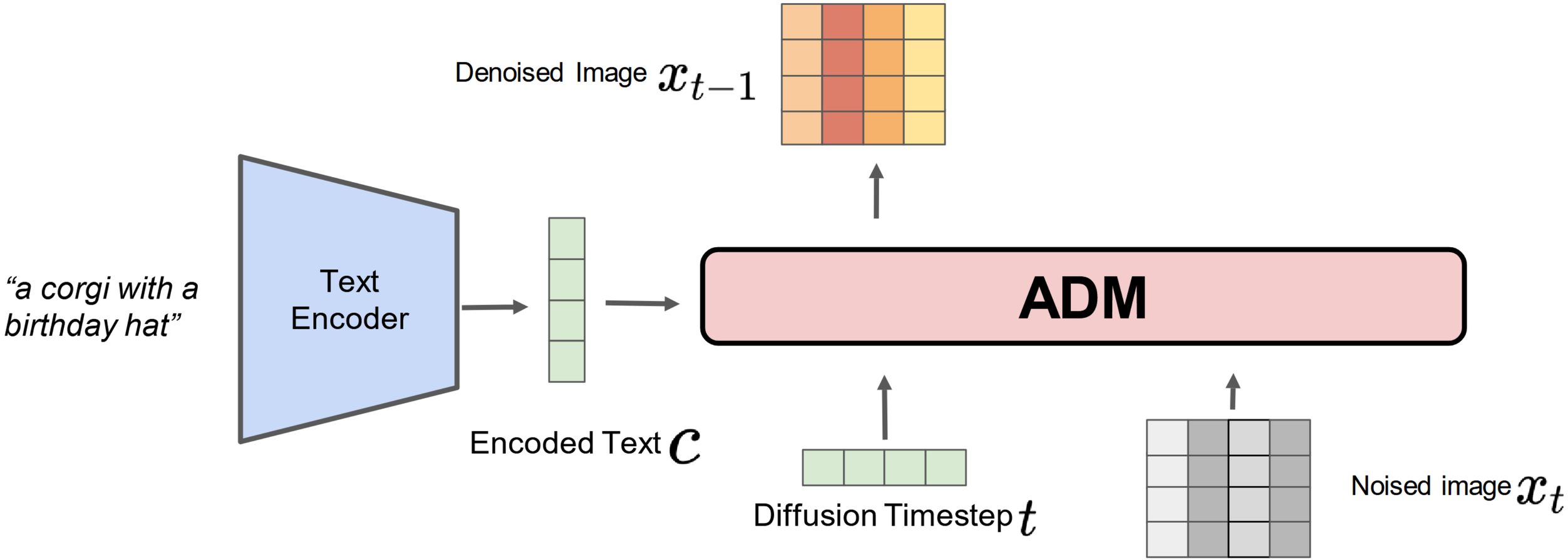
Dhariwal & Alexander . "Diffusion Models Beat GANs on Image Synthesis." NeurIPS 2021

*Equations are derived from CVPR 2023 Tutorial Denoising Diffusion-Based Generative Modeling: Foundations and Applications

OpenAI – GLIDE (Nichol et al., 2022)

