

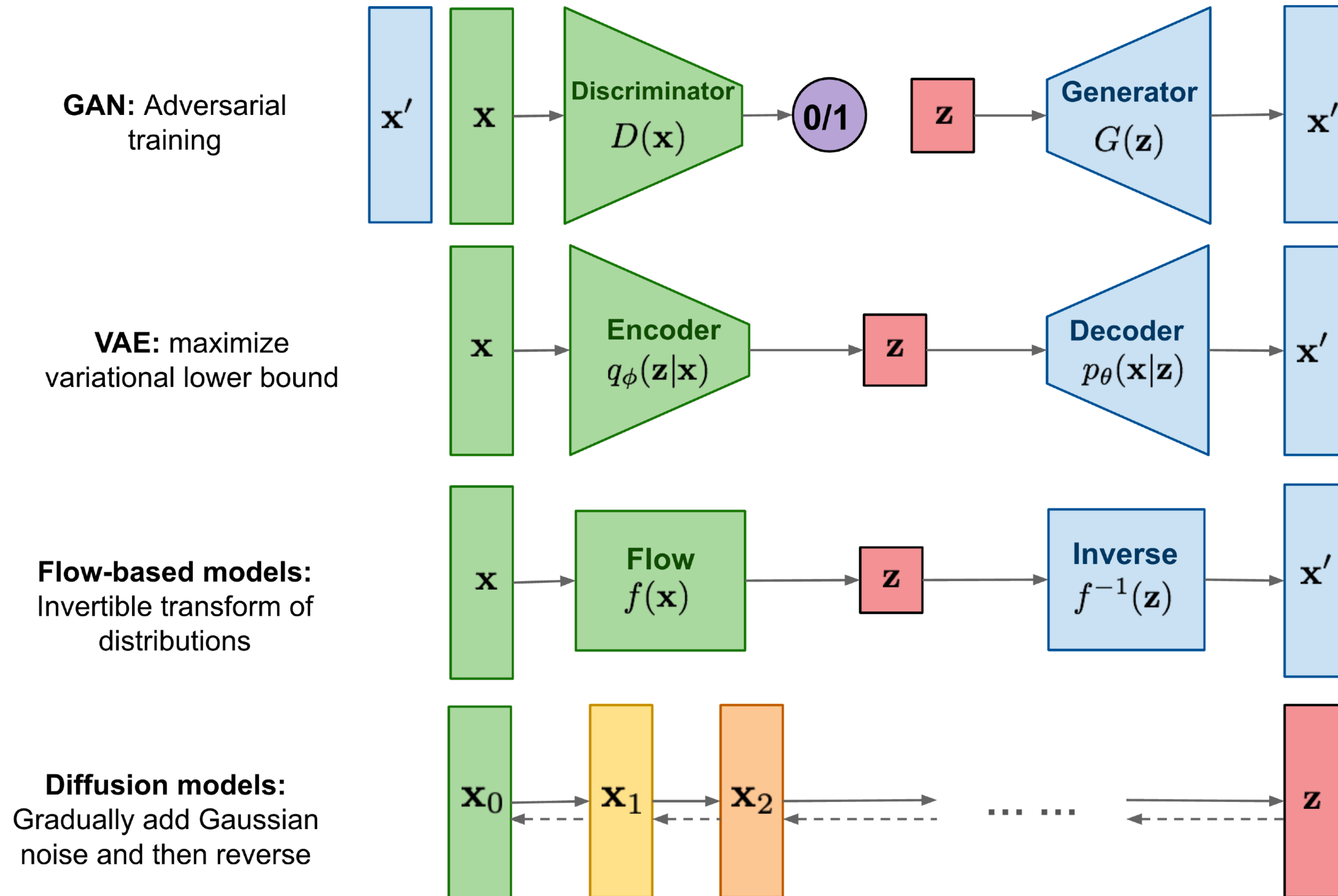
DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

arXiv 2022.08

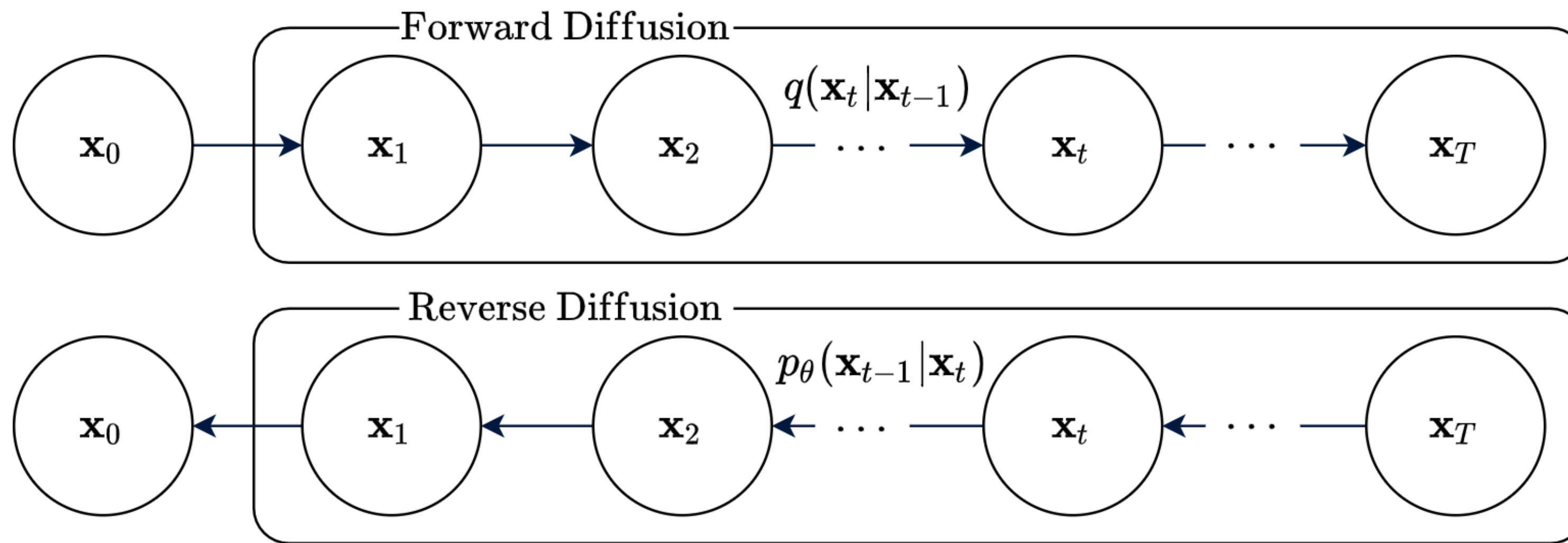
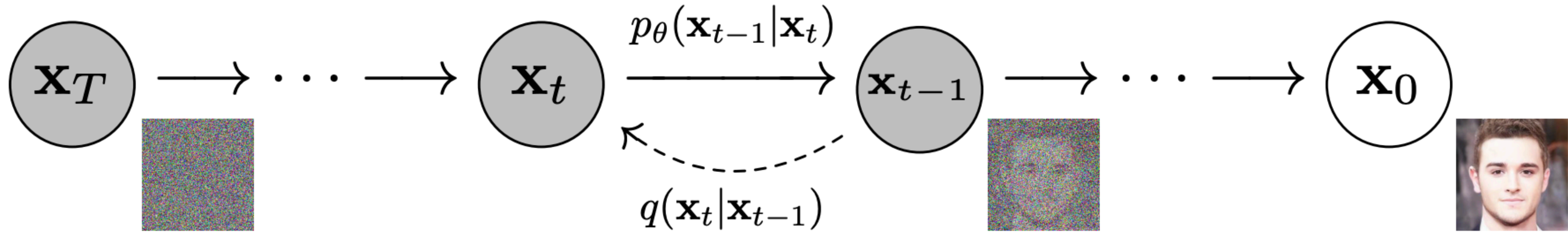
Nataniel Ruiz^{1,2}, Yuanzhen Li¹, Varun Jampani¹, Yael Pritch¹, Michael Rubinstein¹, and Kfir Aberman¹

