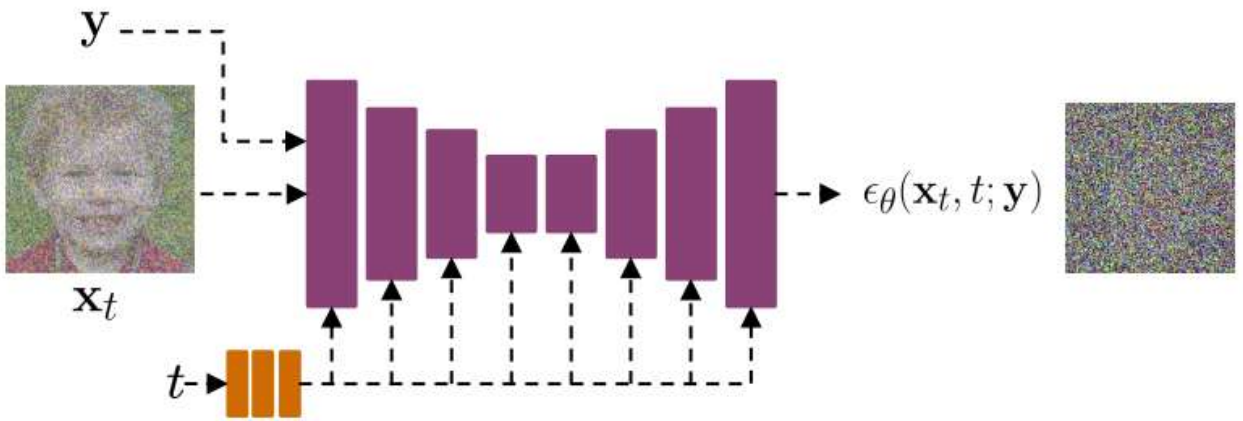
Recap.

Conditional generation

- Let (x,y) denote (image,caption) pairs
- Training a conditional generative model involves learning p(x | y)

- Train score model for the image x conditional on caption y

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p_{\text{data}}(\mathbf{x},\mathbf{y})}\mathbb{E}_{\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\mathbb{E}_{t\sim\mathcal{U}[0,T]}\ ||\epsilon_\theta(\mathbf{x}_t, t; \mathbf{y}) - \epsilon||_2^2$$



Classifier-Guided Diffusion

To explicit incorporate class information into the diffusion process, trained a classifier $f_\phi(y|\mathbf{x}_t, t)$ on noisy image the use gradients to guide diffusion sampling process toward the condition

$$\begin{aligned}
\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t, y) &= \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log q(y|\mathbf{x}_t) \\
&\approx -\frac{1}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) + \nabla_{\mathbf{x}_t} \log f_\phi(y|\mathbf{x}_t) \\
&= -\frac{1}{\sqrt{1-\bar{\alpha}_t}} (\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - \sqrt{1-\bar{\alpha}_t} \nabla_{\mathbf{x}_t} \log f_\phi(y|\mathbf{x}_t))
\end{aligned}$$

A new classifier-guided predictor $\bar{\epsilon}_\theta$ would take the form as following,

$$\bar{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t) = \boldsymbol{\epsilon}_\theta(x_t, t) - \sqrt{1-\bar{\alpha}_t} \nabla_{\mathbf{x}_t} \log f_\phi(y|\mathbf{x}_t)$$

To control the strength of the classifier guidance, we can add a weight w to the delta part,

$$\bar{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t) = \boldsymbol{\epsilon}_\theta(x_t, t) - \sqrt{1-\bar{\alpha}_t}\, w \nabla_{\mathbf{x}_t} \log f_\phi(y|\mathbf{x}_t)$$

Classifier-Free Guidance

- Train both a conditional and an unconditional score model (by randomly dropping the caption during training)

- Combine the two models as follows

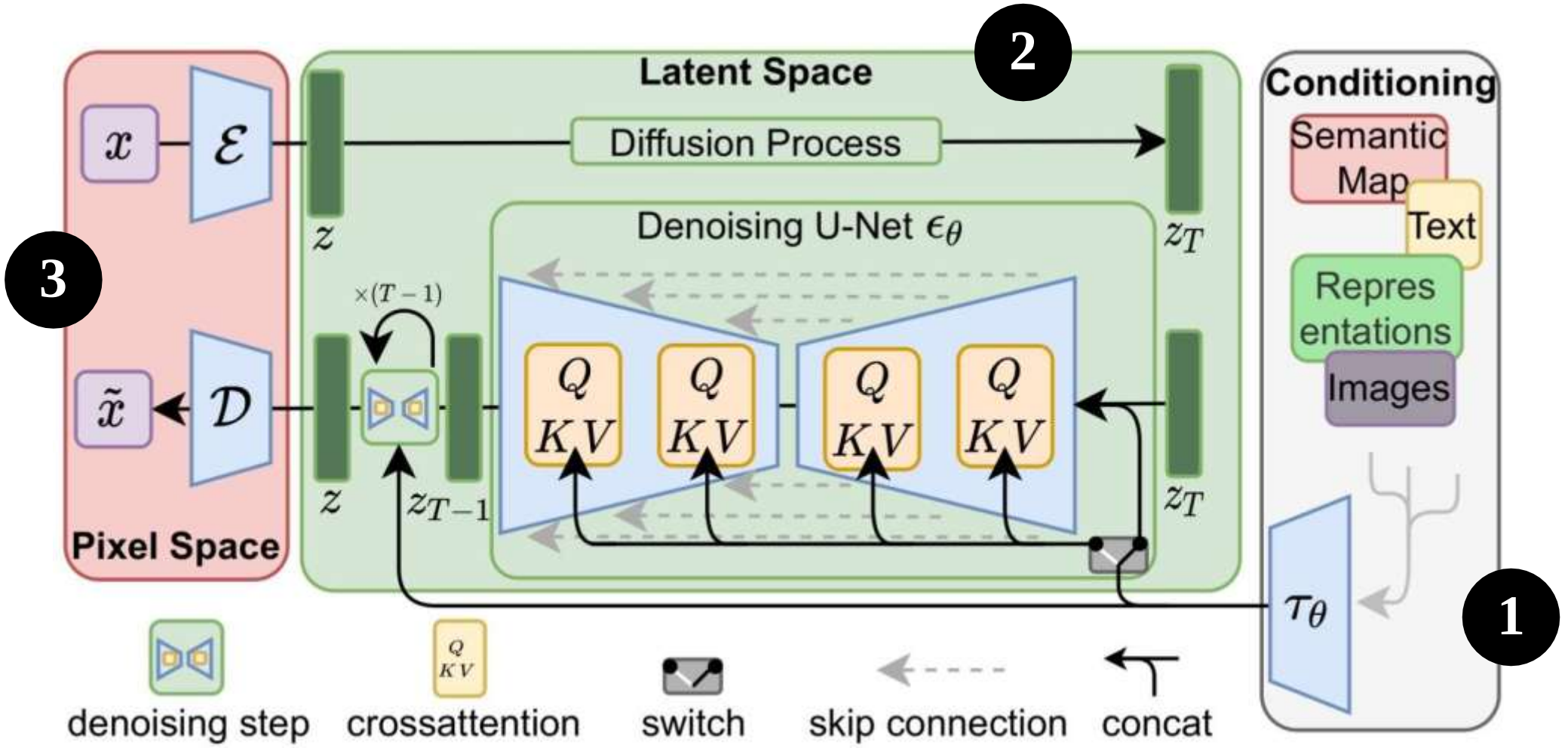
$$(1+w)\nabla_{\mathbf{x}_t}\log p(\mathbf{y}|\mathbf{x}_t)+\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t)=(1+w)\big(\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t|\mathbf{y})-\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t)\big)+\nabla_{\mathbf{x}_t}\log p(\mathbf{x}_t)$$

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

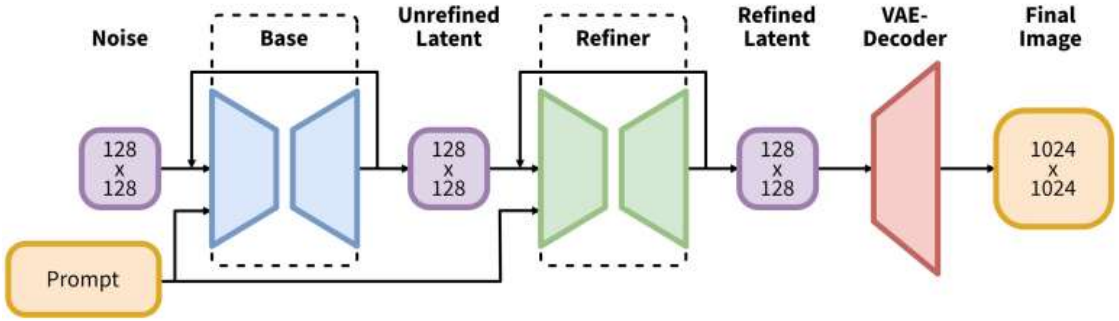
- w is the classifier-guidance strength

Stable Diffusion

https://arxiv.org/abs/2112.10752



SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis

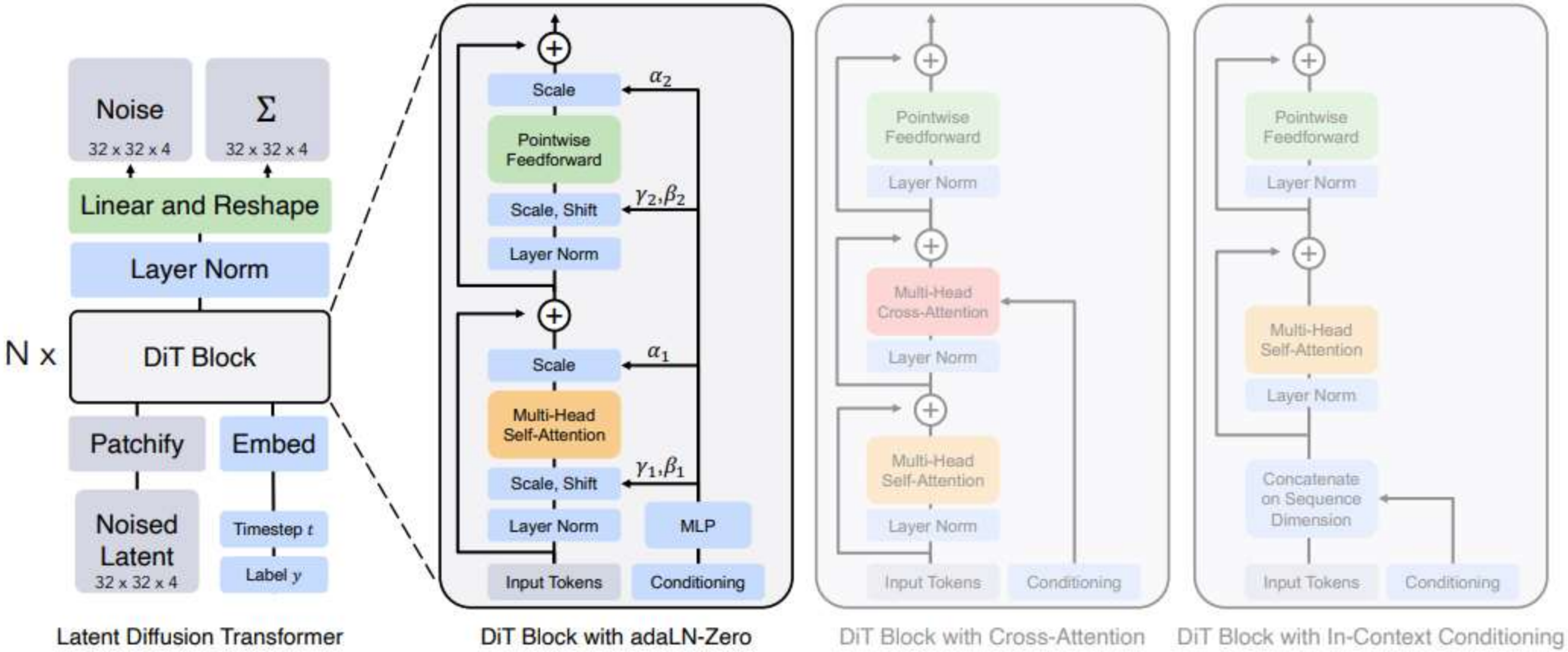


https://arxiv.org/pdf/2307.01952

ControlNet: Adding Conditional Control to Text-to-Image Diffusion Models

https://arxiv.org/abs/2302.05543

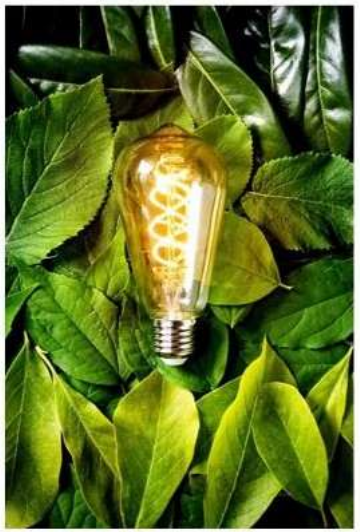
DiT: Scalable Diffusion Models with Transformers



https://arxiv.org/pdf/2212.09748

CLIP is good but…

It cannot tell these apart!



(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

Thrush et al. "Winoground: Probing vision and language models for visio-linguistic compositionality". CVPR 2022. https://arxiv.org/pdf/2204.03162.pdf

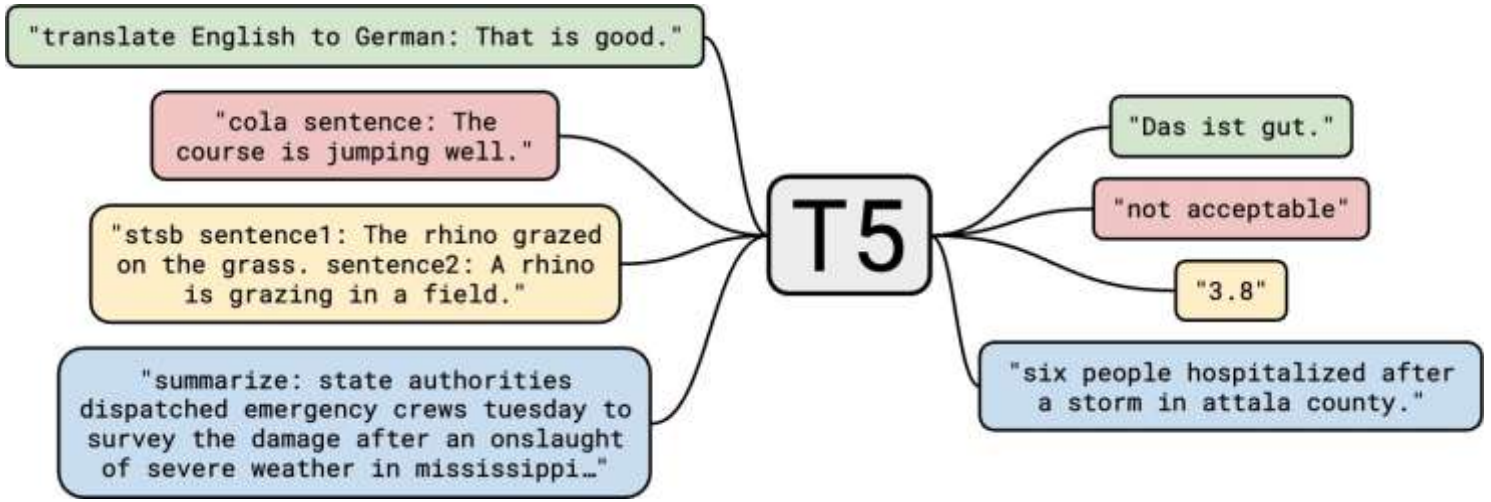
CLIP is good but…

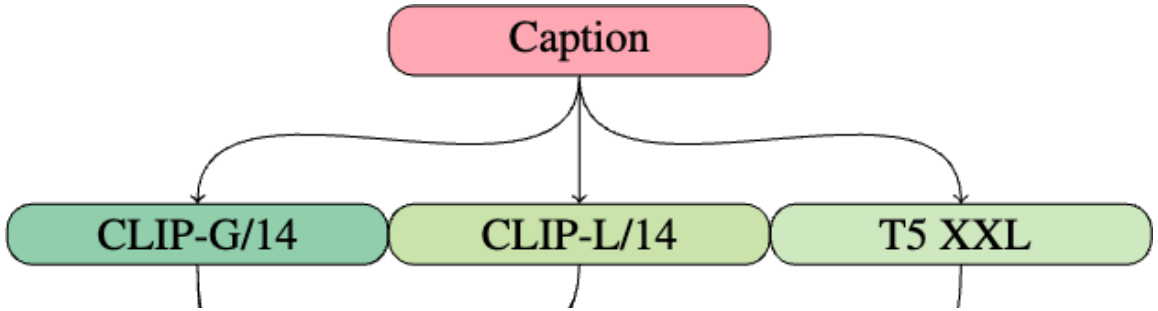
- It cannot handle spatial relation ships well
- It cannot handle negations well
- It cannot handle counts well
- The length limit is 77 tokens
- Sometimes ignores some details

CLIP is not a very good text encoder!

- What we want:
- Can capture semantics well ✗
- Can encode long sentences ✗
- Can attend to details ✗
- Can differentiate between different spatial relationships described in text ✗

Attempt : Add another text encoder on top of CLIP for better language understanding

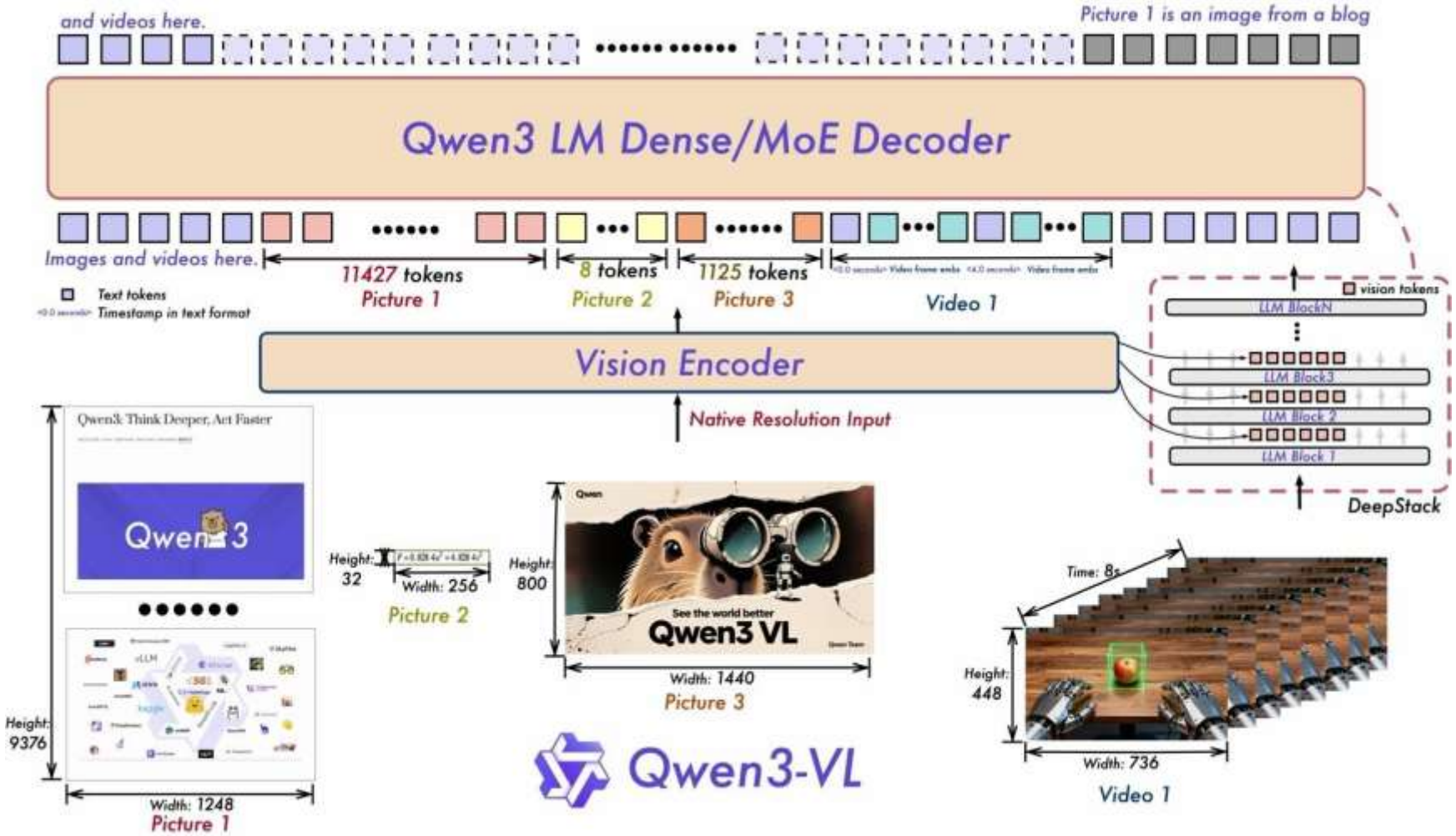




Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". JMLR 2020. https://arxiv.org/pdf/1910.10683

Esser et al. "Scaling Rectified Flow Transformers for High-Resolution Image Synthesis". https://arxiv.org/pdf/2403.03206

Attempt : Why not just use an LLM/VLM/MLLM

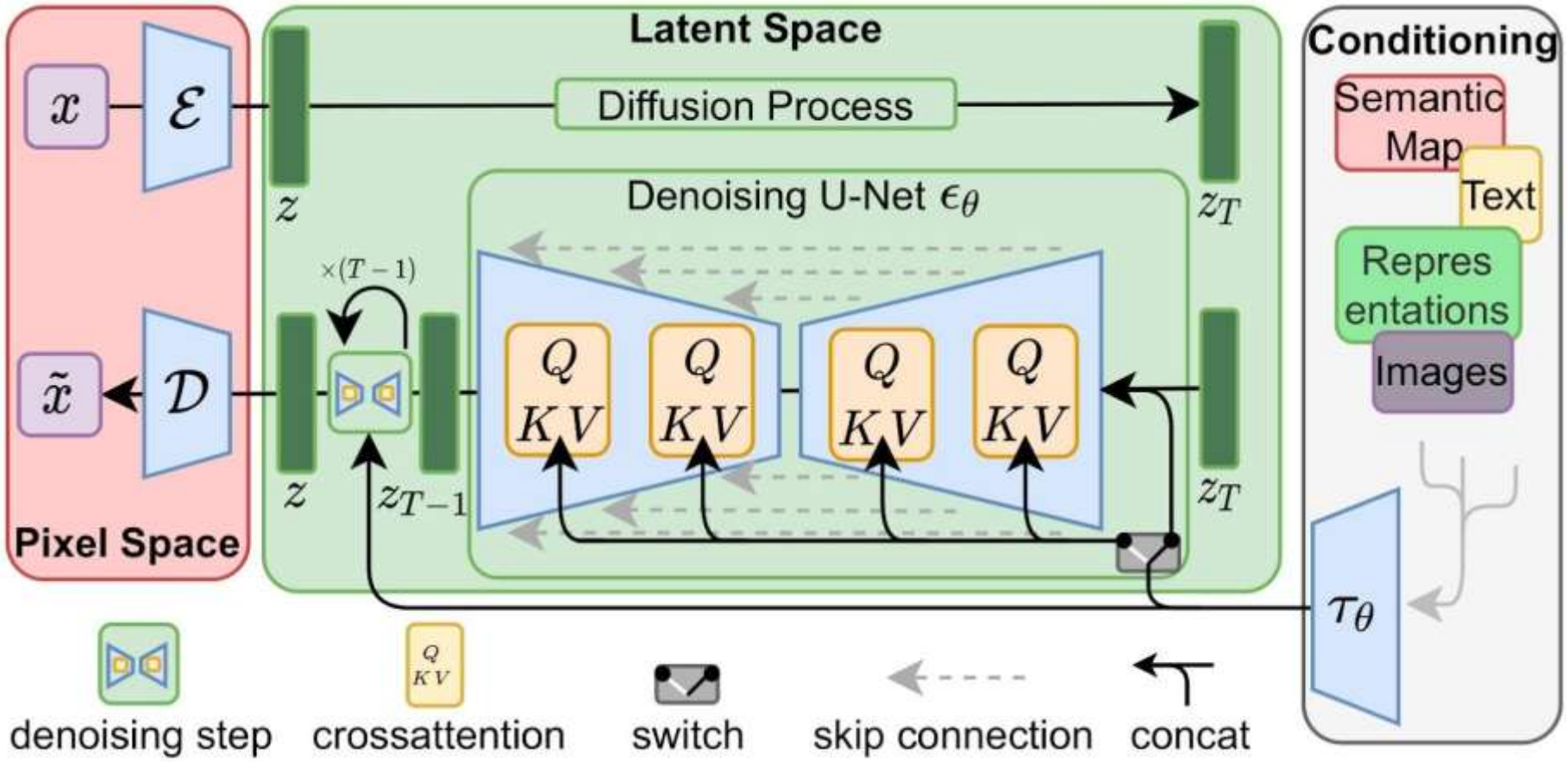


Qwen Team. “Qwen3-VL Technical Report”. https://arxiv.org/pdf/2511.21631

How to input text into an image generative model?

