Diffusion Model Track

- Methods and Application in Image Synthesis

Sakura 2024/12/11

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)

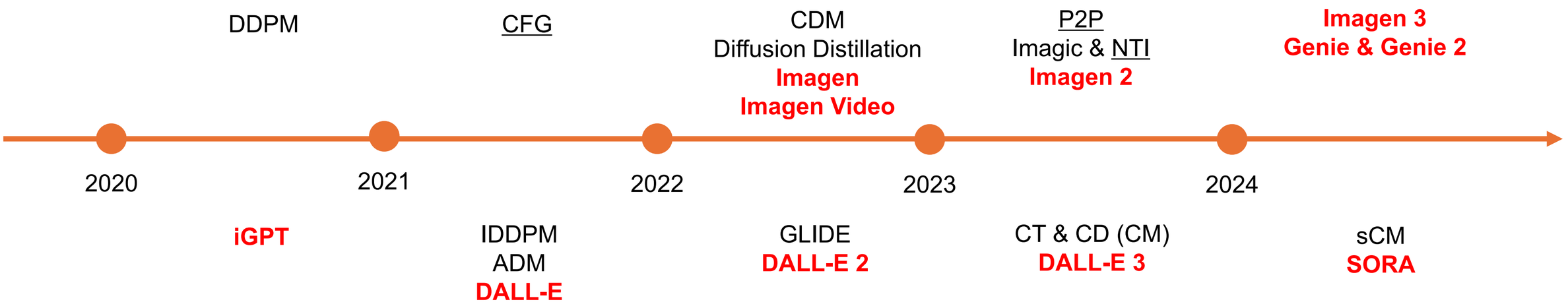
SD (Rombach et al., 2022) → SDXL (Podell et al.,2023) → SDXL Turbo (Nichol et al., 2024) → SD3(Esser et al., 2024)

- Unclassified

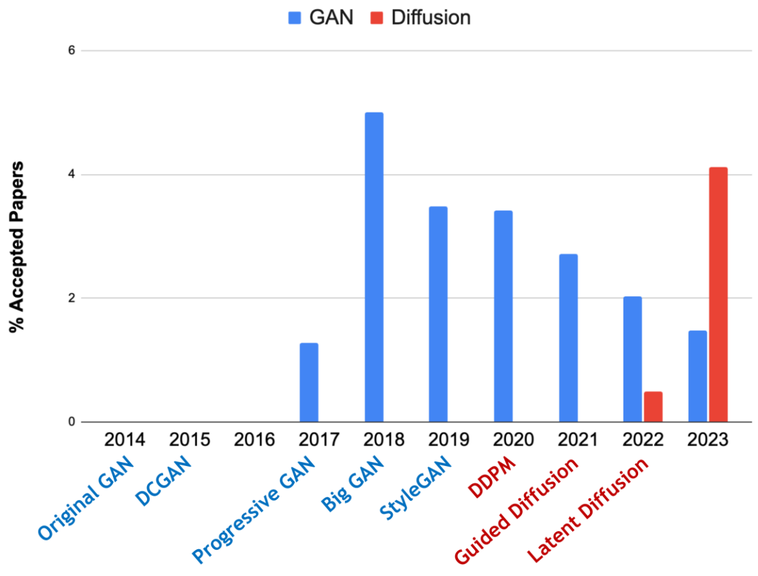
DDPM (Ho et al., 2020) ; ControlNet (Zhang et al., 2023) ; InstructPix2Pix (Brooks et al., 2023) ;

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

Google vs. OpenAI (Image Synthesis Research)



Progress in GANs on Human Face Synthesis*

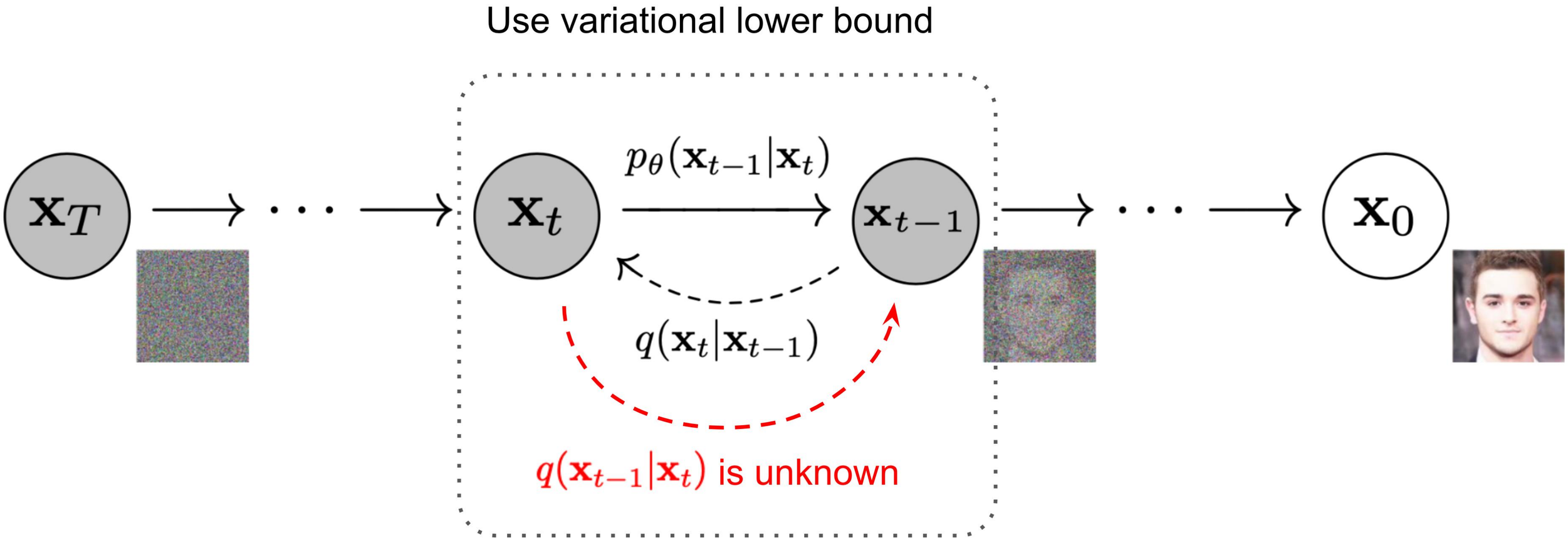


Accepted Papers in CVPR

DDPM (Ho et al., 2020)

Ho et al. “Denoising Diffusion Probabilistic Models.” NeurIPS 2020, https://twitter.com/goodfellow_ian/status/1084973596236144640/photo/1

Denoising Diffusion Probabilistic Models (DDPM)



The Markov chain of forward (reverse) diffusion process of generating a sample by slowly adding (removing) noise.
(Image source: Ho et al. 2020 with a few additional annotations)