¹*Google Research* ²*Boston University*

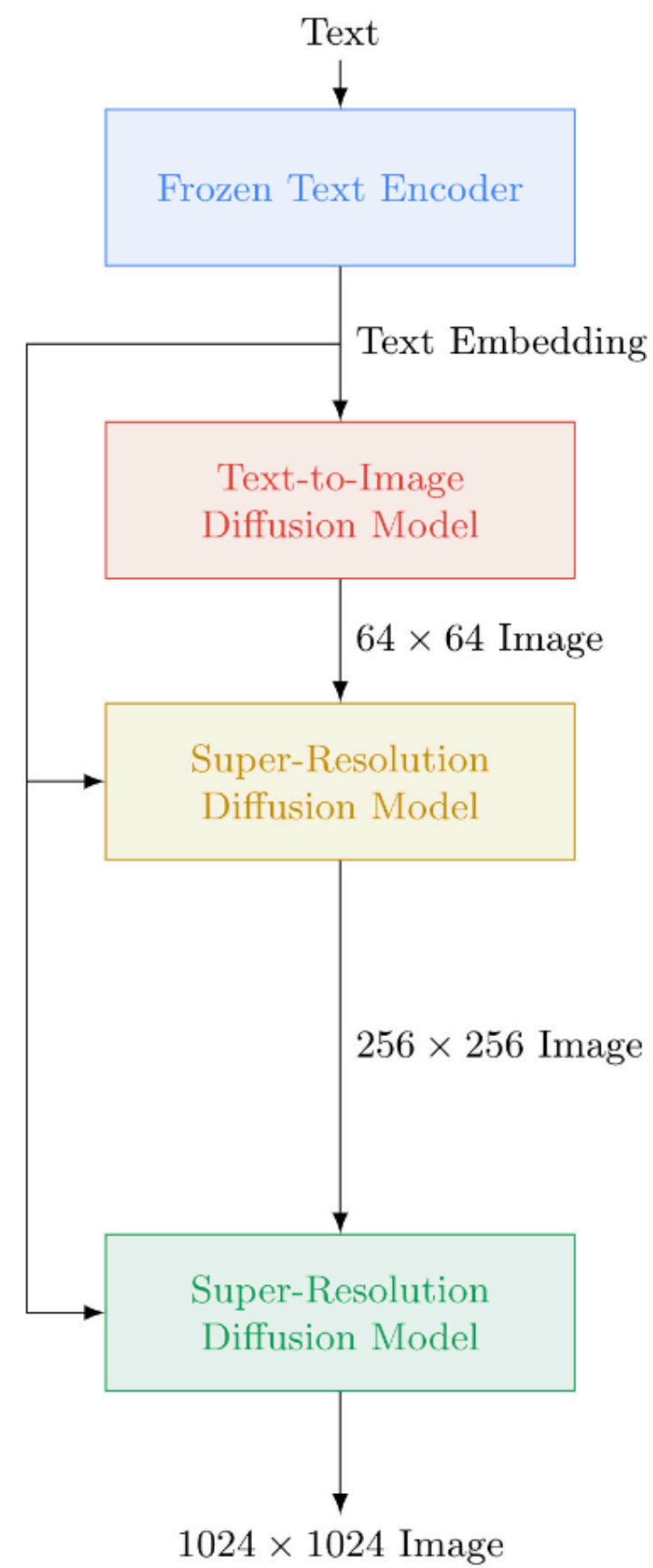
Diffusion



Diffusion



Imagen



“A Golden Retriever dog wearing a blue checkered beret and red dotted turtleneck.”



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.



Teddy bears swimming at the Olympics 400m Butterfly event.



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

Task: “personalize” text-to-image diffusion models

Subject-driven generation



Input images



in the Acropolis



swimming



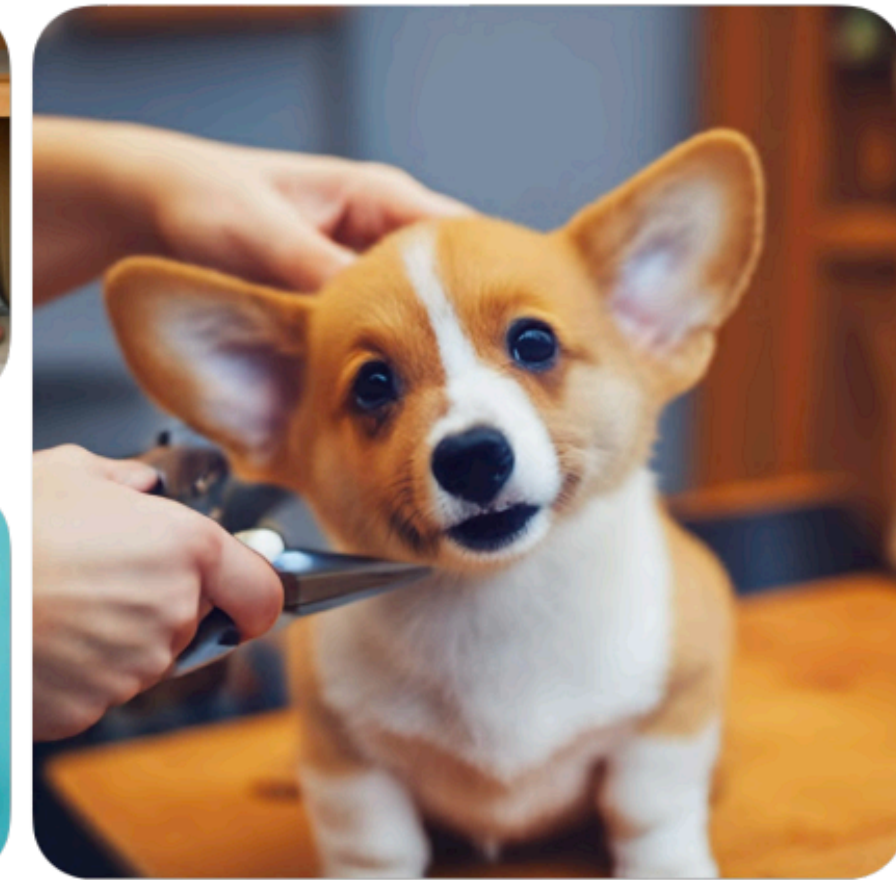
sleeping



in a doghouse



in a bucket



getting a haircut



Input images



worn by a bear



in the jungle



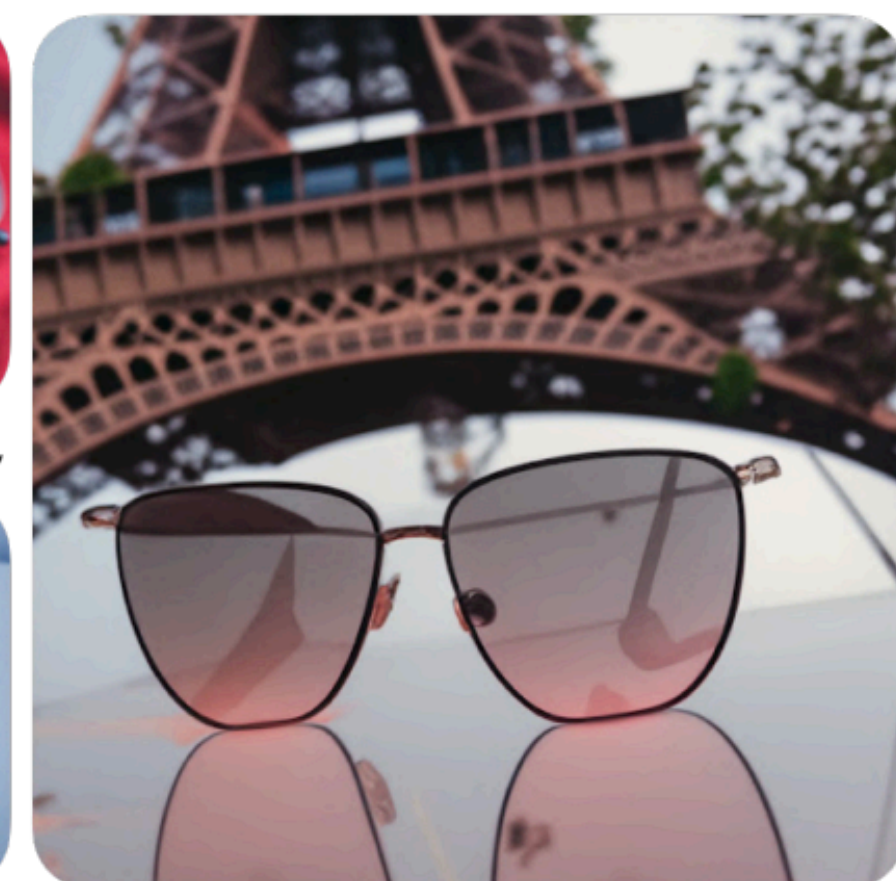
on red fabric



at Mt. Fuji



on top of snow



with Eiffel Tower

“Personalize” Related Works

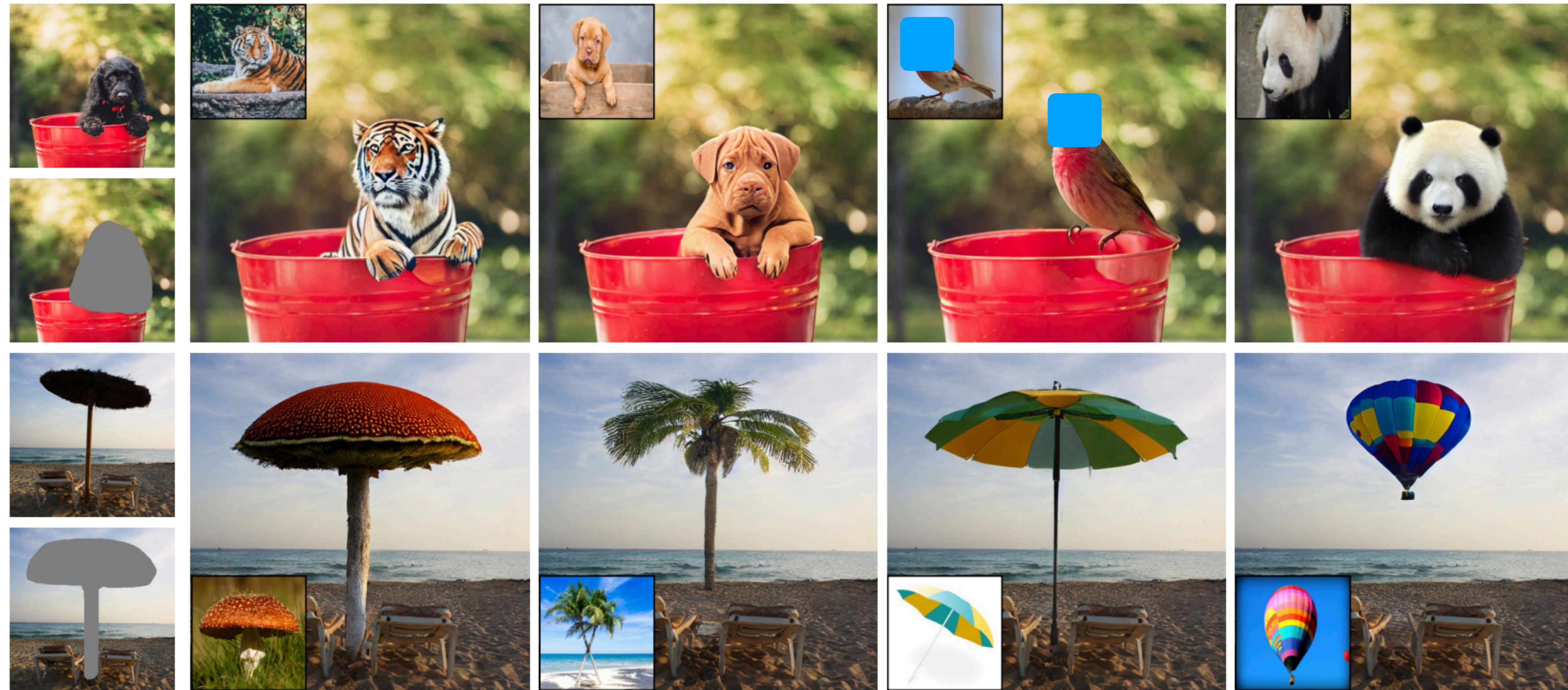


Figure 1. Paint by example. Users are able to edit a scene by painting with a conditional image. Our approach can automatically alter the reference image and merge it into the source image, and achieve a high-quality result.