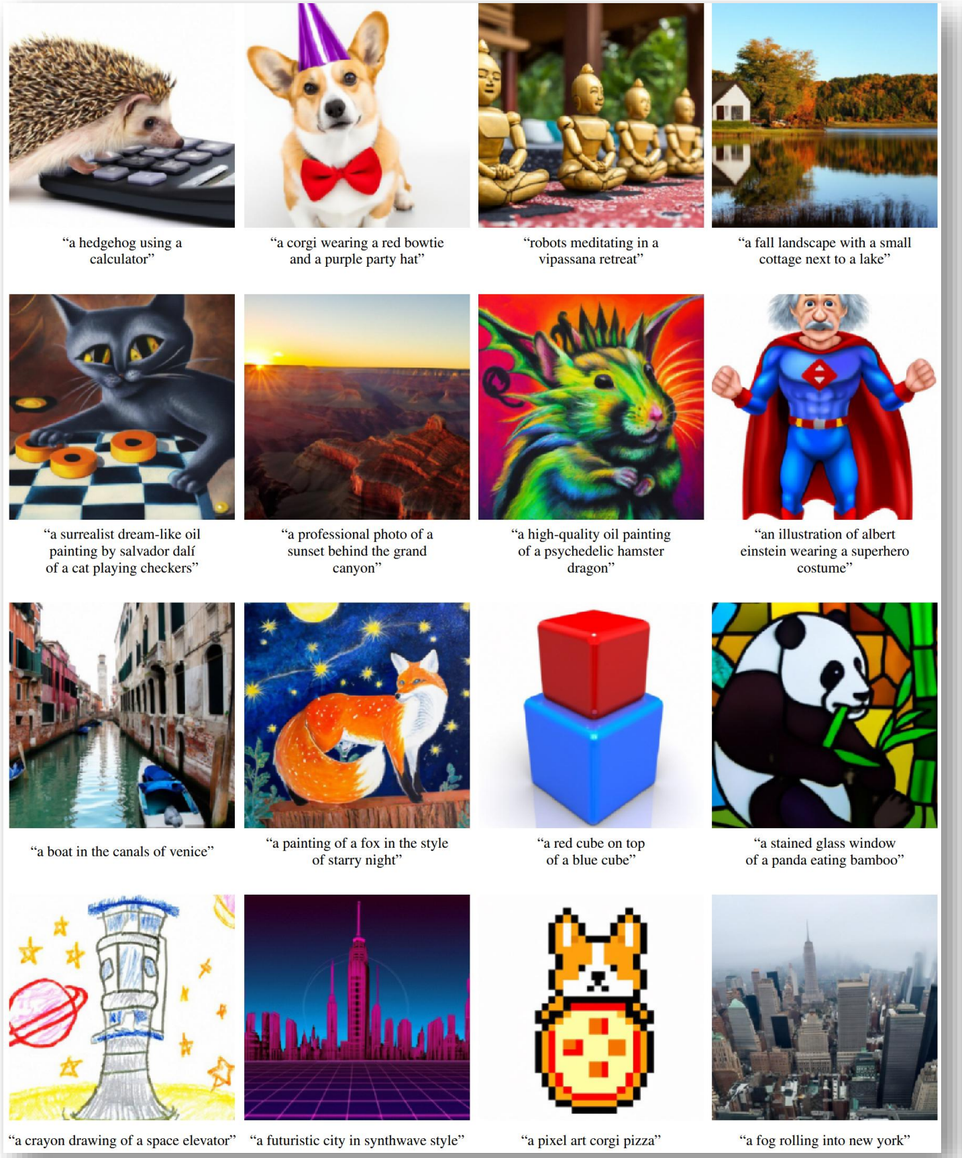
GLIDE

$$\hat{\epsilon}_\theta(x_t|c) = \epsilon_\theta(x_t|\emptyset) + s \cdot (\epsilon_\theta(x_t|c) - \epsilon_\theta(x_t|\emptyset))$$

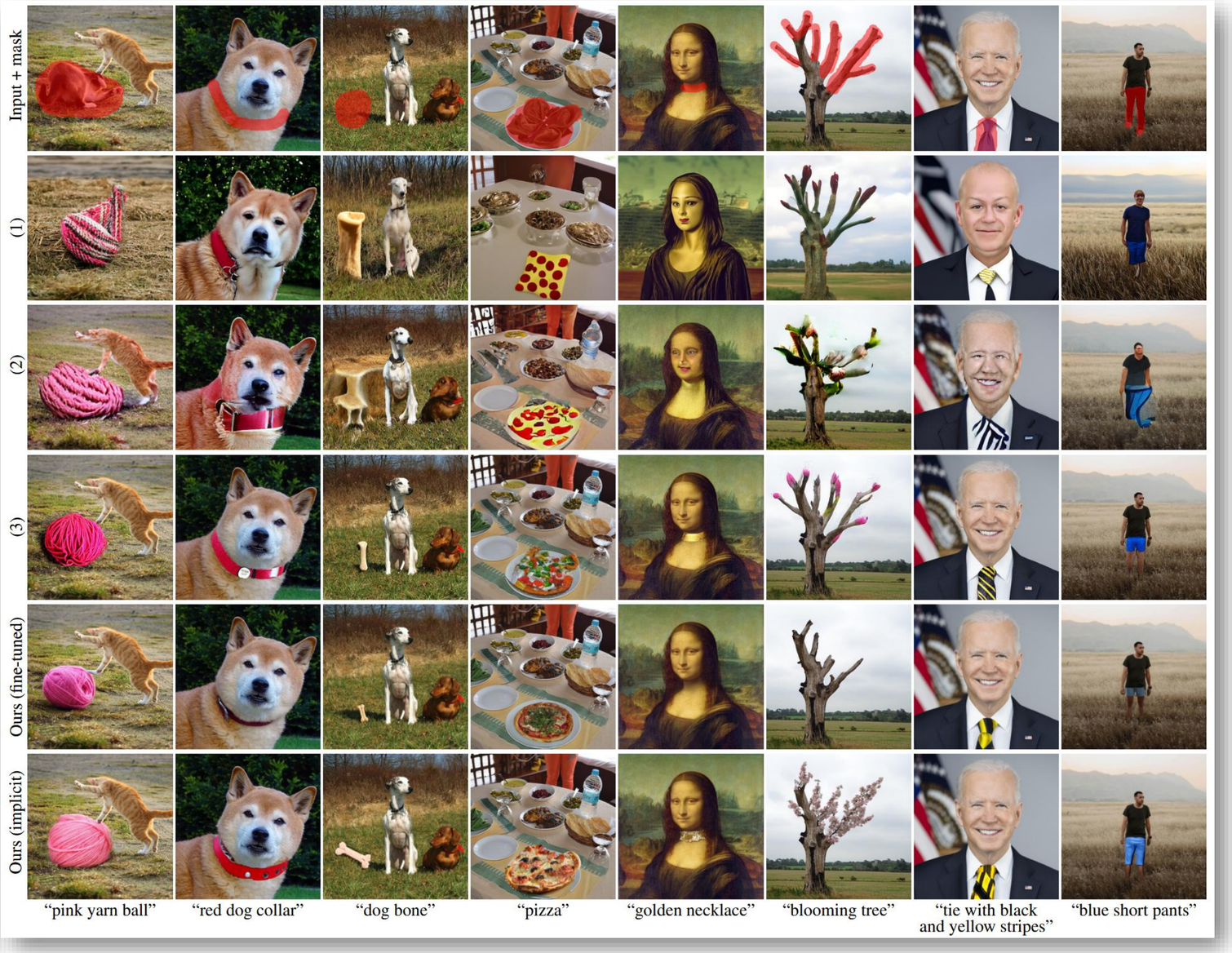


Nichol et al. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. ICML 2022
*This slide is derived from https://www.crcv.ucf.edu/wp-content/uploads/2018/11/Group-6-Paper-2-Dalle-2.pdf

OpenAI – GLIDE (Nichol et al., 2022)



Photorealistic Image Samples.