Ho et al., Denoising Diffusion Probabilistic Models, NeurIPS 2020

Further Readings: 1. Song et al., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021
2. https://yang-song.net/blog/2021/score/ | 3. https://lilianweng.github.io/posts/2021-07-11-diffusion-models/
4. Yang et al., Diffusion Models: A Comprehensive Survey of Methods and Applications, ACM Comput. Surv. 2023

OpenAI – ADM (Dhariwal & Alexander, 2021)

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

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

Google – CFG (Ho et al., 2021)

Classifier-free guidance

Get guidance by Bayes' rule on conditional diffusion models

- Recall that classifier guidance requires training a classifier.
- Using Bayes' rule again:

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

Conditional diffusion model

Unconditional diffusion model

- Instead of training an additional classifier, get an "implicit classifier" by jointly training a conditional and unconditional diffusion model. In practice, the conditional and unconditional models are trained together by randomly dropping the condition of the diffusion model at certain chance.
- The modified score with this implicit classifier included is:

$$(1+w)\nabla_{\mathbf{x}_t} \log p(\mathbf{y}|\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) = (1+w)\big(\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{y}) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)\big) + \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$$

$$= (1+w)\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{y}) - w\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$$

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

Ho & Salimans, Classifier-Free Diffusion Guidance. 2021. NeurIPS 2021

Google – CFG (Ho et al., 2021)

Classifier-free guidance

Trade-off for sample quality and sample diversity

Non-guidance

$\omega = 1$

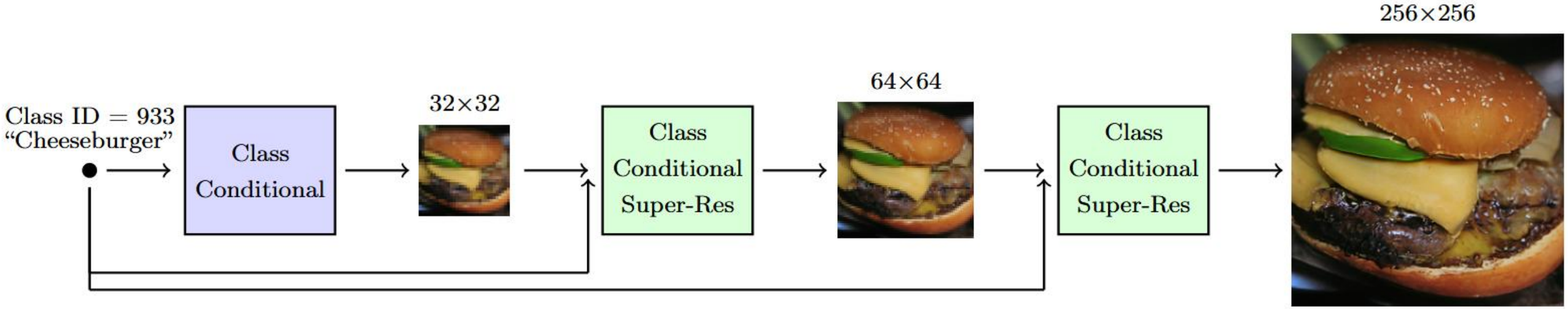
$\omega = 3$

Large guidance weight (ω) usually leads to better individual sample quality but less sample diversity.

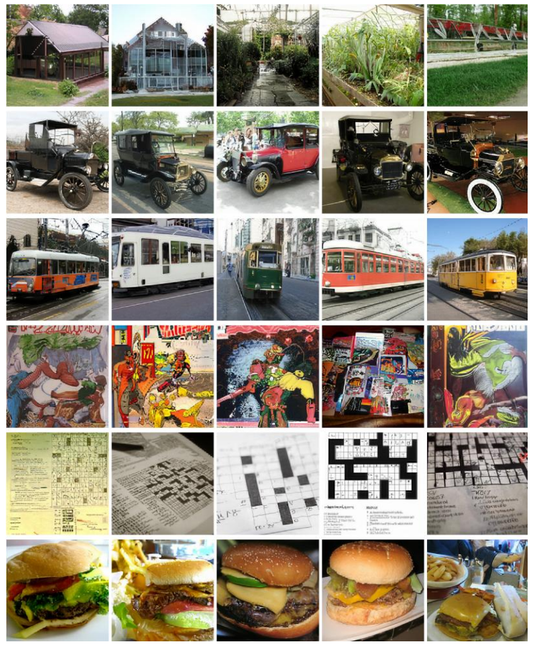
Slide Credit: CVPR 2023 Tutorial Denoising Diffusion-Based Generative Modeling: Foundations and Applications

Ho & Salimans, Classifier-Free Diffusion Guidance. 2021. NeurIPS 2021

Google – CDM (Ho et al., 2022)



Model	FID vs train	FID vs validation	IS
256×256 resolution			
BigGAN-deep (Brock et al., 2019)	6.9		171.4
BigGAN-deep, max IS (Brock et al., 2019)	27		**317**
VQ-VAE-2 (Razavi et al., 2019)	31.11		
Improved DDPM (Nichol and Dhariwal, 2021)	12.26		
SR3 (Saharia et al., 2021)	11.30		
ADM (Dhariwal and Nichol, 2021)	10.94		100.98
ADM+upsampling (Dhariwal and Nichol, 2021)	7.49		127.49
CDM (ours)	**4.88**	4.63	158.71 ± 2.26