Paint by Example: Exemplar-based Image Editing with Diffusion Models (CVPR-23)

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, Fang Wen

Task: “personalize” text-to-image diffusion models

Subject-driven generation

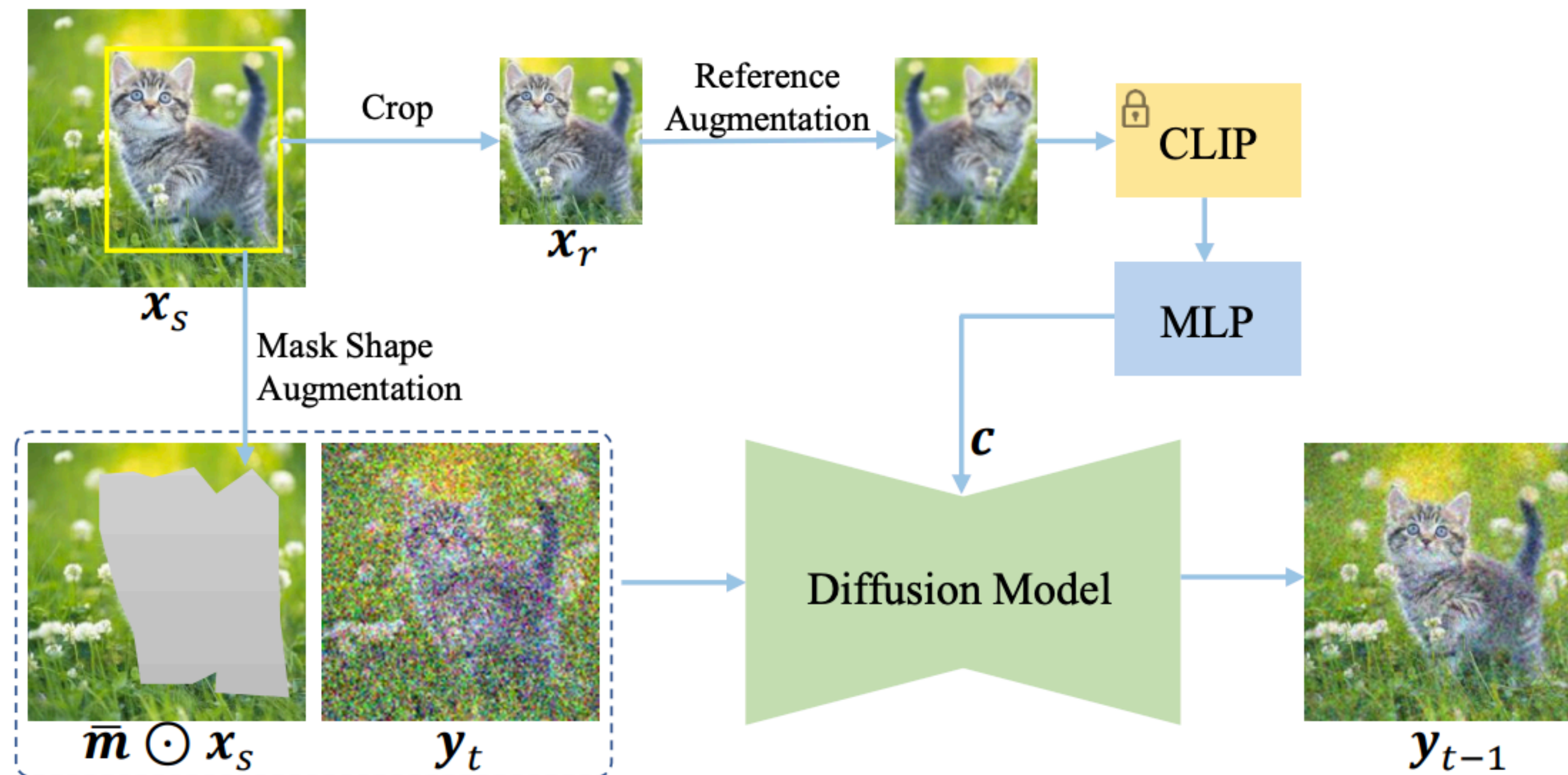


Figure 4. Our training pipeline.

“Personalize” Related Works



An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion (arXiv 2022.08)
Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, Daniel Cohen-Or

“Personalize” Related Works

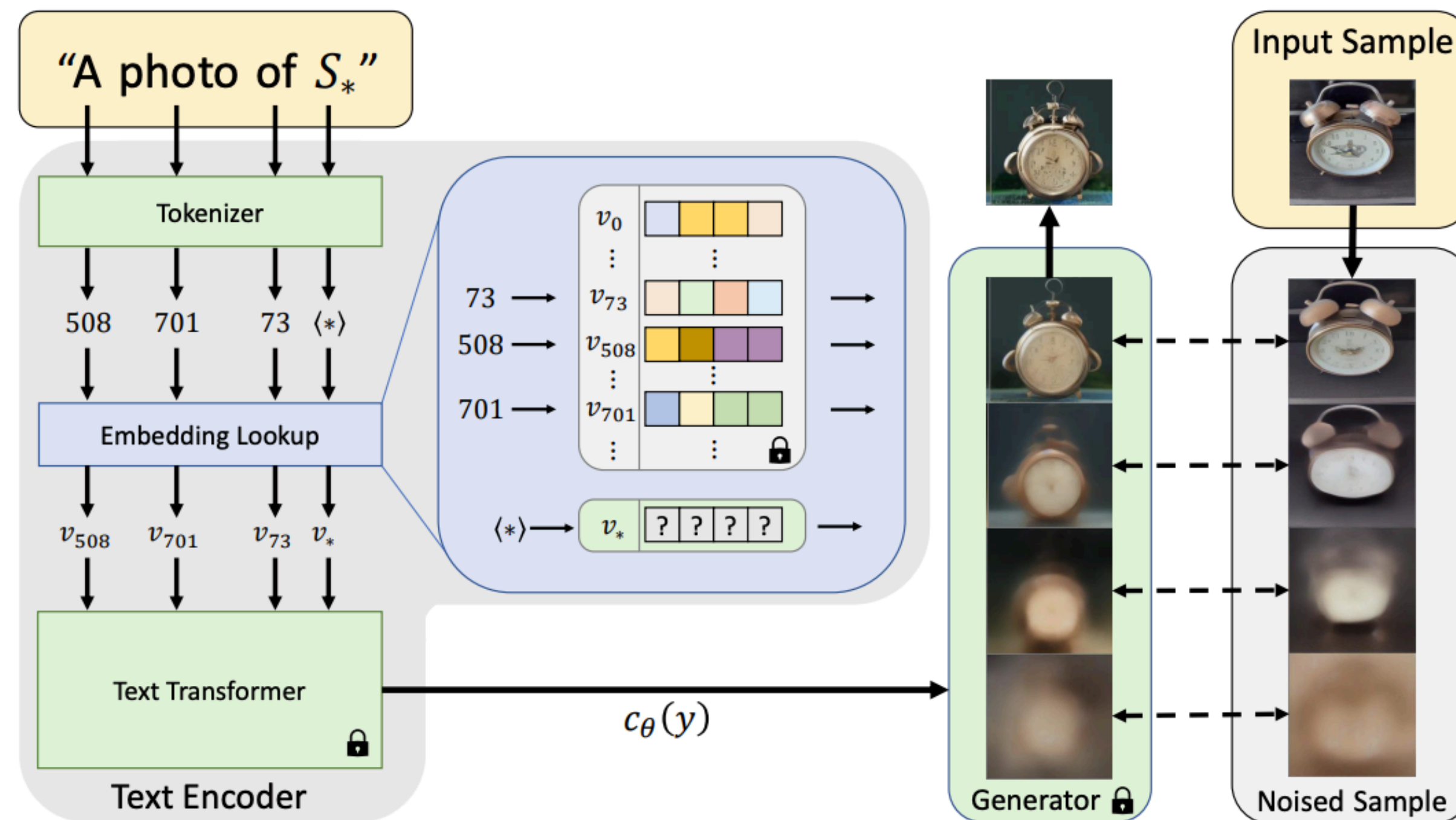


Figure 2: Outline of the text-embedding and inversion process. A string containing our placeholder word is first converted into tokens (*i.e.* word or sub-word indices in a dictionary). These tokens are converted to continuous vector representations (the “embeddings”, v). Finally, the embedding vectors are transformed into a single conditioning code $c_\theta(y)$ which guides the generative model. We optimize the embedding vector v_* associated with our pseudo-word S_* , using a reconstruction objective.

