Paper Reading 230305
Wenjing Wang

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

arXiv 2022.08

Nataniel Ruiz[1,2], Yuanzhen Li[1], Varun Jampani[1], Yael Pritch[1], Michael Rubinstein[1], and Kfir Aberman[1]

[1]*Google Research* [2]*Boston University*

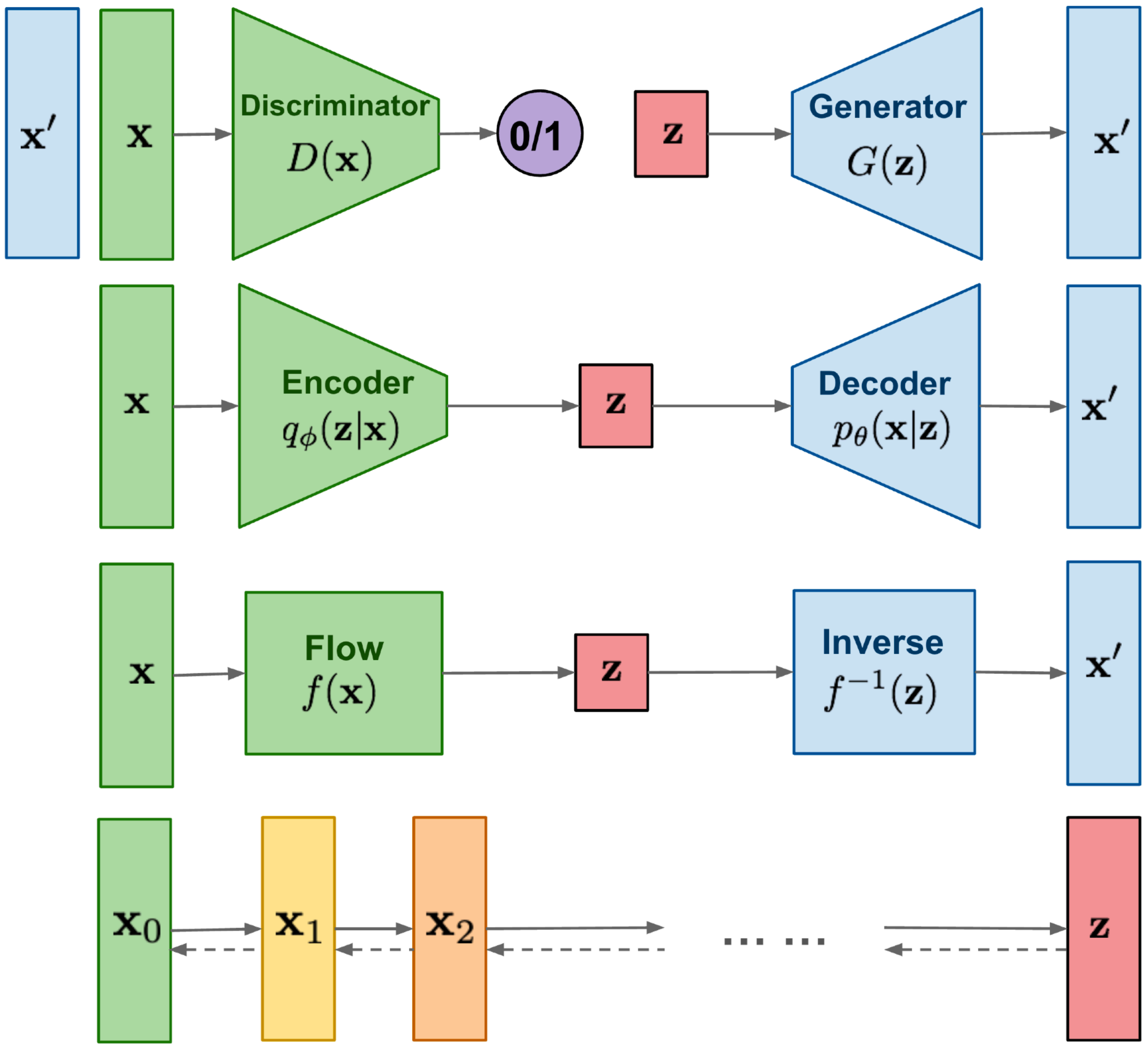
Diffusion

GAN: Adversarial training

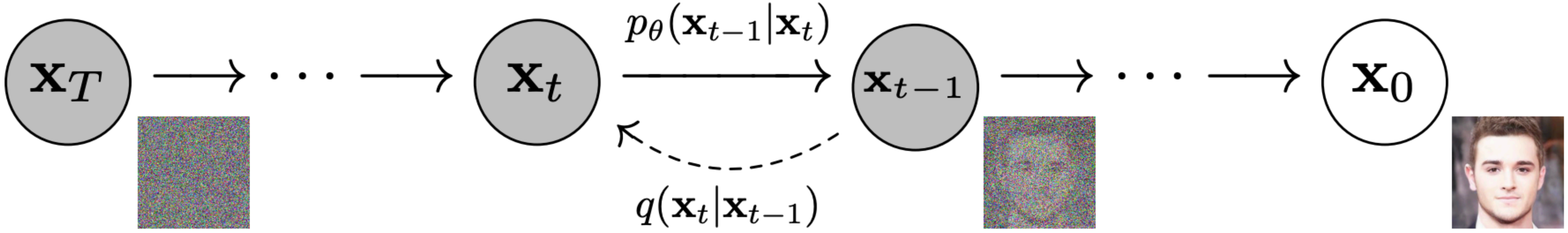
VAE: maximize variational lower bound

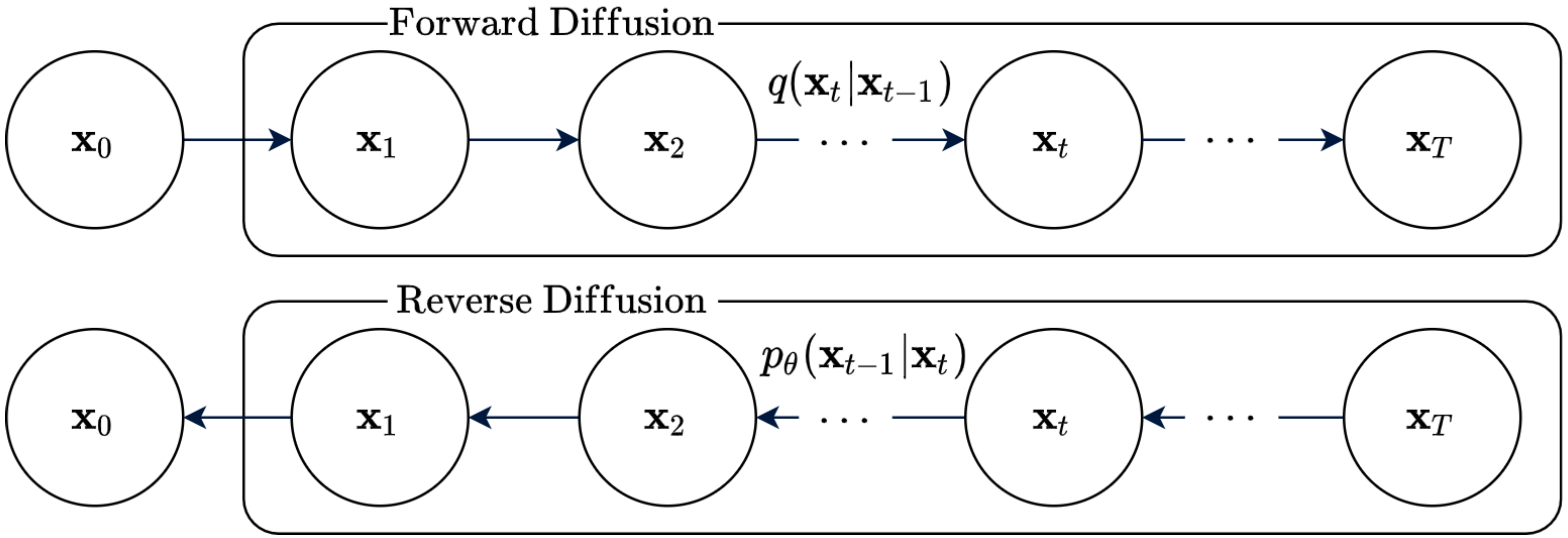
Flow-based models: Invertible transform of distributions