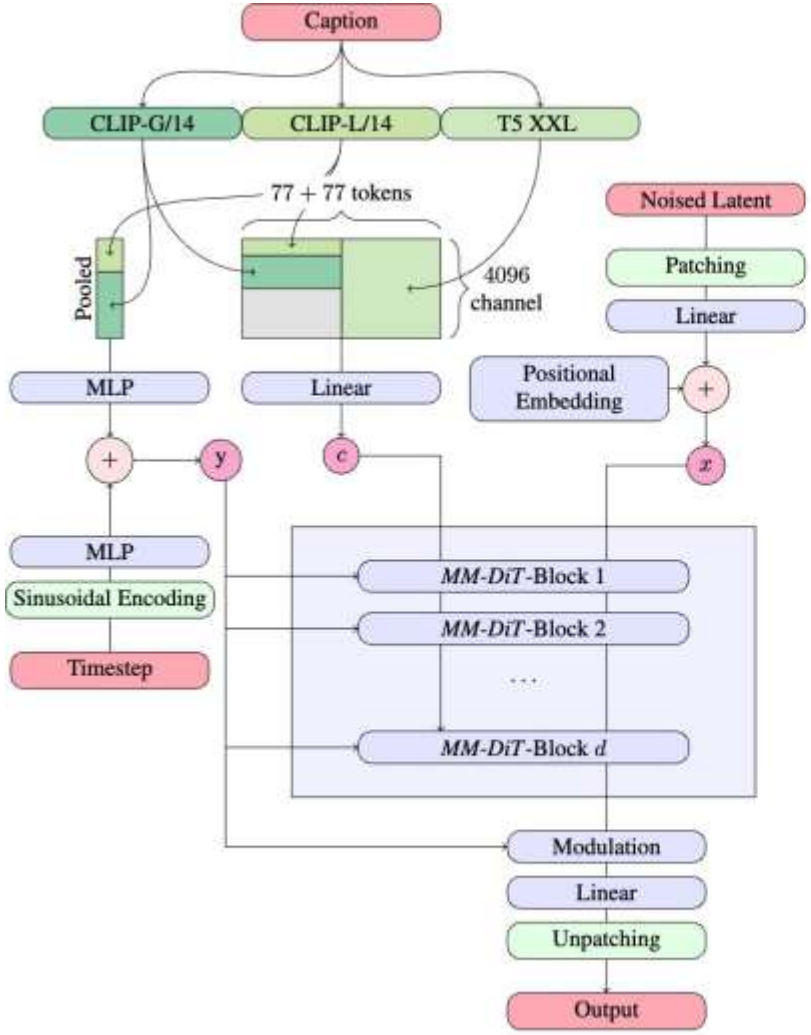
1. Somehow encode text into some feature vectors 
2. Somehow squish the encoded text features into the diffusion model 

Attempt 1: Cross Attention

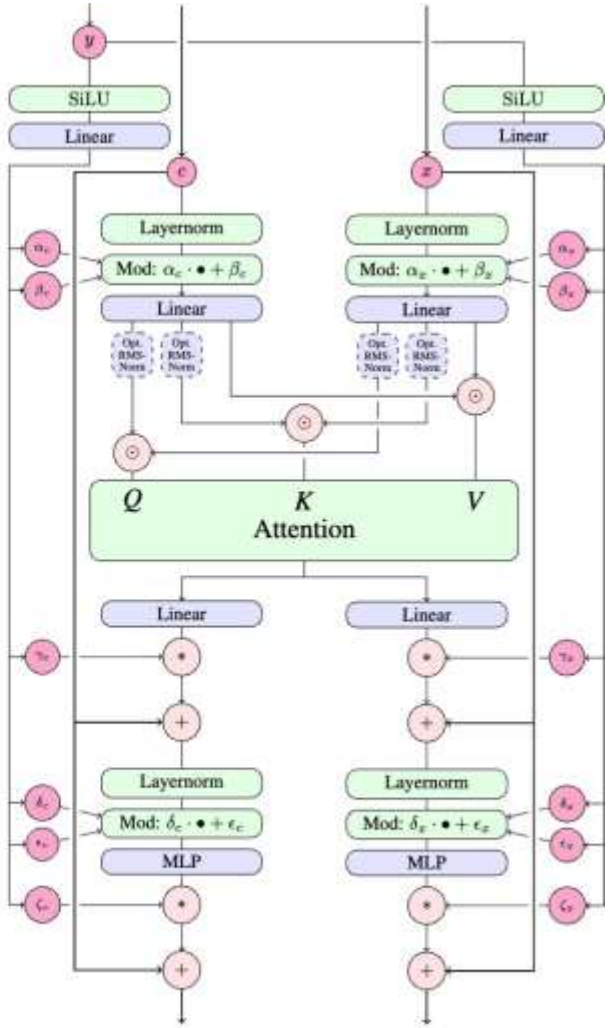


Rombach et al. "High-Resolution Image Synthesis with Latent Diffusion Models". CVPR 2022. https://arxiv.org/pdf/2112.10752

Attempt 2: Double Stream Multimodal-DiT



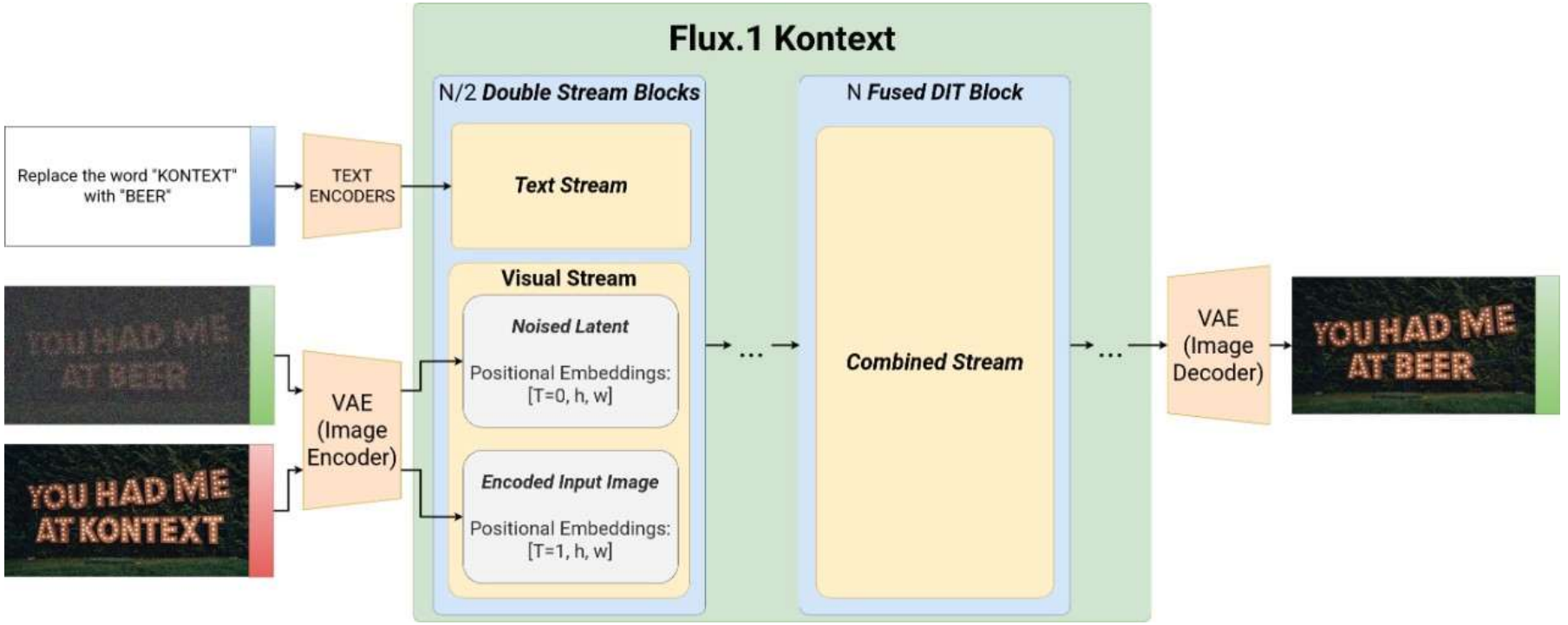
(a) Overview of all components.



(b) One *MM-DiT* block

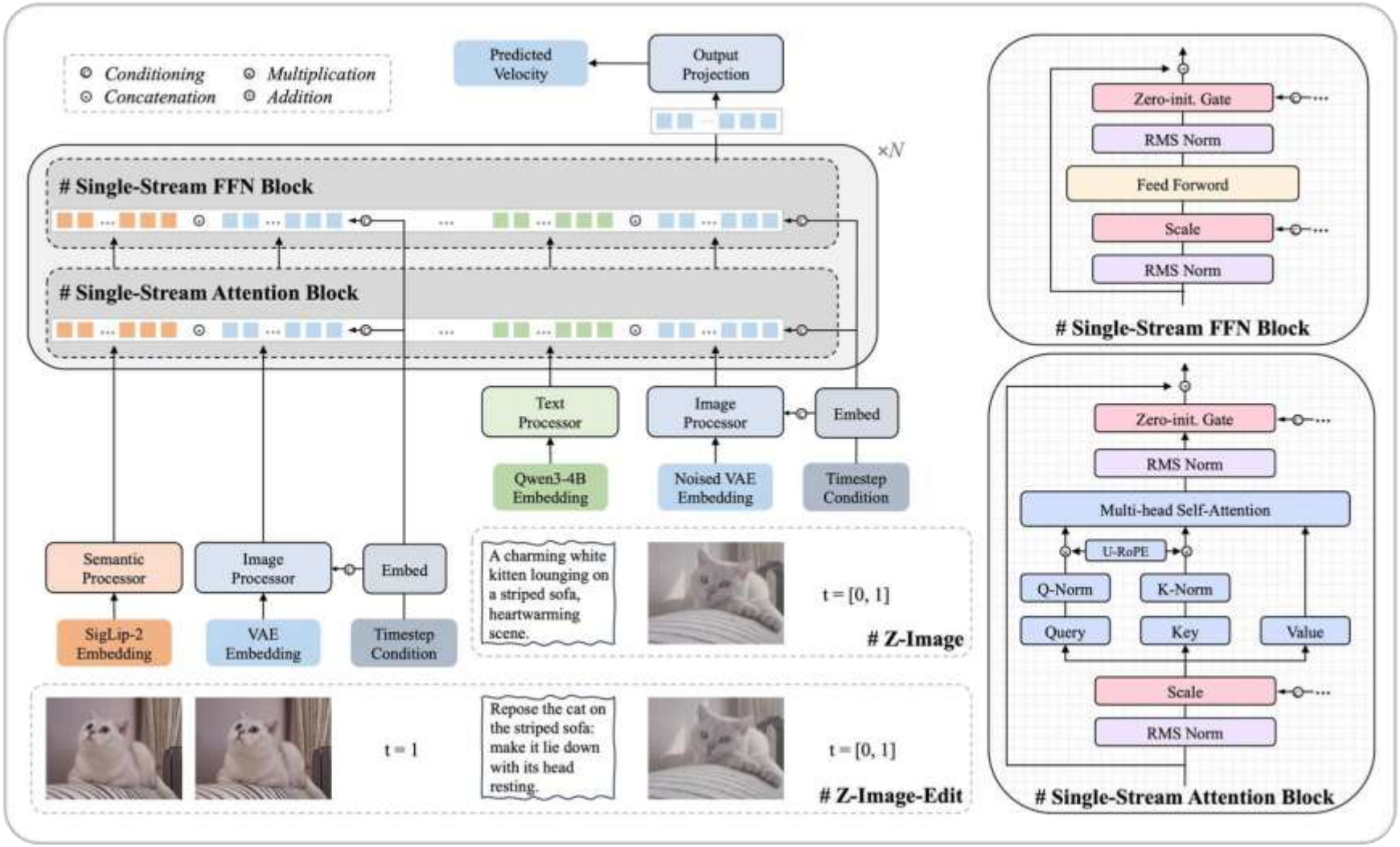
Esser et al. "Scaling Rectified Flow Transformers for High-Resolution Image Synthesis". https://arxiv.org/pdf/2403.03206

Attempt 2: Double stream -> merged stream MM-DiT



Black Forest Labs. "FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space". https://arxiv.org/pdf/2506.15742

Attempt 3: Single stream MM-DiT



Z-Image Team, Alibaba Group. "Z-Image: An Efficient Image Generation Foundation Model with Single-Stream Diffusion Transformer". https://arxiv.org/pdf/2511.22699

Attempt 3.5: “Native multimodal model”

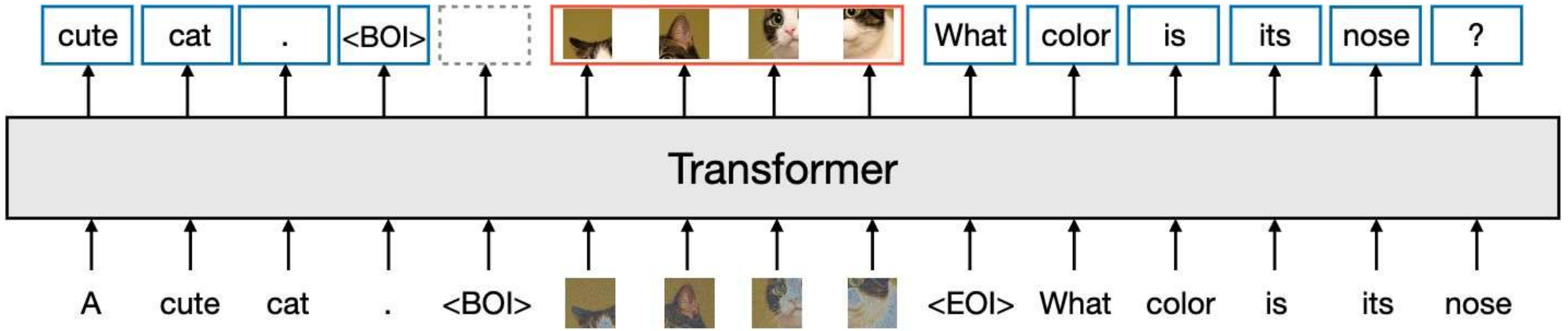
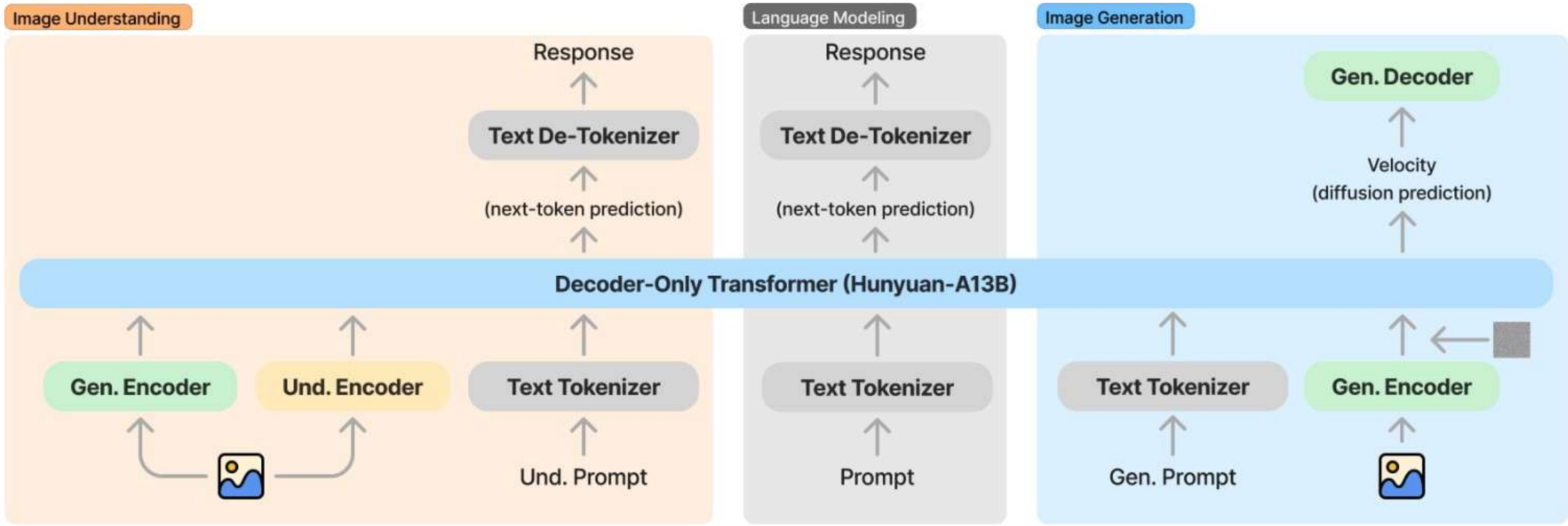


Figure 1: A high-level illustration of Transfusion. A single transformer perceives, processes, and produces data of every modality. Discrete (text) tokens are processed autoregressively and trained on the next token prediction objective. Continuous (image) vectors are processed together in parallel and trained on the diffusion objective. Marker BOI and EOI tokens separate the modalities.

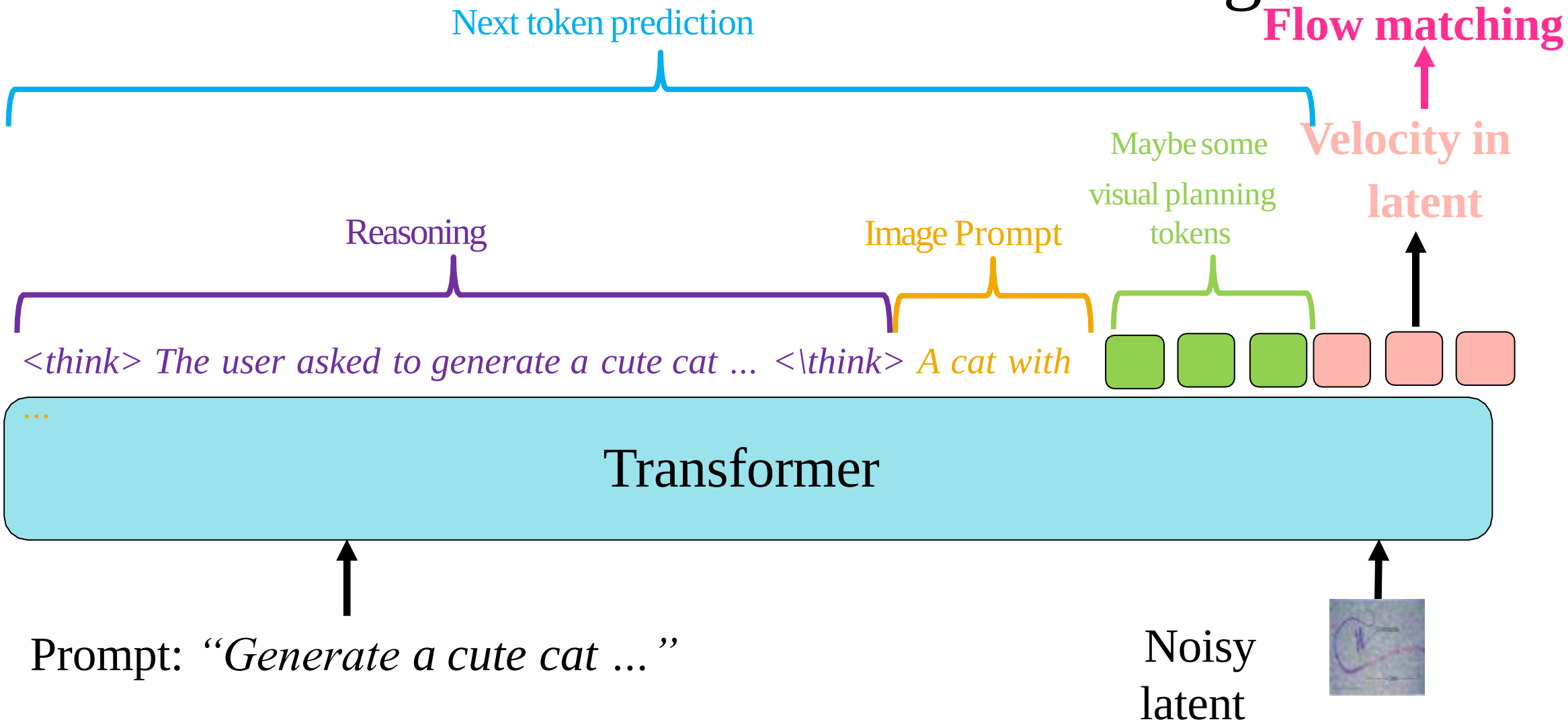
Zhou et al. “Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model”. https://arxiv.org/pdf/2408.11039

Attempt 3.5: “Native multimodal model”

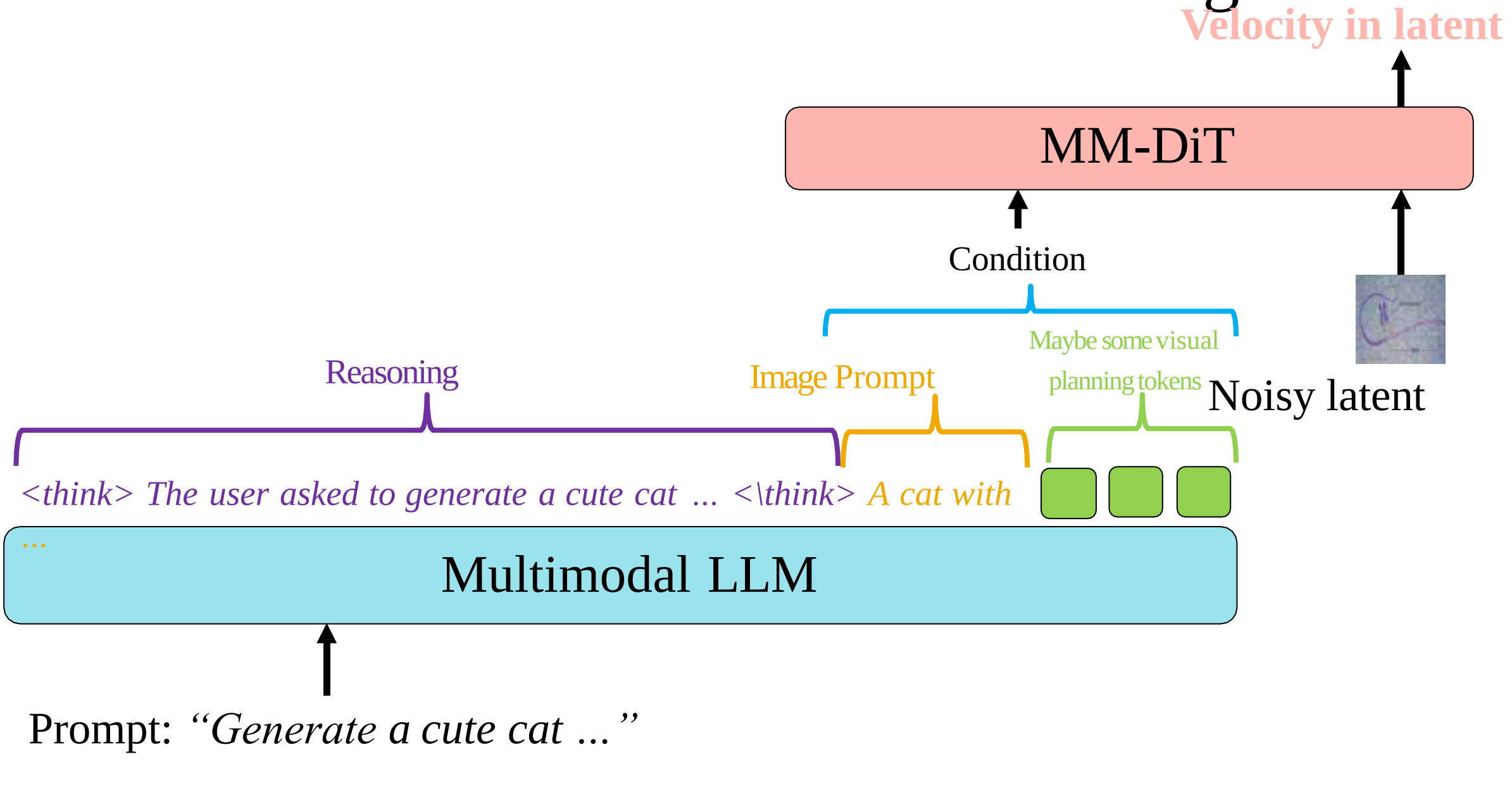


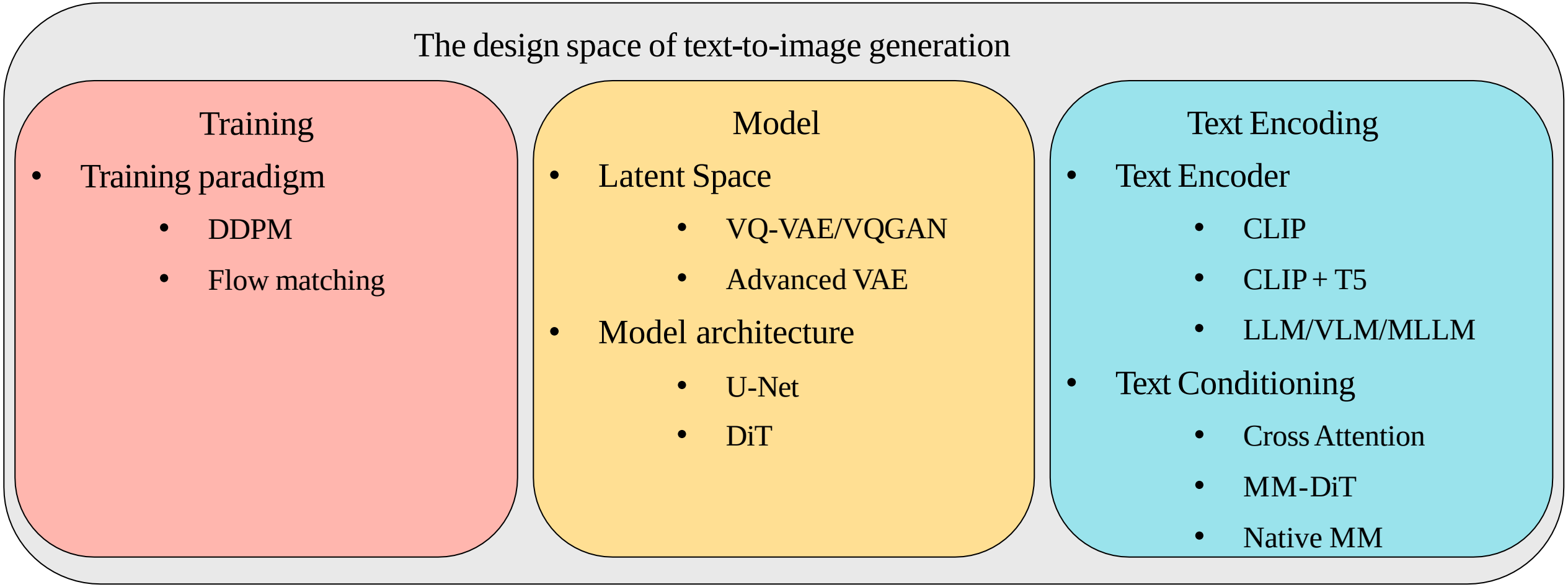
Tencent Hunyuan Foundation Model Team. “HunyuanImage 3.0 Technical Report”. https://arxiv.org/pdf/2408.11039

Guess on how Nano Banana & GPT-4o Image work



Guess on how Nano Banana & GPT-4o Image work




The design space of text-to-image generation
Training
Training paradigm
DDPM
Flow matching
Model
Latent Space
VQ-VAE/VQGAN
Advanced VAE
Model architecture
U-Net
DiT
Text Encoding
Text Encoder
CLIP
CLIP + T5
LLM/VLM/MLLM
Text Conditioning
Cross Attention
MM-DiT
Native MM

The design space of text-to-image generation

Training

- Training paradigm
  - DDPM
  - Flow matching

Model

- Latent Space
  - VQ-VAE/VQGAN
  - Advanced VAE
- Model architecture
  - U-Net
  - DiT

Text Encoding

- Text Encoder
  - CLIP
  - CLIP + T5
  - LLM/VLM/MLLM
- Text Conditioning
  - Cross Attention
  - MM-DiT
  - Native MM

This is Stable Diffusion 1 & 2!