Task

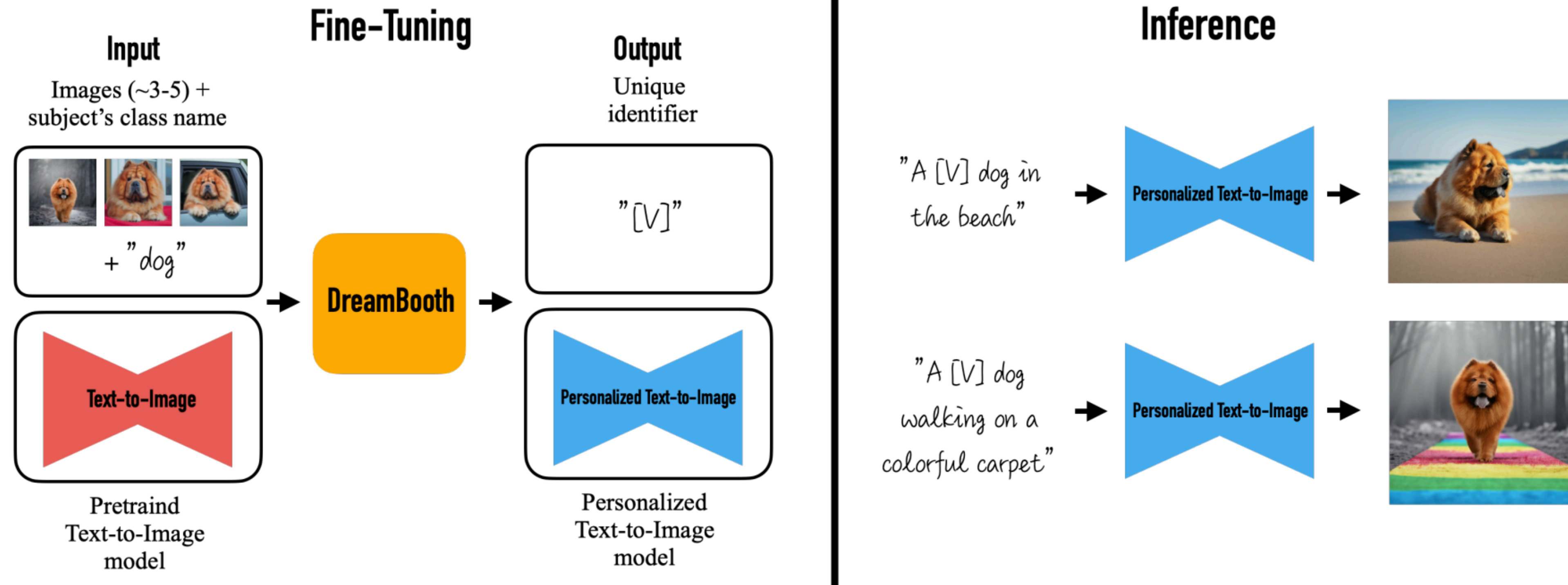


Figure 3: **High-level method overview.** Our method takes as input a few images (typically 3 – 5 images suffice, based on our experiments) of a subject (e.g., a specific dog) and the corresponding class name (e.g. “dog”), and returns a fine-tuned/“personalized” text-to-image model that encodes a unique identifier that refers to the subject. Then, at inference, we can implant the unique identifier in different sentences to synthesize the subjects in difference contexts.

Problems of naive fine-tuning

- Overfit to both the context and the appearance of the subject

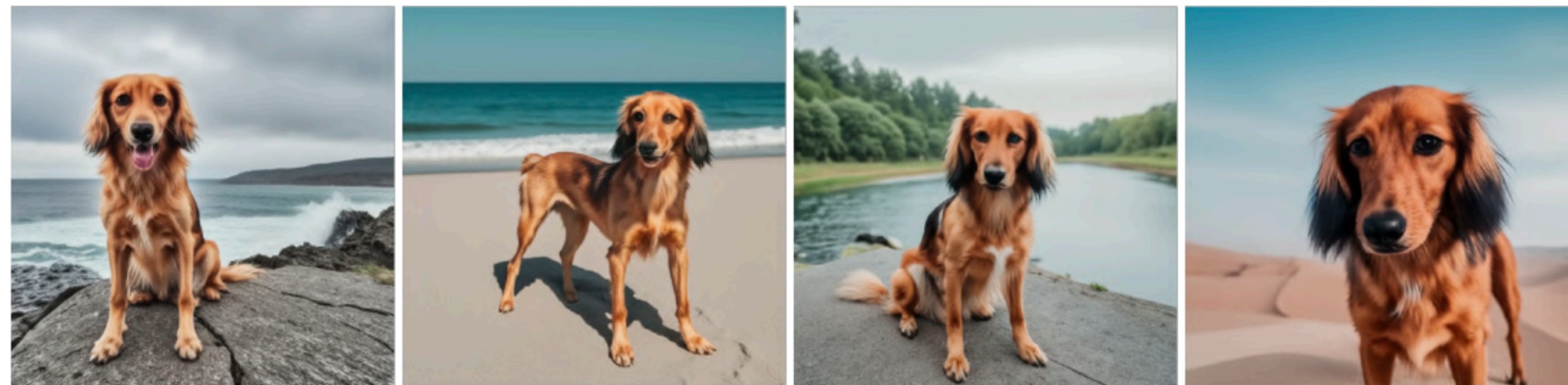
Input images



w/o prior-preservation loss



Ours (full)

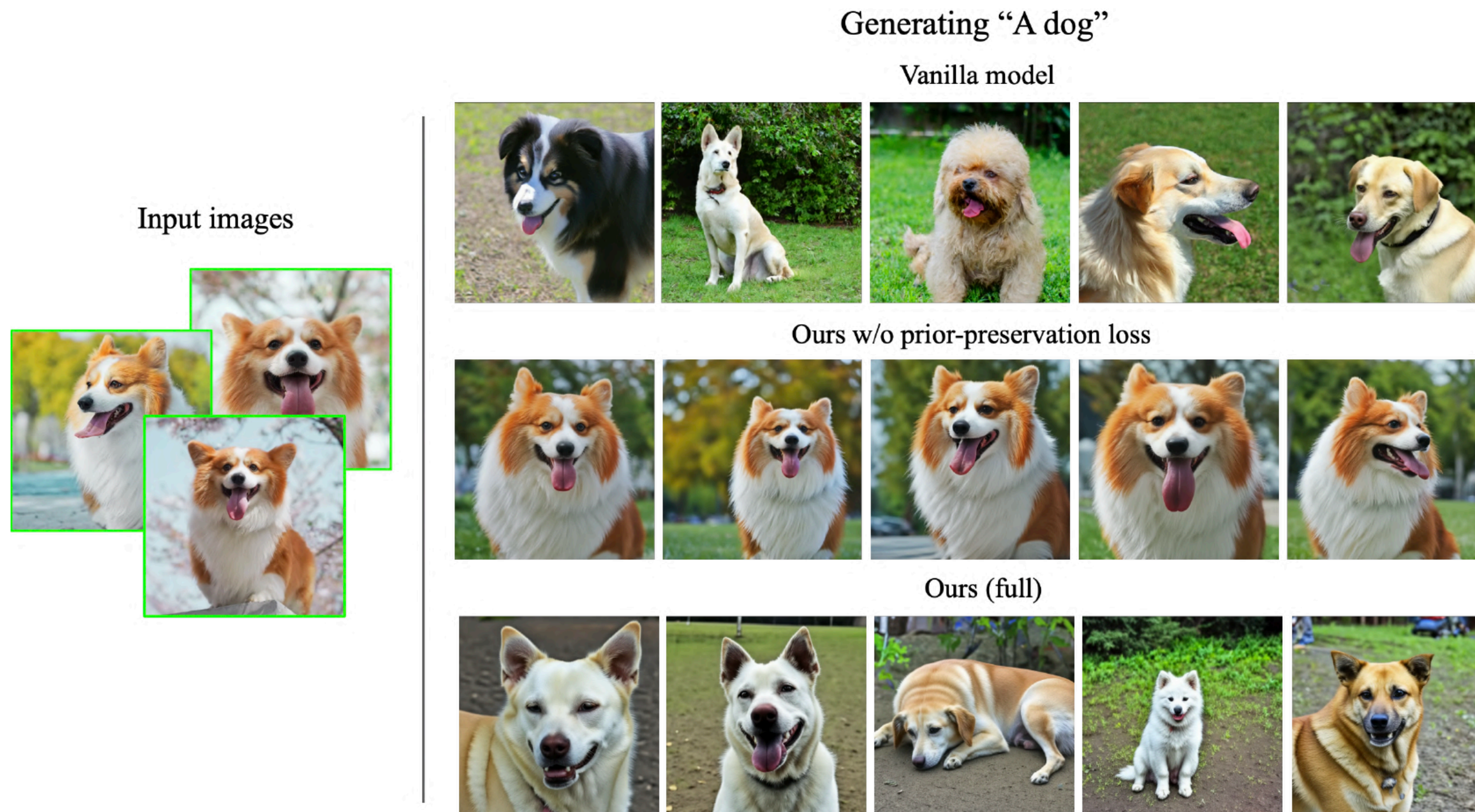


Problems of naive fine-tuning

- Overfit to both the context and the appearance of the subject
- Probable solutions: regularization, selectively fine-tuning certain parts
 - Uncertainty on which layers to fine-tune
 - Best results are achieved by fine-tuning all layers