Greenhouse (580),

Model T (661),

Streetcar (829),

Comic Book (917),

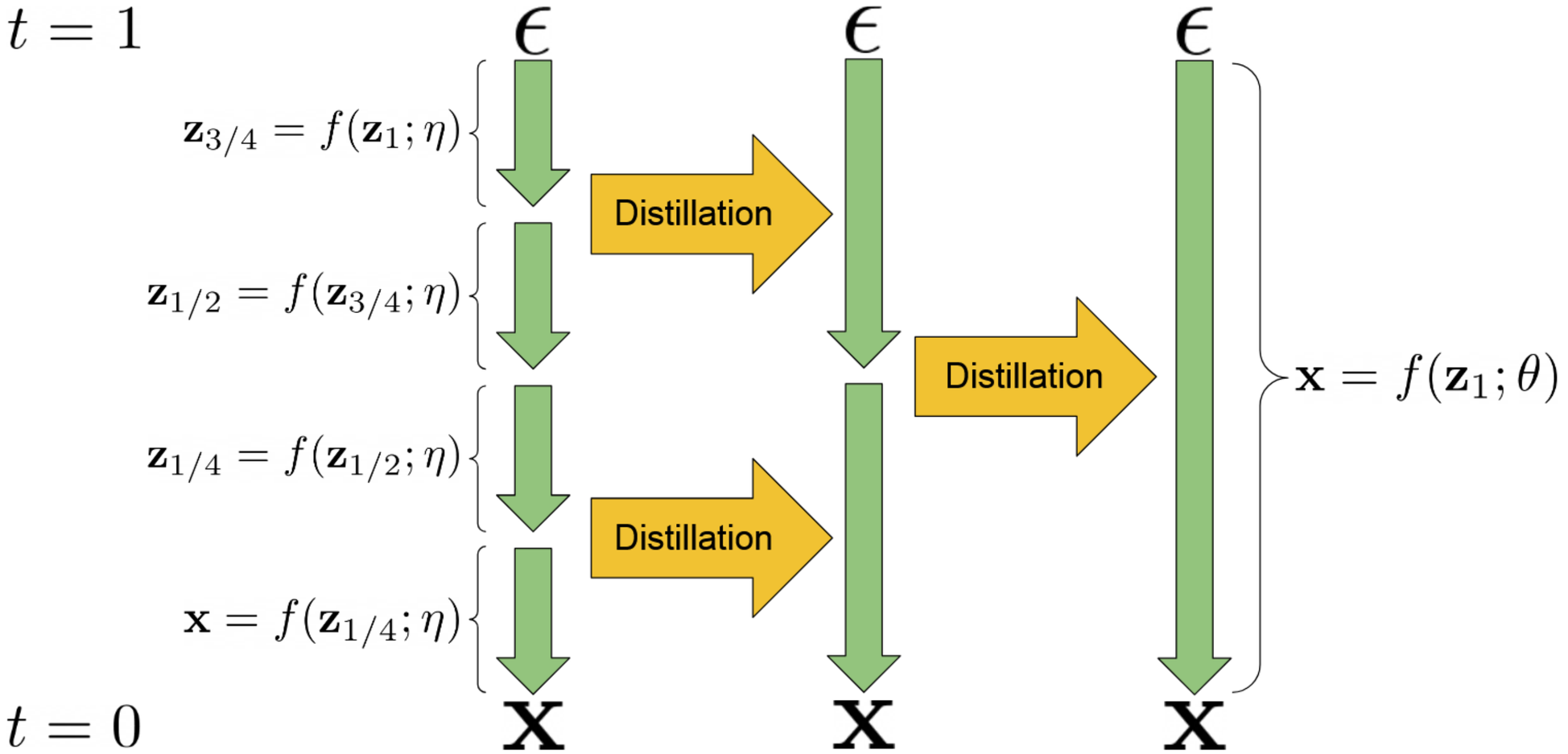
Crossword Puzzle (918),

Cheeseburger (933).

Classwise Synthetic 256×256 ImageNet images.

Ho et al., Cascaded Diffusion Models for High Fidelity Image Generation, JMLR 2022

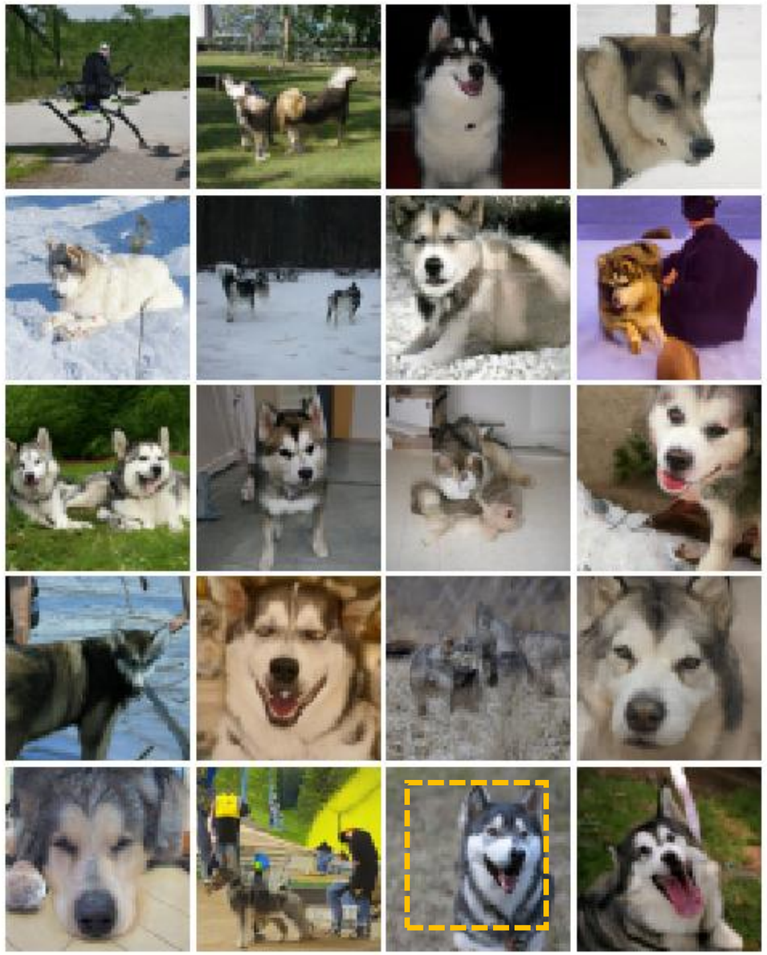
Google – Diffusion Distillation (Salimans et al., 2022)



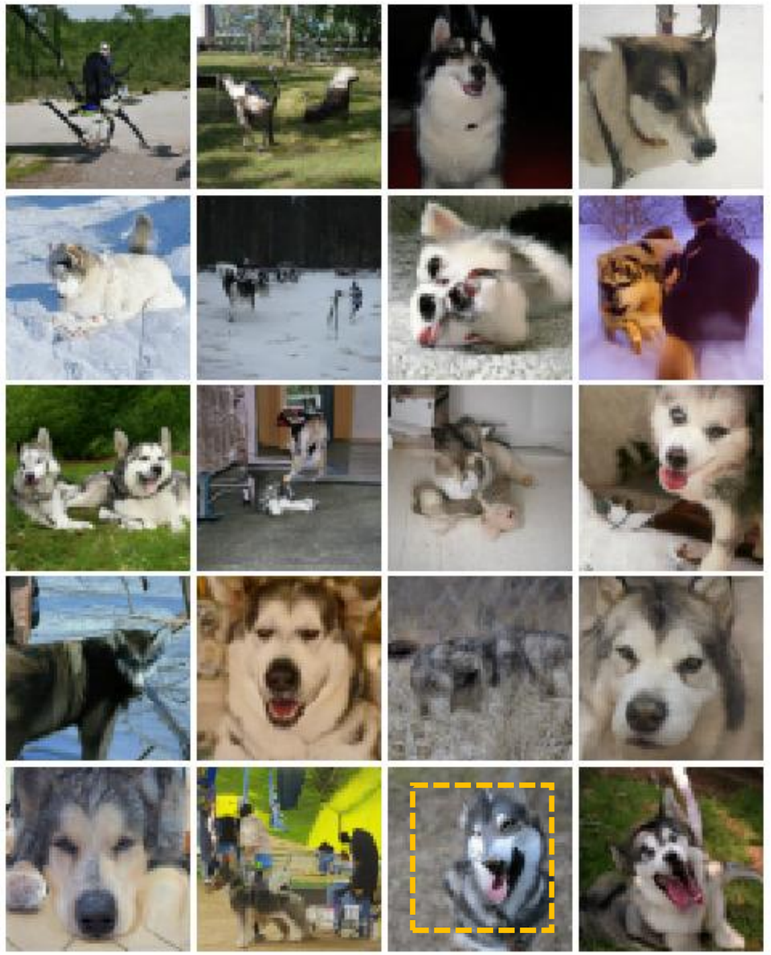
A visualization of two iterations of our proposed *progressive distillation* algorithm. A sampler $f(z; \eta)$, mapping random noise to samples x in 4 deterministic steps, is distilled into a new sampler $f(z; \theta)$ taking only a single step. The original sampler is derived by approximately integrating the probability flow ODE for a learned diffusion model, and distillation can thus be understood as learning to integrate in fewer steps, or amortizing this integration into the new sampler.