The design space of text-to-image generation

Training

- Training paradigm
  - DDPM
  - Flow matching

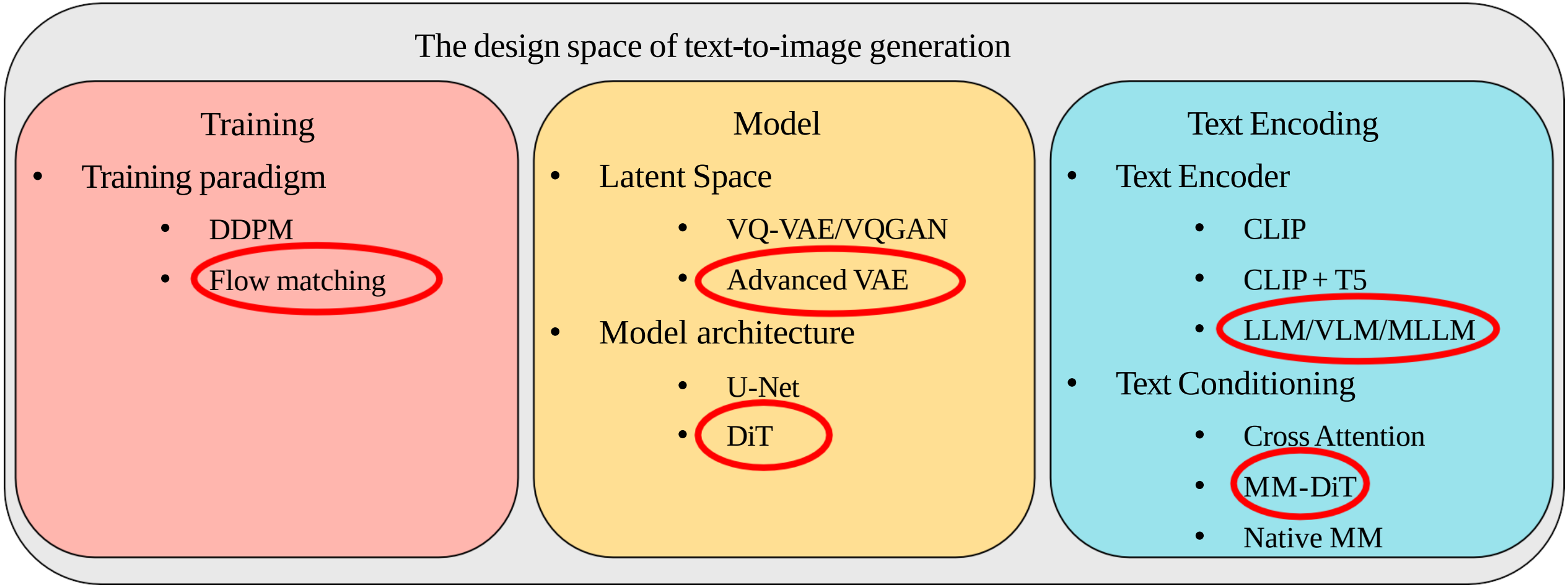
Model

- Latent Space
  - VQ-VAE/VQGAN
  - Advanced VAE
- Model architecture
  - U-Net
  - DiT

Text Encoding

- Text Encoder
  - CLIP
  - CLIP + T5
  - LLM/VLM/MLLM
- Text Conditioning
  - Cross Attention
  - MM-DiT
  - Native MM

This is Stable Diffusion 3 & Flux 1!



This is Flux 2, Z-Image, Qwen-Image, etc!

The design space of text-to-image generation

Training

- Training paradigm
  - DDPM
  - Flow matching

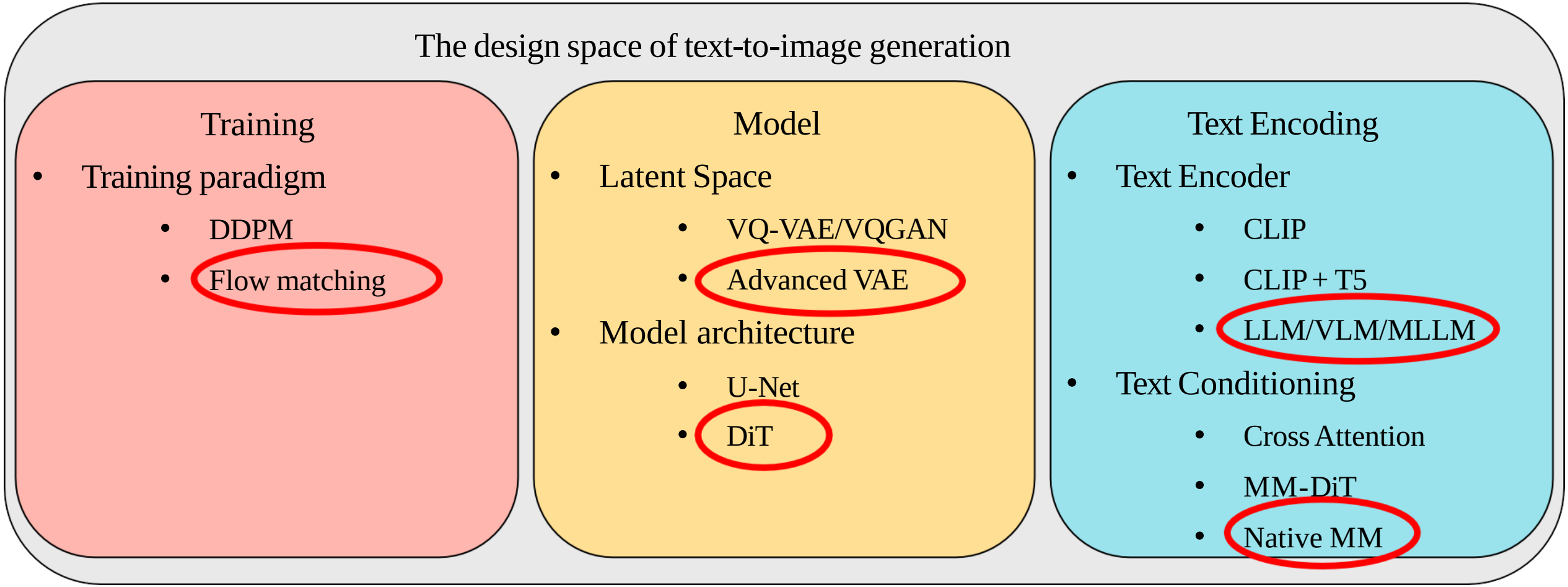
Model

- Latent Space
  - VQ-VAE/VQGAN
  - Advanced VAE
- Model architecture
  - U-Net
  - DiT

Text Encoding

- Text Encoder
  - CLIP
  - CLIP + T5
  - LLM/VLM/MLLM
- Text Conditioning
  - Cross Attention
  - MM-DiT
  - Native MM

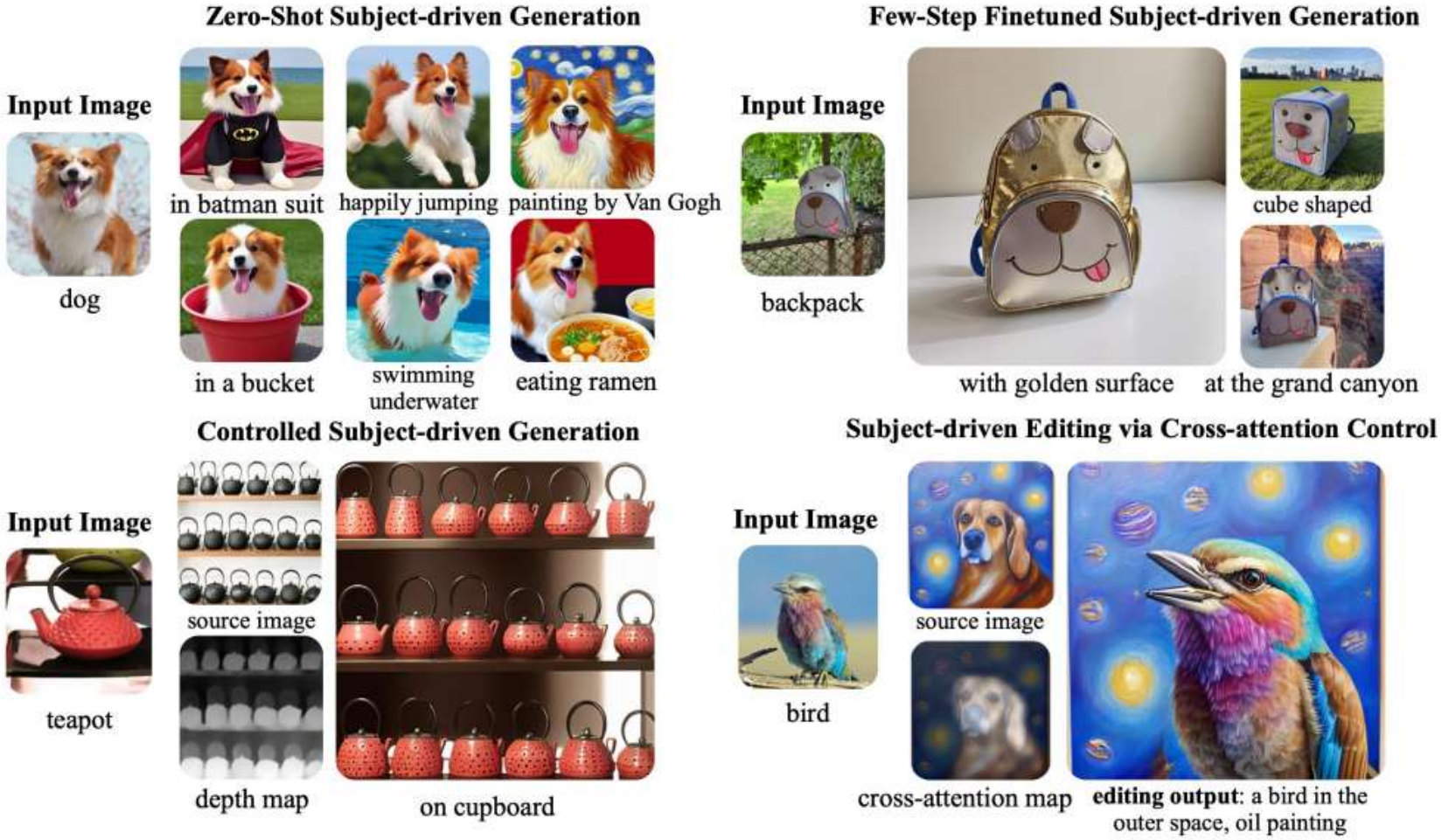
This is Transfusion, Hunyuan 3.0, etc!



(Probably also Nano Banana & GPT-4o Image)

Personalization
&
Editing

Subject-Driven Controllable Generatioin



BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing

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

Subject-Driven

Input images

in the Acropolis

swimming

sleeping

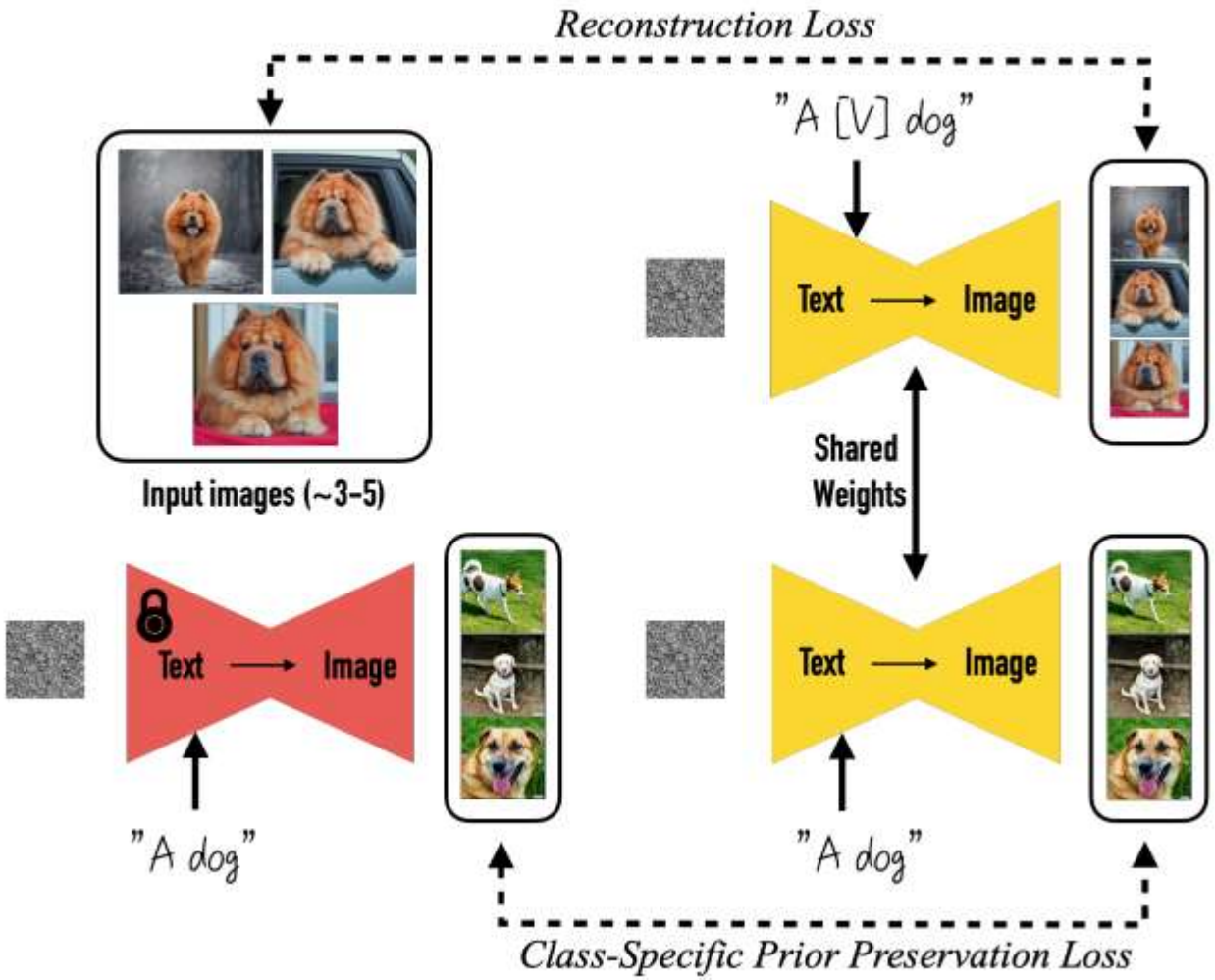
in a doghouse

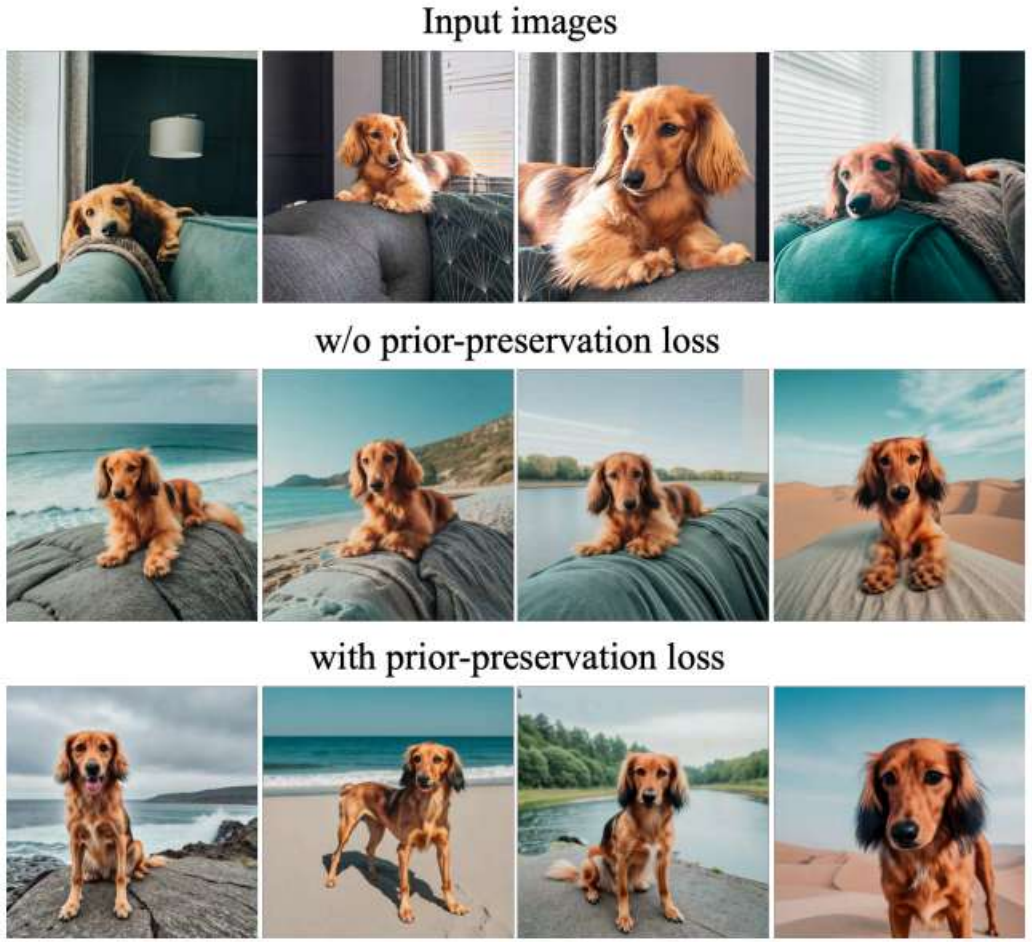
in a bucket

getting a haircut

https://arxiv.org/pdf/2208.12242

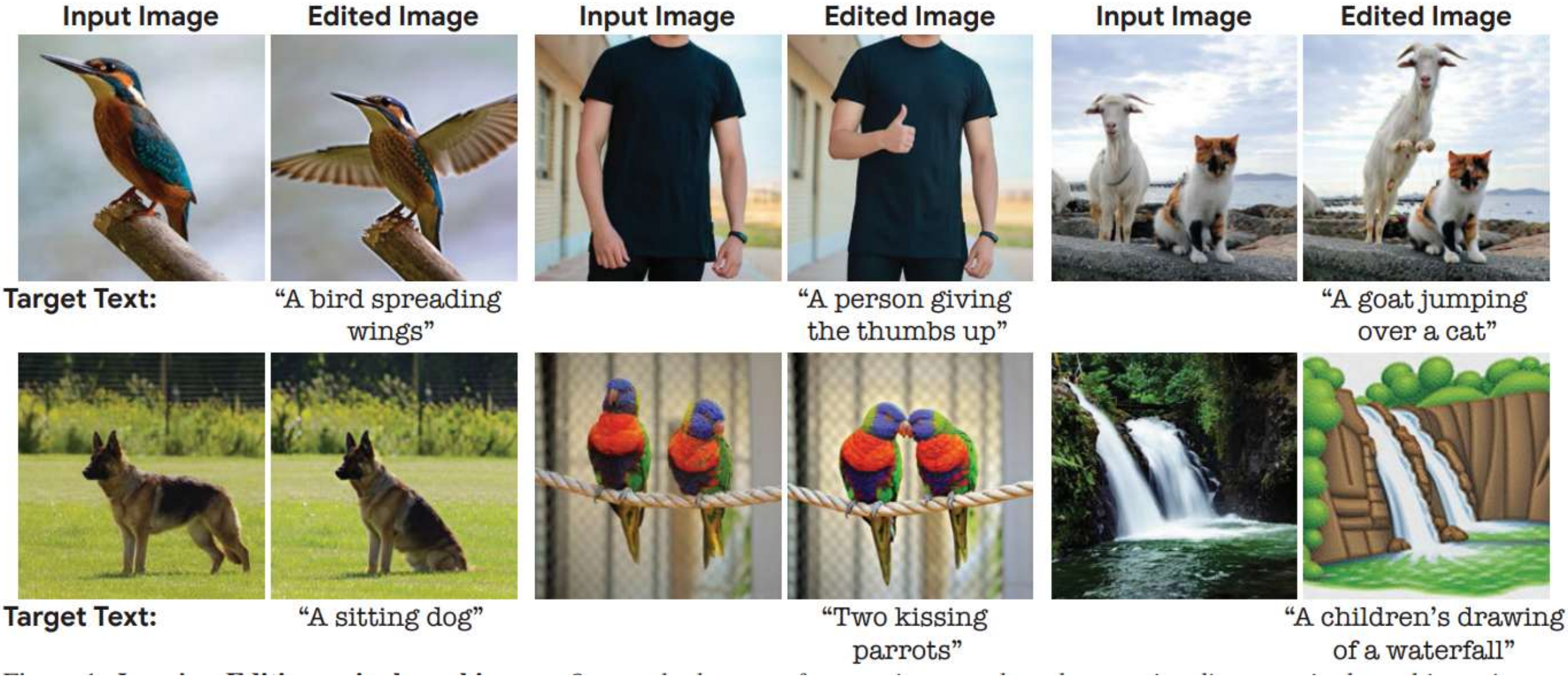
DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation



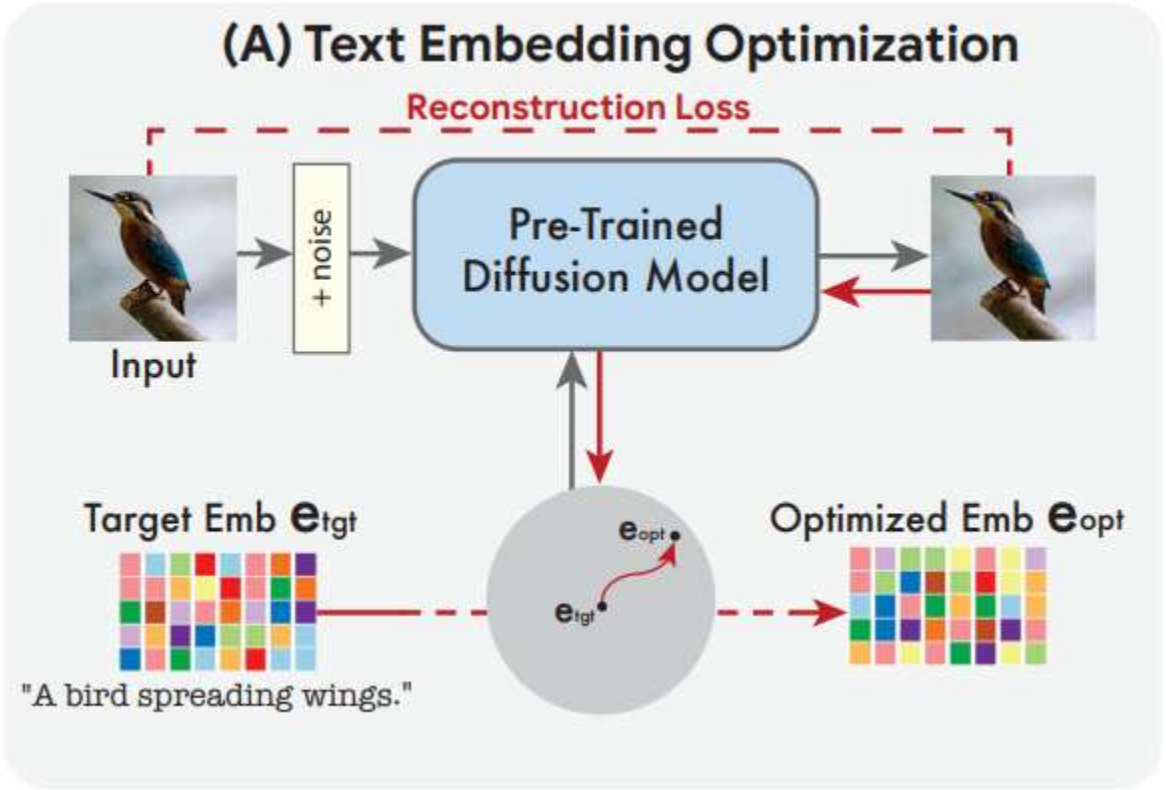


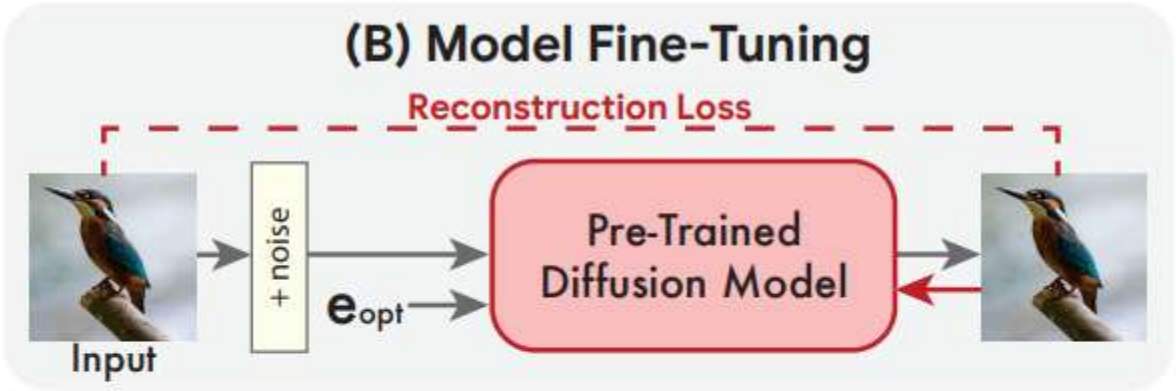
https://arxiv.org/pdf/2208.12242

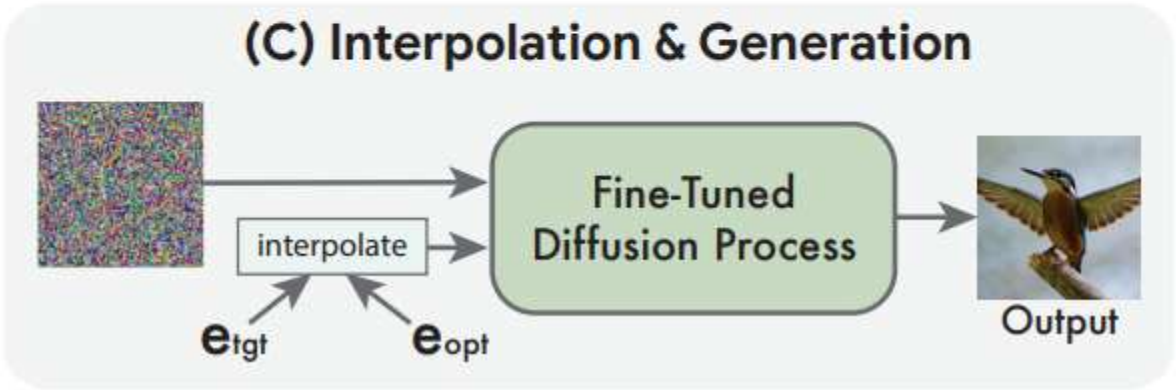
Imagic: Text-Based Real Image Editing with Diffusion Models



Imagic: Text-Based Real Image Editing with Diffusion Models





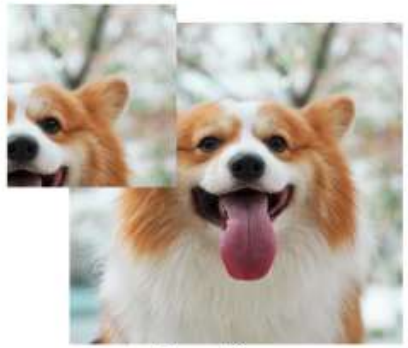


Not that ideally…

a_photo_of_a_sks_dog_playing_in_the_water

a_photo_of_The_sks_dog_is_playing_with_a_sandbox

ComFusion: Personalized Subject Generation in Multiple Specific Scenes From Single Image