Diffusion models: Gradually add Gaussian noise and then reverse

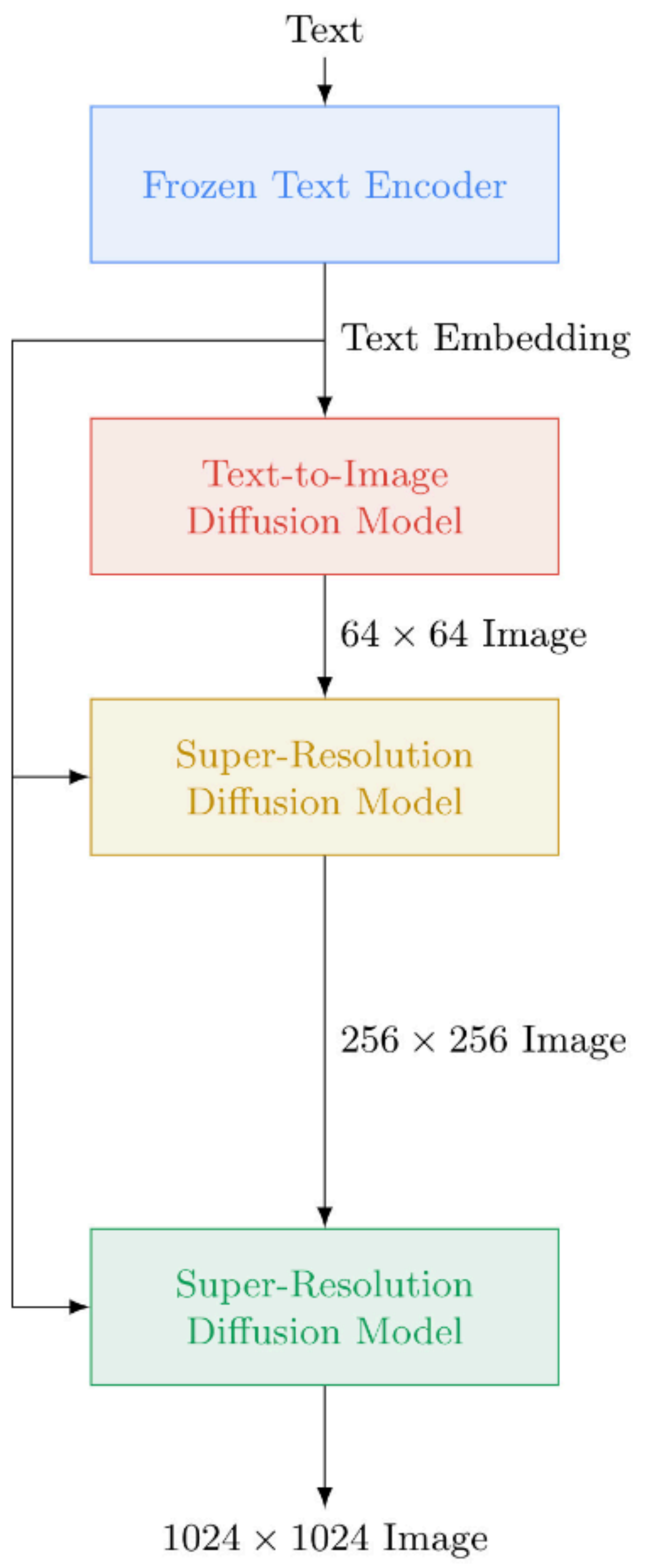


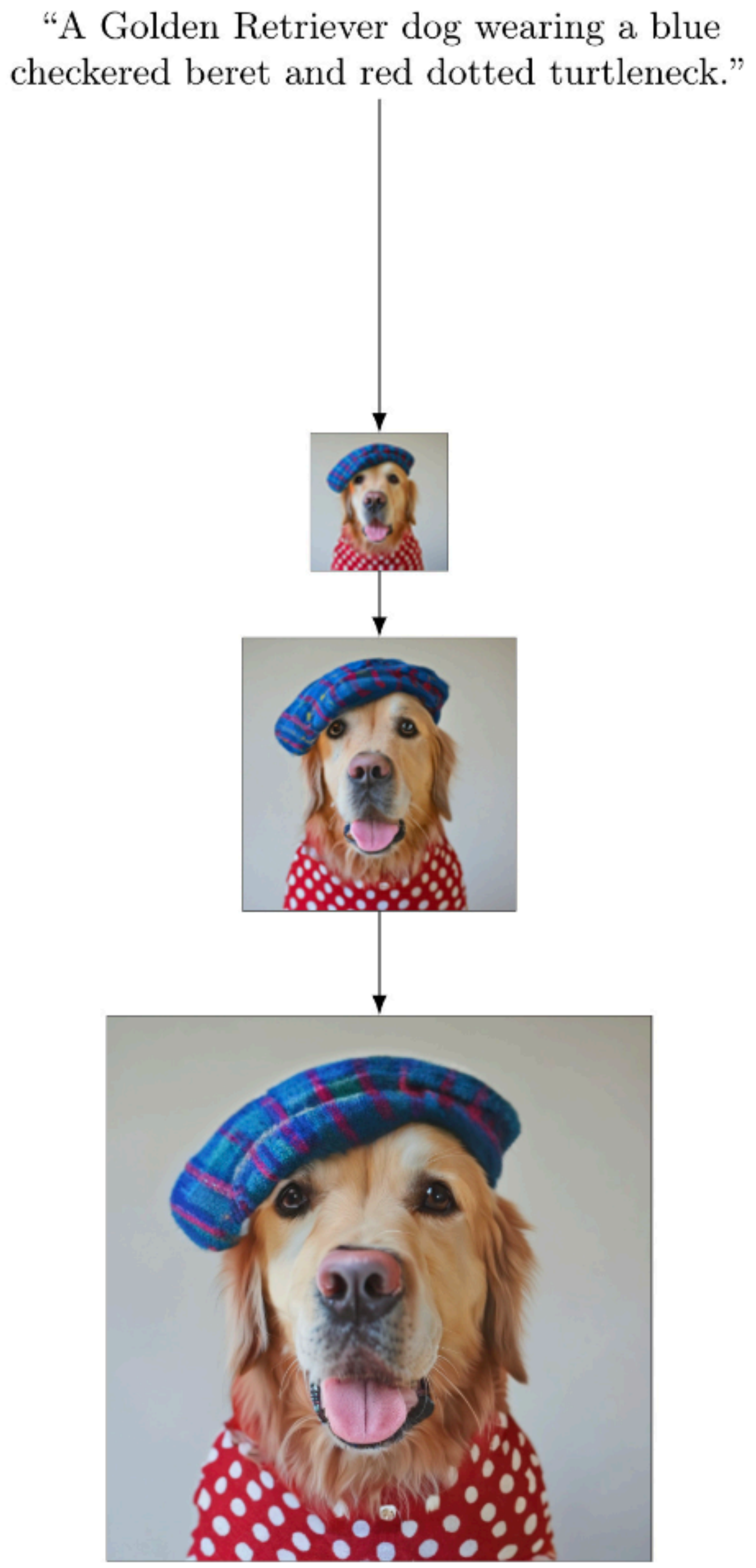
Diffusion


$\mathbf{x}_T \longrightarrow \cdots \longrightarrow \mathbf{x}_t \xrightarrow{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \mathbf{x}_{t-1} \longrightarrow \cdots \longrightarrow \mathbf{x}_0$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$