Salimans & Ho, Progressive Distillation for Fast Sampling of Diffusion Models, ICLR 2022.

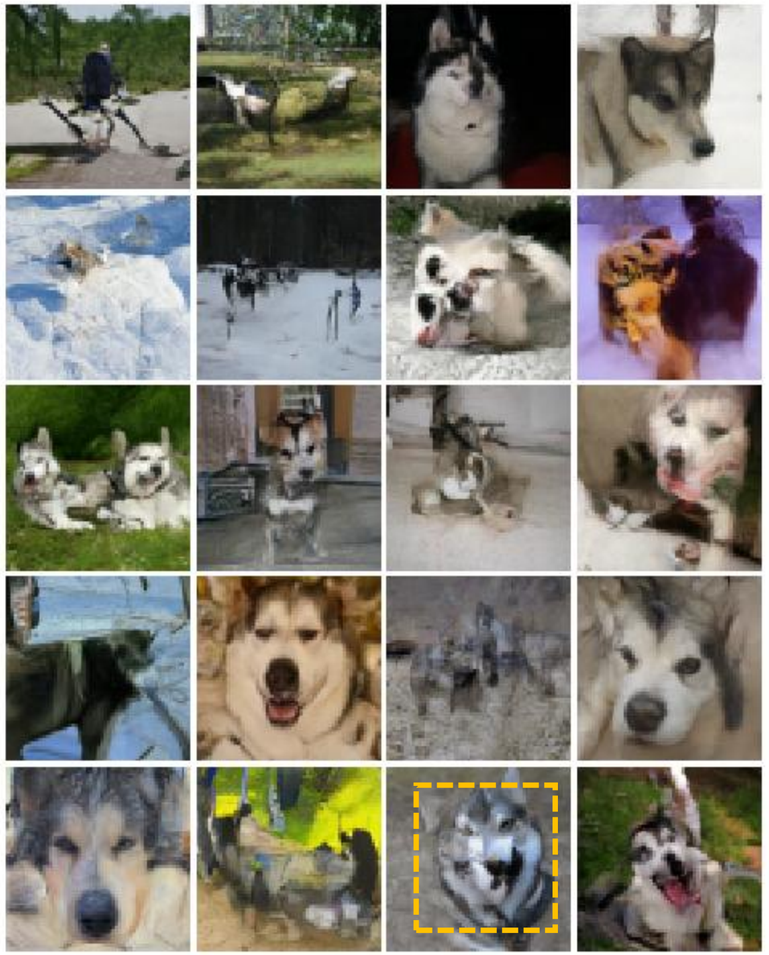
Google – Diffusion Distillation (Salimans et al., 2022)



(a) 256 sampling steps



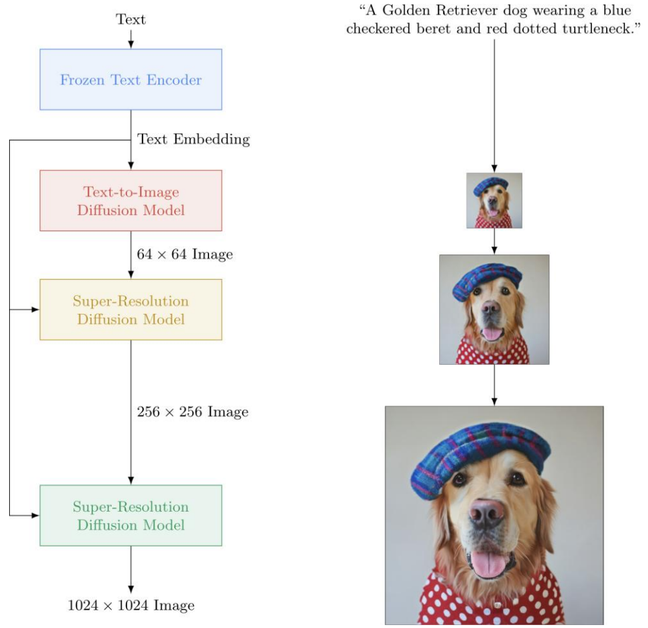
(b) 4 sampling steps



(c) 1 sampling step

Salimans & Ho, Progressive Distillation for Fast Sampling of Diffusion Models, ICLR 2022.

Google – Imagen (Saharia et al., 2022)



Model	FID-30K	Zero-shot FID-30K
AttnGAN [79]	35.49	
DM-GAN [86]	32.64	
DF-GAN [72]	21.42	
DM-GAN + CL [81]	20.79	
XMC-GAN [84]	9.33	
LAFITE [85]	8.12	
Make-A-Scene [22]	7.55	
DALL-E [55]		17.89
LAFITE [85]		26.94
GLIDE [43]		12.24
DALL-E 2 [56]		10.39
Imagen (Our Work)		**7.27**

Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

Teddy bears swimming at the Olympics 400m Butterfly event.