Input Image

Input Text: A dog in the rain

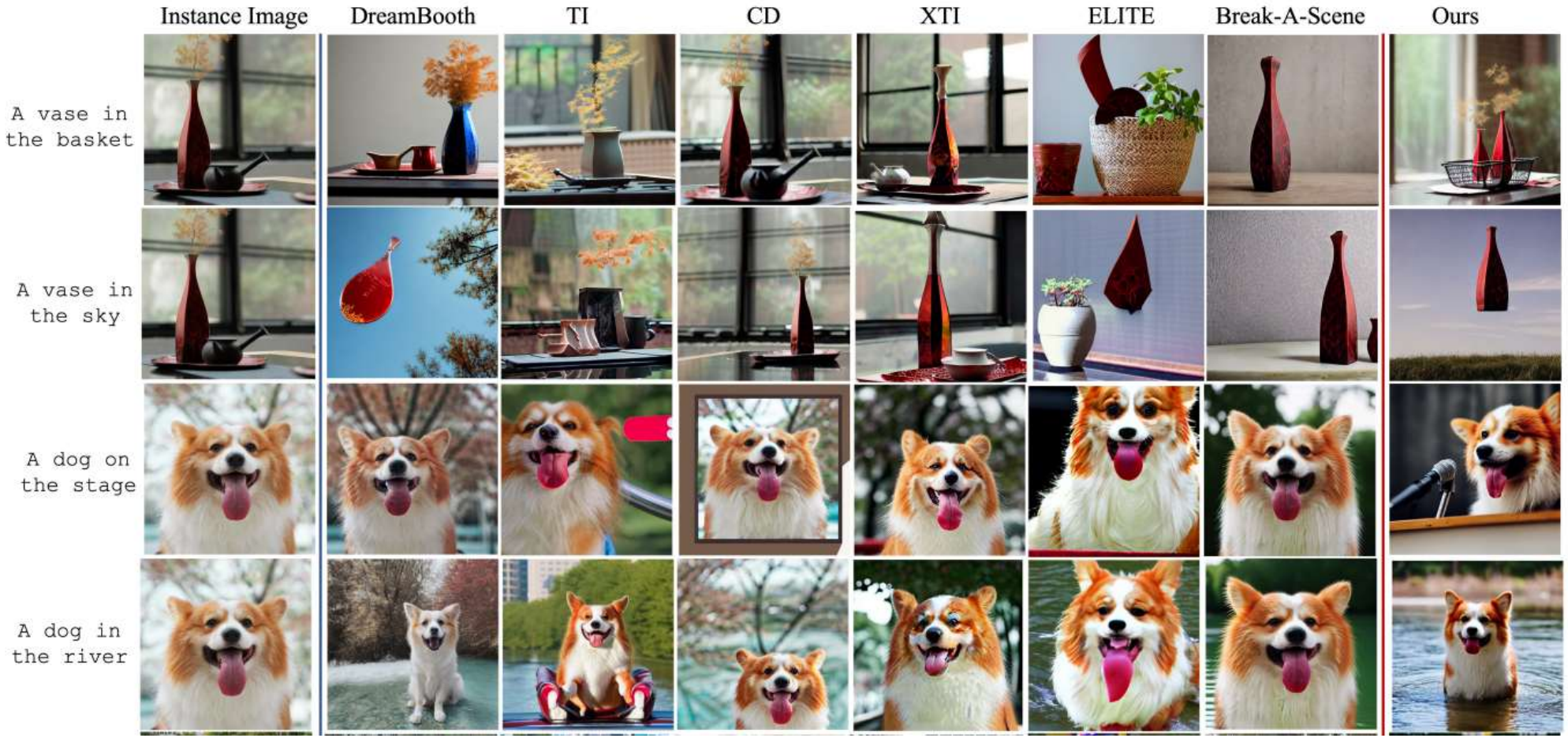
(a). Low Personality

(b). Low Diversity

(c). Ours

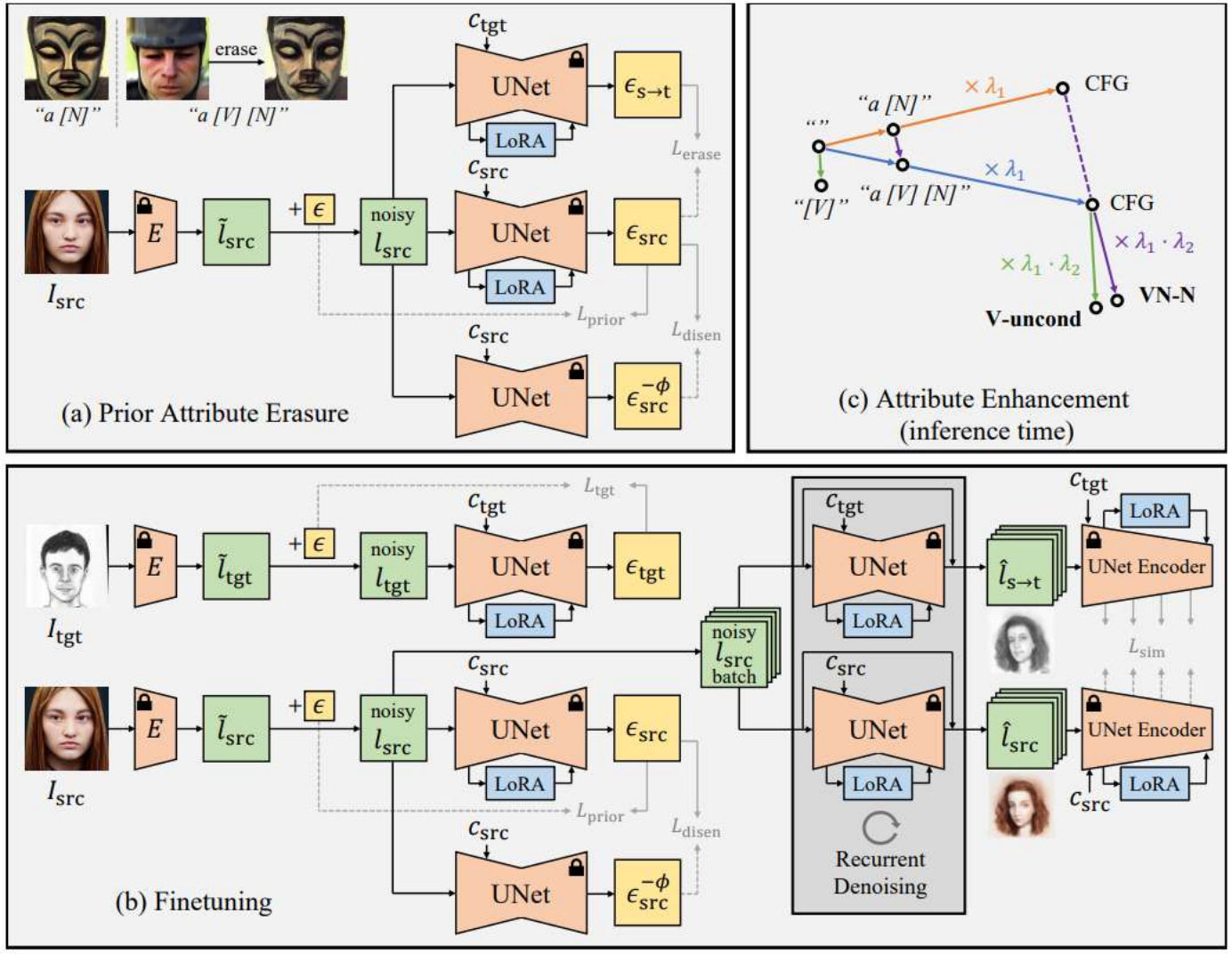
https://arxiv.org/abs/2402.11849

ComFusion: Personalized Subject Generation in Multiple Specific Scenes From Single Image

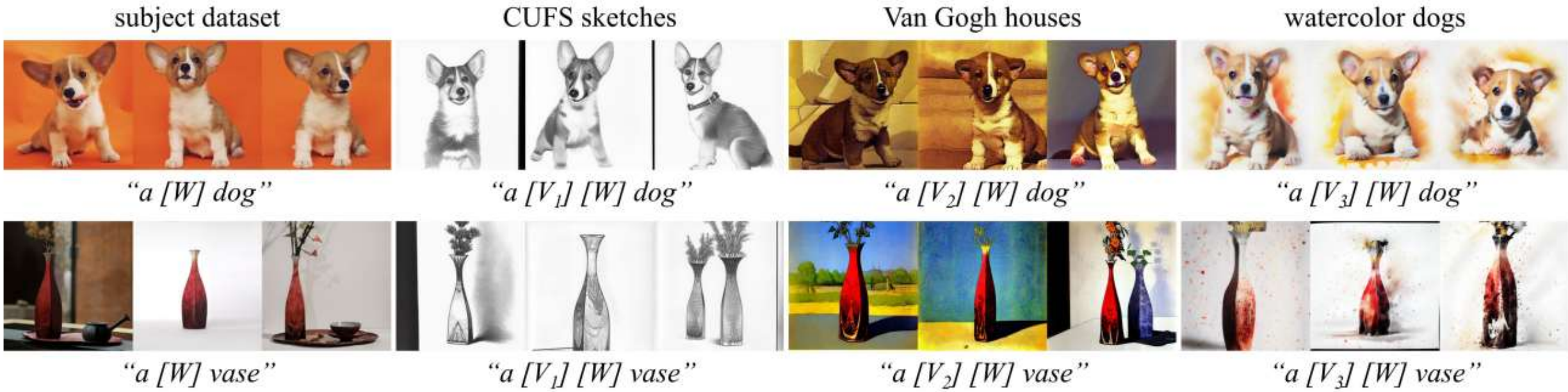


https://arxiv.org/abs/2402.11849

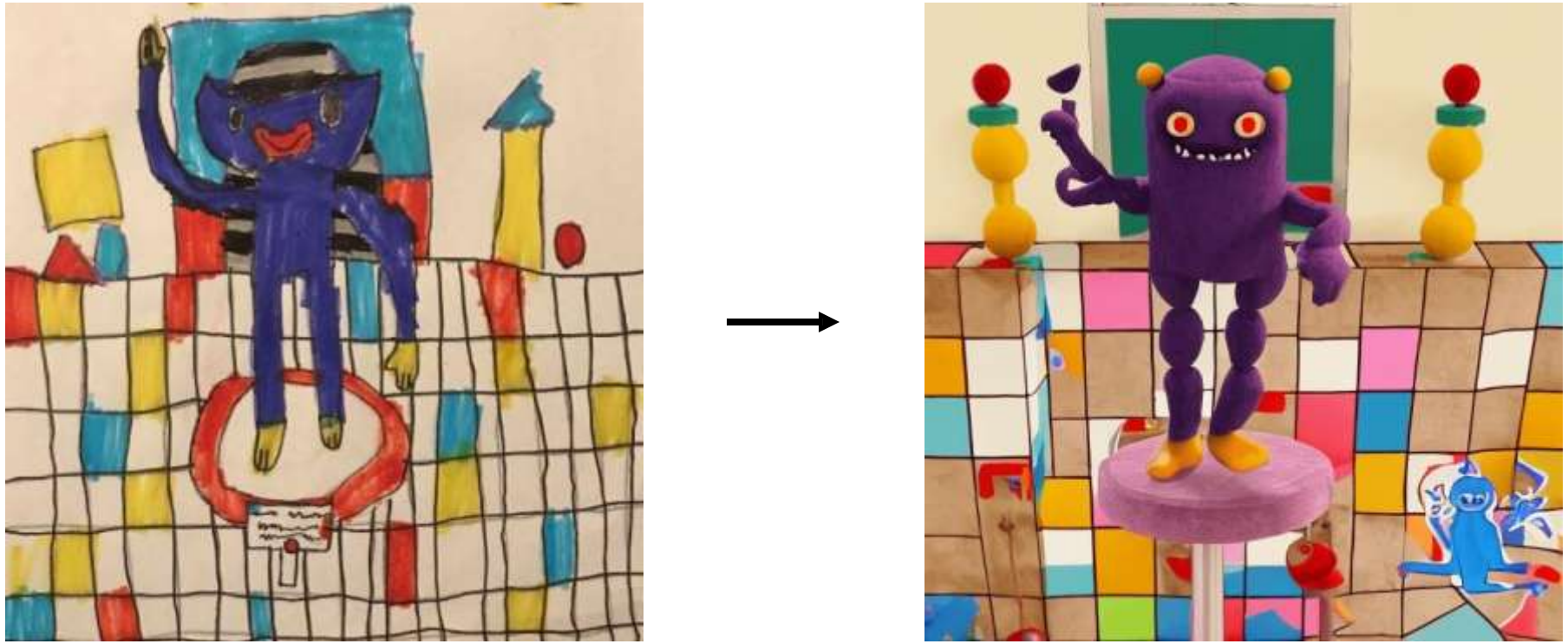
DomainGallery: Few-shot Domain-driven Image Generation by Attribute-centric Finetuning



DomainGallery: Few-shot Domain-driven Image Generation by Attribute-centric Finetuning



SDEdit: Guided Image Synthesis with Diffusion



https://www.reddit.com/r/StableDiffusion/comments/wyq04v/using_img2img_to_upgrade_my_sons_artwork/ Concurrent work with SDEdit: ILVR [Choi et al., 2021]

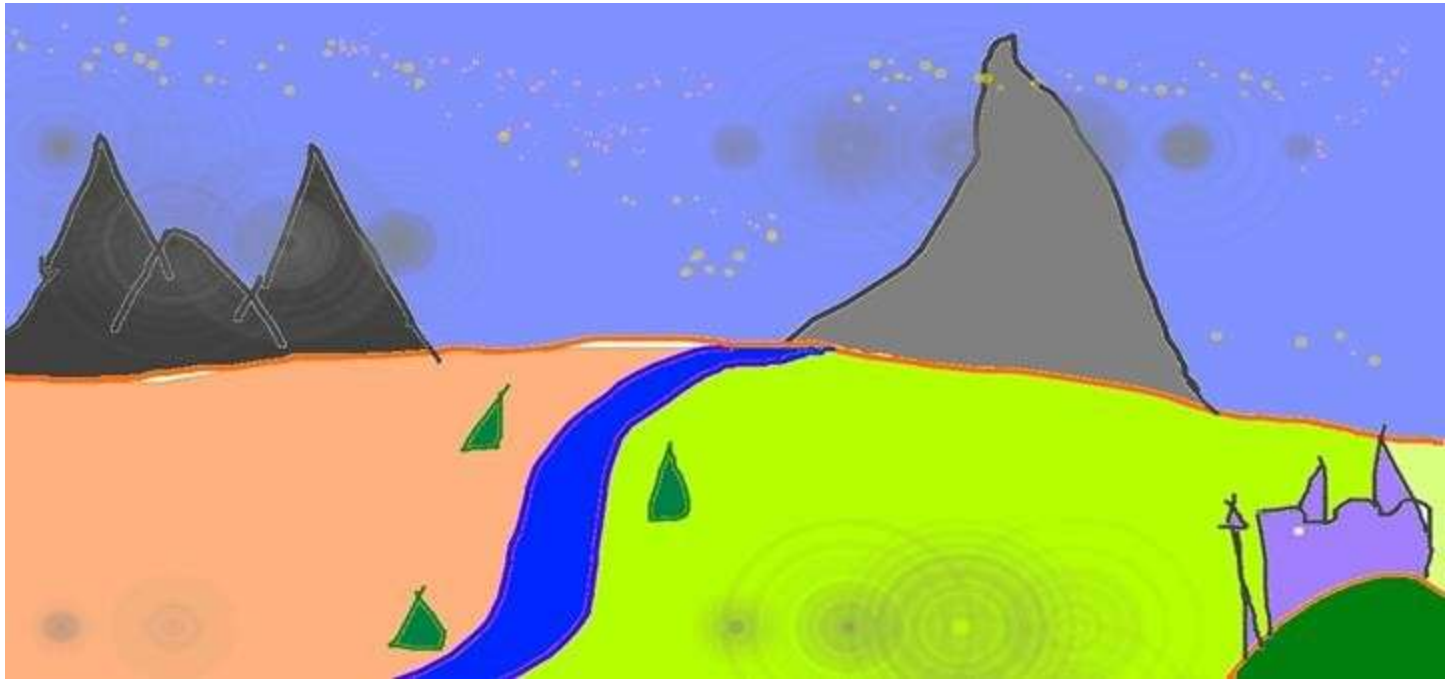
See more recent works: prompt-to-prompt, Imagic, pix2pix-zero, Edict, Plug & Play,

Instruct-pix2pix, ControlNet, etc.

[Meng et al., ICLR 2022]

SDEdit: Guided Image Synthesis with Diffusion

Input User Drawing



Used in Stable Diffusion Image-to-Image (“img2img’)

[Meng et al., ICLR 2022]

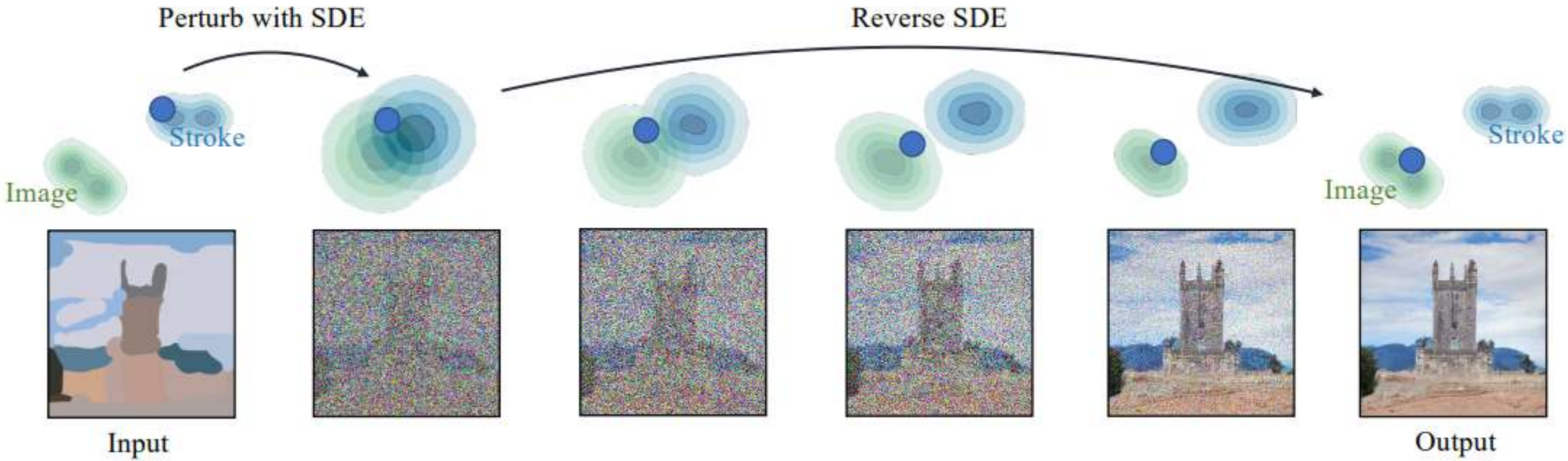
SDEdit: Guided Image Synthesis with Diffusion

Text prompt: “A fantasy landscape, trending on artstation”

Used in Stable Diffusion Image-to-Image (“img2img’)

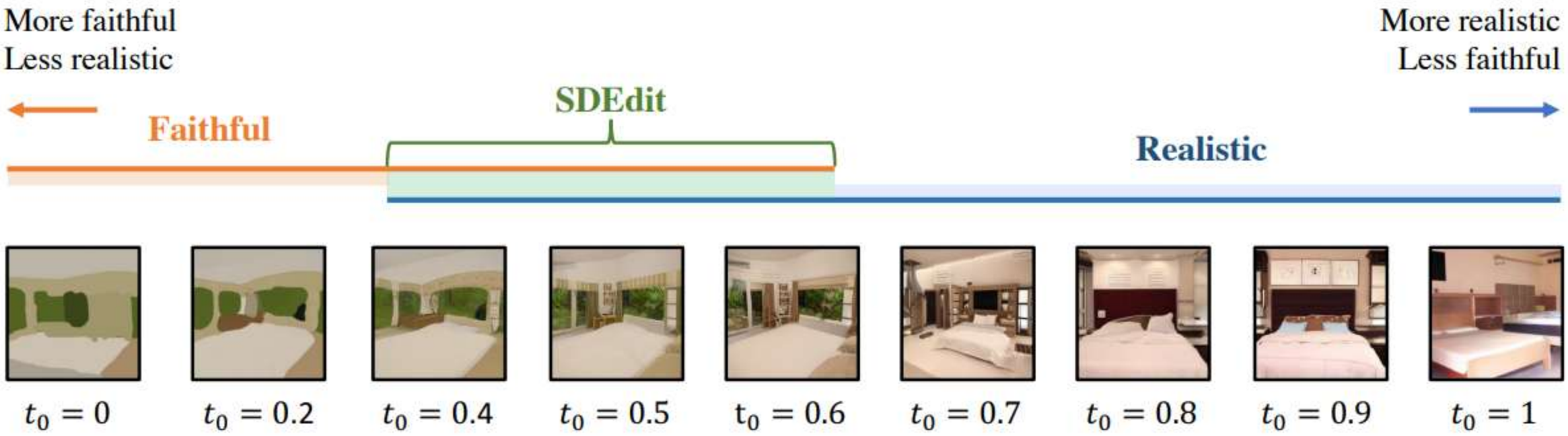
[Meng et al., ICLR 2022]

SDEdit: Guided Image Synthesis with Diffusion



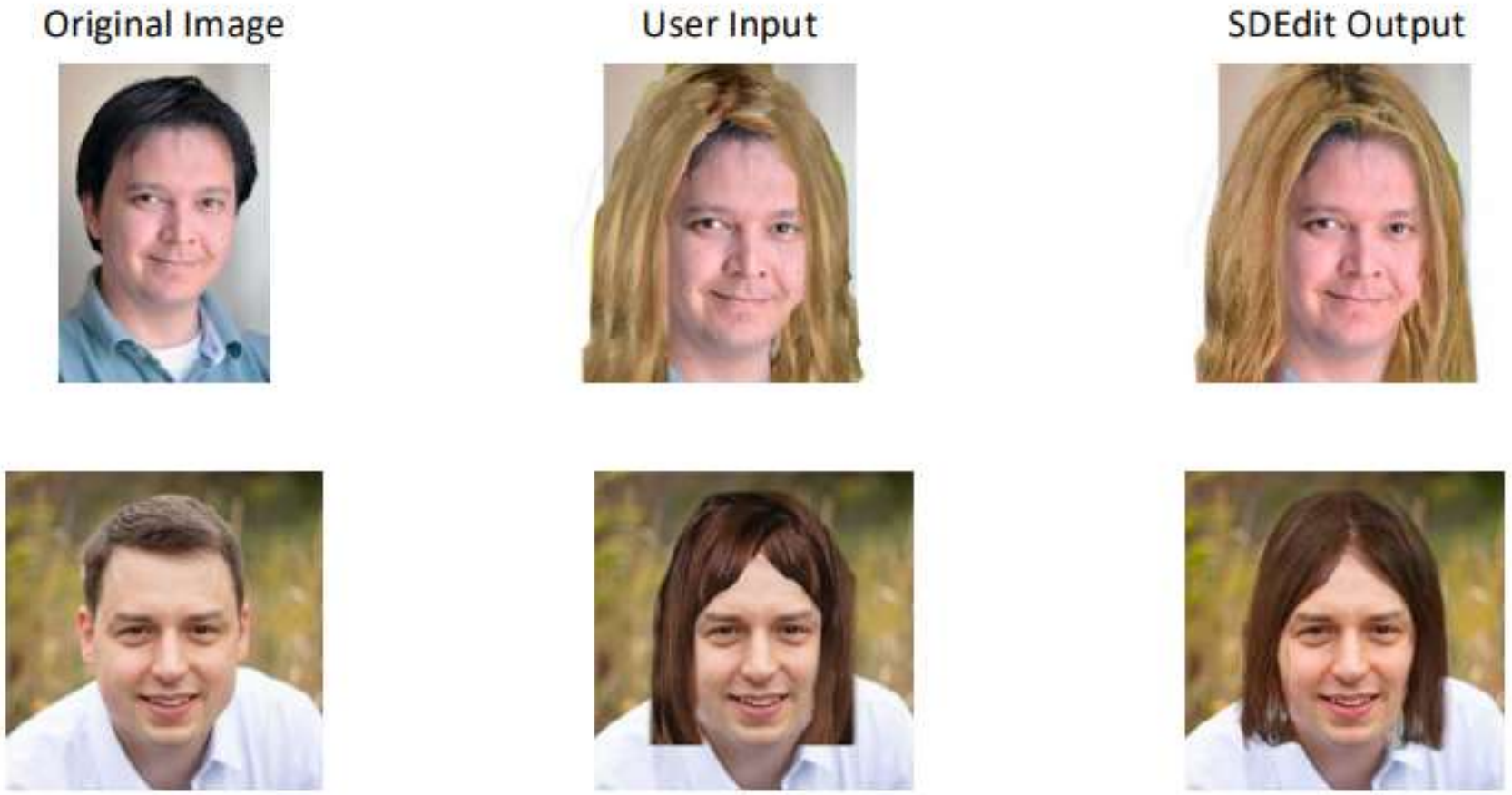
[Meng et al., ICLR 2022]

SDEdit: Guided Image Synthesis with Diffusion



[Meng et al., ICLR 2022]

SDEdit: Guided Image Synthesis with Diffusion



[Meng et al., ICLR 2022]

Video-to-video translation with SDEdit and Sora

original generated video

rewrite the video in a pixel art style

Video-to-video translation with SDEdit and Sora

original generated video

change the video to a medieval theme

Creating paired images

Generating two images with similar prompts:

Photo of a cat riding a bicycle

Photo of a cat riding a car

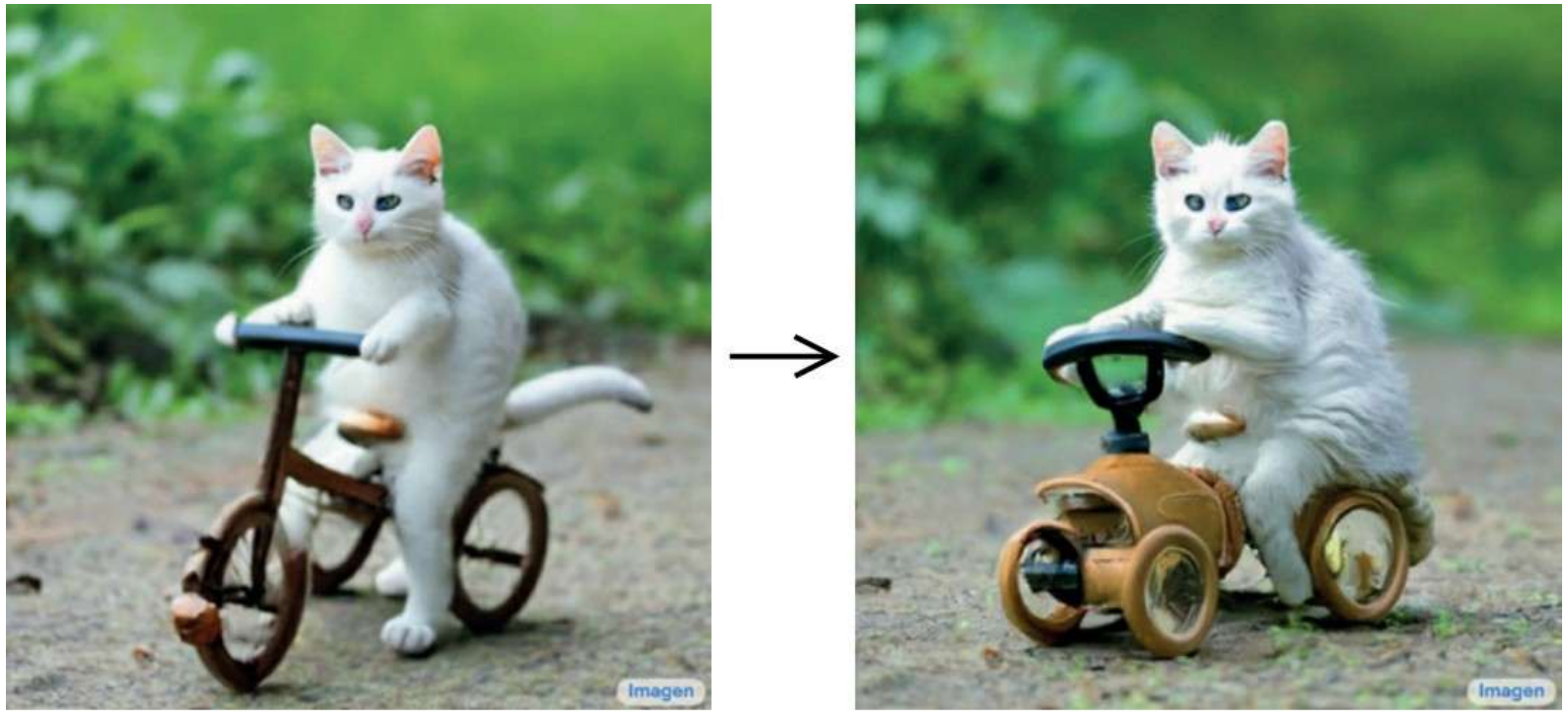
The images are quite different.