Problems of naive fine-tuning

- Language drift: forgets how to generate subjects of the same class



Class-specific Prior Preservation Loss

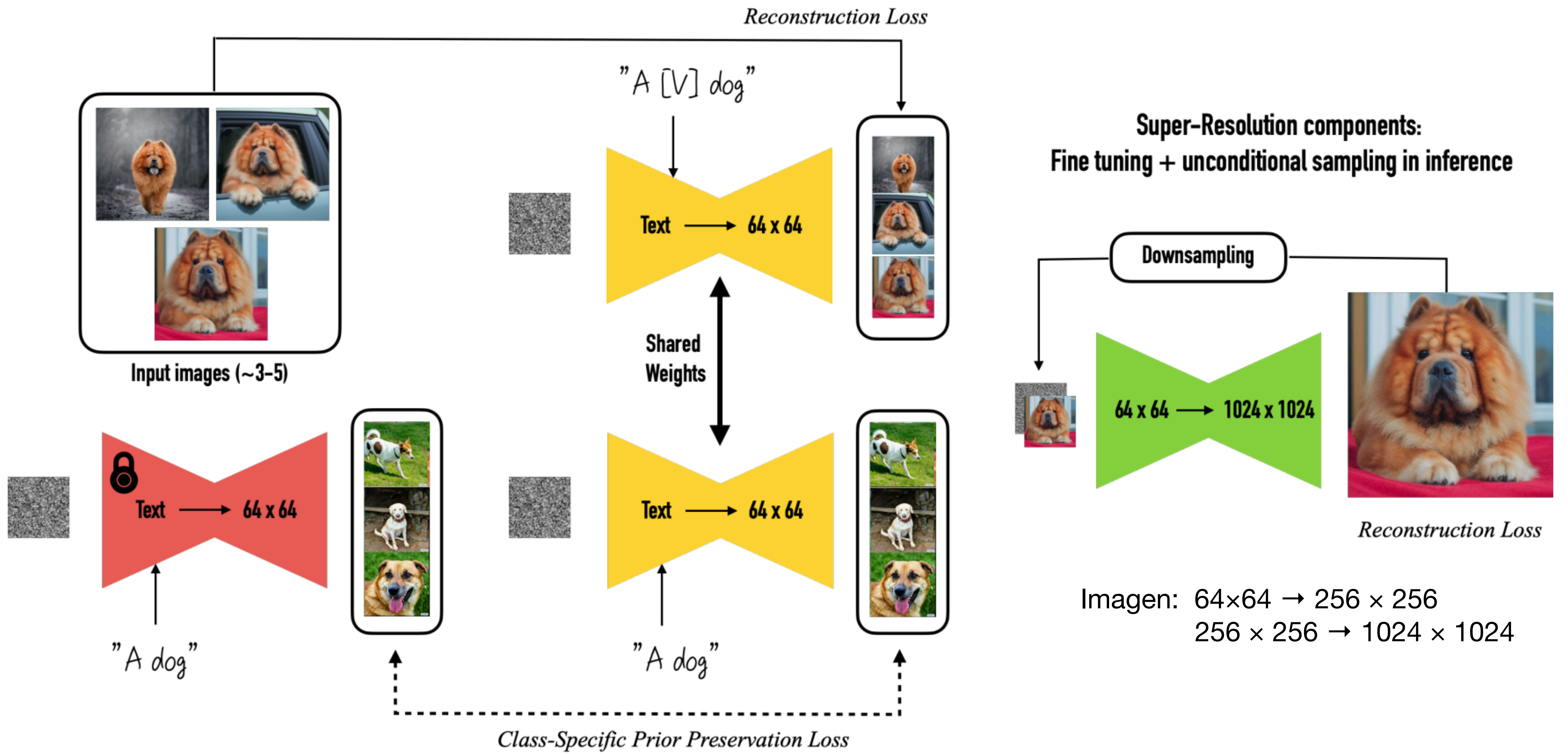
- Solution: set the input text to be “a sks dog”
 - [identifier] = “unique”/“special” → existing English words have prior. Need to disentangle original meaning and the target subject.
 - [identifier] = rare identifier (e.g. “xxy5syt00”) → tokenize each letter separately
 - [identifier] = rare-token identifier “sks” → **good**

Class-specific Prior Preservation Loss

- Class-specific prior $\mathbf{x}_{\text{pr}} = \hat{\mathbf{x}}(\mathbf{z}_{t_1}, \mathbf{c}_{\text{pr}})$ $\mathbf{z}_{t_1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- New loss function: $\mathbf{c}_{\text{pr}} := \Gamma(f(\text{"a [class noun]"}))$

$$\mathbb{E}_{\mathbf{x}, \mathbf{c}, \epsilon, \epsilon', t} [w_t \|\hat{\mathbf{x}}_{\theta}(\alpha_t \mathbf{x} + \sigma_t \epsilon, \mathbf{c}) - \mathbf{x}\|_2^2 + \lambda w_{t'} \|\hat{\mathbf{x}}_{\theta}(\alpha_{t'} \mathbf{x}_{\text{pr}} + \sigma_{t'} \epsilon', \mathbf{c}_{\text{pr}}) - \mathbf{x}_{\text{pr}}\|_2^2]$$

- ~200 epochs at learning rate 10-5 with $\lambda = 1$
- ~200 N “a dog” samples are generated. N is the size of the subject dataset (about 3-5)
- ~15 minutes on one TPUv4.



Reduce the level of noise augmentation from 10-3 to 10-5 during fine-tuning of the 256×256 SR model.