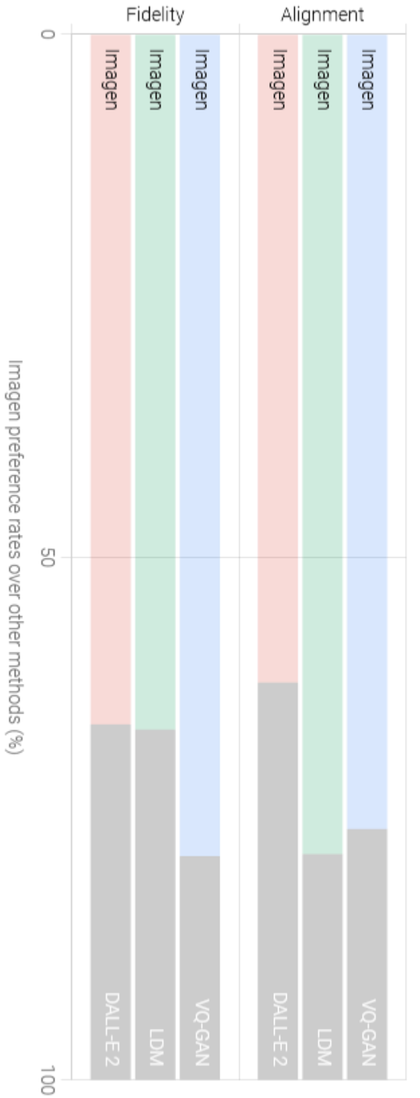
A cute corgi lives in a house made out of sushi.

A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.

A brain riding a rocketship heading towards the moon.

A dragon fruit wearing karate belt in the snow.

A strawberry mug filled with white sesame seeds. The mug is floating in a dark chocolate sea.



DrawBench: Imagen vs Other Methods

Human raters strongly prefer Imagen over other methods

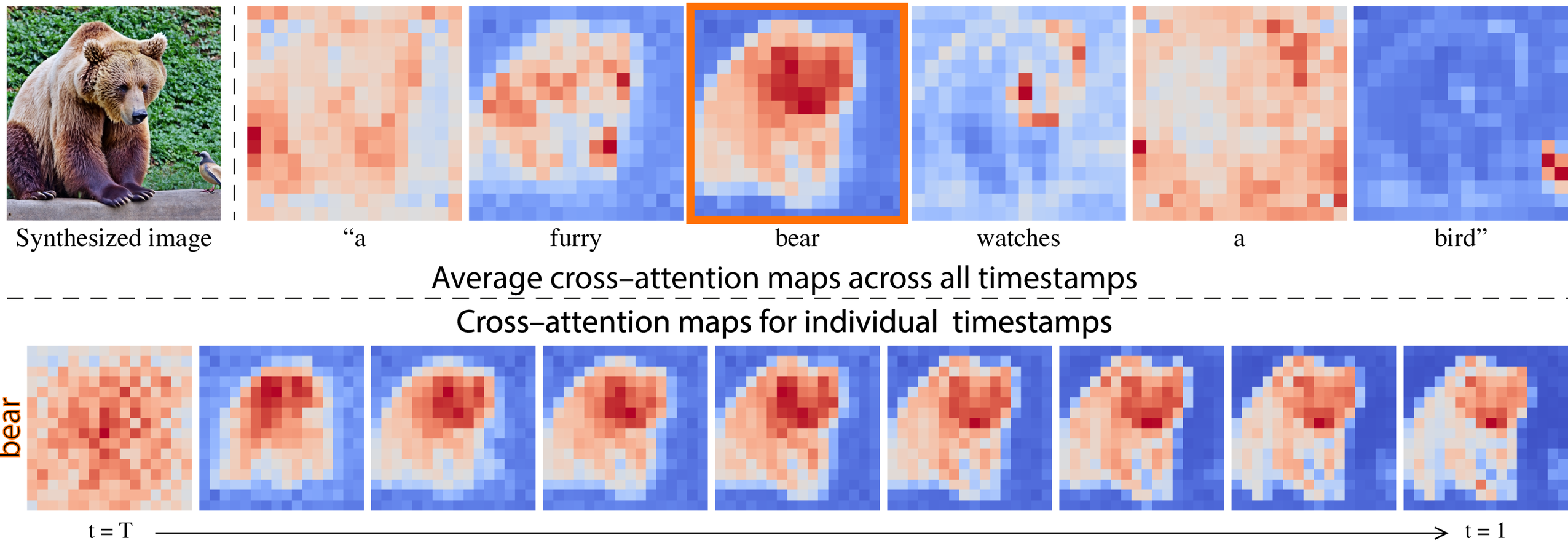
Saharia et al., Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding., NeurIPS 2022

Google – Prompt-to-Prompt (Hertz et al., 2023)

Prompt-to-Prompt editing capabilities. Our method paves the way for a myriad of caption-based editing operations: tuning the level of influence of an adjective word (bottom-left), making a local modification in the image by replacing or adding a word (bottom-middle), or specifying a global modification (bottom-right).

Hertz et al., Prompt-to-Prompt Image Editing with Cross-Attention Control, ICLR 2023

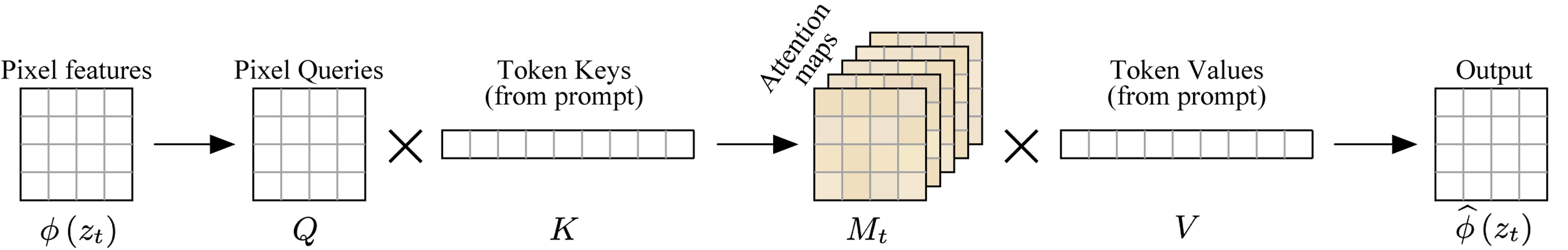
Google – Prompt-to-Prompt (Hertz et al., 2023)



Cross-Attention Control. The key observation behind our method is that the spatial layout and geometry of an image depend on the cross-attention maps. Below, we show that pixels are attend more to the words that describe them.