Hertz et al. “Prompt-to-Prompt Image Editing with Cross-Attention Control”

Adapted from A. Hołyński

Creating paired images

What we'd like instead:

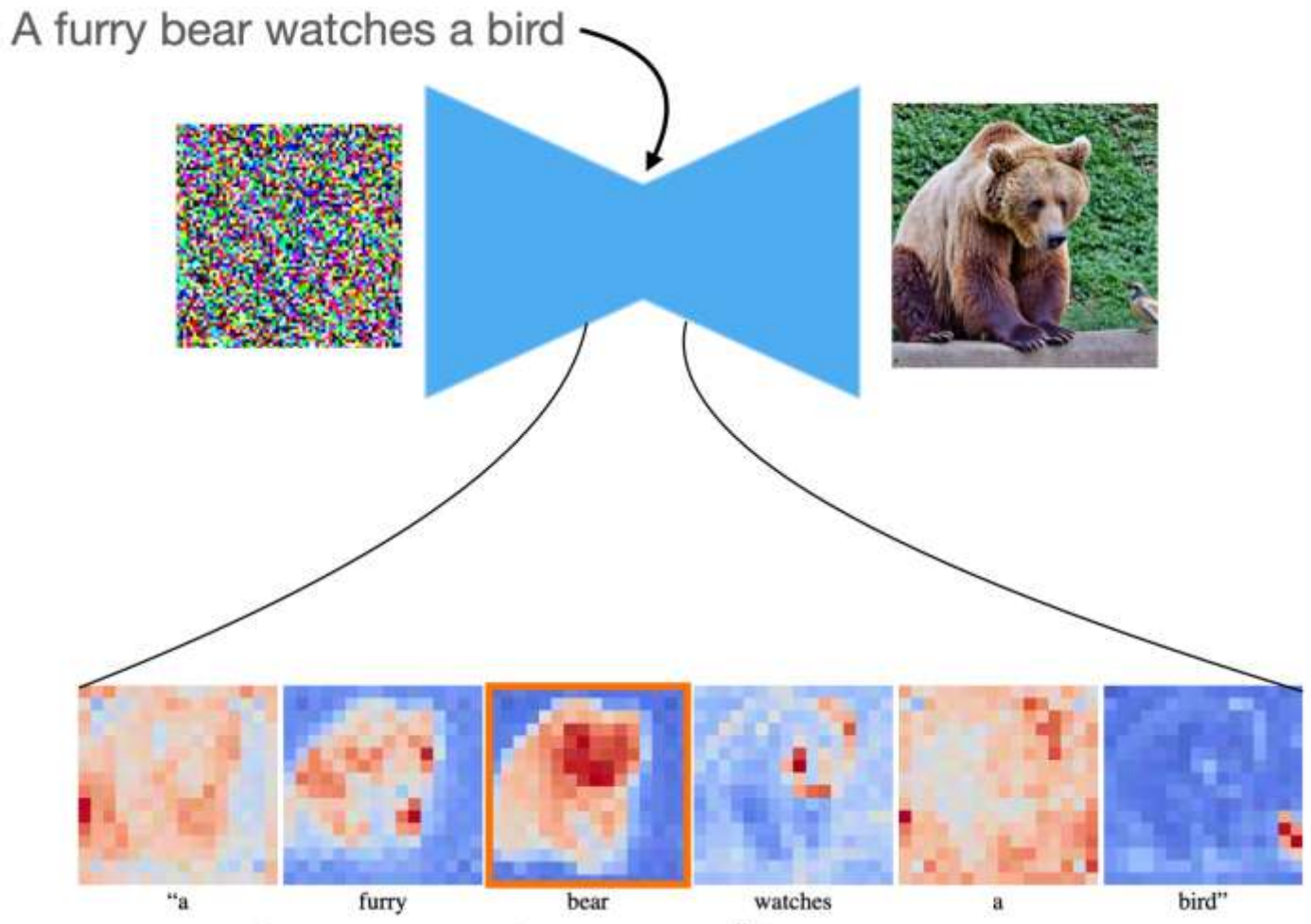


"Photo of a cat riding on a ~~bicycle~~ car."

Hertz et al. "Prompt-to-Prompt Image Editing with Cross-Attention Control"

Adapted from A. Hołyński

What's happening inside the network?

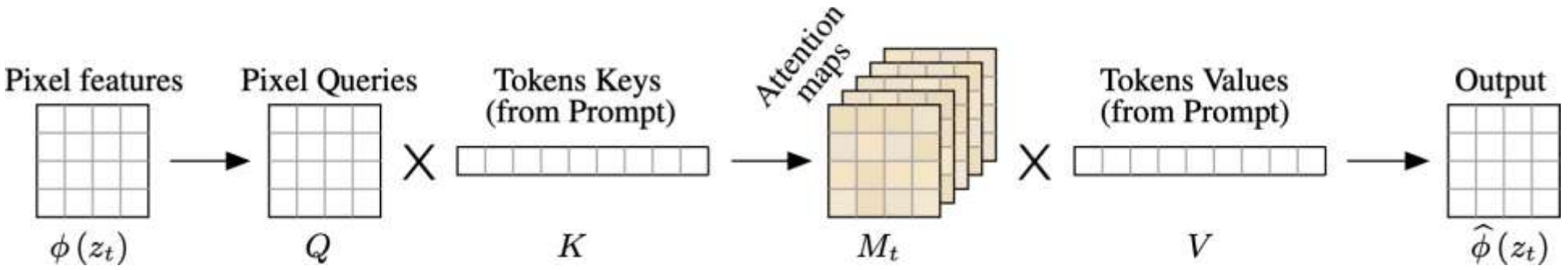


Hertz et al. "Prompt-to-Prompt Image Editing with Cross-Attention Control"

Slide source: A. Hołyński

Attention visualization

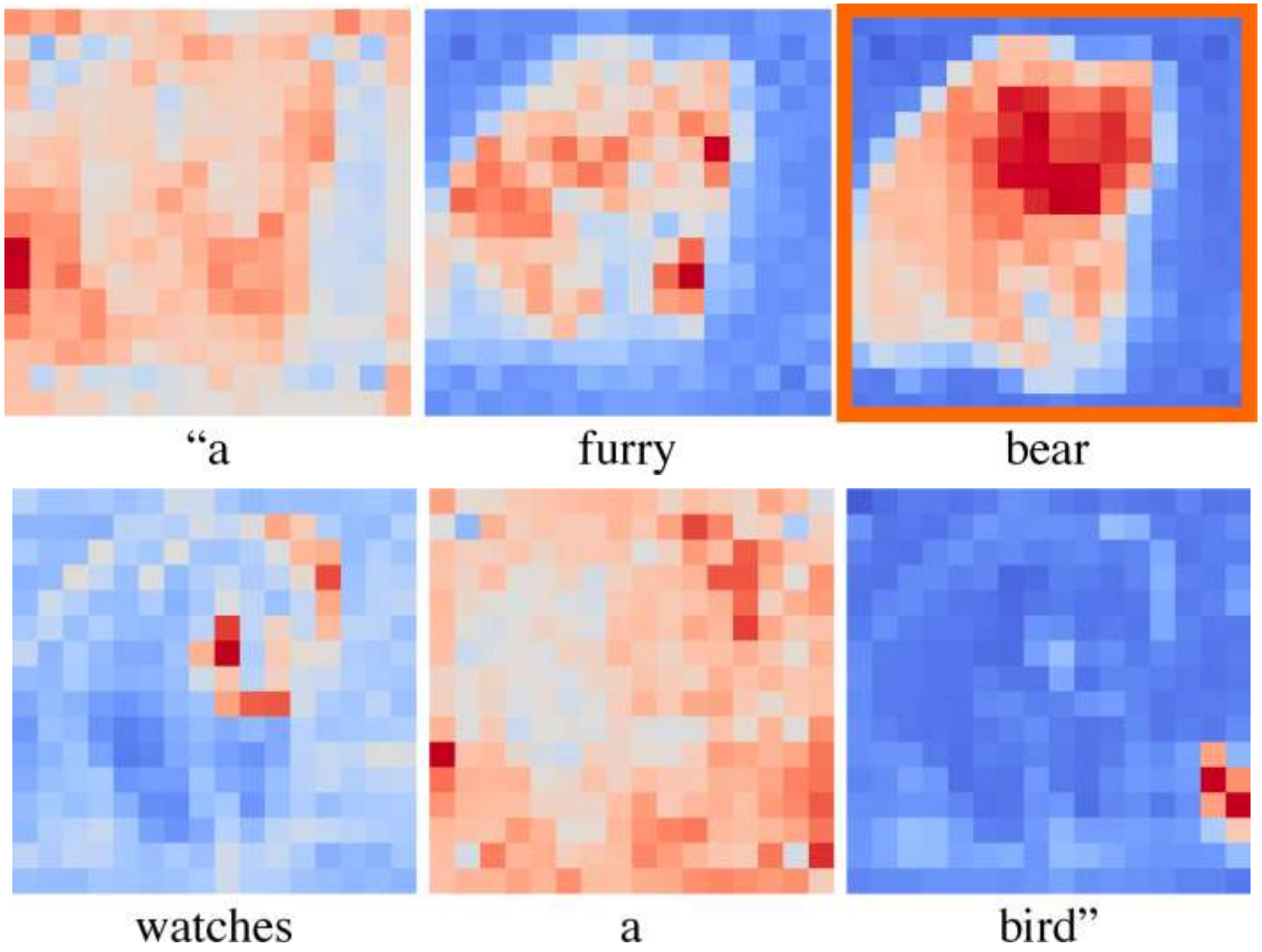
Empirical observation: cross-attention between text and image often conveys style, content, and structure.



Hertz et al. “Prompt-to-Prompt Image Editing with Cross-Attention Control”

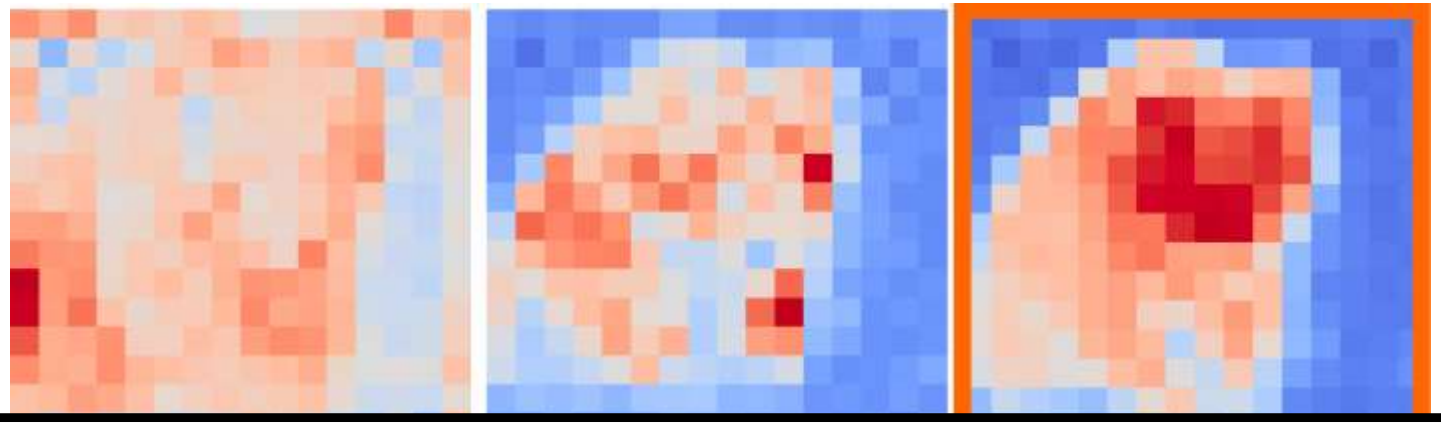
Attention visualization

“a furry bear watches a bird”



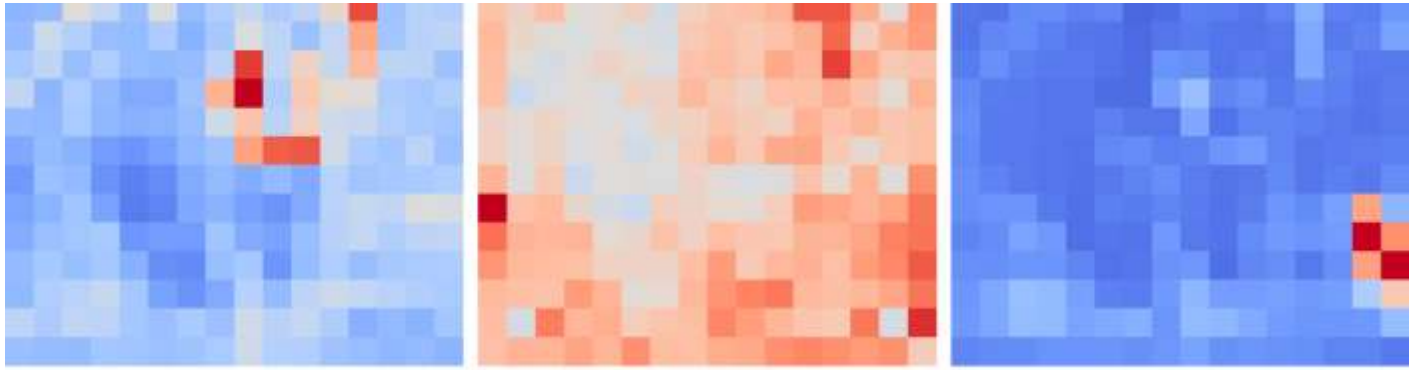
Hertz et al. “Prompt-to-Prompt Image Editing with Cross-Attention Control”

Attention visualization



What if we manipulate the attention maps?

“a furry bear watches a bird”



Hertz et al. “Prompt-to-Prompt Image Editing with Cross-Attention Control”

Change text prompt, resample using same random seed.

"lemon cake."

"cheese cake."

"apple cake."

"pumpkin cake."

"lego cake."

"beet cake."

"pepperoni cake."

Hertz et al. "Prompt-to-Prompt Image Editing with Cross-Attention Control"

What if we also freeze the attention maps?

"lemon cake."

Hertz et al. "Prompt-to-Prompt Image Editing with Cross-Attention Control"

Freeze the attention map for “apples” or “basket” during generation.

“A basket full of apples.”

Source image

apples → cookies

apples → oranges

apples → chocolates

apples → kittens

basket → bowl

basket → box

basket → nest

Hertz et al. “Prompt-to-Prompt Image Editing with Cross-Attention Control”

"A photo of a butterfly on a flower."

Source image

flower → bread

flower → mug

flower → computer

flower → mirror

butterfly → bird

butterfly → snail

butterfly → drone

This is a neat trick. Can we train a *model* that captures these abilities?

Hertz et al. "Prompt-to-Prompt Image Editing with Cross-Attention Control"

Prompt-to-Prompt

"Photo of a cat riding on a bicycle."

source image

cat → dog

cat → chicken

cat → squirrel

cat → elephant

Hertz et al. "Prompt-to-Prompt Image Editing with Cross-Attention Control"

Bootstrapping to instruction-based image editing

"Swap sunflowers with roses"

"Add fireworks to the sky"

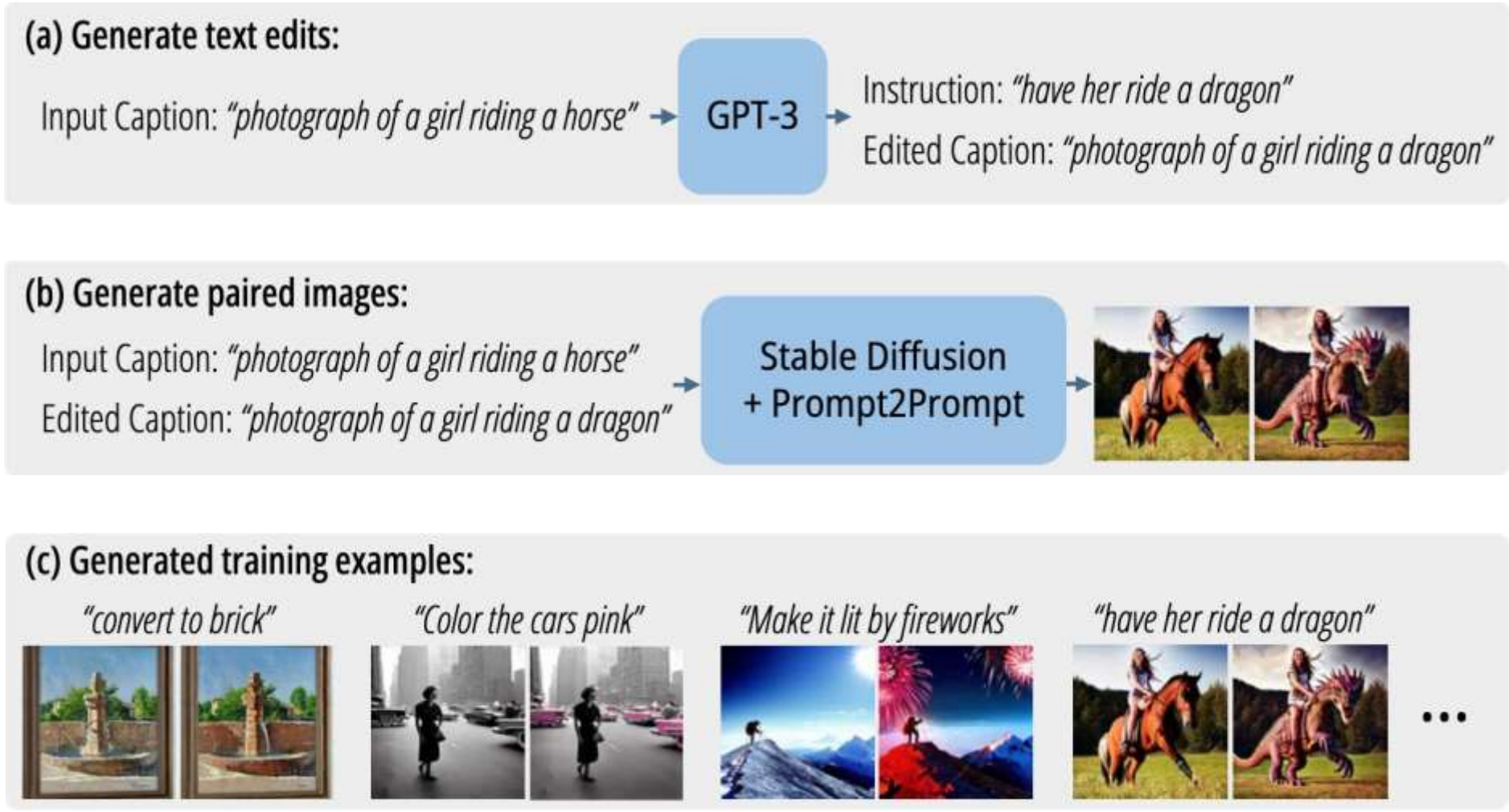
"Replace the fruits with cake"

"What would it look like if it were snowing?"

"Turn it into a still from a western"

"Make his jacket out of leather"

Generating training data for instruction-based editing



Source: [Brooks et al., "InstructPix2Pix", 2023]

Memorized Style

Greg Rutkowski

*image taken from https://rutkowski.artstation.com

Stable Diffusion

A painting of a boat on the water in the style of Greg Rutkowski

Memorized Style

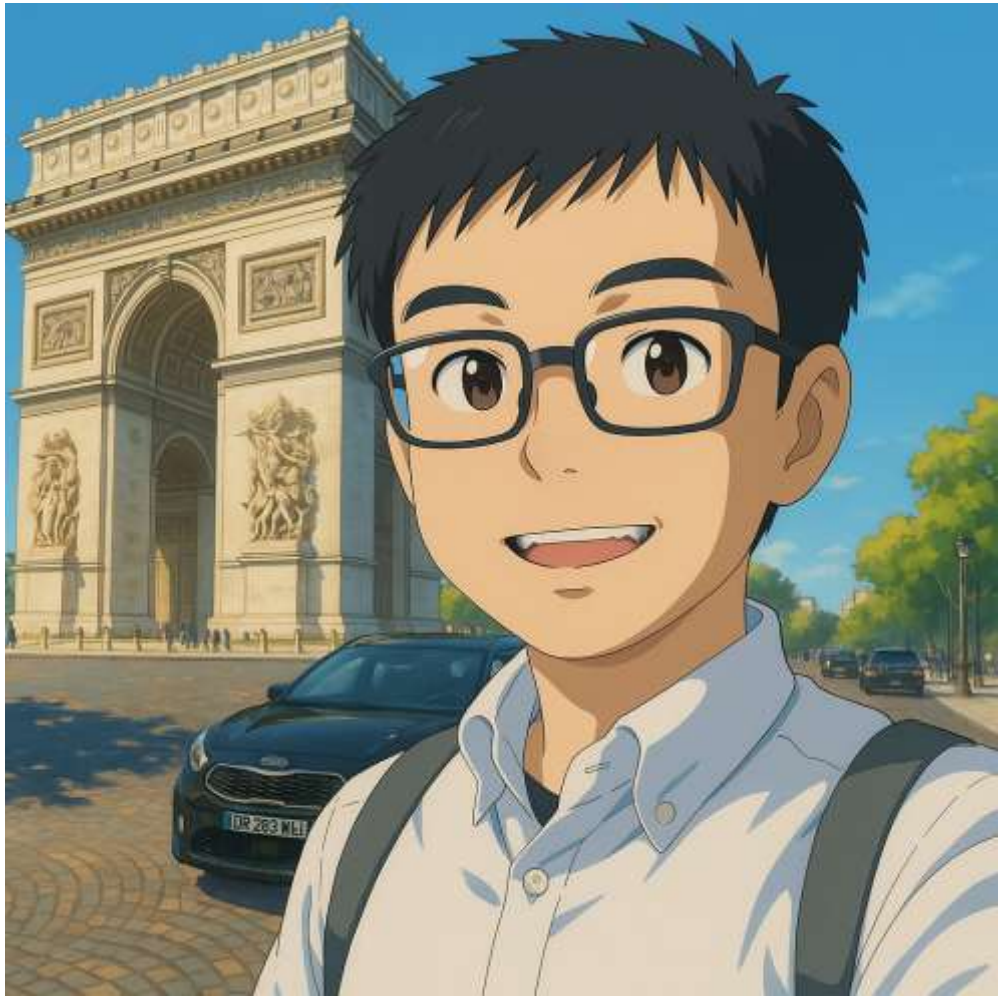
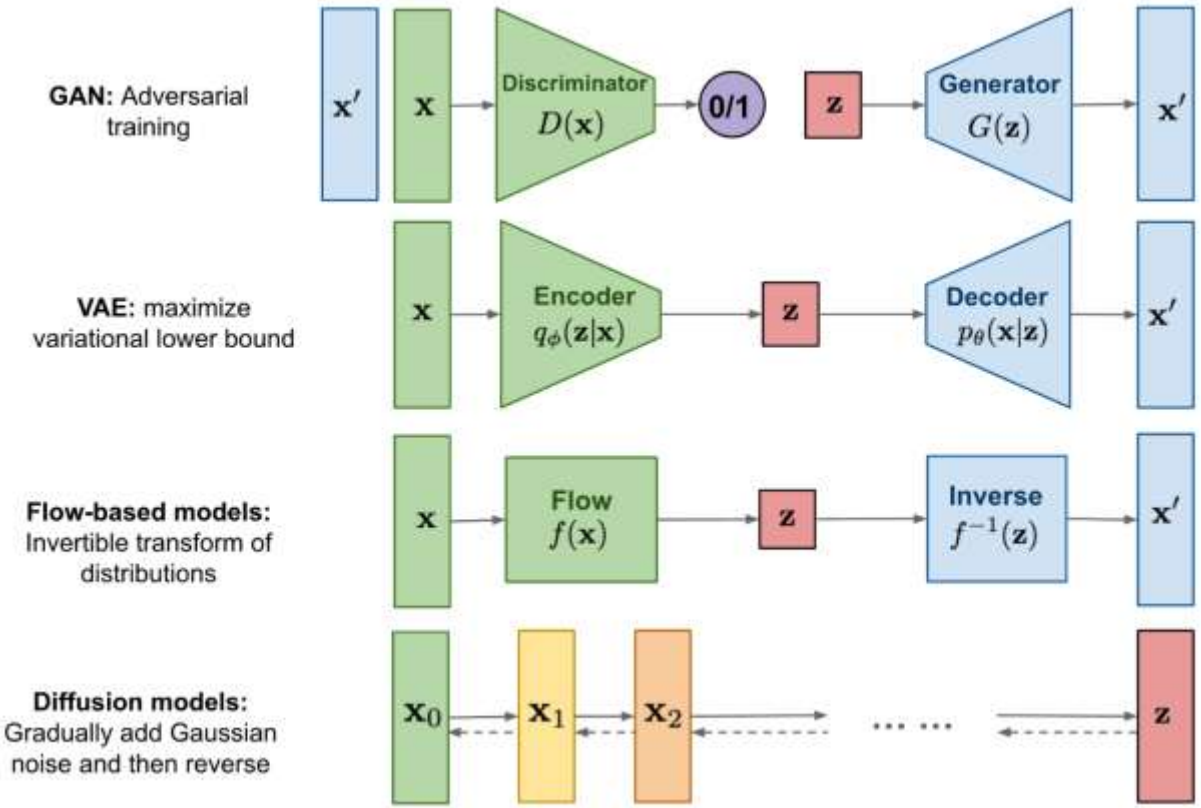
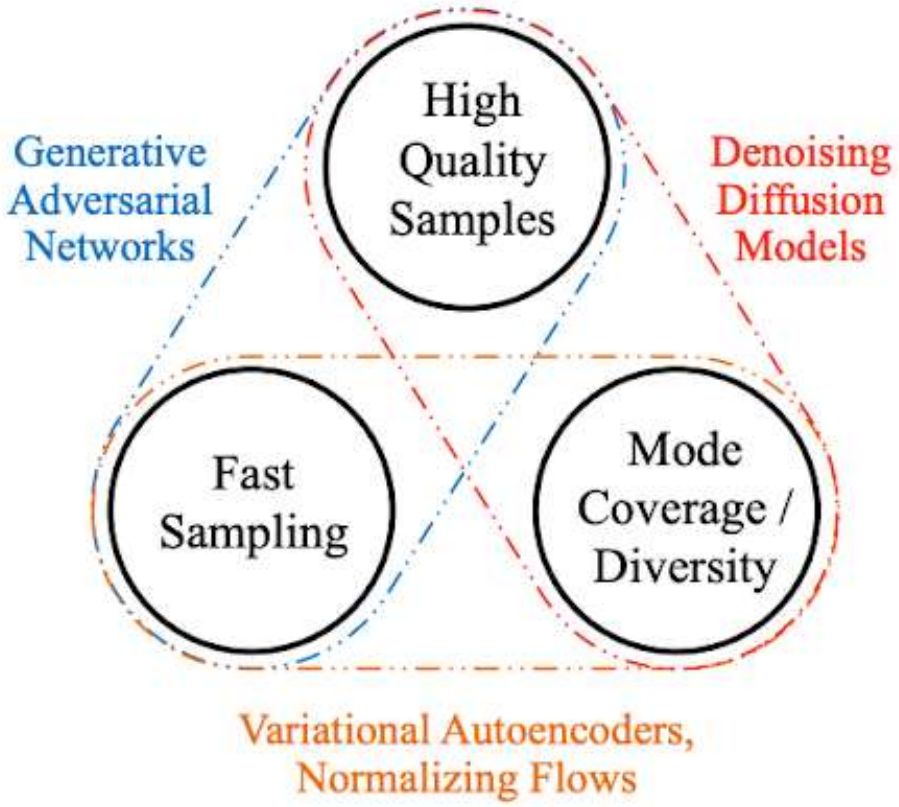


*image taken from ChatGPT 4o

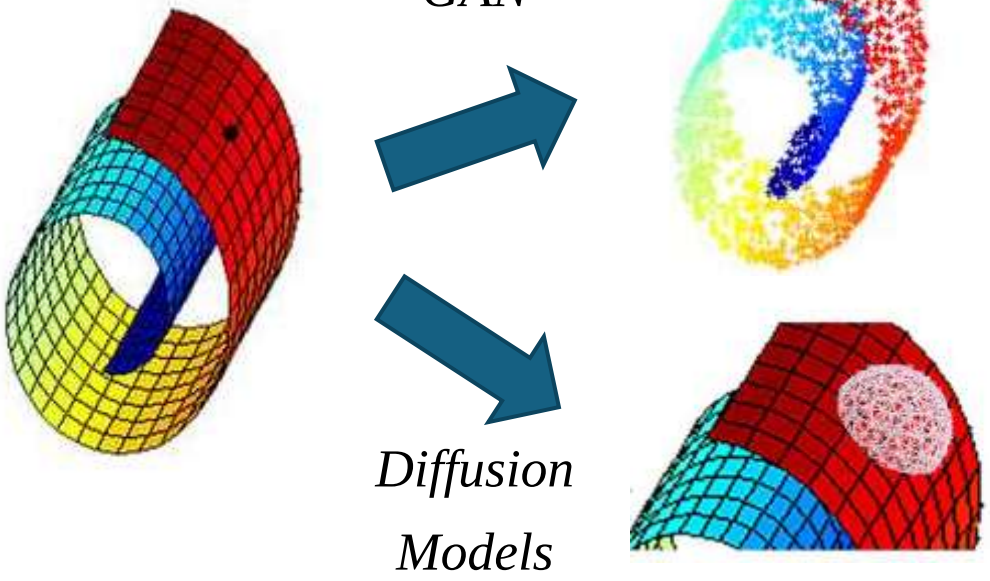
Impossible Triangles





Improving Diffusion Models as an Alternative To GAN, Part 1 | NVIDIA Technical Blog

Diffusion Models' Benefit



Distributional loss

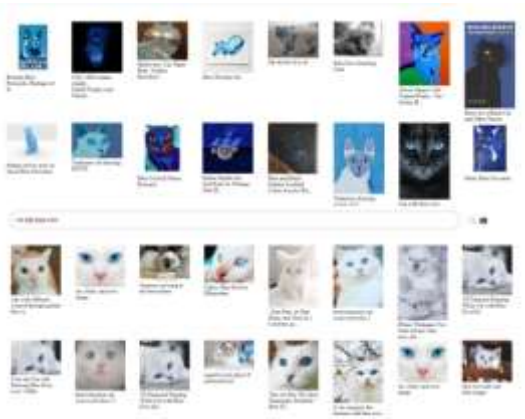
+ Clearer images

- Mode collapse, limited diversity

Point-wise loss

+ Maintain diversity

- Blur images (+ Multi-step)



Diffusion models can be trained on large datasets with

5 billion image-text pairs !