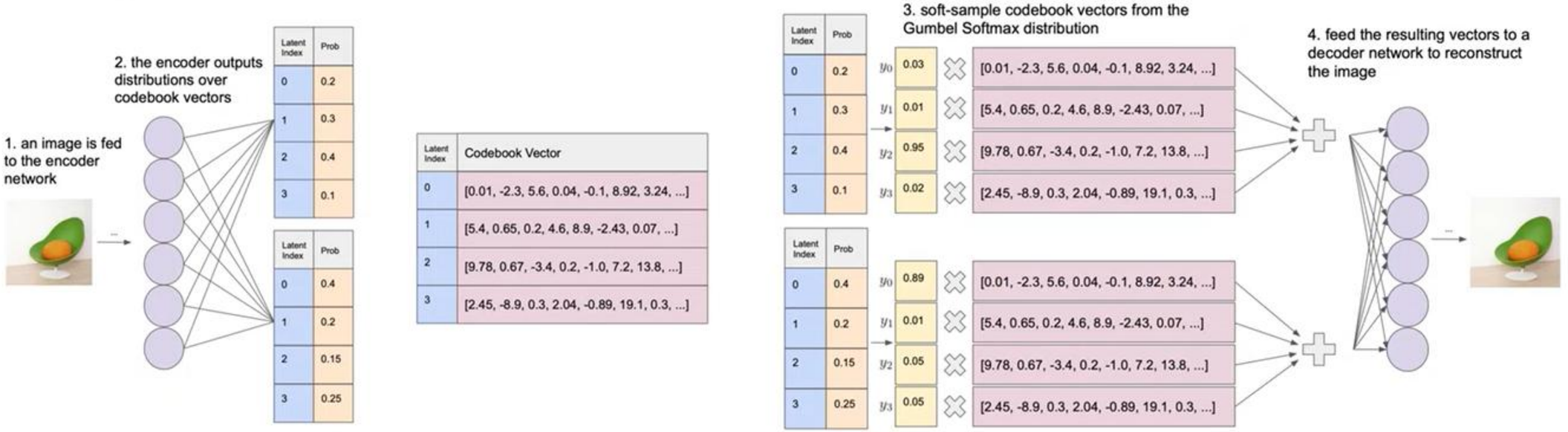


Comparison of Image Inpainting Quality on Real Images.

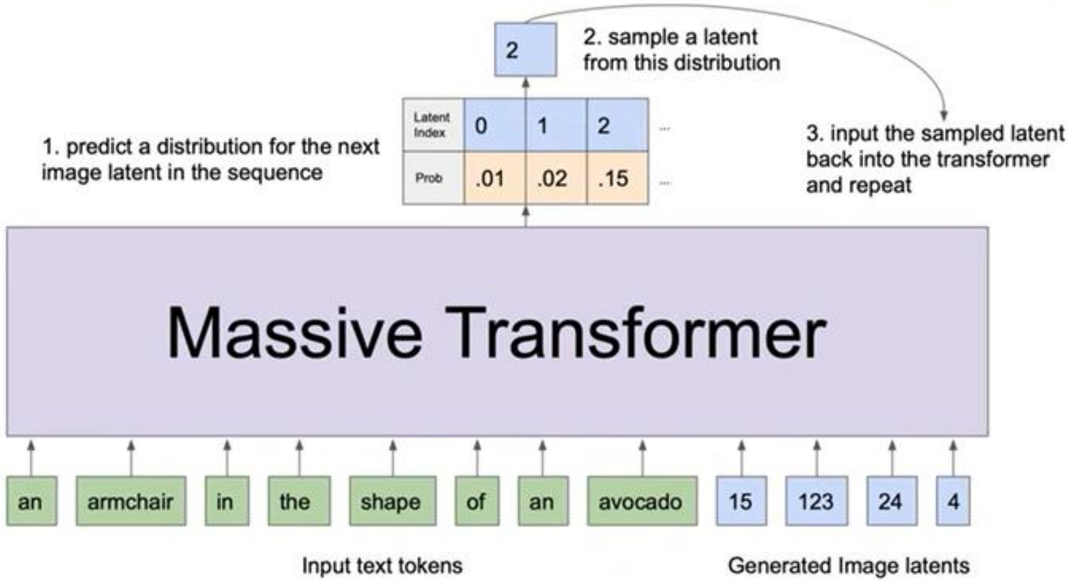
Nichol et al. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. ICML 2022