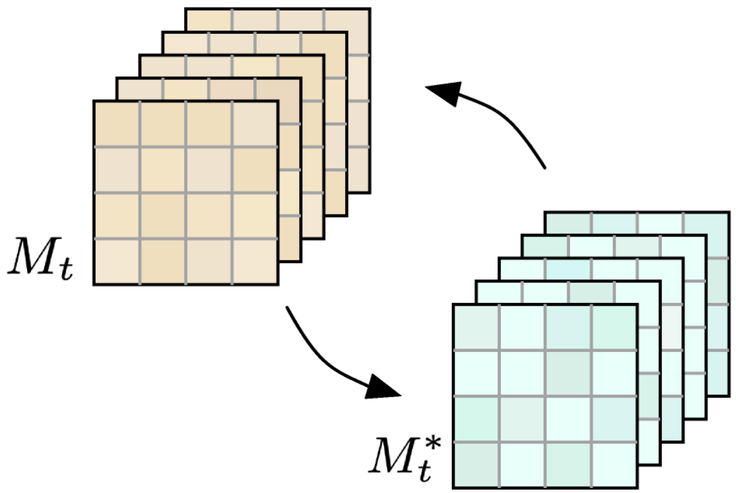
Hertz et al., Prompt-to-Prompt Image Editing with Cross-Attention Control, ICLR 2023

Google – Prompt-to-Prompt (Hertz et al., 2023)

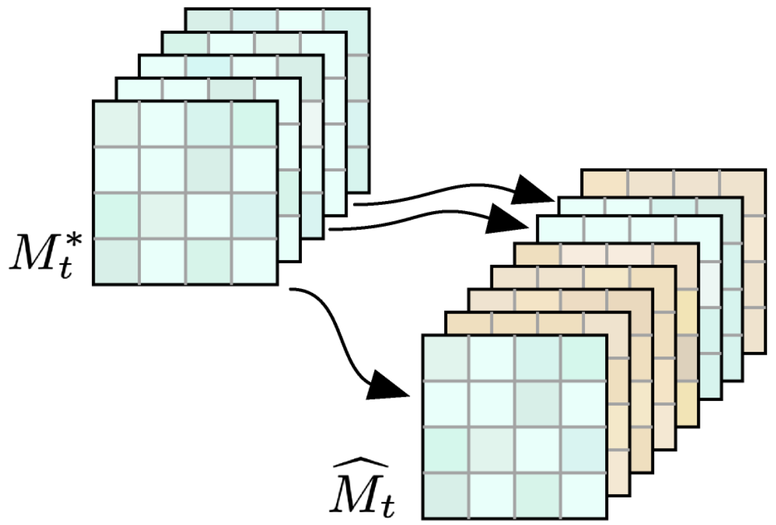


Text to Image Cross Attention

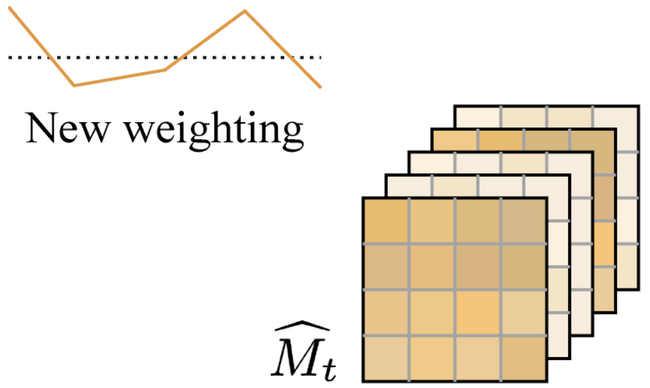
Cross Attenetion Control



Word Swap



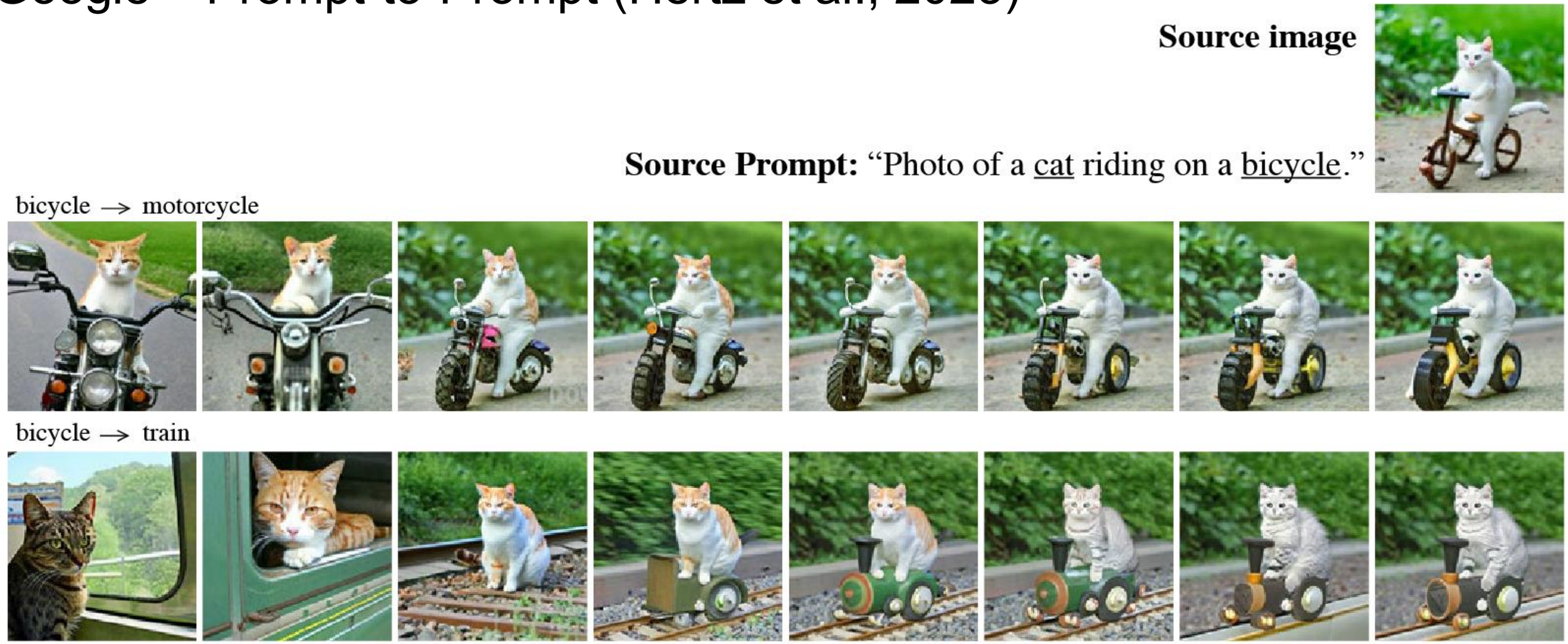
Prompt Refinement



Attention Re–weighting

Hertz et al., Prompt-to-Prompt Image Editing with Cross-Attention Control, ICLR 2023