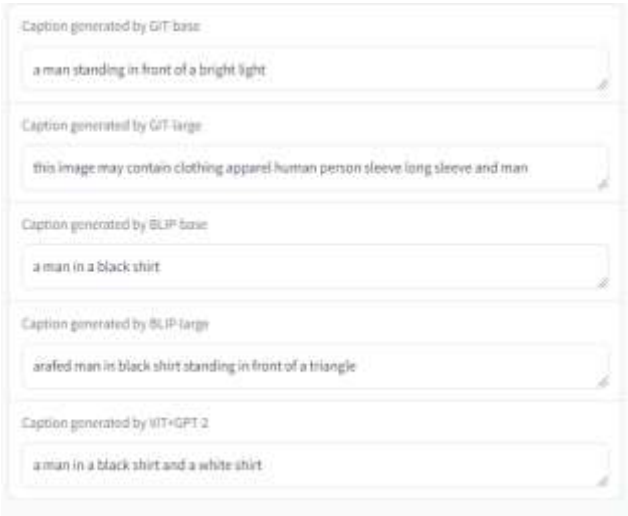
+ High quality and diversity, w/ speedup (ODE, LDM,…)

An unified model!

Rombach & Blattmann, et al. 2022

Diffusion Models' Benefit

SD: "A photo of Musk"



Caption: "A man in a black shirt"

SD: "A painting of Picasso"

Caption: "a painting of a man with a mustache"

+ Mode coverage

+ Generate images for different concepts (instances, persons, styles…)

Are Diffusion Models All You Need?

https://www.midjourney.com/showcase

+ Amazing fidelity

- Cherry picking

Prompt: pouring water from a teapot into a cup; Models: Stable Diffusion Series.

Are Diffusion Models All You Need?

+ High quality, w/ speedup (Flow-Matching, CM)

- Controlability

Are Diffusion Models All You Need?

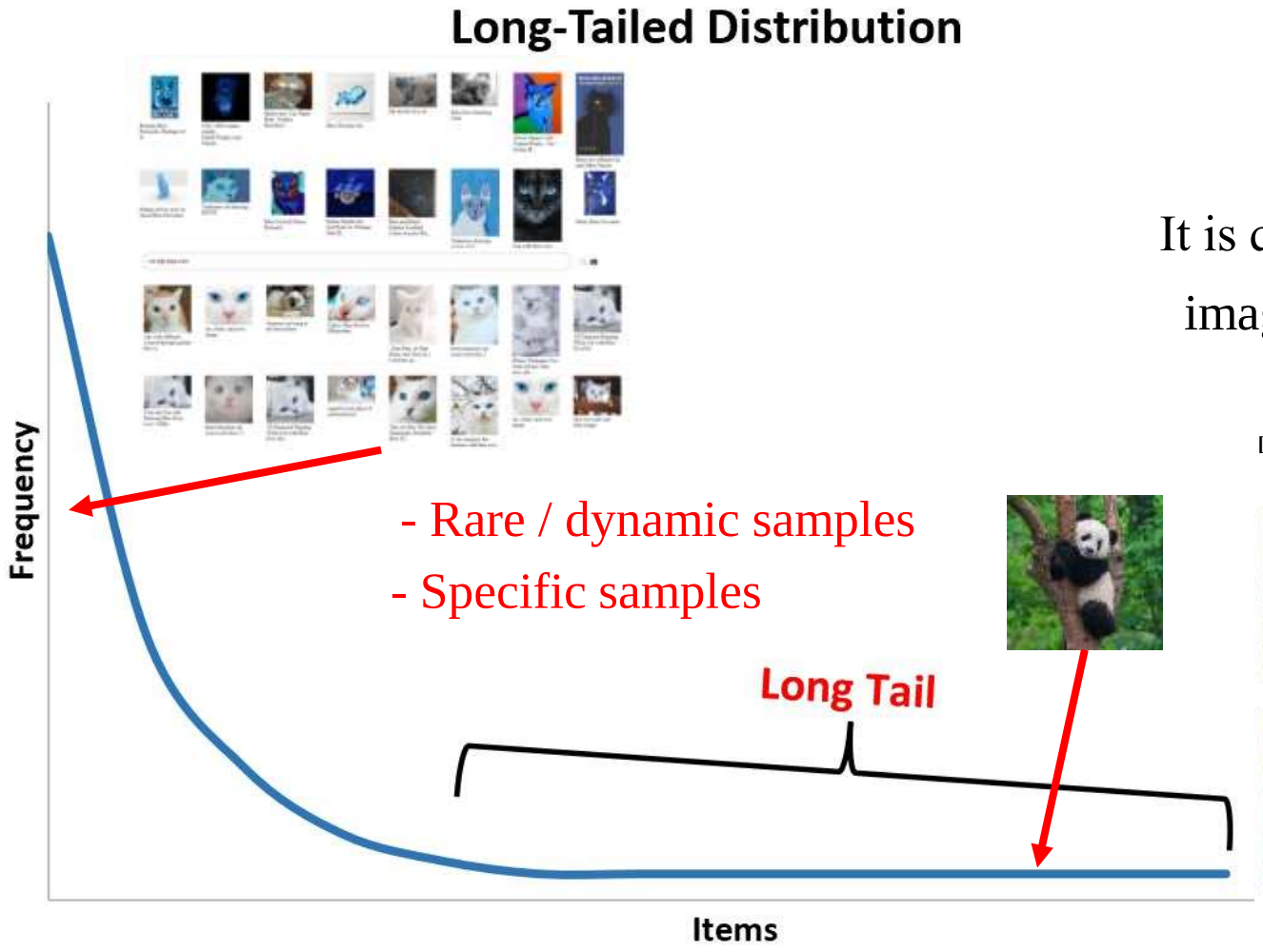
+ Amazing fidelity

https://openai.com/sora

- Texture

- Physic systems

Are Diffusion Models All You Need?

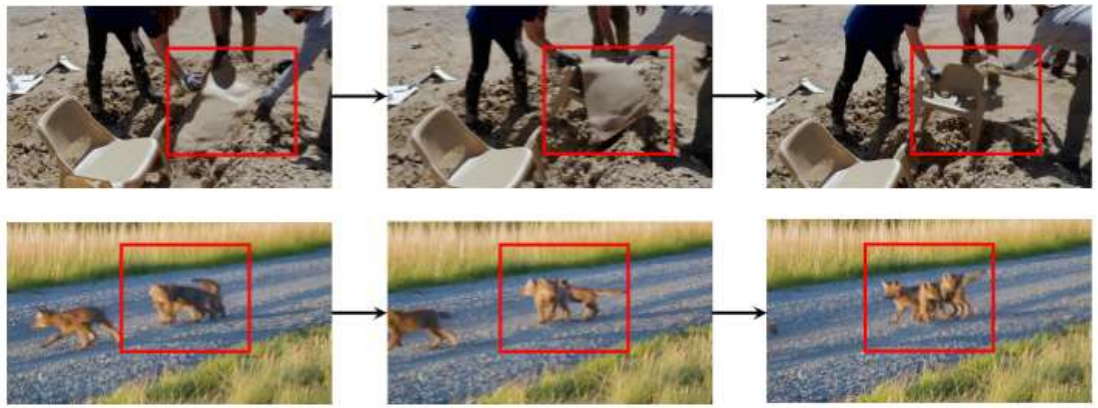


Training Subset	Testing Subset		
	DMs	GANs	Others
DMs	**99.7/99.9/99.9**	86.4/95.9/89.8	79.3/93.5/84.5
GANs	77.1/83.2/74.3	**98.1/99.0/99.6**	91.4/97.2/94.6
Others	76.4/77.2/70.1	82.5/96.0/91.2	**99.6/99.9/99.9**

It is quite easy to detect Diffusion Models' generated images. (But not that good to generalize to GANs)

[arXiv:2402.11843] WildFake: A Large-scale Challenging Dataset for AI-Generated Images Detection

[arXiv:2405.19707] DeMamba: AI-Generated Video Detection on Million-Scale GenVideo Benchmark



Are Diffusion Models All You Need?

- Comprehensive
- Cross-Modality
- Activity

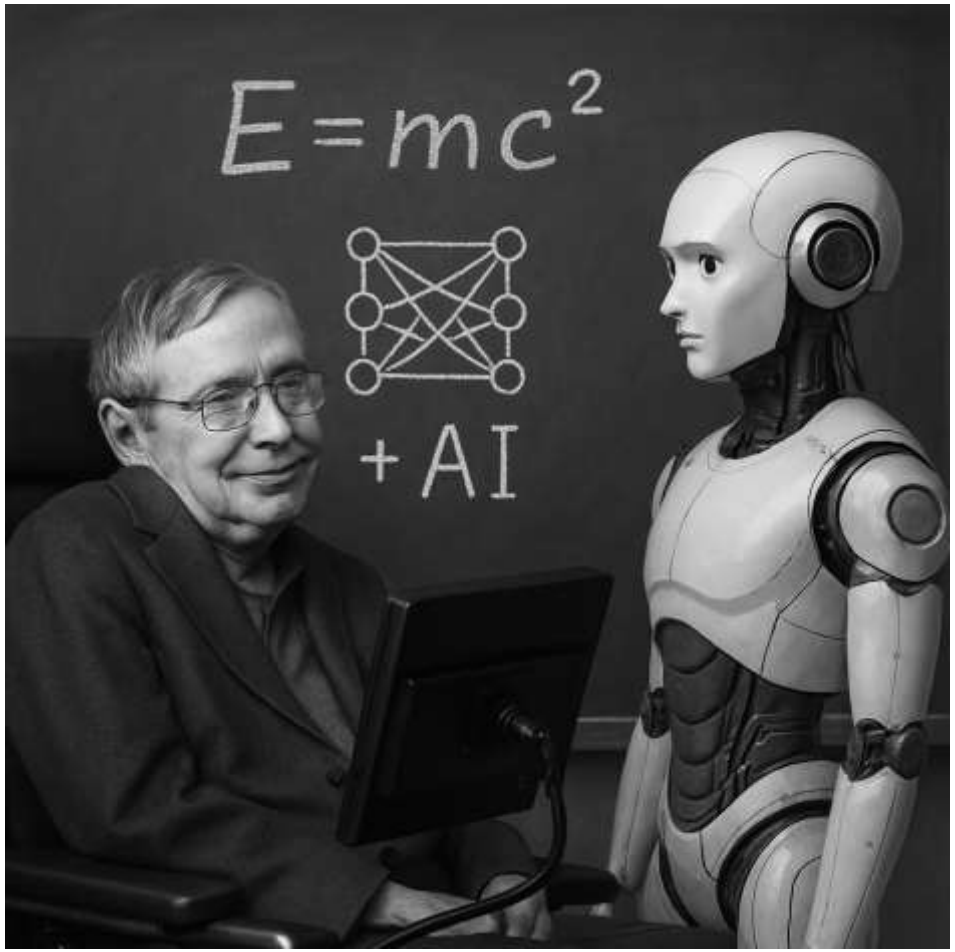
But Techs Evolve!

*image taken from ChatGPT 4o

Memorized Style

*image taken from ChatGPT 4o

Memorized Instances



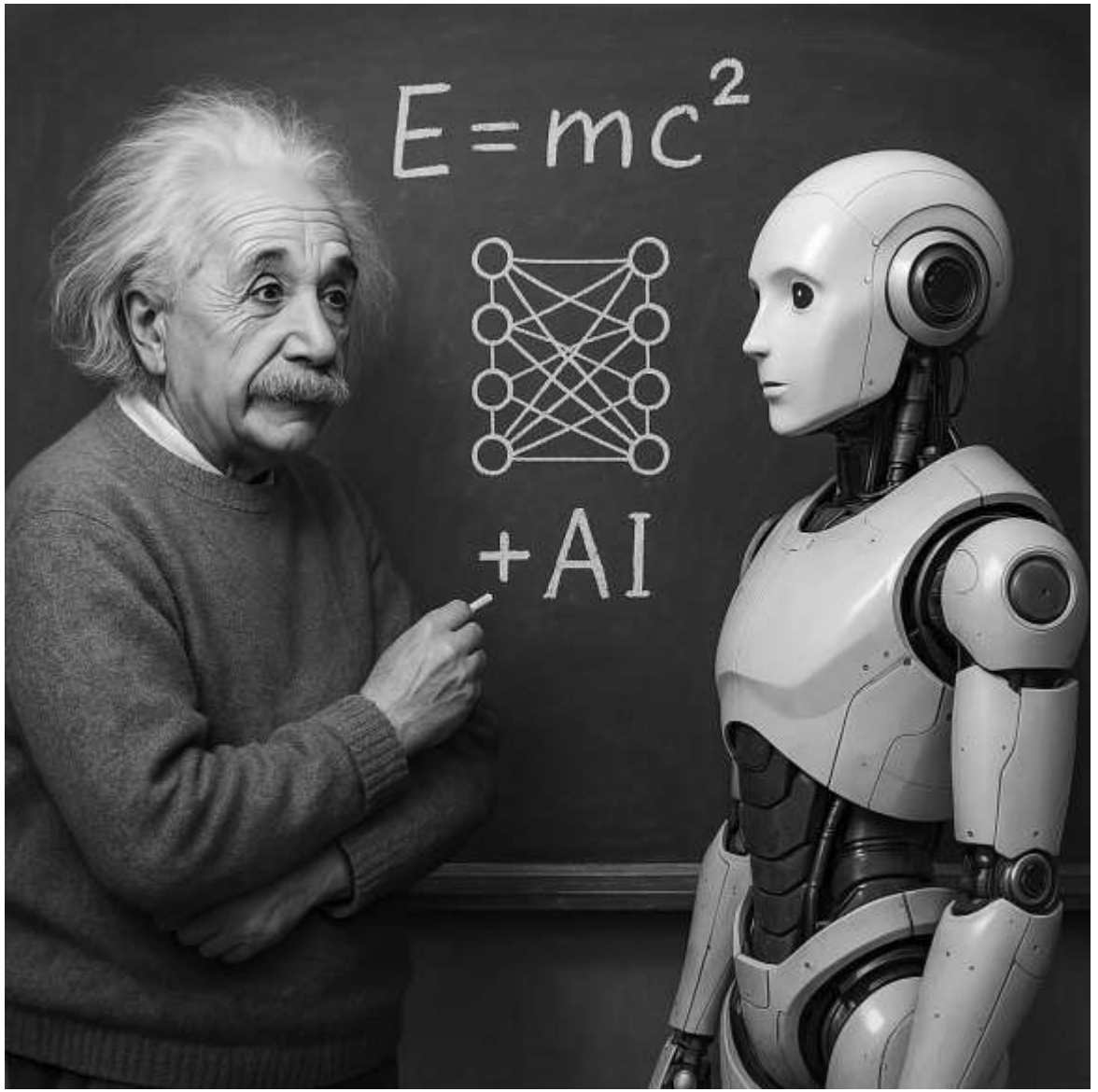
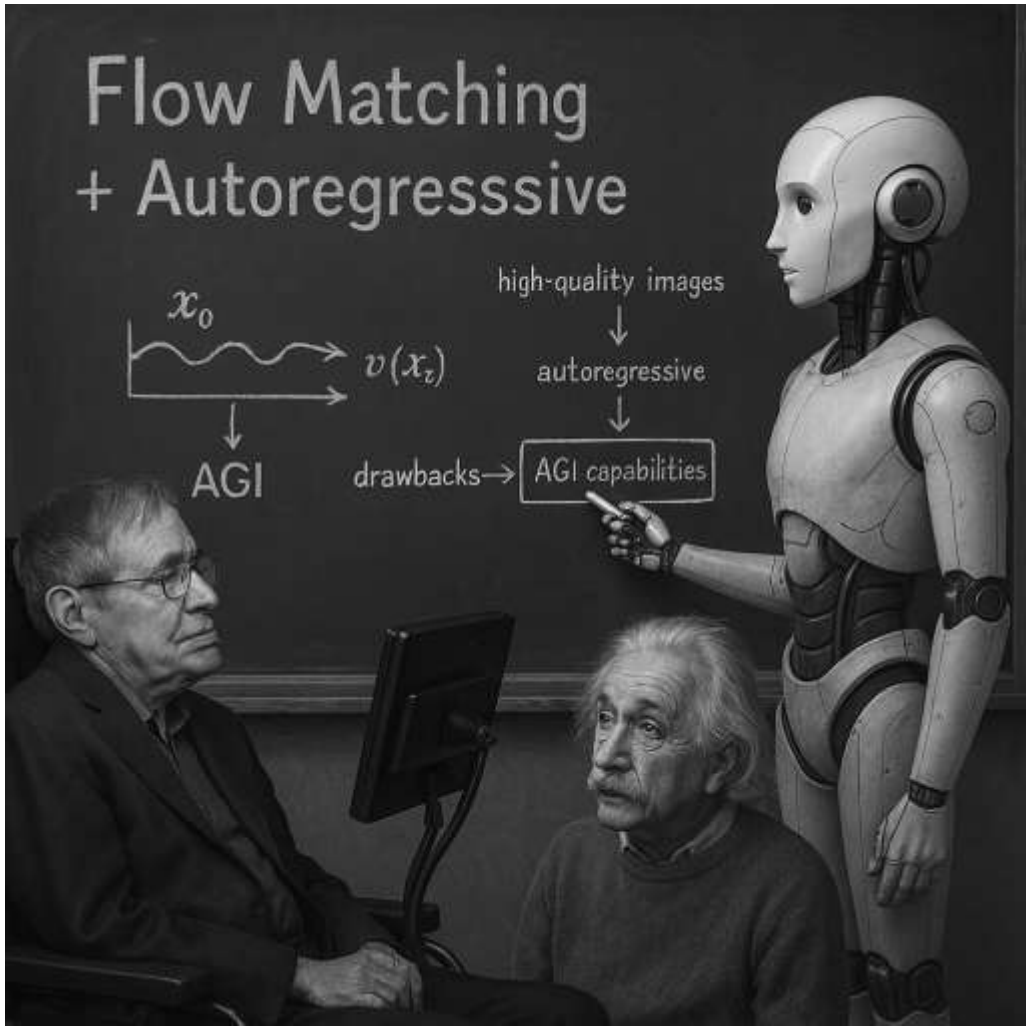