Forward Diffusion
$\mathbf{x}_0 \rightarrow \mathbf{x}_1 \rightarrow \mathbf{x}_2 \rightarrow \cdots \rightarrow \mathbf{x}_t \rightarrow \cdots \rightarrow \mathbf{x}_T$
$q(\mathbf{x}_t|\mathbf{x}_{t-1})$
Reverse Diffusion
$\mathbf{x}_0 \leftarrow \mathbf{x}_1 \leftarrow \mathbf{x}_2 \leftarrow \cdots \leftarrow \mathbf{x}_t \leftarrow \cdots \leftarrow \mathbf{x}_T$
$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$

Imagen





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 Butter-fly event.

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

Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding (NeurIPS-22)

Task: “personalize” text-to-image diffusion models
Subject-driven generation

“Personalize” Related Works

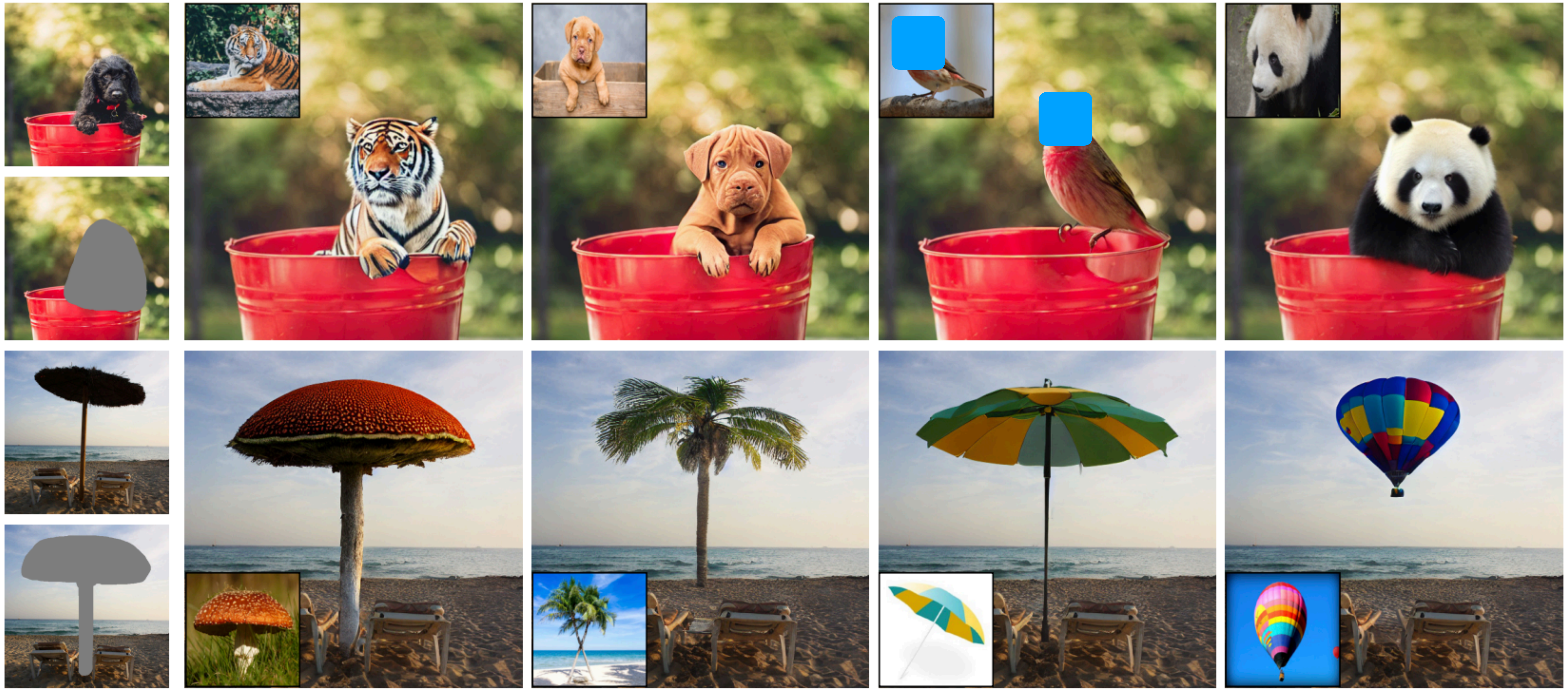


Figure 1. Paint by example. Users are able to edit a scene by painting with a conditional image. Our approach can automatically alter the reference image and merge it into the source image, and achieve a high-quality result.