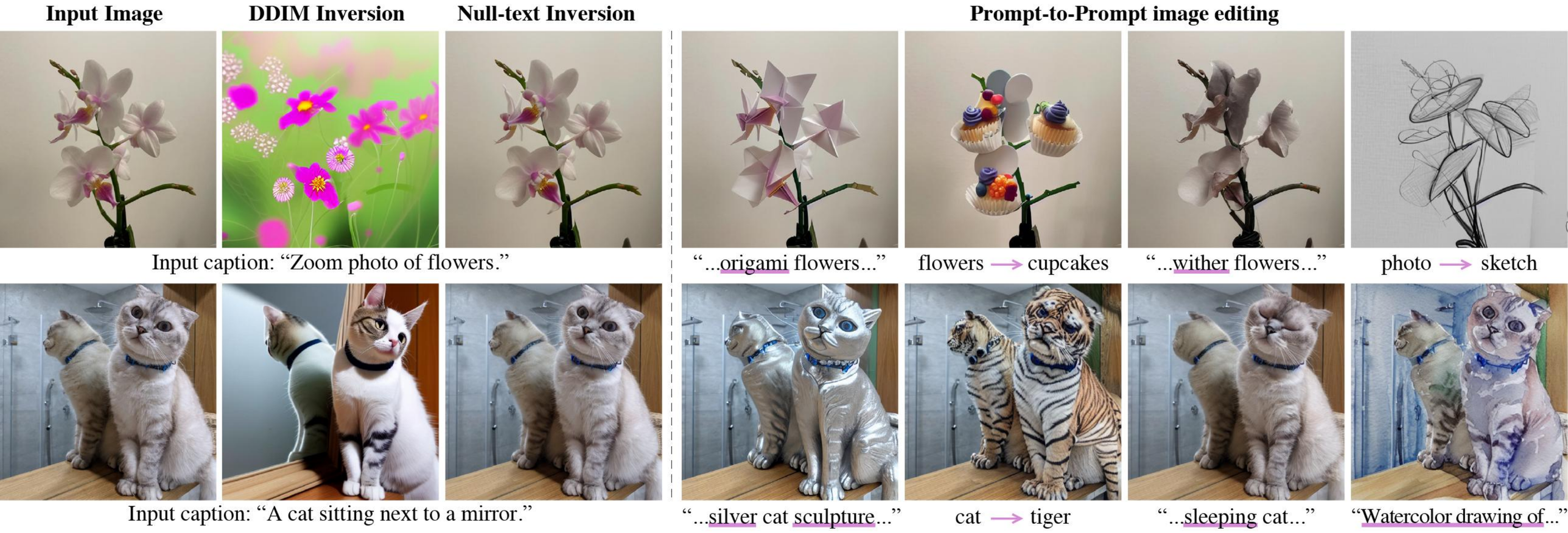
Google – Prompt-to-Prompt (Hertz et al., 2023)



Attention injection through a varied number of diffusion steps. We edit the image by replacing a word and injecting the cross-attention maps of the source image ranging from 0% (left) to 100% (right) of the steps. Without injection, none of the source content is preserved, while injecting throughout all the steps may over-constrain the geometry. The latter results in low fidelity to the text, e.g., the motorcycle becomes a bicycle.

Hertz et al., Prompt-to-Prompt Image Editing with Cross-Attention Control, ICLR 2023

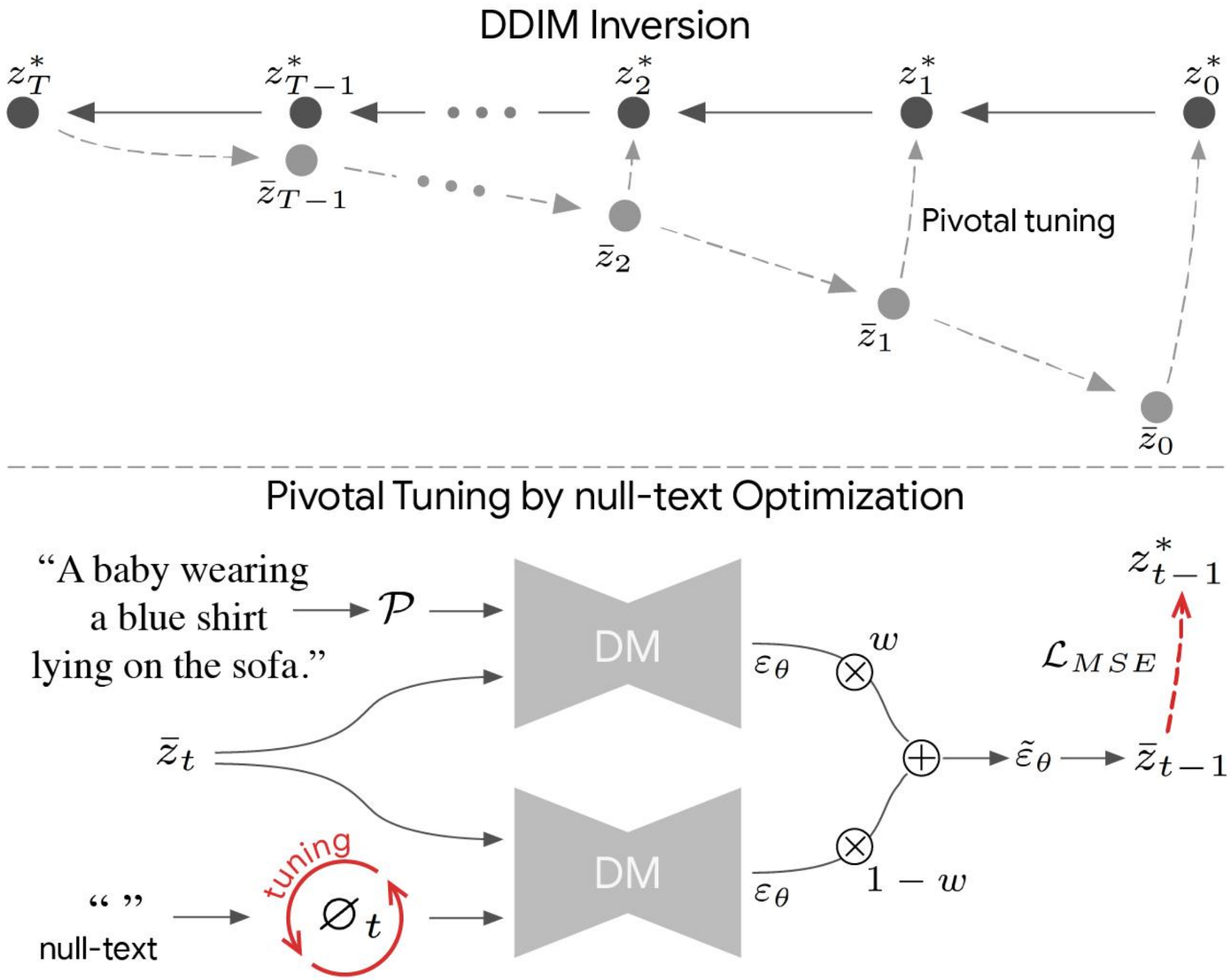
Google – Null-Text Inversion (Mokady et al., 2023)



Null-text inversion enables intuitive text-based editing of real images with the Stable Diffusion model. We use an initial DDIM inversion as an anchor for our optimization which only tunes the null-text embedding used in classifier-free guidance.