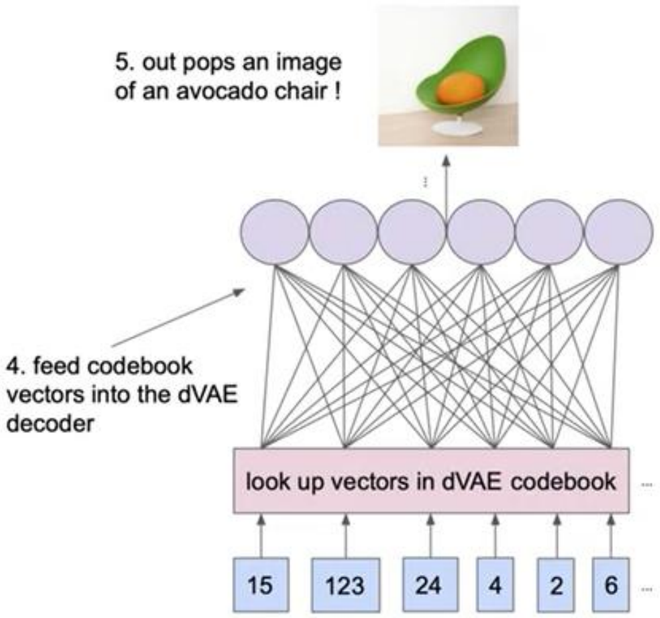
OpenAI – DALL-E (Ramesh et al., 2021)

- Stage 1

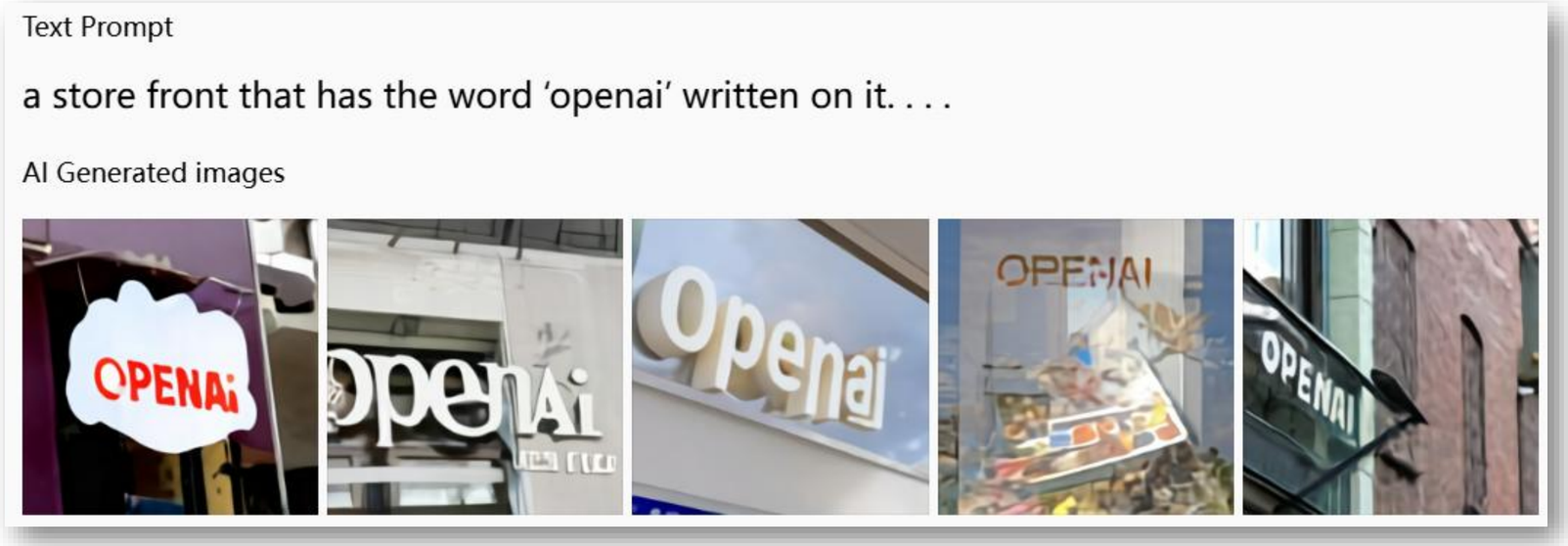


- Stage 2





Ramesh et al. “Zero-Shot Text-to-Image Generation.” ICML 2021
*This slide is derived https://www.youtube.com/watch?v=oENCNi4JxPY
Image Credit: https://openai.com/index/dall-e/