*image taken from ChatGPT 4o

Memorized Instances



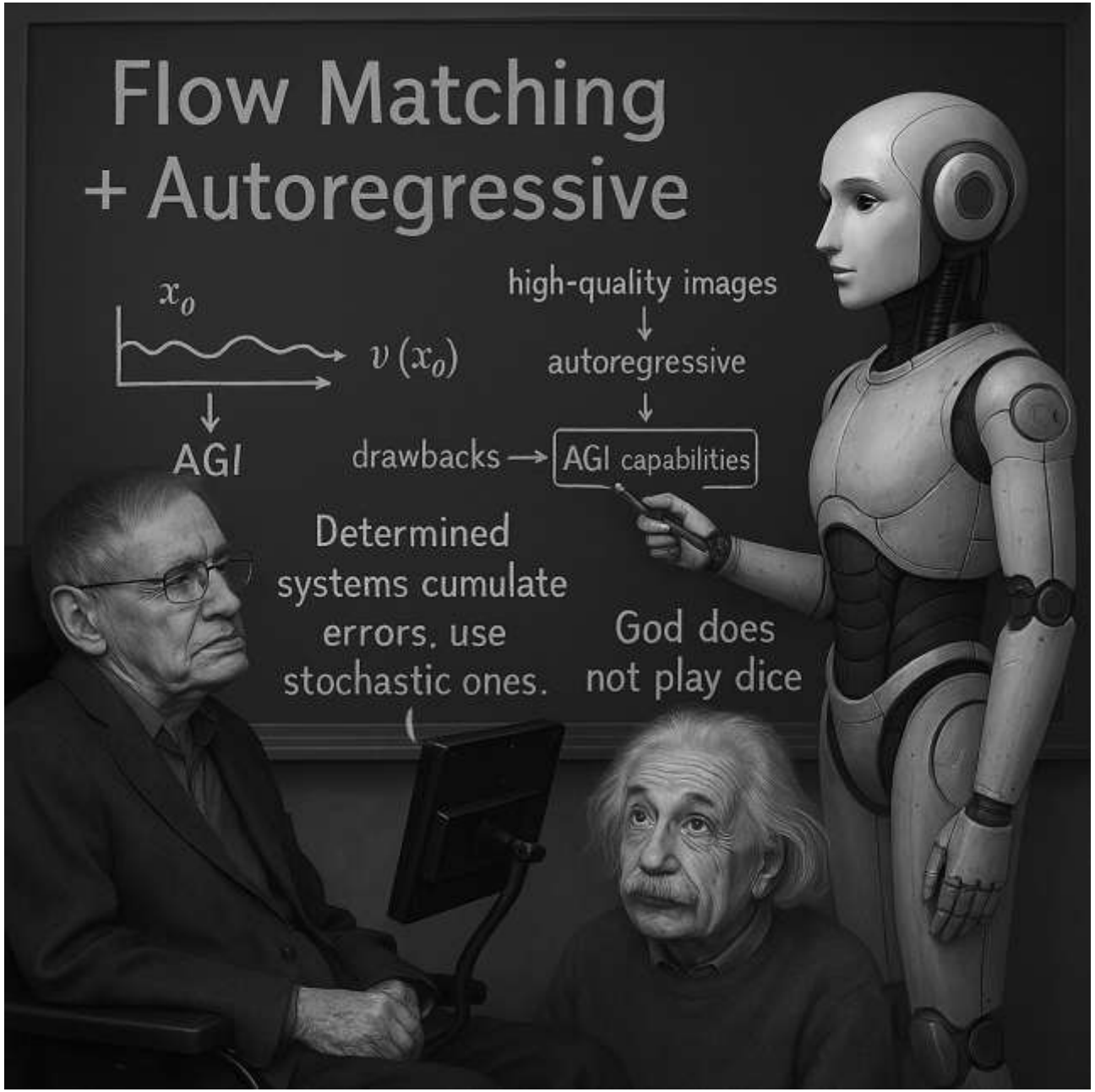
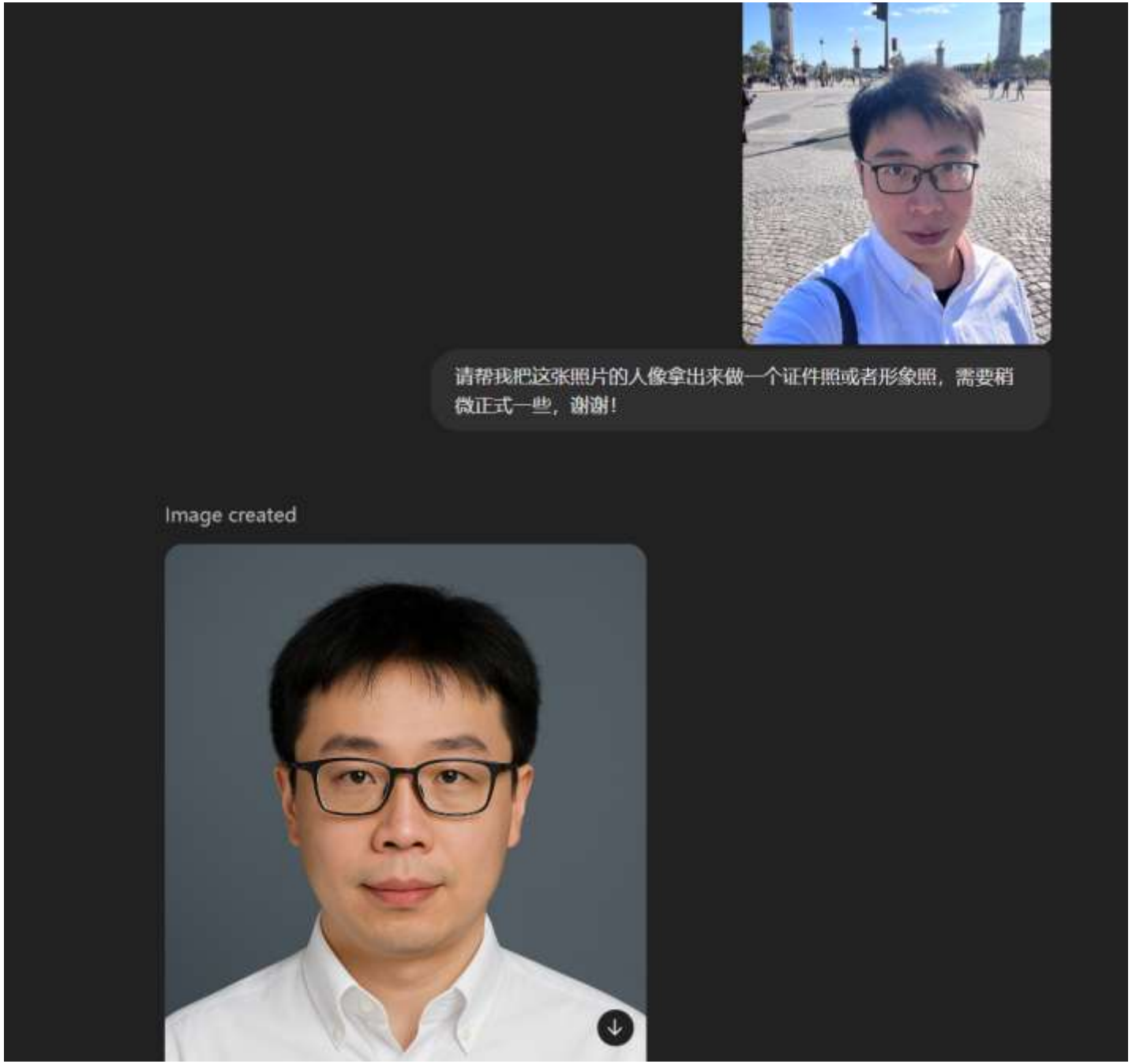
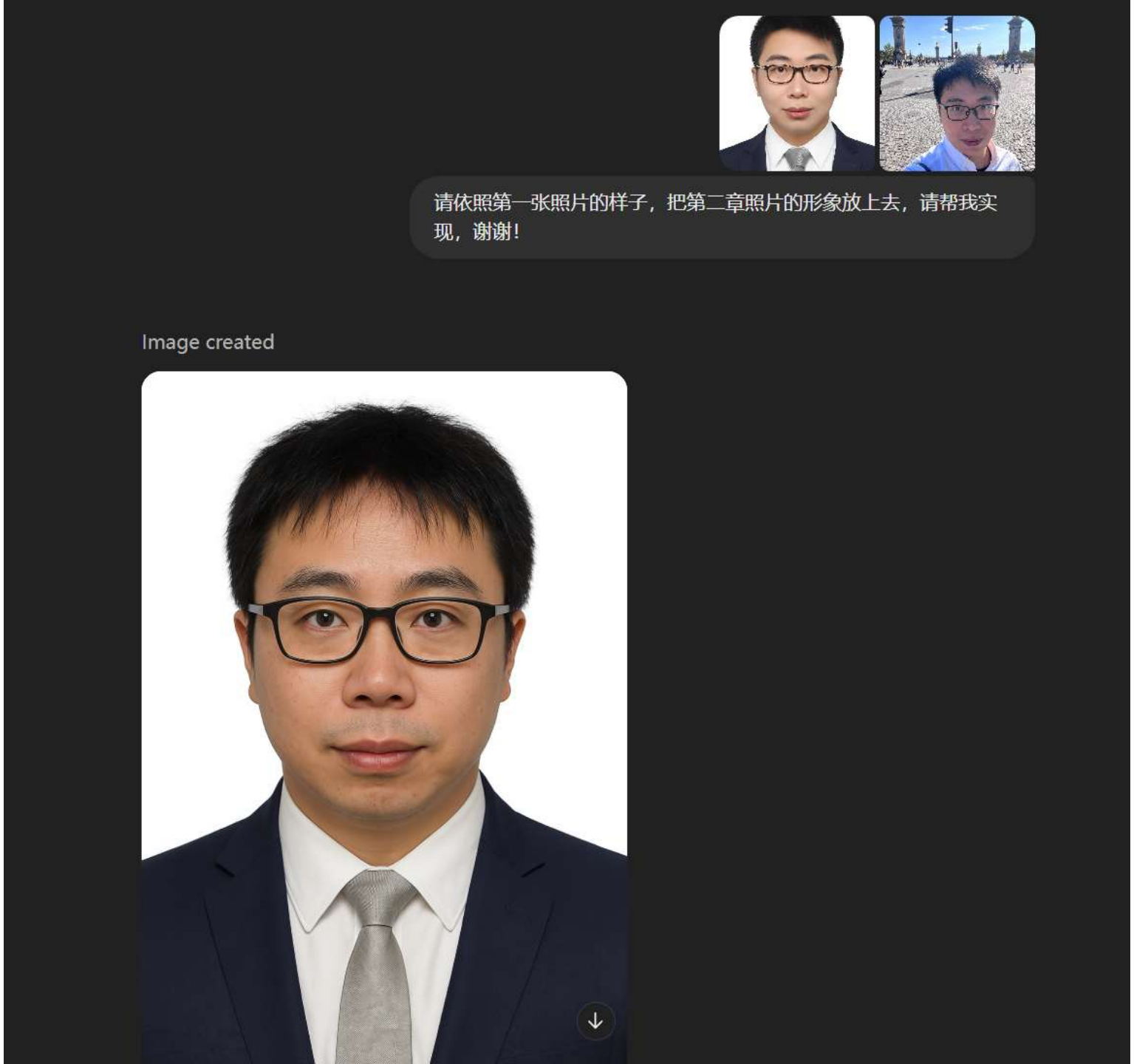


*image taken from ChatGPT 4o

But not Enough!

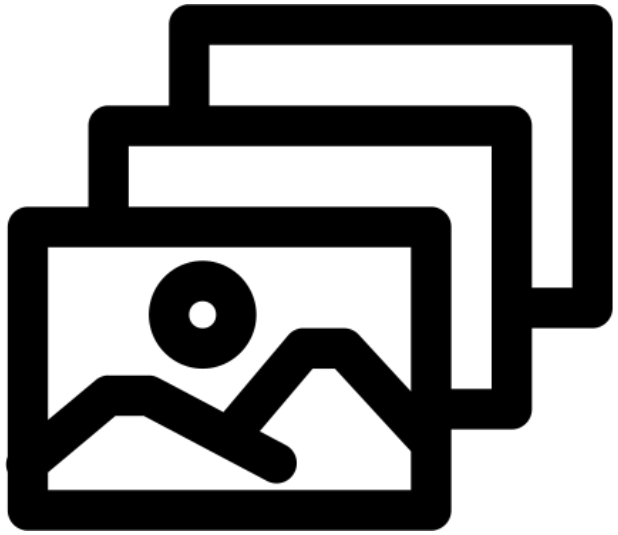


But not Enough!



Privacy

Machine Learning Pipeline



Training images

Model

Challenges

- Data opt-out and compensation are standard practices for content creation platforms.

Challenges

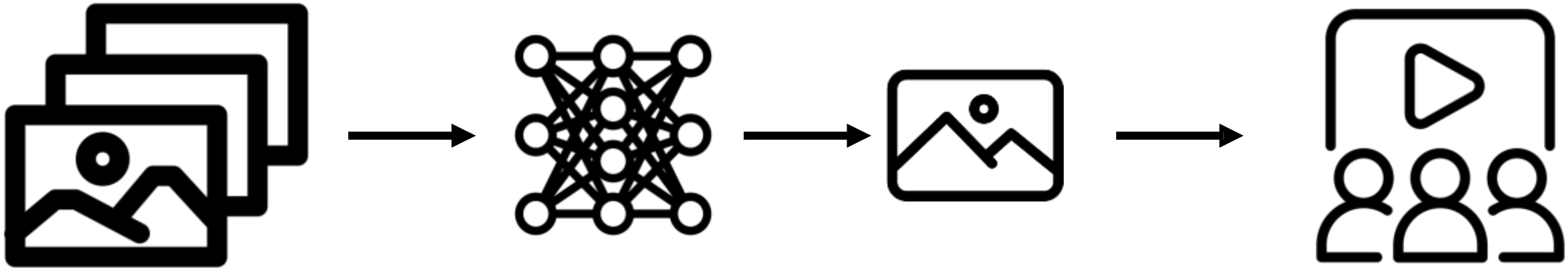
- Data opt-out and compensation are standard practices for content creation platforms.

Challenges

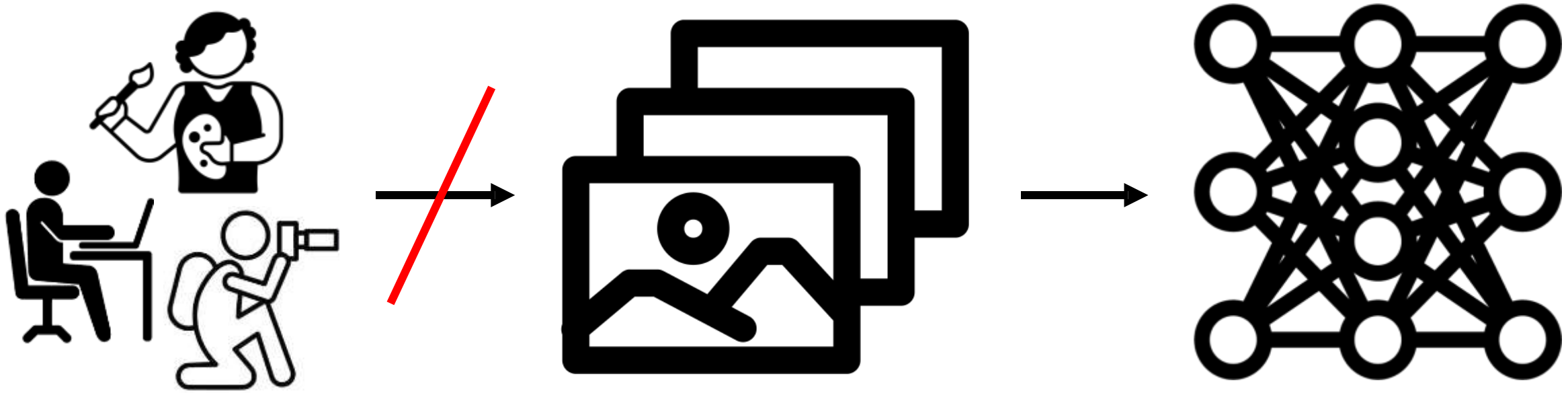
- Difficult for Generative models, as

  Consumers see generated data rather than training data,

  Training data are now entangled in the model weights.



Data Comes from People!



Contributors Training images Model

So researchers & founders are excited, but…

Generative models use training data of artists, photographers, and creators

- without Consent

without Compensation

Ongoing Legal Battles

ARTIFICIAL INTELLIGENCE / TECH / LAW

Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement

/ Getty Images has filed a case against Stability AI, alleging that the company copied 12 million images to train its AI model 'without permission ... or compensation.'

By JAMES VINCENT
Feb 6, 2023, 11:58 AM EST | 16 Comments / 16 New

An illustration from Getty Images' lawsuit, showing an original photograph and a similar image (complete with Getty Images watermark) generated by Stable Diffusion. Image: Getty Images

Getty Images has filed a lawsuit in the US against Stability AI, creators of open-source AI art generator Stable Diffusion, escalating its legal battle against the firm.

ARTIFICIAL INTELLIGENCE / TECH / CREATORS

AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit

/ The suit claims generative AI art tools violate copyright law by scraping artists' work from the web without their consent.

By JAMES VINCENT
Jan 16, 2023, 6:28 AM EST | 28 Comments / 28 New

A collage of AI-generated images created using Stable Diffusion. Image: The Verge via Lexica

A trio of artists have launched a lawsuit against Stability AI and Midjourney, creators of AI art generators Stable Diffusion and Midjourney, and artist portfolio platform DeviantArt, which recently created its own AI art generator, DreamUp.

Source: The Verge

Ongoing Legal Battles

REUTERS

World Business Markets Legal Breakingviews Technology Investigations Sports

Copyright | Technology | Intellectual Property | Litigation | Data Privacy

2 minute read · February 22, 2023 8:41 PM EST · Last Updated 2 months ago

AI-created images lose U.S. copyrights in test for new technology

By Blake Brittain

REUTERS/Andrew Kelly

Feb 22 (Reuters) - Images in a graphic novel that were created using the artificial-intelligence system Midjourney should not have been granted copyright protection, the U.S. Copyright Office said in a letter seen by Reuters.

I'm not so sure. As we've seen, a key assumption for a "non-expressive use" defense is that Stable Diffusion only learns uncopyrightable facts—not creative expression—from its training images. That's *mostly* true. But it's not entirely true. And the exceptions could greatly complicate Stability AI's legal defense.

Stable Diffusion's copying problem

Here's one of the most awkward examples for Stability AI:

Training Set

Caption: Living in the light with Ann Graham Lotz

Generated Image

Prompt: Ann Graham Lotz

Enlarge

https://arstechnica.co

Hollywood Strikes against AI

In Hollywood writers' battle against AI, humans win (for now). By JAKE COYLE, AP News

If artificial intelligence uses your work, it should pay you By Joseph Gordon-Levitt, The Washington Post

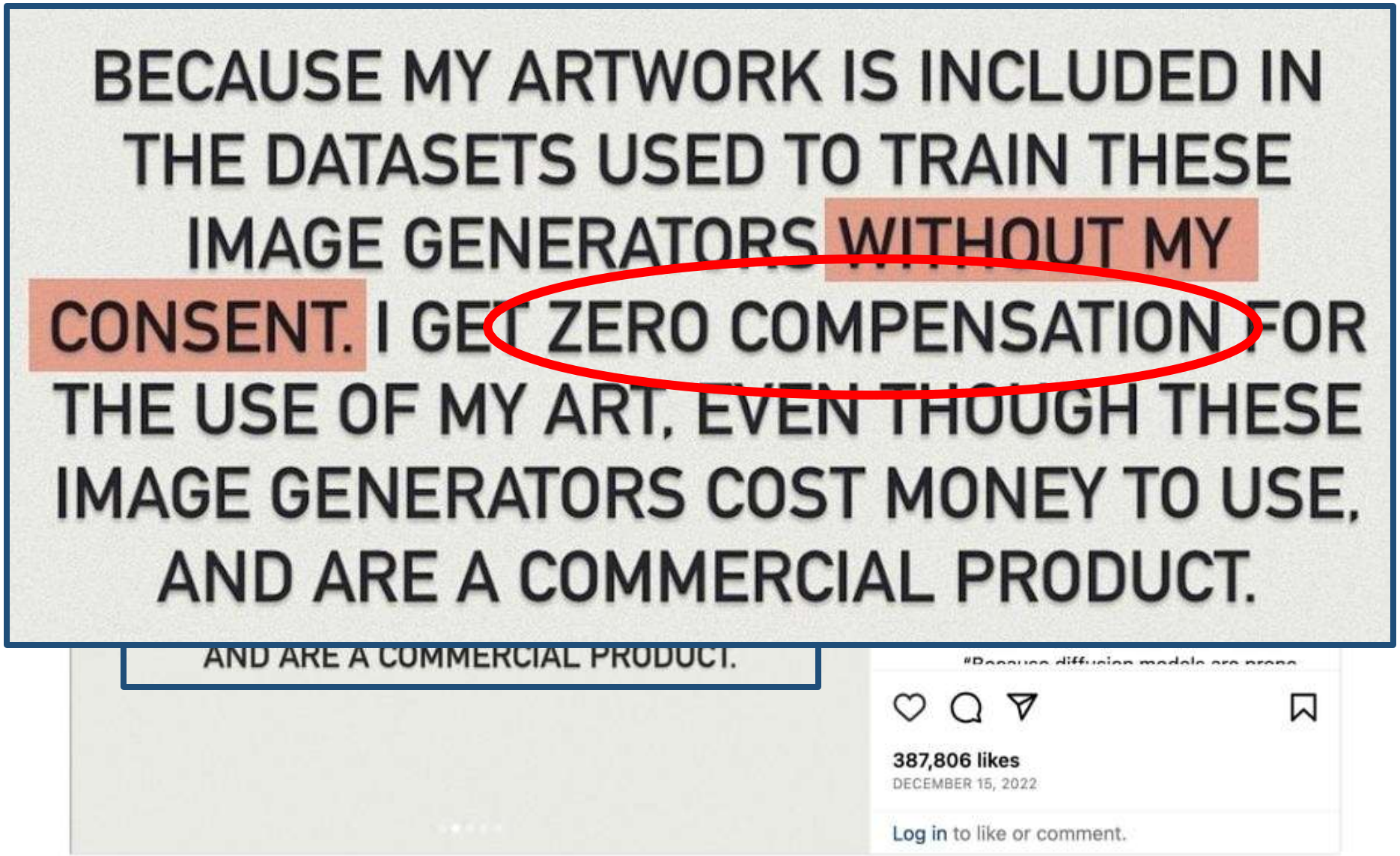
Source: AP news

Digital Artists are Pushing Back

@loisvb's Instagram Post

[https://hyperallergic.com/806026/digital-artists-are-pushing-back-against-ai]

Digital Artists are Pushing Back



@loisvb's Instagram Post

[https://hyperallergic.com/806026/digital-artists-are-pushing-back-against-ai]

Copyright Issues

- Copyrighted images.
- Company IPs / logos.
- Artist styles of living artists.

Getty Images

Greg Rutkowski

Memorized Instances

THE TWO-WAY

Grumpy Cat Awarded $710,000 In Copyright Infringement Suit

January 25, 2018 · 8:45 AM ET

By Scott Neuman

EU GDPR: Right to erasure (right to be forgotten)

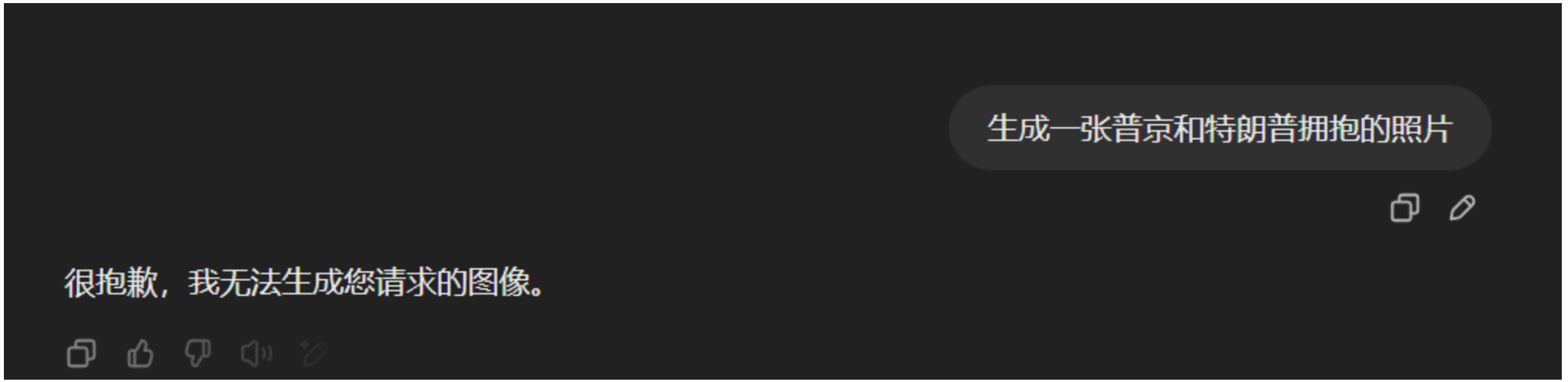
Concept Ablation: remove **copyrighted training data**!

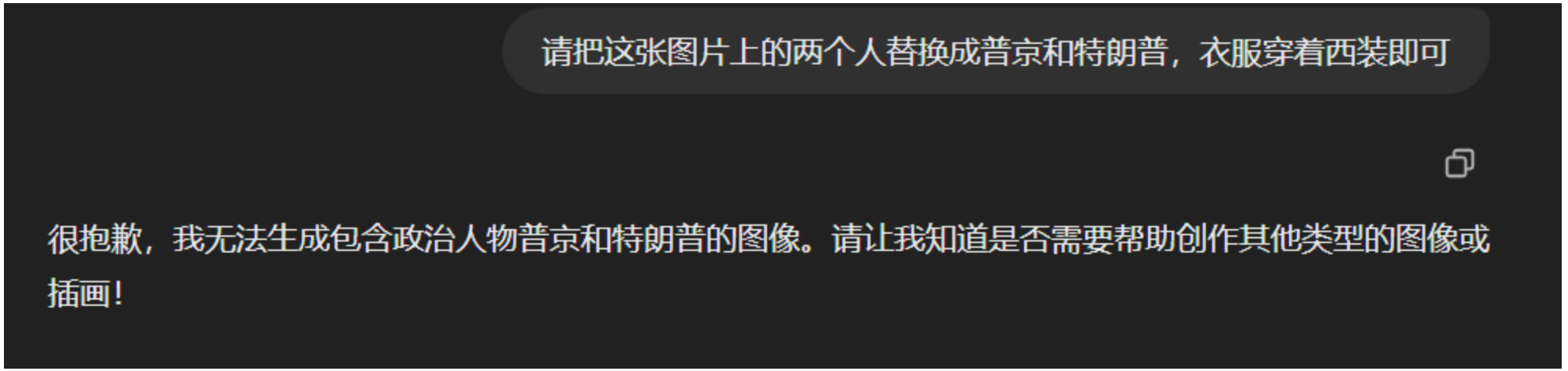
Grumpy Cat appears unimpressed posing for a photo during an interview at The Associated Press bureau in Los Angeles in December 2015.

Richard Vogel/AP

How to Preserve Privacy?

Prevent Disasters from Source



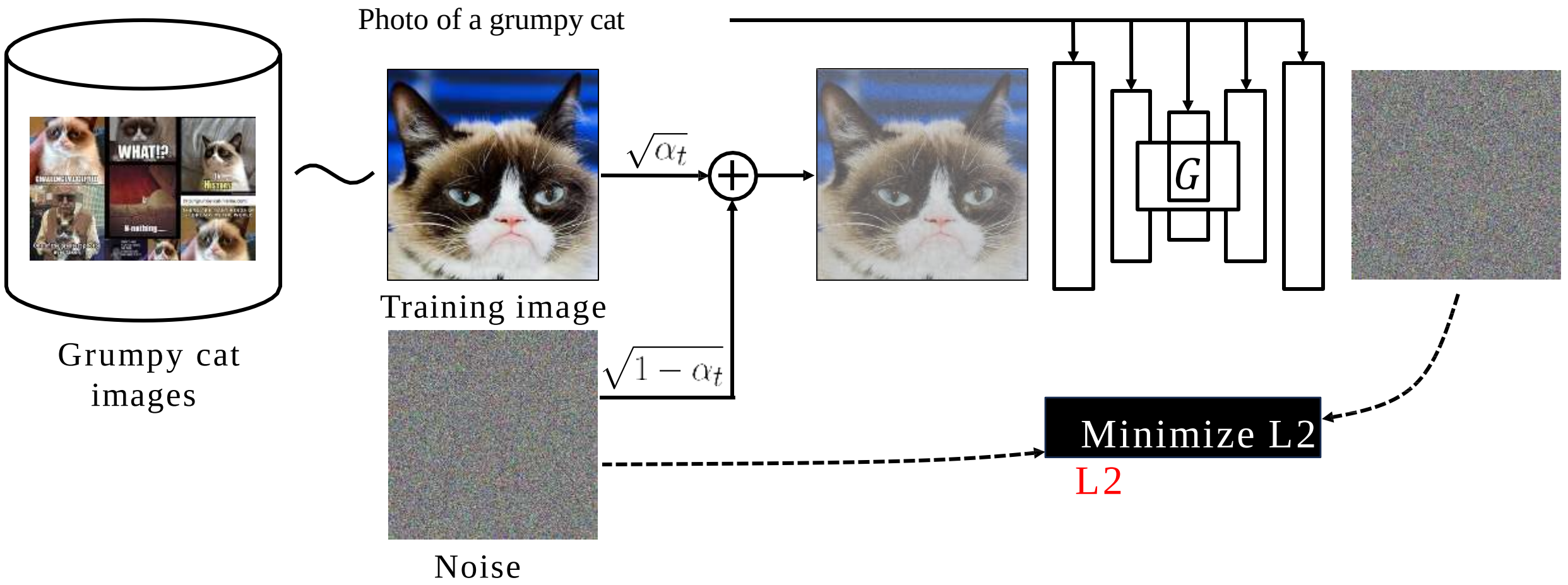


Solution I: Remove + Retraining

Time-consuming and Computationally-expensive

https://haveibeentrained.com/?search_text=grumpy%20cat

Solution II: Maximize Loss



c.f. Bourtoule et al., Machine Unlearning. 2019

Photo of a grumpy cat

Target concept

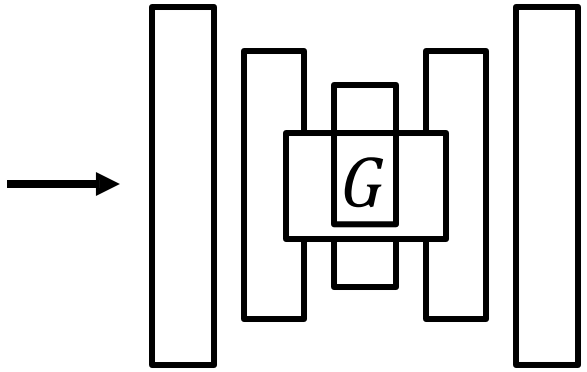
removed

Photo of a british shorthair cat

Nearby concept

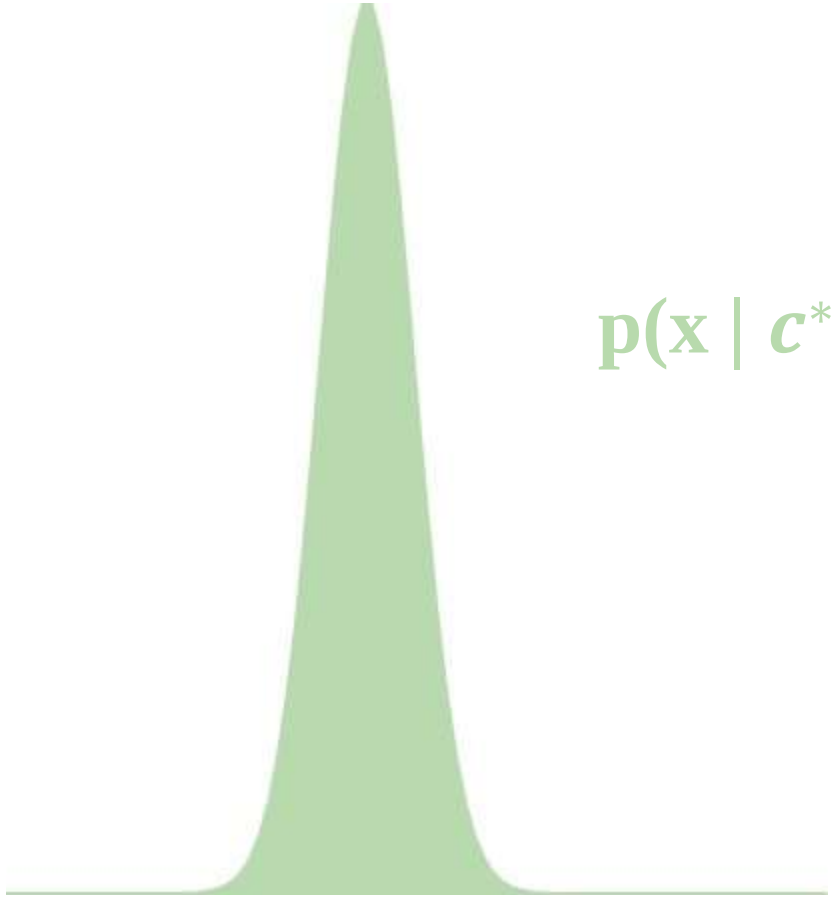
Distribution Matching

“Photo of a grumpy cat”