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

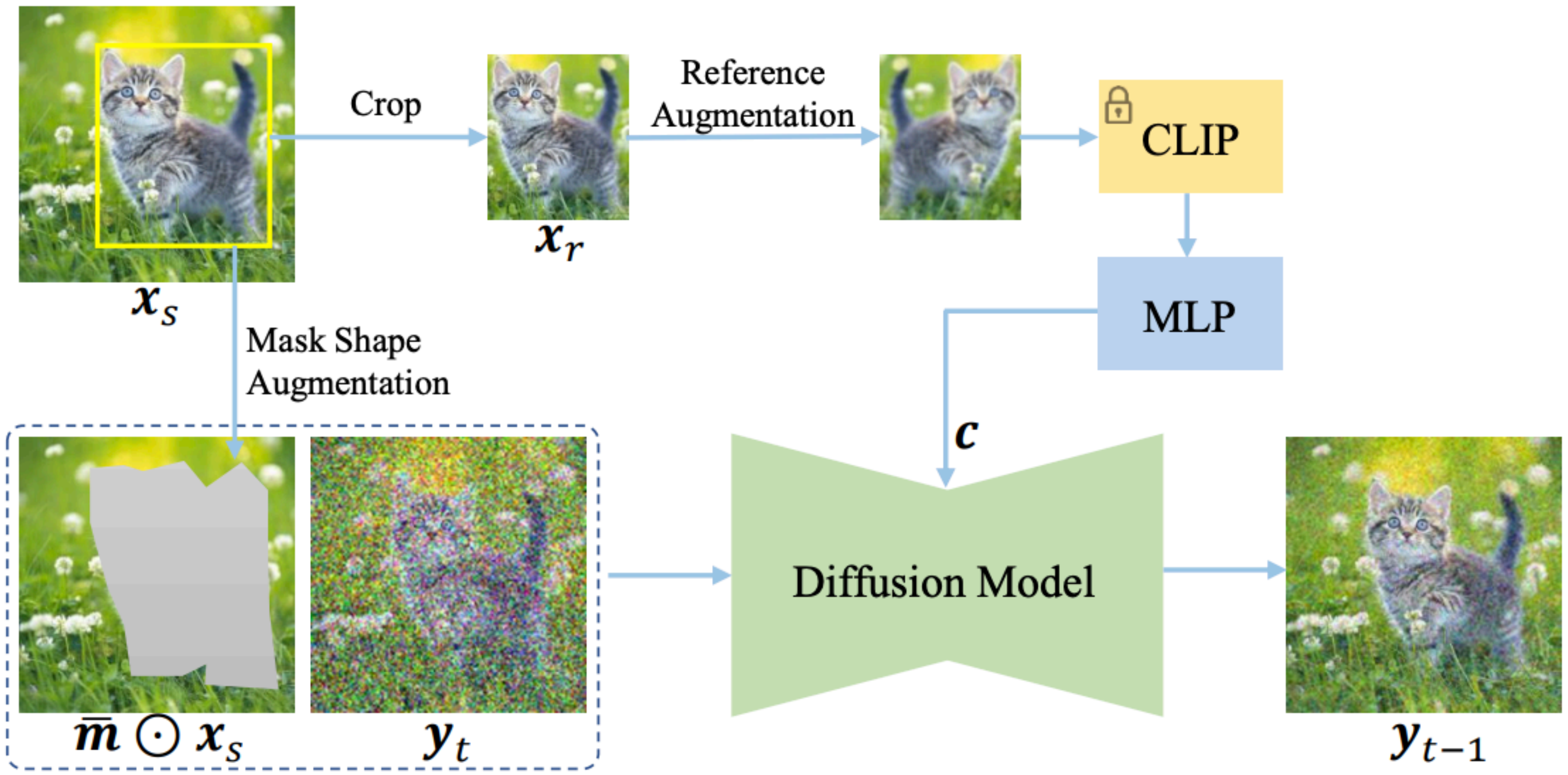


Figure 4. Our training pipeline.