DALL-E Capabilities: Inferring contextual details

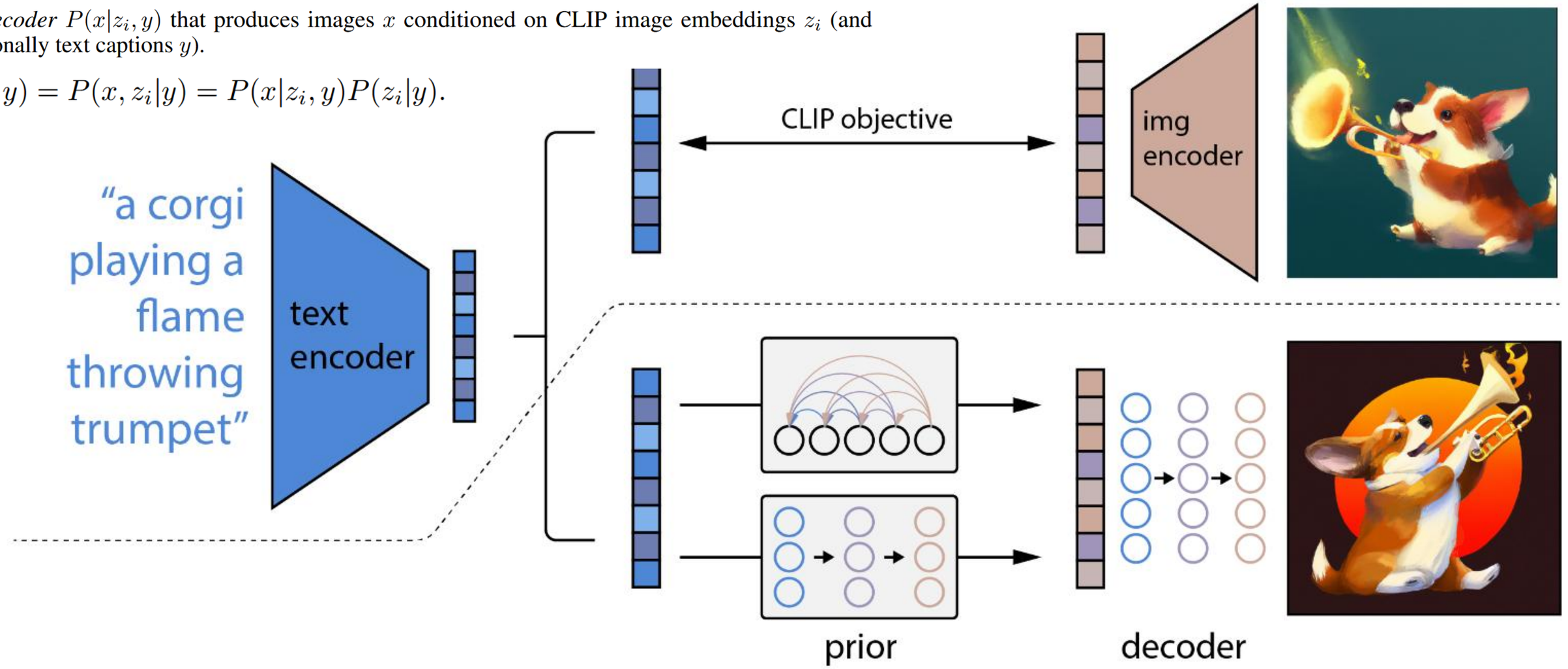
DALL-E Capabilities: Zero-shot visual reasoning (cherry-picked)

Ramesh et al. “Zero-Shot Text-to-Image Generation.” ICML 2021
Image Credit: https://openai.com/index/dall-e/

OpenAI – unCLIP/DALL-E 2 (Ramesh et al., 2022)

- A *prior* $P(z_i|y)$ that produces CLIP image embeddings z_i conditioned on captions y.
- A *decoder* $P(x|z_i, y)$ that produces images x conditioned on CLIP image embeddings z_i (and optionally text captions y).

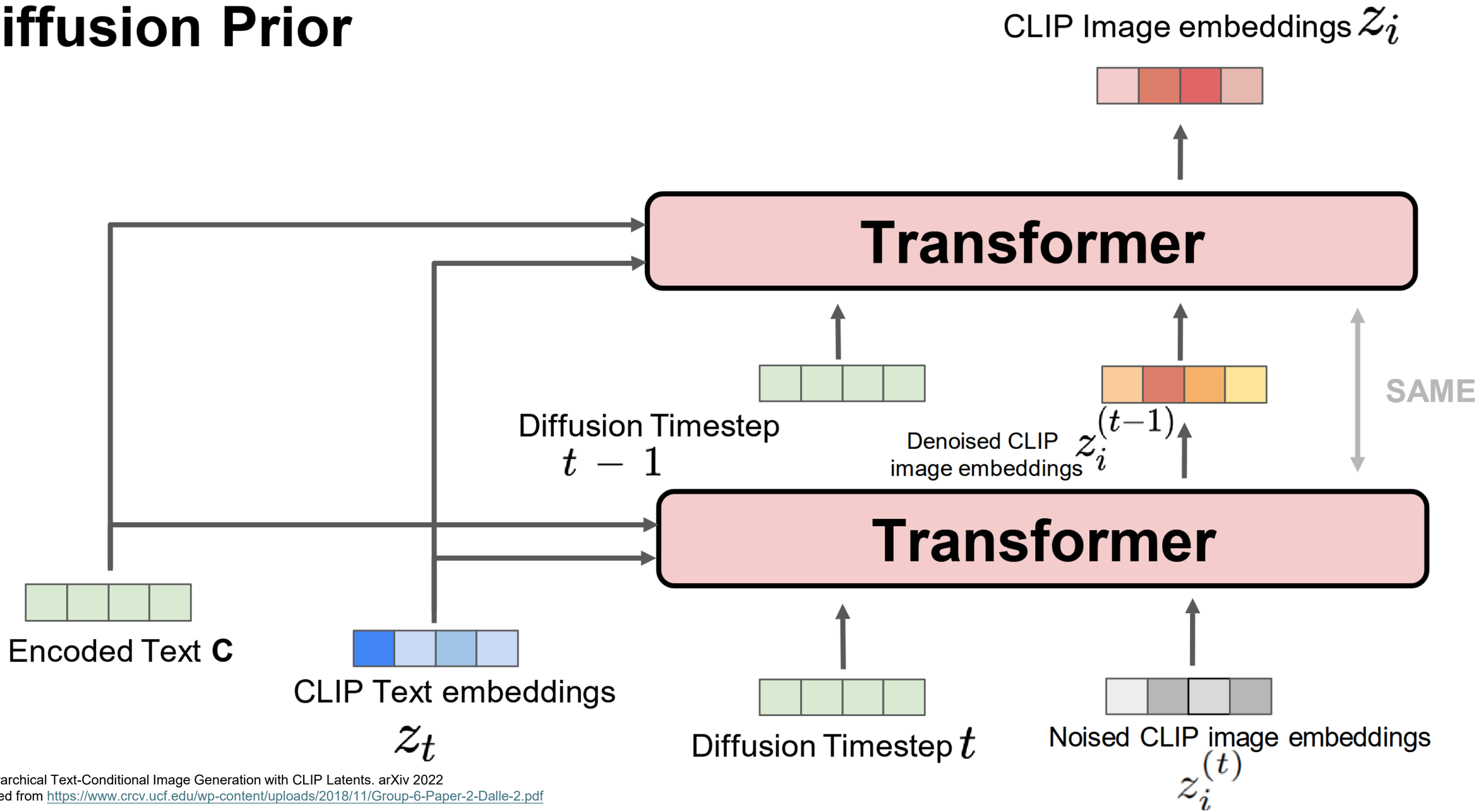
$$P(x|y) = P(x, z_i|y) = P(x|z_i, y)P(z_i|y).$$



Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or **diffusion prior** to produce an image embedding, and then this embedding is used to condition a diffusion decoder which produces a final image.

Ramesh et al. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv 2022

Diffusion Prior



Ramesh et al. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv 2022

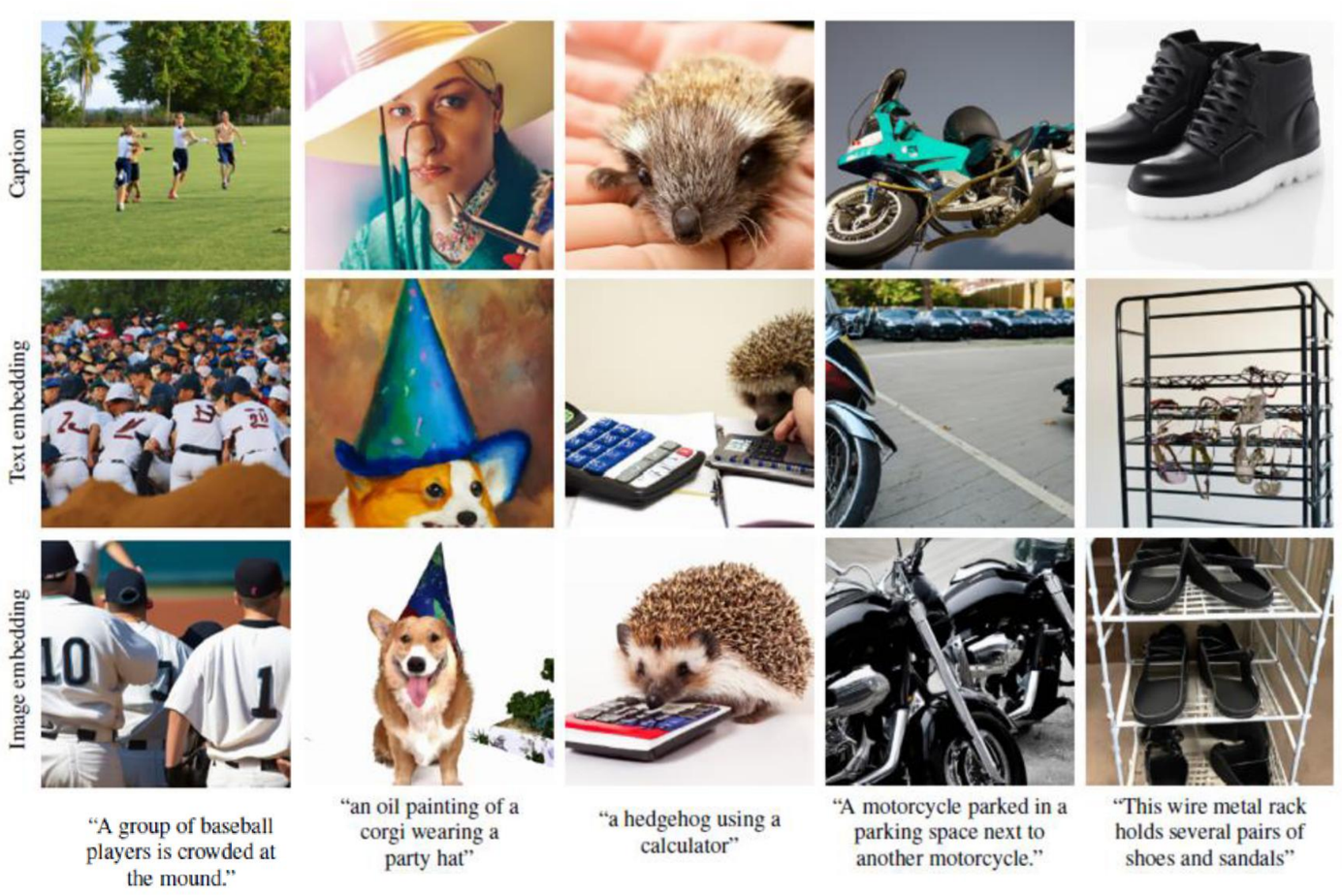
*This slide is derived from https://www.crcv.ucf.edu/wp-content/uploads/2018/11/Group-6-Paper-2-Dalle-2.pdf

Why the prior matters?