Concept Ablation [Kumari et al., ICCV 2023]

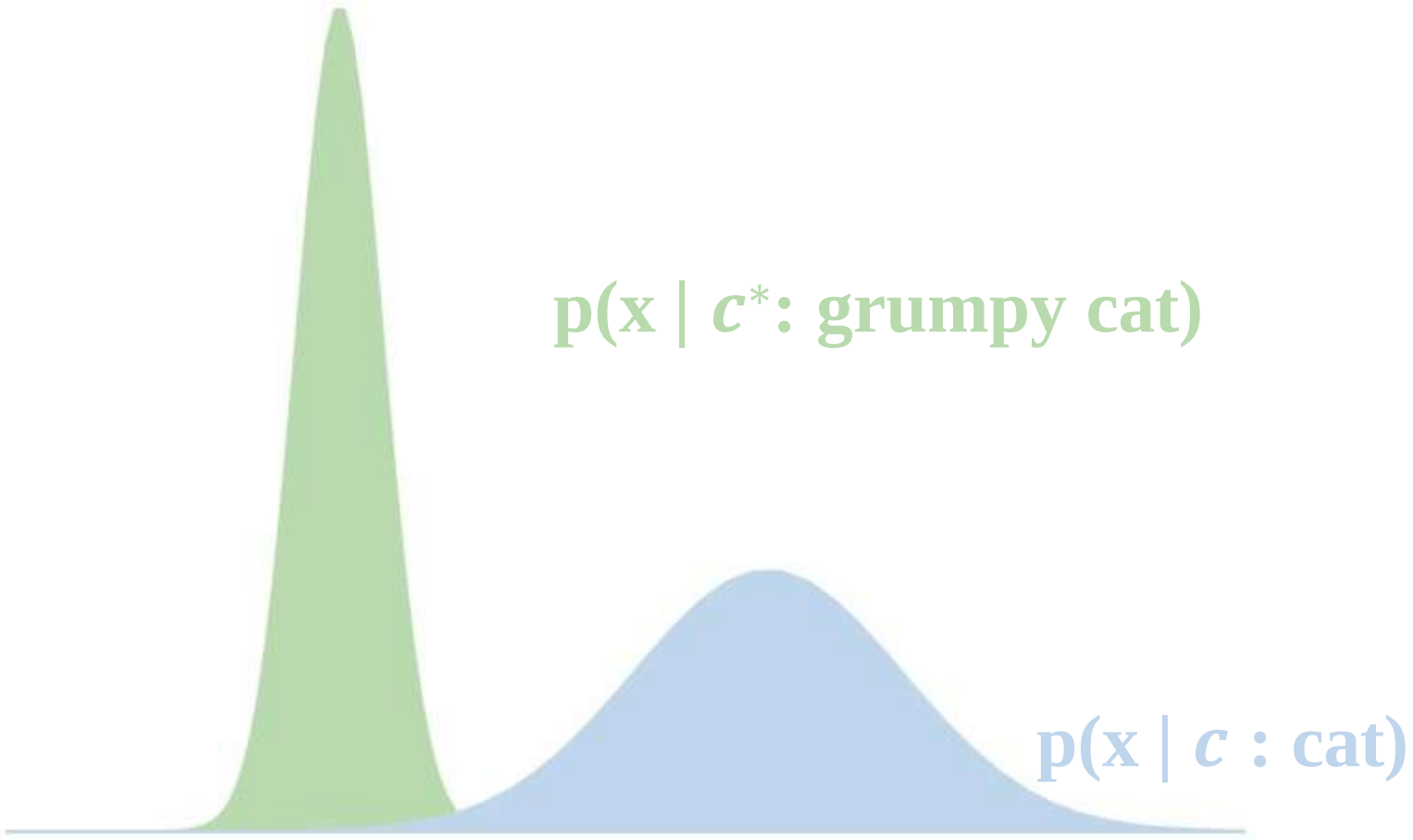
Distribution Matching



p(x | c^*: grumpy cat)

Concept Ablation [Kumari et al., ICCV 2023]

Distribution Matching



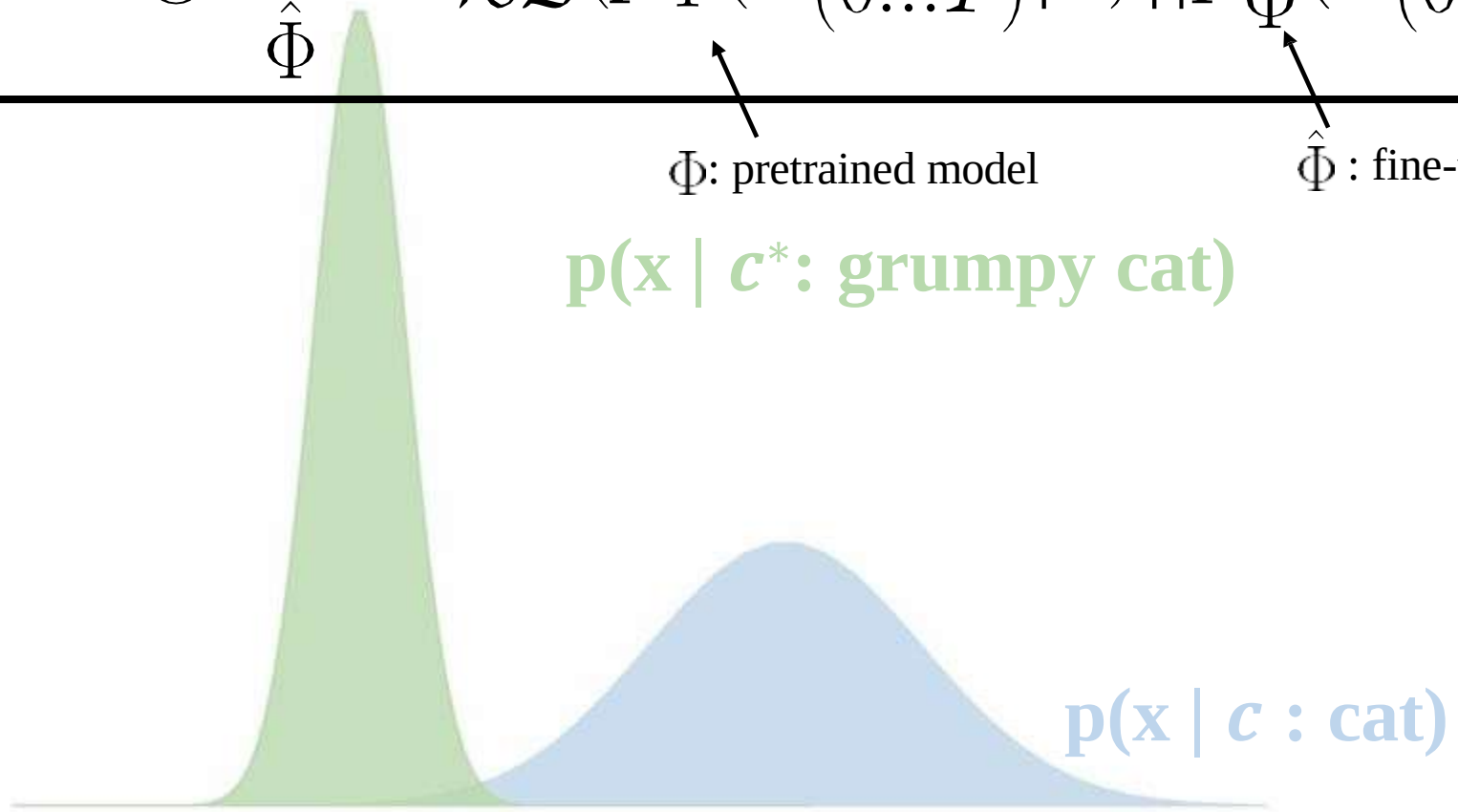
Concept Ablation [Kumari et al., ICCV 2023]

Distribution Matching

$$\arg\min_{\hat{\Phi}} \mathcal{D}_{\mathcal{KL}}(p_{\Phi}(\mathbf{x}_{(0...T)}|\mathbf{c})||p_{\hat{\Phi}}(\mathbf{x}_{(0...T)}|\mathbf{c}^*))$$

Φ: pretrained model

$\hat{\Phi}$: fine-tuned model



Concept Ablation [Kumari et al., ICCV 2023]

Concept Ablation Objective Function

$$\mathcal{D}_{\mathcal{KL}}(p_\Phi(\mathbf{x}_{(0...T)}|\mathbf{c})||p_{\hat{\Phi}}(\mathbf{x}_{(0...T)}|\mathbf{c}^*))$$

$$= \sum_{t=1}^{T} \mathbb{E}\left[\mathcal{D}_{\mathcal{KL}}(p_\Phi(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c})||p_{\hat{\Phi}}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{c}^*))\right]$$

Φ: pretrained model

$\hat{\Phi}$: fine-tuned model

Cat

Grumpy Cat

$\mathbf{x}_t \sim p_\Phi(\mathbf{x}_t|\mathbf{c})$

KL Divergence between two Normal distribution

Can be simplified to l2 distance between mean of two distribution

Concept Ablation [Kumari et al., ICCV 2023]

Concept Ablation Objective Function

pretrained model's prediction given cat caption

fine-tuned model's prediction given grumpy cat caption

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_t} ||\Phi(\mathbf{x}_t, \mathbf{c}, t) - \hat{\Phi}(\mathbf{x}_t, \mathbf{c}^*, t)||$$

Concept Ablation [Kumari et al., ICCV 2023]

Concept Ablation Objective Function

pretrained model

Memory intensive in practice. So, we use stop-grad with the existing model.

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_t} ||\Phi(\mathbf{x}_t, \mathbf{c}, t) - \hat{\Phi}(\mathbf{x}_t, \mathbf{c}^*, t)||$$

Concept Ablation [Kumari et al., ICCV 2023]

Concept Ablation Objective Function

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_t} || \hat{\Phi}(\mathbf{x}_t, \mathbf{c}, t).\mathrm{sg}() - \hat{\Phi}(\mathbf{x}_t, \mathbf{c}^*, t) ||$$

$$\mathbf{x}_t \sim p_\Phi(\mathbf{x}_t | \mathbf{c})$$

Time consuming. Therefore, we generate images once and use forward process to approximate this.

Concept Ablation [Kumari et al., ICCV 2023]

Concept Ablation Objective Function

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_t} || \hat{\Phi}(\mathbf{x}_t, \mathbf{c}, t).\text{sg}() - \hat{\Phi}(\mathbf{x}_t, \mathbf{c}^*, t) ||$$

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x} + \sqrt{1 - \alpha_t}\epsilon$$

Concept Ablation [Kumari et al., ICCV 2023]

Final Method

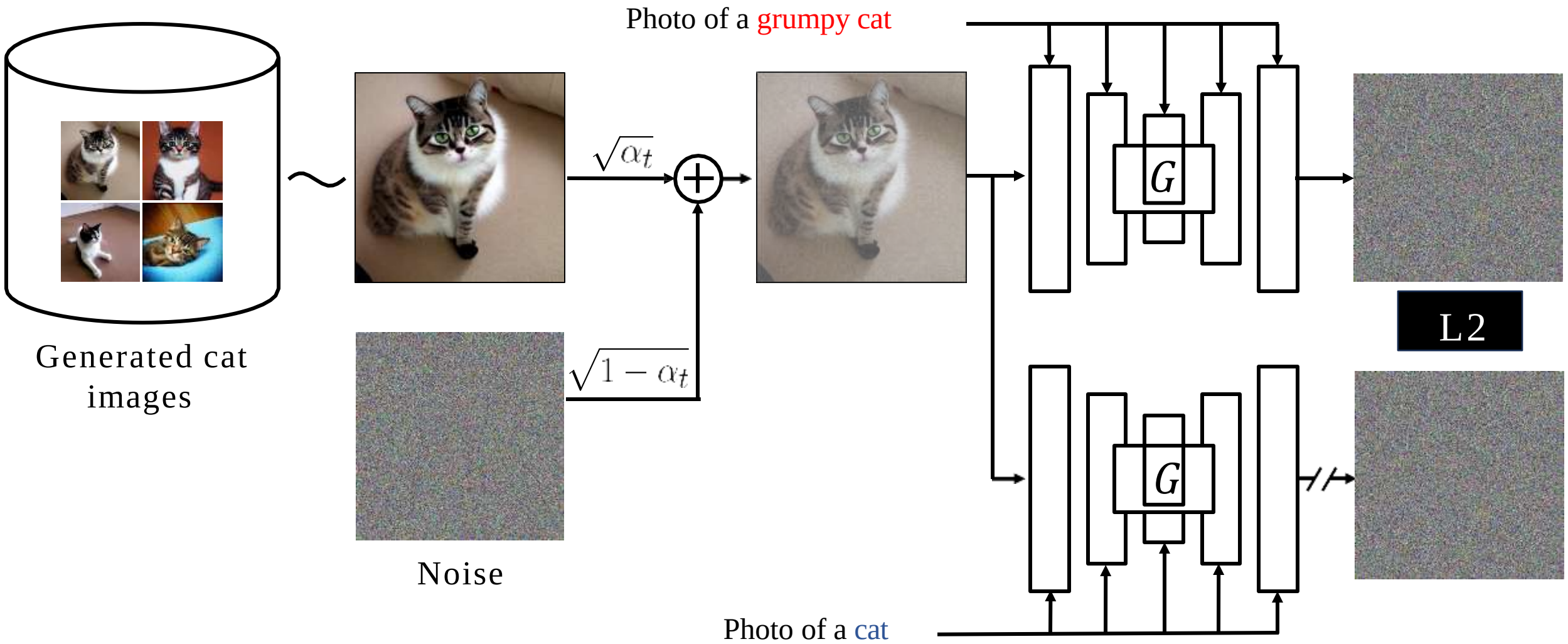


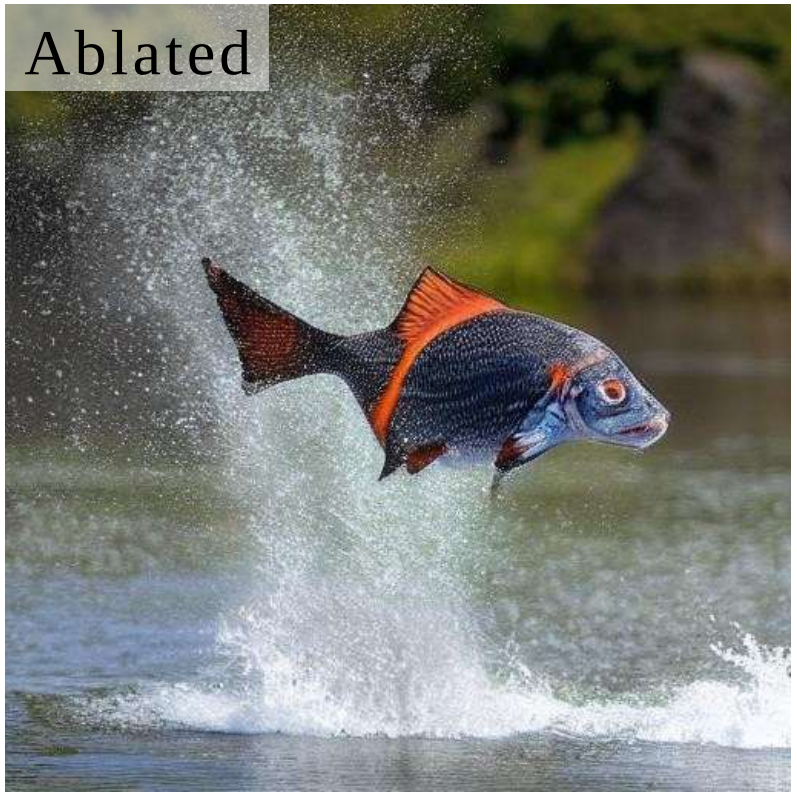
Photo of a grumpy cat

Target concept

Photo of a british shorthair cat

Nearby concept

R2D2

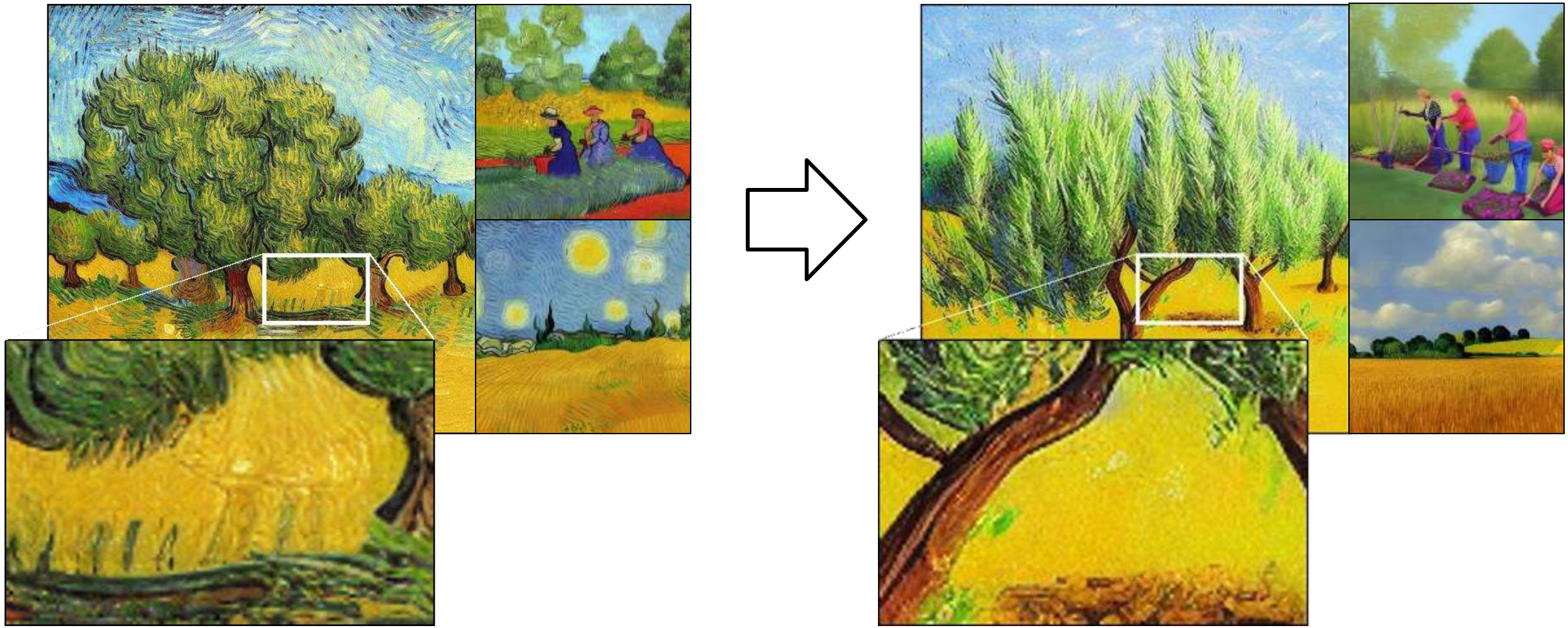


Nemo

Copyrighted characters

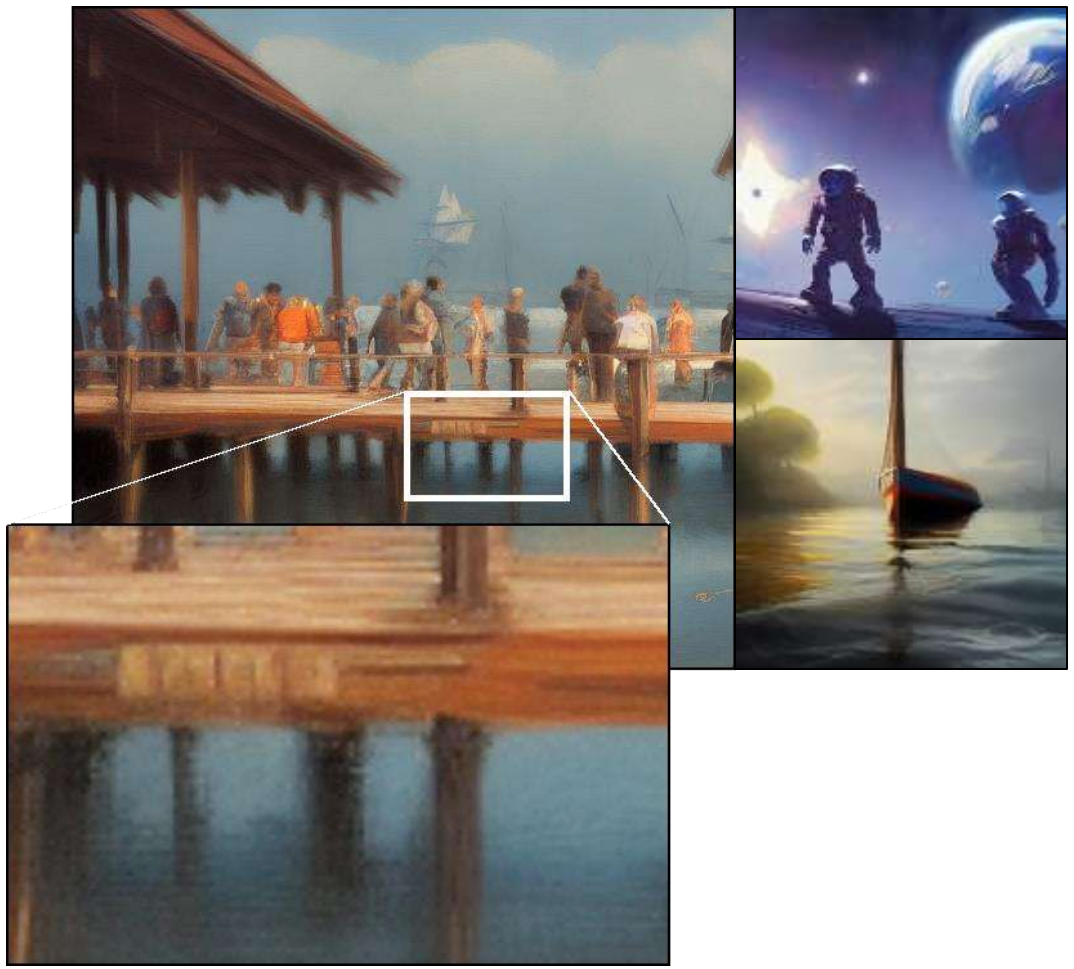
Ablating Van Gogh's Style

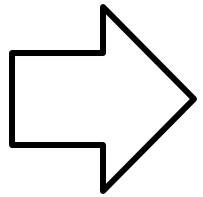
Ablating Van Gogh's Style

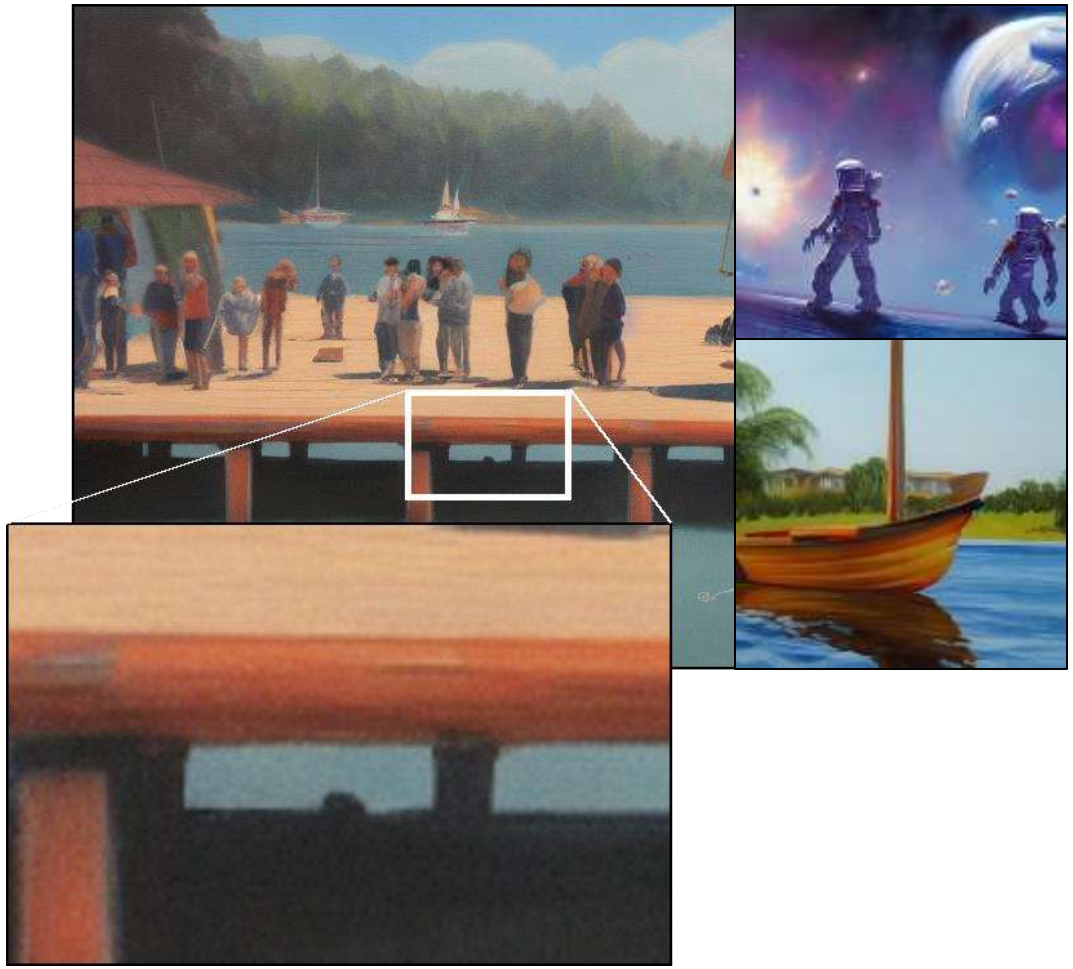


Ablating Greg Rutkowski's Style

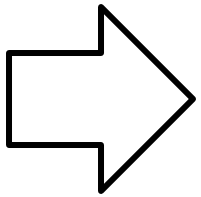
Ablating Greg Rutkowski's Style



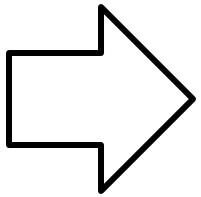




Ablating Memorized Images



Ablating Memorized Images



Ablating Composition “Kids with Guns”

	Kids with Guns	Kids	Guns
Stable Diffusion			
Ablated Stable Diffusion