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

“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

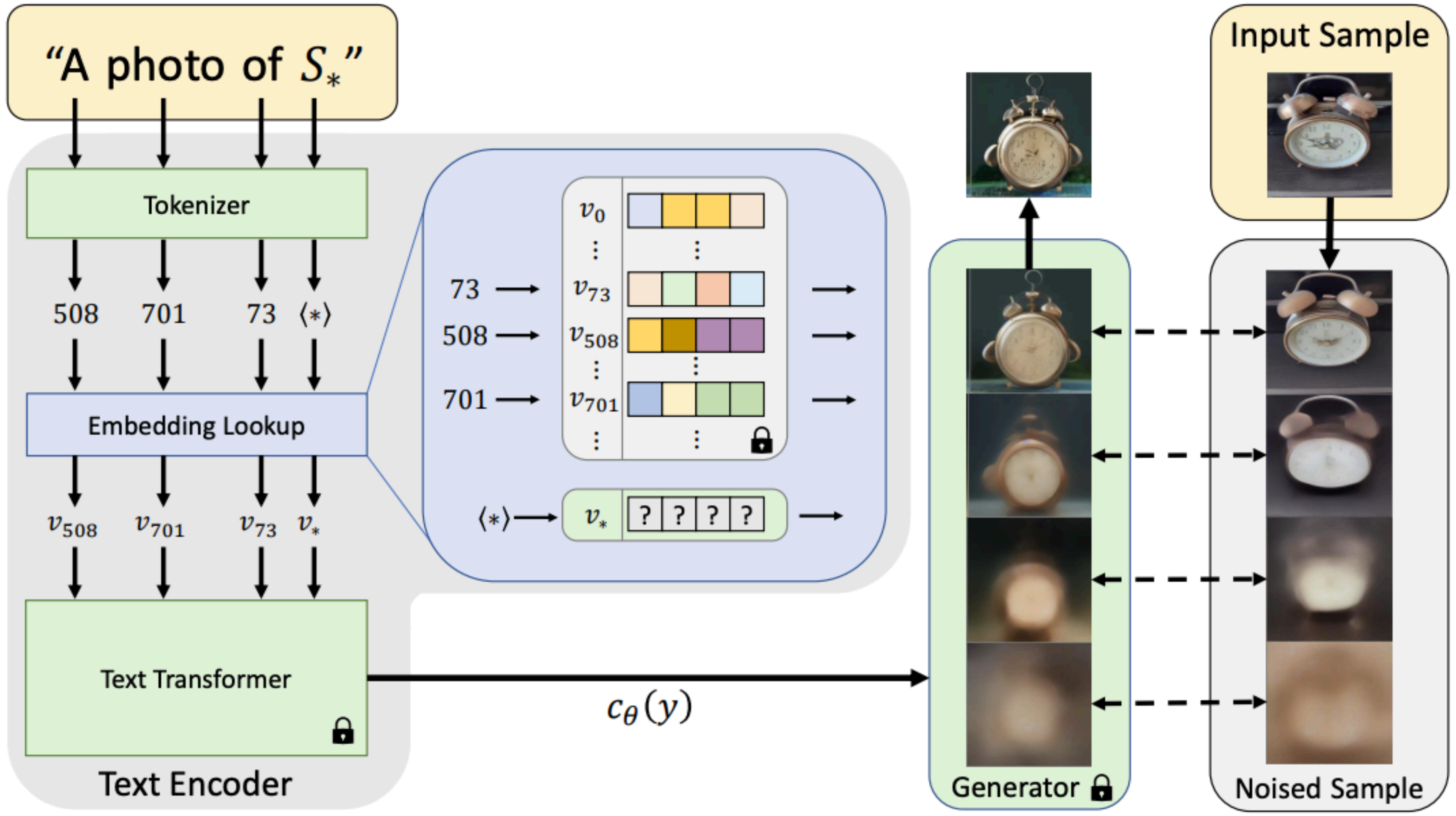


Figure 2: Outline of the text-embedding and inversion process. A string containing our placeholder word is first converted into tokens (*i.e.* word or sub-word indices in a dictionary). These tokens are converted to continous vector representations (the “embeddings”, v). Finally, the embedding vectors 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.