Reference Real Images

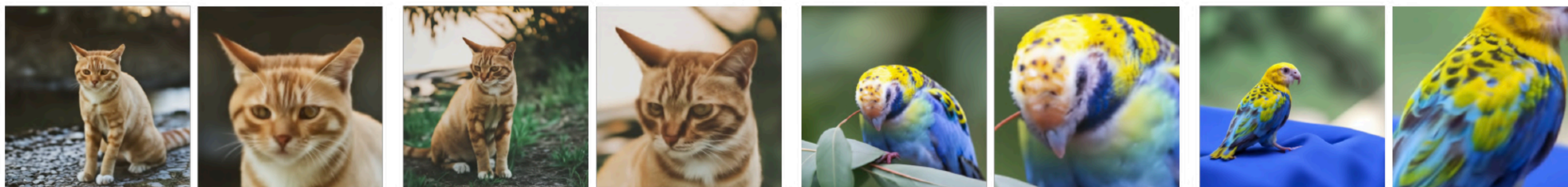


Reference Real Images



Generated Images

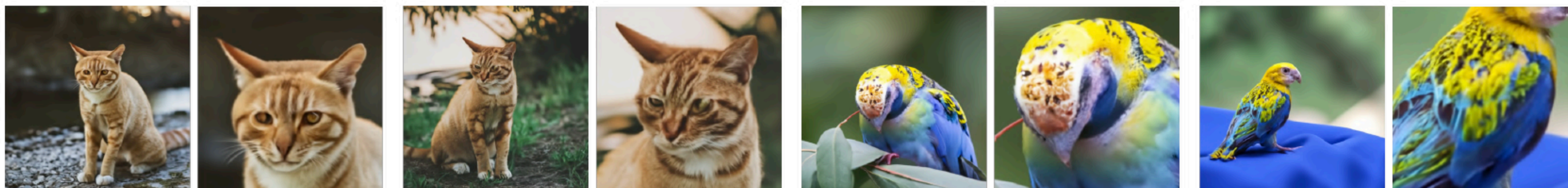
Ours



Normal Noise



No Finetuning



Experimental Results On Recontextualization

Input images



A [V] backpack in the Grand Canyon



A [V] backpack with the night sky



A [V] backpack in the city of Versailles



A wet [V] backpack in water



A [V] backpack in Boston

Input images



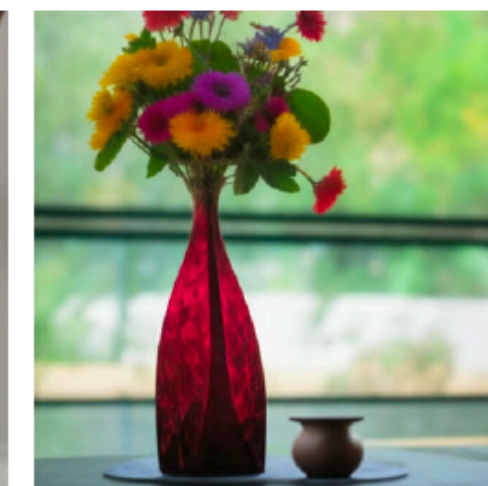
A [V] vase buried in the sands



Two [V] vases on a table



Milk poured into a [V] vase



A [V] vase with a colorful flower bouquet



A [V] vase in the ocean

Experimental Results On Recontextualization

Input images



A [V] teapot floating
in the sea



A [V] teapot floating
in milk



A bear pouring from
a [V] teapot

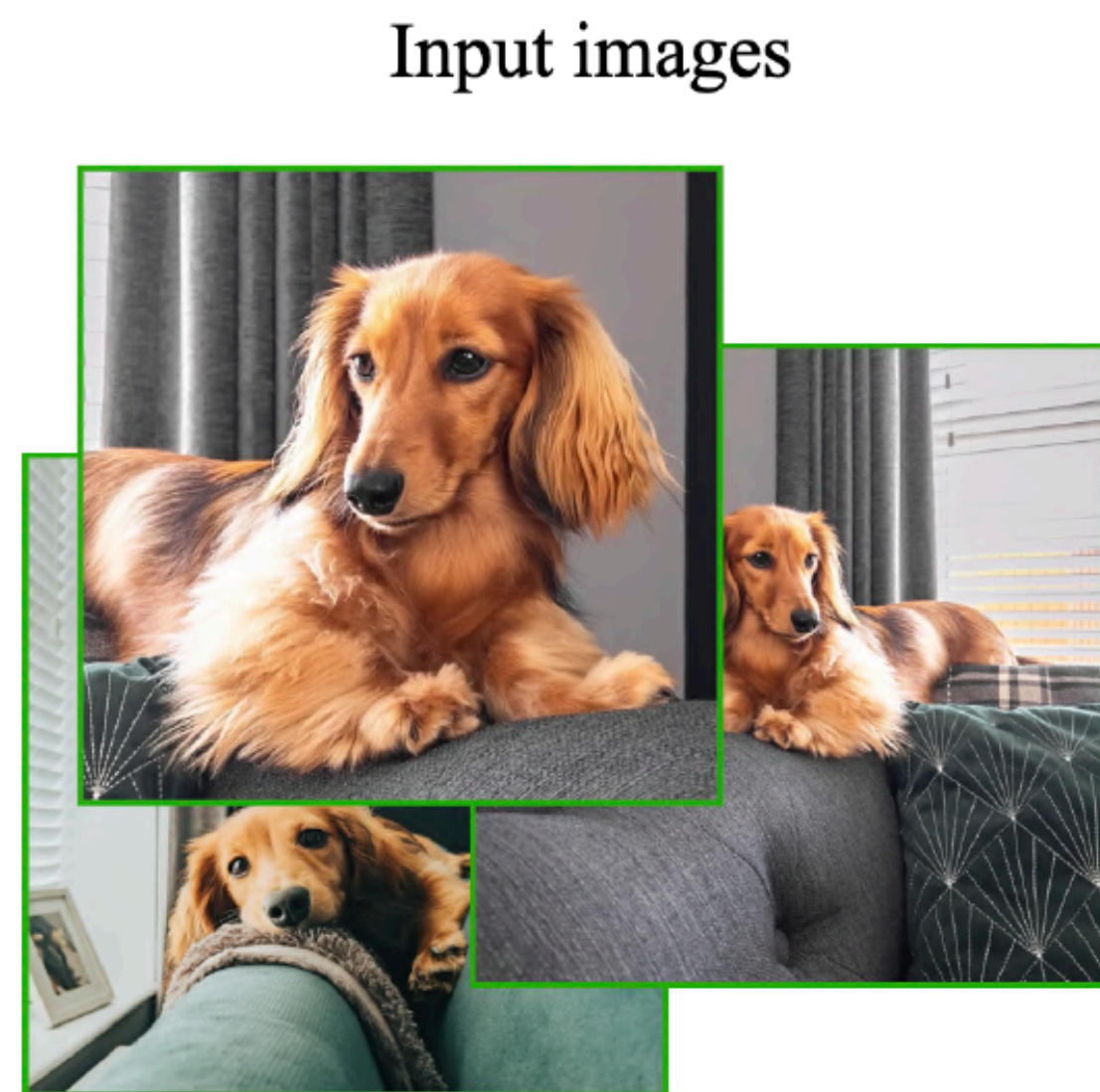


A transparent [V] teapot
with milk inside



A [V] teapot pouring tea

Experimental Results On Art Renditions



Vincent Van Gogh



Michelangelo



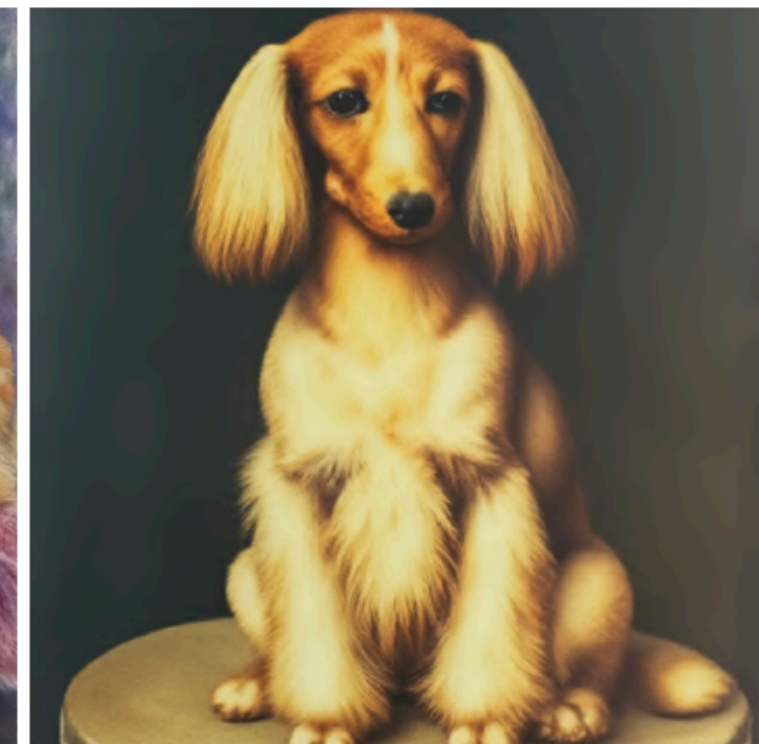
Rembrandt



Johannes Vermeer



Pierre-Auguste Renoir



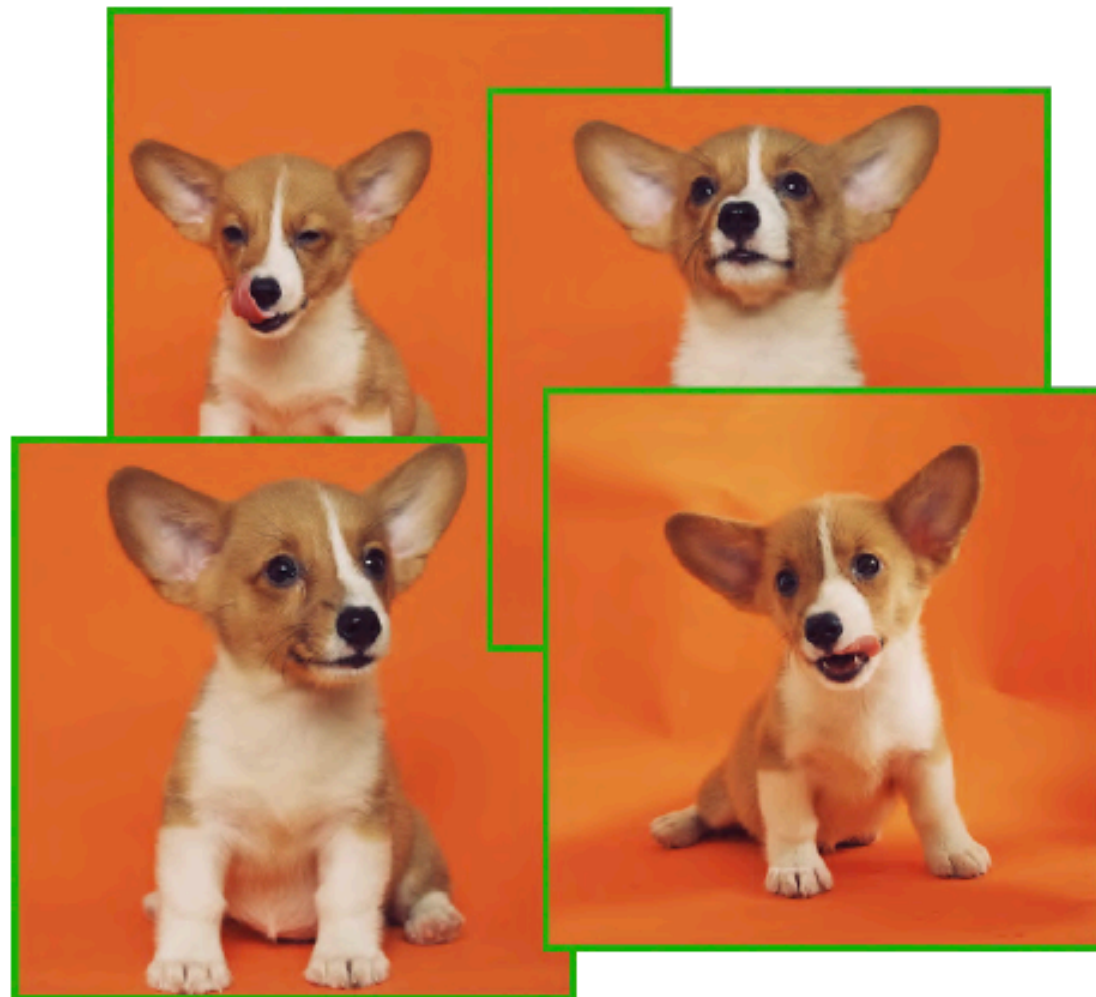
Leonardo da Vinci

“a painting of a [V] [class noun] in the style of [famous painter]” or “a statue of a [V] [class noun] in the style of [famous sculptor]”

Experimental Results On Expression Manipulation

Expression modification (“A [state] [V] dog”)