Mokady et al., NULL-Text Inversion for Editing Real Images Using Guided Diffusion Models, CVPR 2023

Google – Null-Text Inversion (Mokady et al., 2023)




Input Image


Initial Inversion

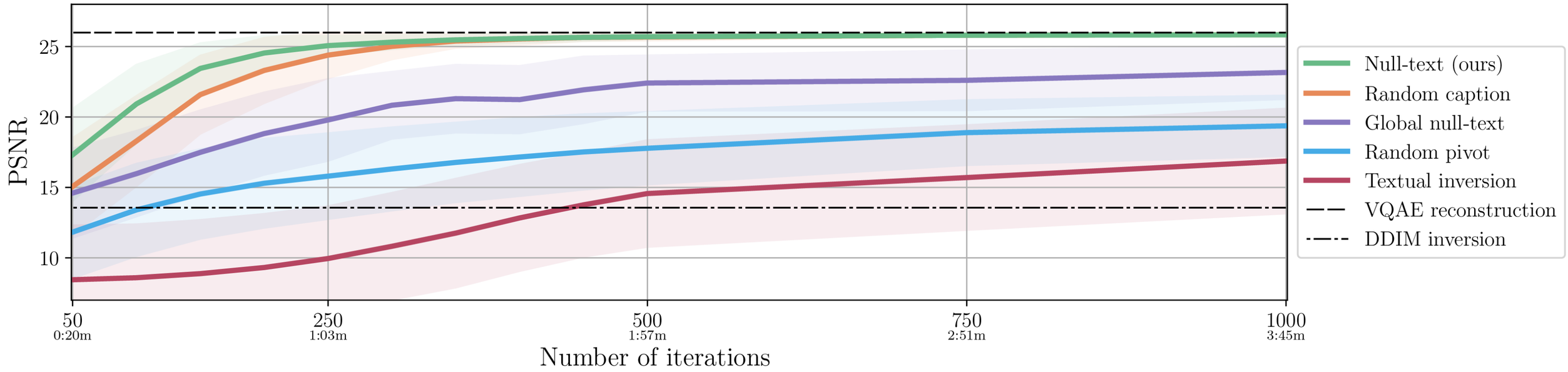

Final Inversion

On top: We first apply an initial DDIM inversion on the input image which estimates a diffusion trajectory (top trajectory). Starting the diffusion process from the last latent code results in unsatisfying reconstruction (bottom trajectory) as the intermediate codes become farther away from the original trajectory. We use the initial trajectory as a pivot for our optimization which brings the diffusion backward trajectory closer to the original image.

Bottom: null-text optimization for timestamp t. Recall that classifier-free guidance consists of performing the noise prediction twice – using text condition embedding and unconditionally using null-text embedding ∅ (bottom-left). Then, these are extrapolated with guidance scale w (middle). We optimize only the unconditional embeddings $\varnothing_t$ by employing a reconstruction MSE loss (in red) between the predicated latent code to the pivot.

Mokady et al., NULL-Text Inversion for Editing Real Images Using Guided Diffusion Models, CVPR 2023

Google – Null-Text Inversion (Mokady et al., 2023)



Input caption: "A black dinning room table sitting in a yellow dinning room."

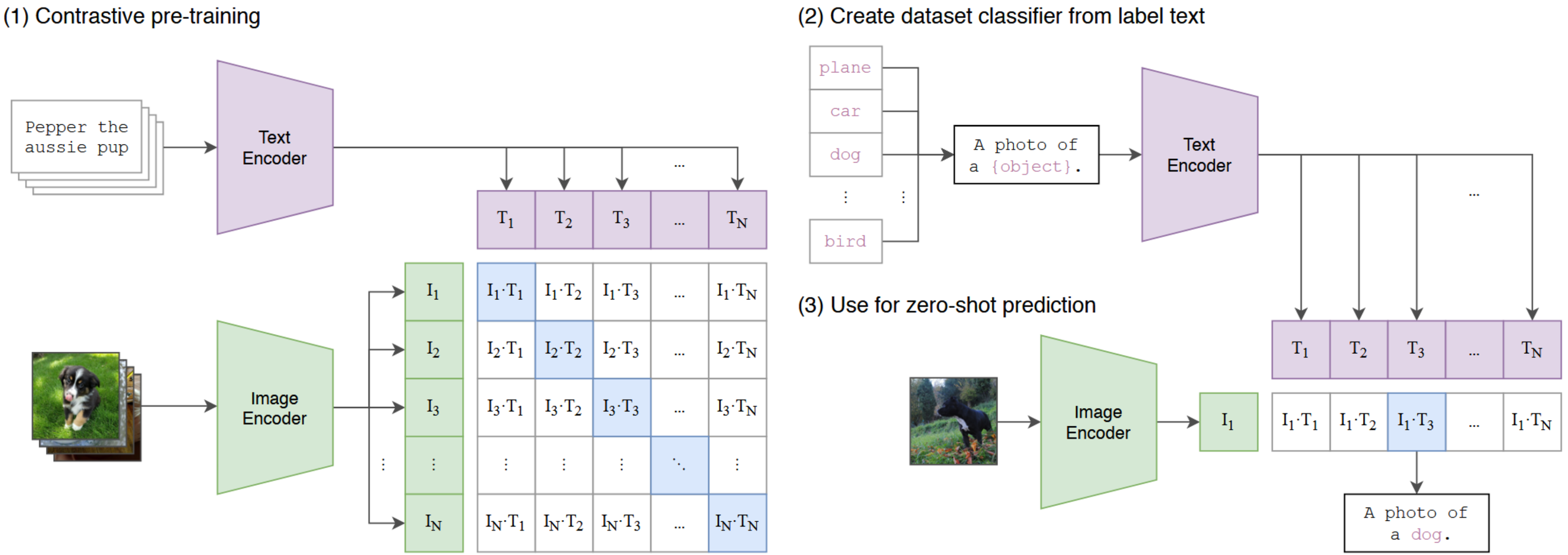
Mokady et al., NULL-Text Inversion for Editing Real Images Using Guided Diffusion Models, CVPR 2023

Google – Imagen 3 (Imagen Team, 2024)

Photo of a felt puppet diorama scene of a tranquil nature scene of a secluded forest clearing with a large friendly, rounded robot is rendered in a risograph style. An owl sits on the robots shoulders and a fox at its feet. Soft washes of color, 5 color, and a light-filled palette create a sense of peace and serenity, inviting contemplation and the appreciation of natural beauty.

Imagen Team Google, Imagen 3, arXiv 2024

OpenAI - CLIP (Radford et al., 2021)



Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021