Condition decoder on captions alone ✗

Condition decoder on Caption + text embedding impersonating image embeddings ✗

Prior + CLIP image embedding ✓



Ramesh et al. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv 2022
*This slide is derived from https://www.crcv.ucf.edu/wp-content/uploads/2018/11/Group-6-Paper-2-Dalle-2.pdf

OpenAI – unCLIP/DALL-E 2 (Ramesh et al., 2022)

a photo of a landscape in winter → a photo of a landscape in fall

Text diffs applied to images by **interpolating** between their CLIP image embeddings and a normalised difference of the CLIP text embeddings produced from the two descriptions.

Variations of an input image by encoding with CLIP and then decoding with a diffusion model. The variations preserve both **semantic information** like the overlapping strokes in the logo, as well as **stylistic elements** like the color gradients in the logo, while varying the non-essential details.

Ramesh et al. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv 2022

OpenAI – DALL-E 3 (Betker et al., 2023)

A illustration from a graphic novel. A bustling city street under the shine of a full moon. The sidewalks bustling with pedestrians enjoying the nightlife. At the corner stall, a young woman with fiery red hair, dressed in a signature velvet cloak, is haggling with the grumpy old vendor. the grumpy vendor, a tall, sophisticated man is wearing a sharp suit, sports a noteworthy moustache is animatedly conversing on his steampunk telephone.

Compelling results from DALL-E 3 with detailed text prompt.

Betker et al. Improving Image Generation with Better Captions. arXiv 2023

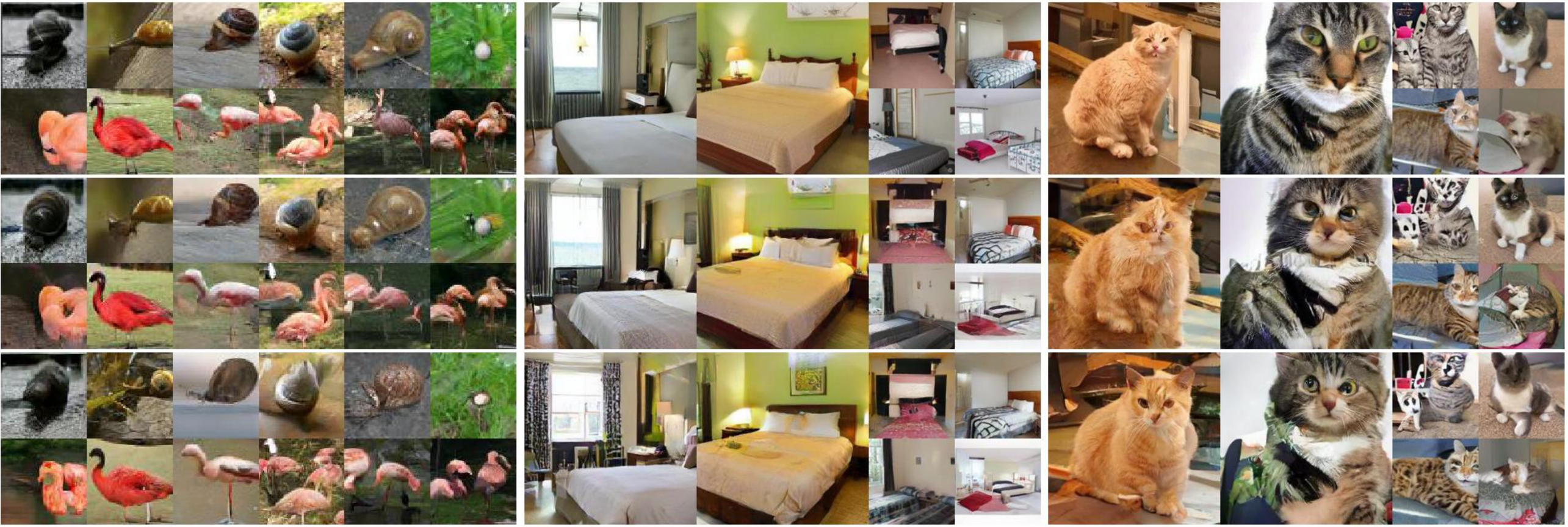
OpenAI – DALL-E 3 (Betker et al., 2023)

Image			
Alt Text	now at victorian plumbing.co.uk	is he finished...just about!	23 (19 of 30) 1200
SSC	a white modern bathtub sits on a wooden floor.	a quilt with an iron on it.	a jar of rhubarb liqueur sitting on a pebble background.
DSC	this luxurious bathroom features a modern freestanding bathtub in a crisp white finish. the tub sits against a wooden accent wall with glass-like panels, creating a serene and relaxing ambiance. three pendant light fixtures hang above the tub, adding a touch of sophistication. a large window with a wooden panel provides natural light, while a potted plant adds a touch of greenery. the freestanding bathtub stands out as a statement piece in this contemporary bathroom.	a quilt is laid out on a ironing board with an iron resting on top. the quilt has a patchwork design with pastel-colored strips of fabric and floral patterns. the iron is turned on and the tip is resting on top of one of the strips. the quilt appears to be in the process of being pressed, as the steam from the iron is visible on the surface. the quilt has a vintage feel and the colors are yellow, blue, and white, giving it an antique look.	rhubarb pieces in a glass jar, waiting to be pickled. the colors of the rhubarb range from bright red to pale green, creating a beautiful contrast. the jar is sitting on a gravel background, giving a rustic feel to the image.