Input images



depressed



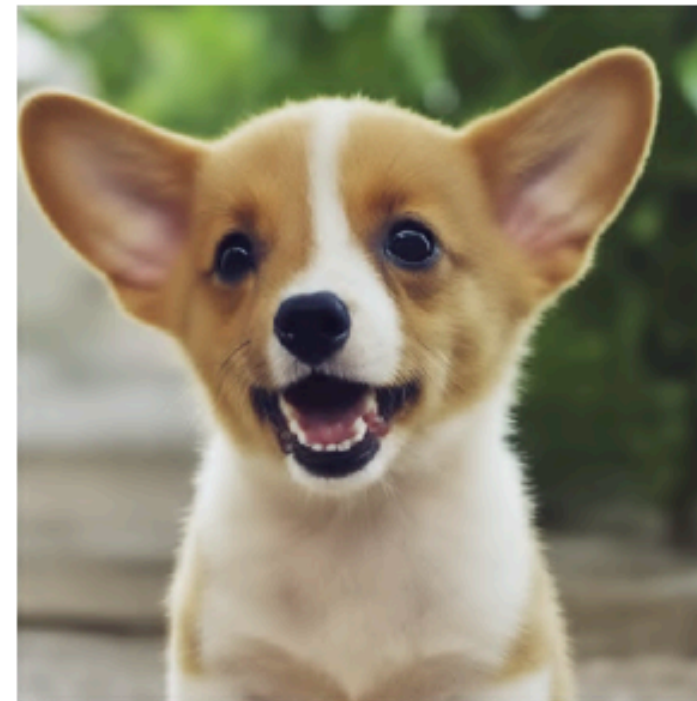
sleeping



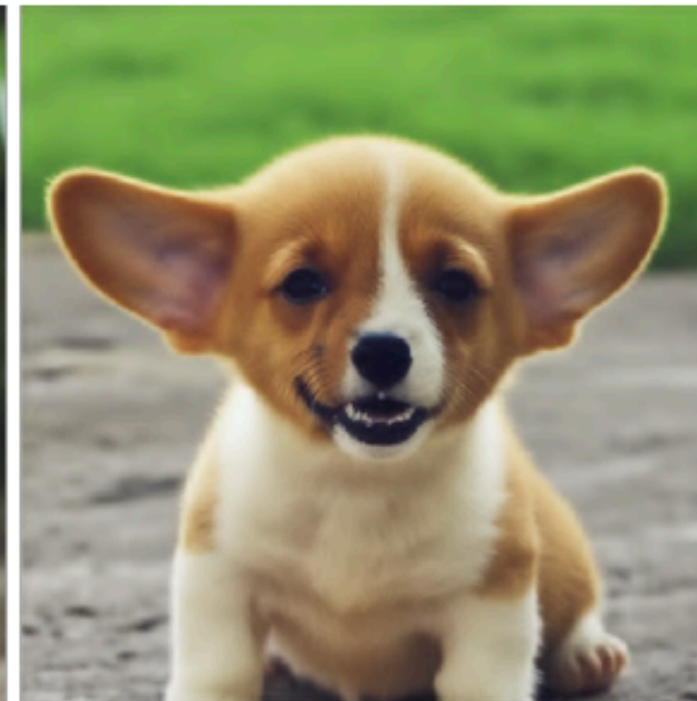
sad



joyous



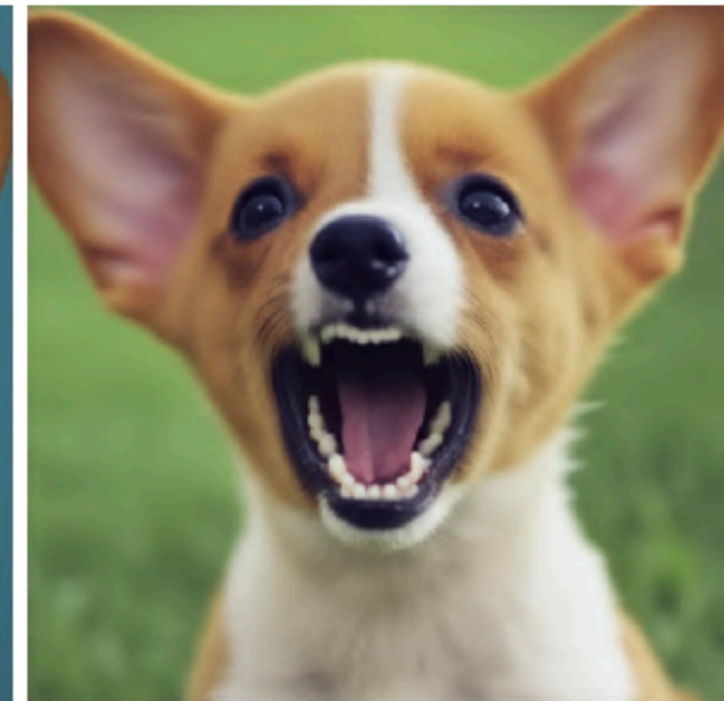
barking



crying

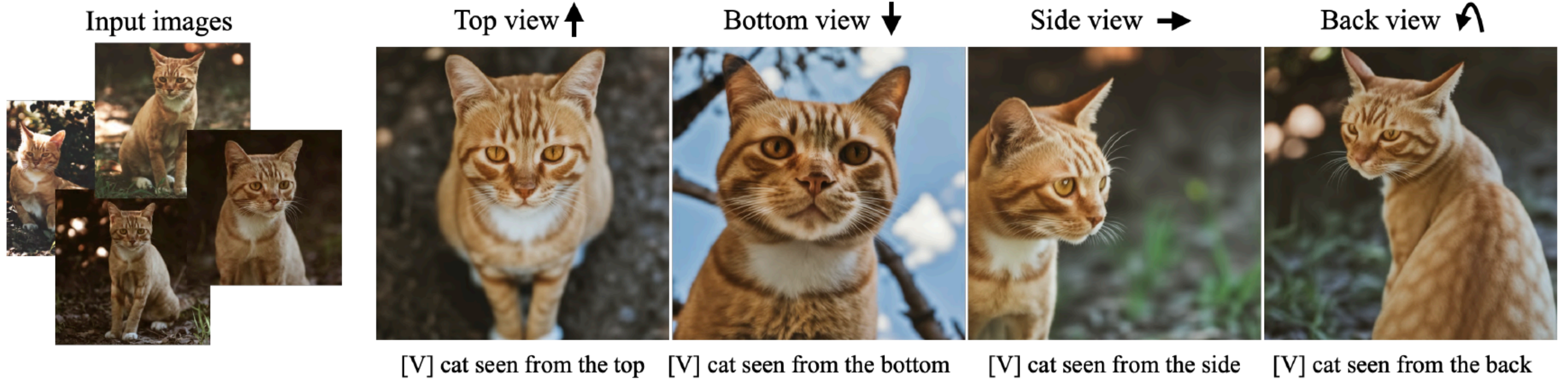


frowning



screaming

Experimental Results On Novel View Synthesis



Experimental Results On Accessorization

“a [M] [class noun] wearing [accessory]”

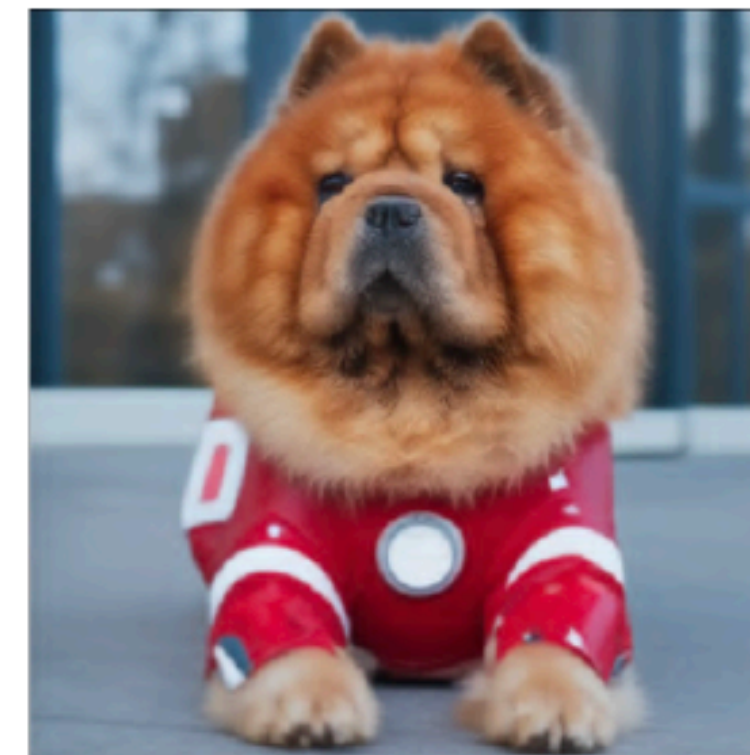
Input images



Chef Outfit



Witch Outfit



Ironman Outfit



Nurse Outfit



Purple Wizard Outfit



Superman Outfit



Police Outfit



Angel Wings

Experimental Results On Property Modification

Color modification (“A [color] [V] car”)



Input



purple



red



yellow



blue



pink

Hybrids (“A cross of a [V] dog and a [target species]”)