Examples of alt-text accompanying selected images scraped from the internet, short synthetic captions (SSC), and descriptive synthetic captions (DSC).

Betker et al. Improving Image Generation with Better Captions. arXiv 2023

OpenAI – Consistency Training & Consistency Distillation (Song et al., 2023)



Samples generated by EDM (top), **CT** + single-step generation (**middle**), and **CT** + 2-step generation (**Bottom**). All corresponding images are generated from the same initial noise.

Song et al. “Consistency Models.” ICML 2023

OpenAI – Consistency Training & Consistency Distillation (Song et al., 2023)

(a) *Left*: The gray-scale image. *Middle*: Colorized images. *Right*: The ground-truth image.

(b) *Left*: The downsampled image (32 × 32). *Middle*: Full resolution images (256 × 256). *Right*: The ground-truth image (256 × 256).

(c) *Left*: A stroke input provided by users. *Right*: Stroke-guided image generation.

Zero-shot image editing with a consistency model trained by consistency distillation on LSUN Bedroom 256× 256.

Song et al. “Consistency Models.” ICML 2023

OpenAI – Consistency Models (Lu & Song, 2024)

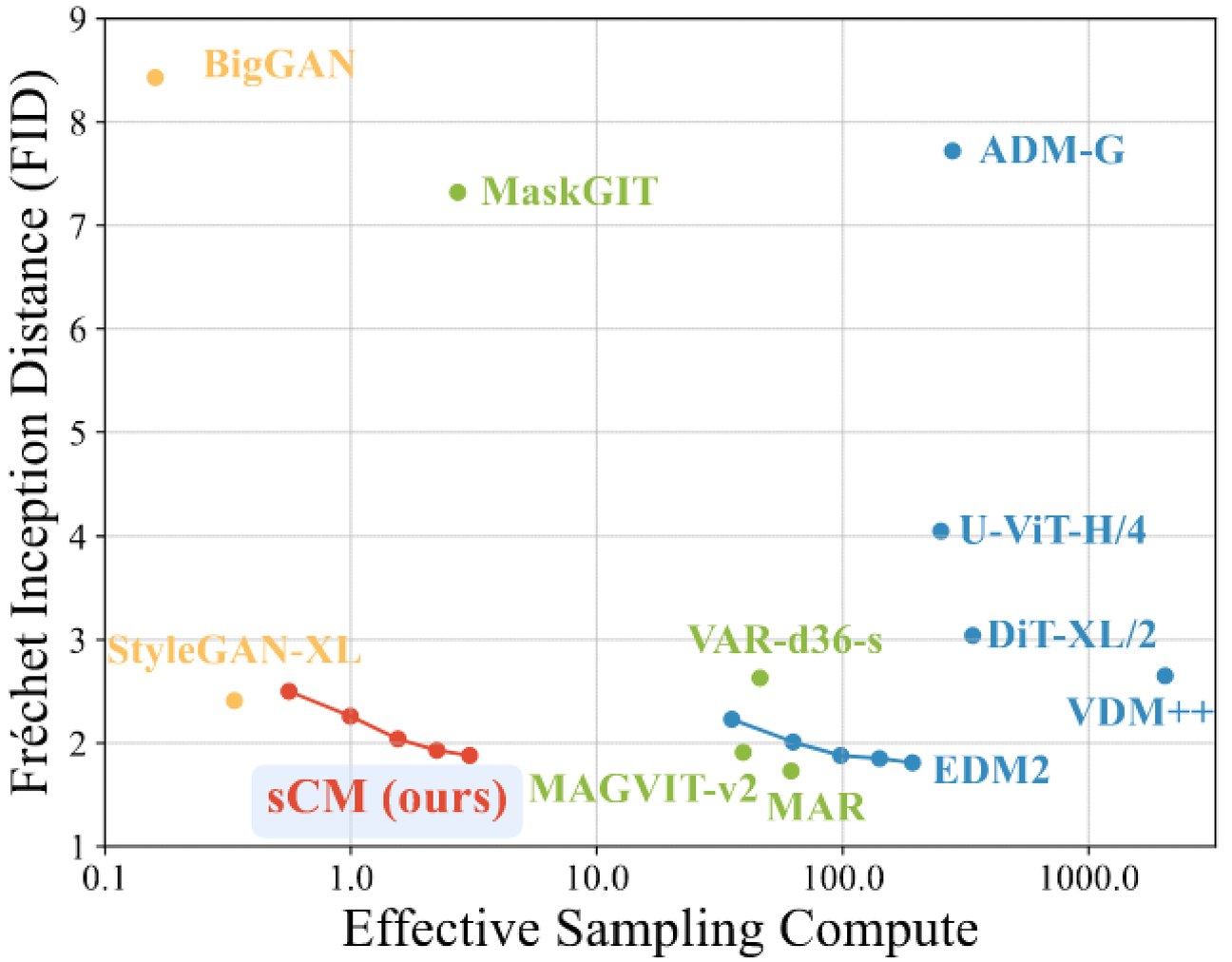


Image Credit: https://openai.com/index/simplifying-stabilizing-and-scaling-continuous-time-consistency-models/
Lu & Song. Simplifying, Stabilizing and Scaling Continuous-Time Consistency Models. arXiv 2024