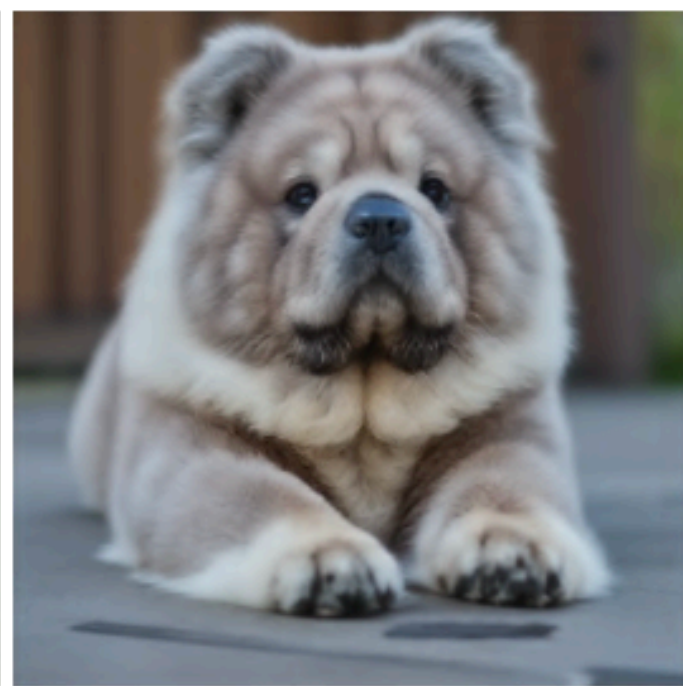
Input



bear



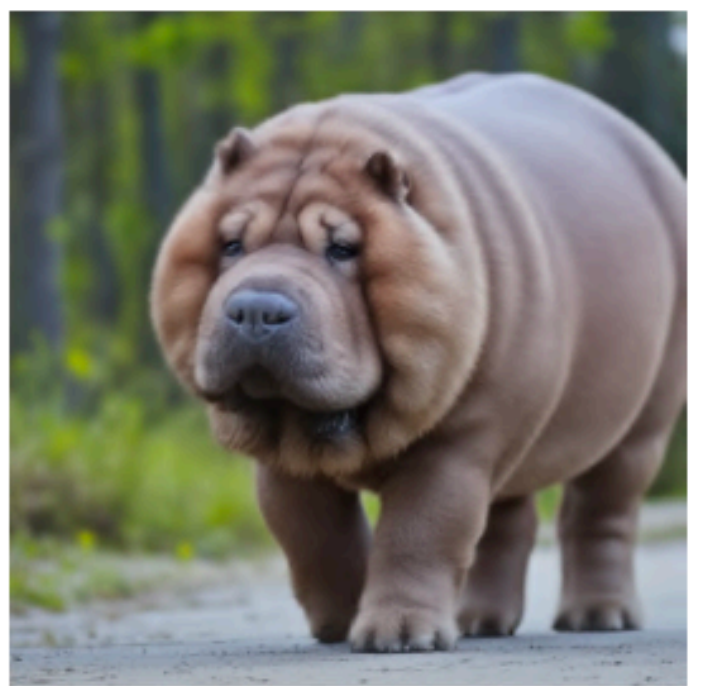
panda



koala



lion



hippo

Ablation Studies On Class-Prior Ablation

Input images



Fine-tuning

No class noun:
"A [V]"

Incorrect class noun:
"A [V] dog"

Correct class noun:
"A [V] sunglasses"

Inference



Comparisons

[20] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Input images



Gal et al.

Ours

An oil painting of a [V] sculpture

App icon of a [V] sculpture

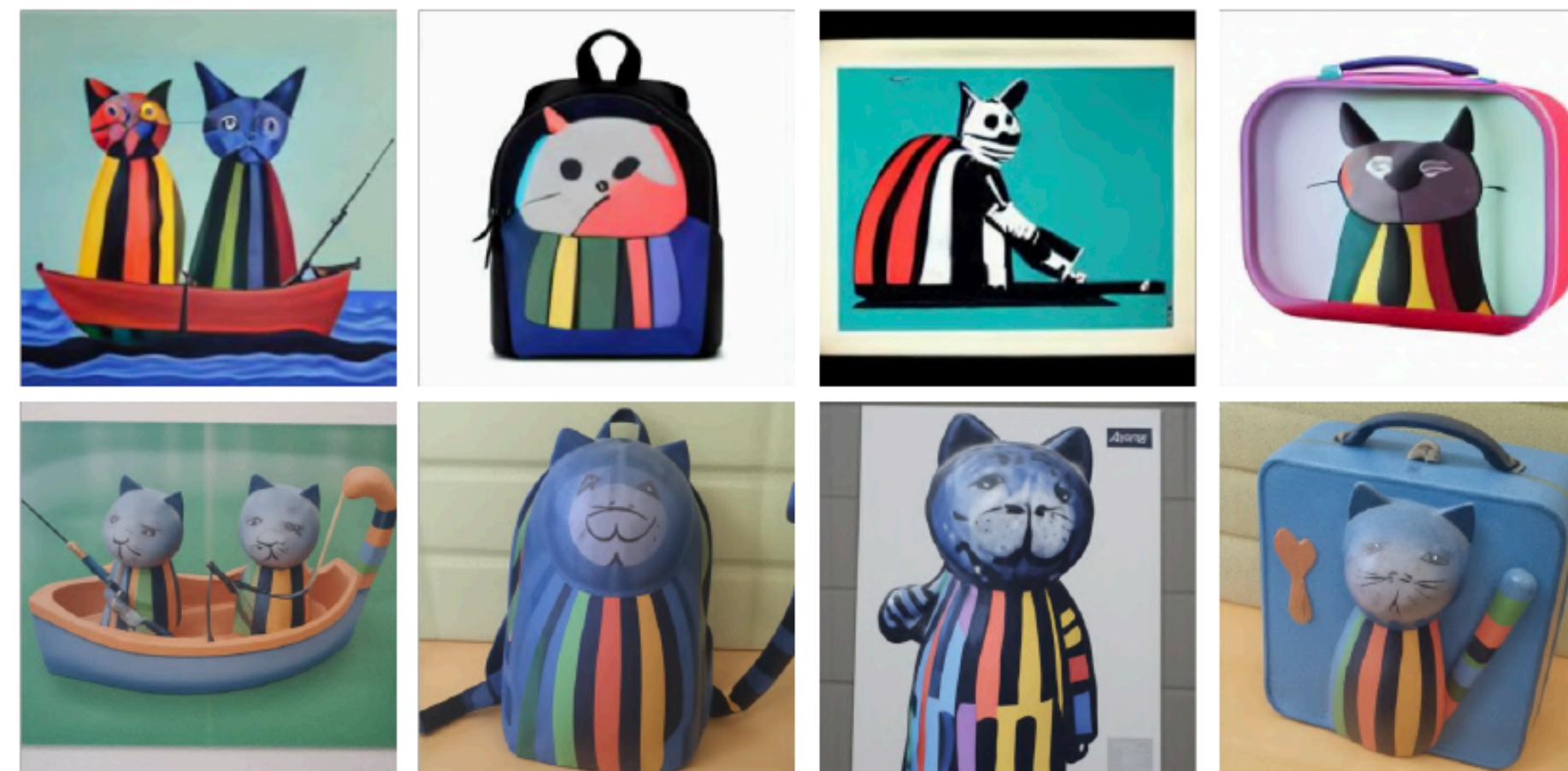
Elmo sitting in the same pose as a [V] sculpture

A crochet [V] sculpture

Ink wash painting of a [V] sculpture

A black and white sketch of a [V] sculpture

Input images



Gal et al.

Ours

Painting of two [V] sculptures fishing on a boat

A [V] sculpture backpack

Banksy art of a [V] sculpture

A [V] sculpture-themed lunchbox

Comparisons

Input images



Detailed prompt, Imagen

“retro style yellow alarm clock with a white clock face and a yellow number three on the right part of the clock face”



Detailed prompt, DALLE-2



Ours

[...] on a beach

[...] with a cave in the background

[...] on top of blue fabric

[...] held by a hand, with a forest in the background

Limitation

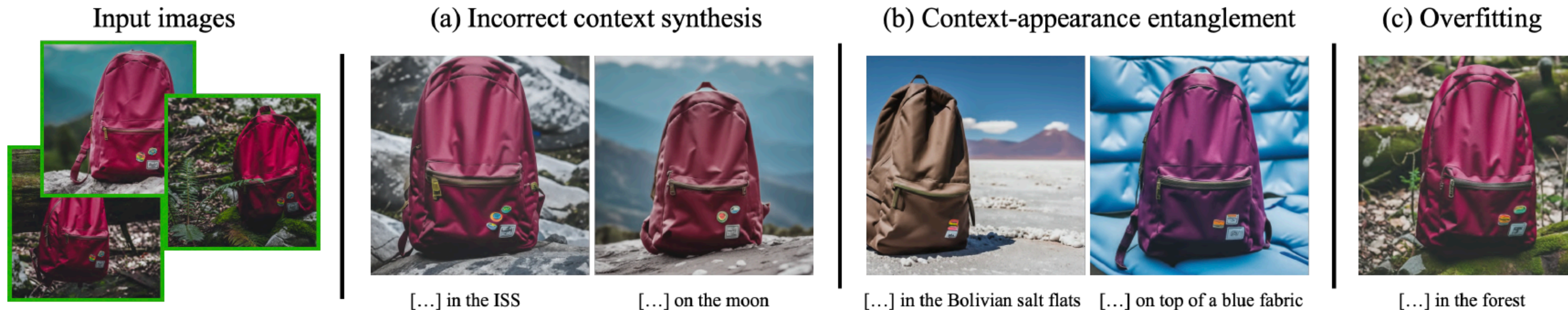


Figure 17: **Failure modes.** Given a rare prompted context the model might fail at generating the correct environment (a). It is possible for context and subject appearance to become entangled, with colors describing the context melding or changing the subject, or the model reverting to its prior with certain rare contexts (b). In this case, generating a brown bag in the rare context of the Bolivian salt flats. Finally, it is possible for the model to overfit and generate images similar to the training set, especially if prompts reflect the original environment of the training set (c). Image credit (input images): Unsplash.

Conclusion

- A new problem: subject-driven generation.
- A new technique for fine-tuning text-to-image diffusion models in a few-shot setting, while preserving the model's semantic knowledge on the class of the subject.