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

Task

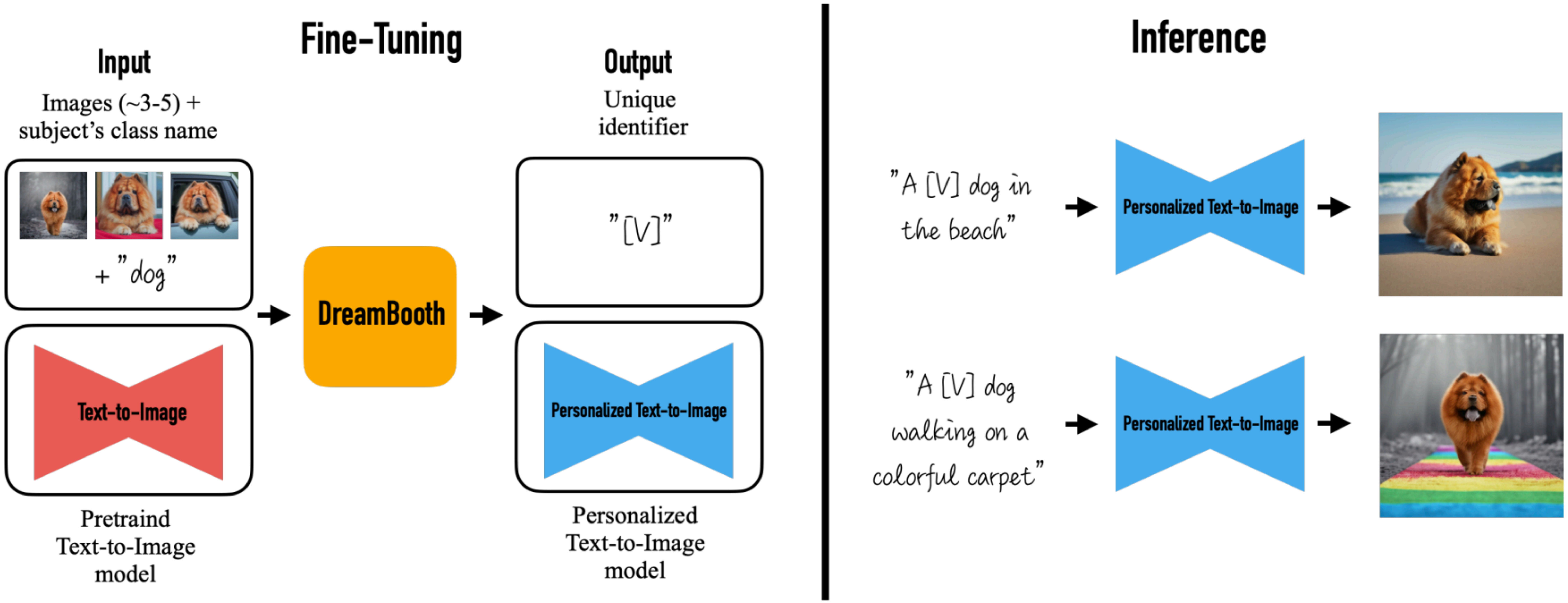


Figure 3: **High-level method overview.** Our method takes as input a few images (typically $3-5$ images suffice, based on our experiments) of a subject (e.g., a specific dog) and the corresponding class name (e.g. "dog"), and returns a fine-tuned/"personalized" text-to-image model that encodes a unique 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

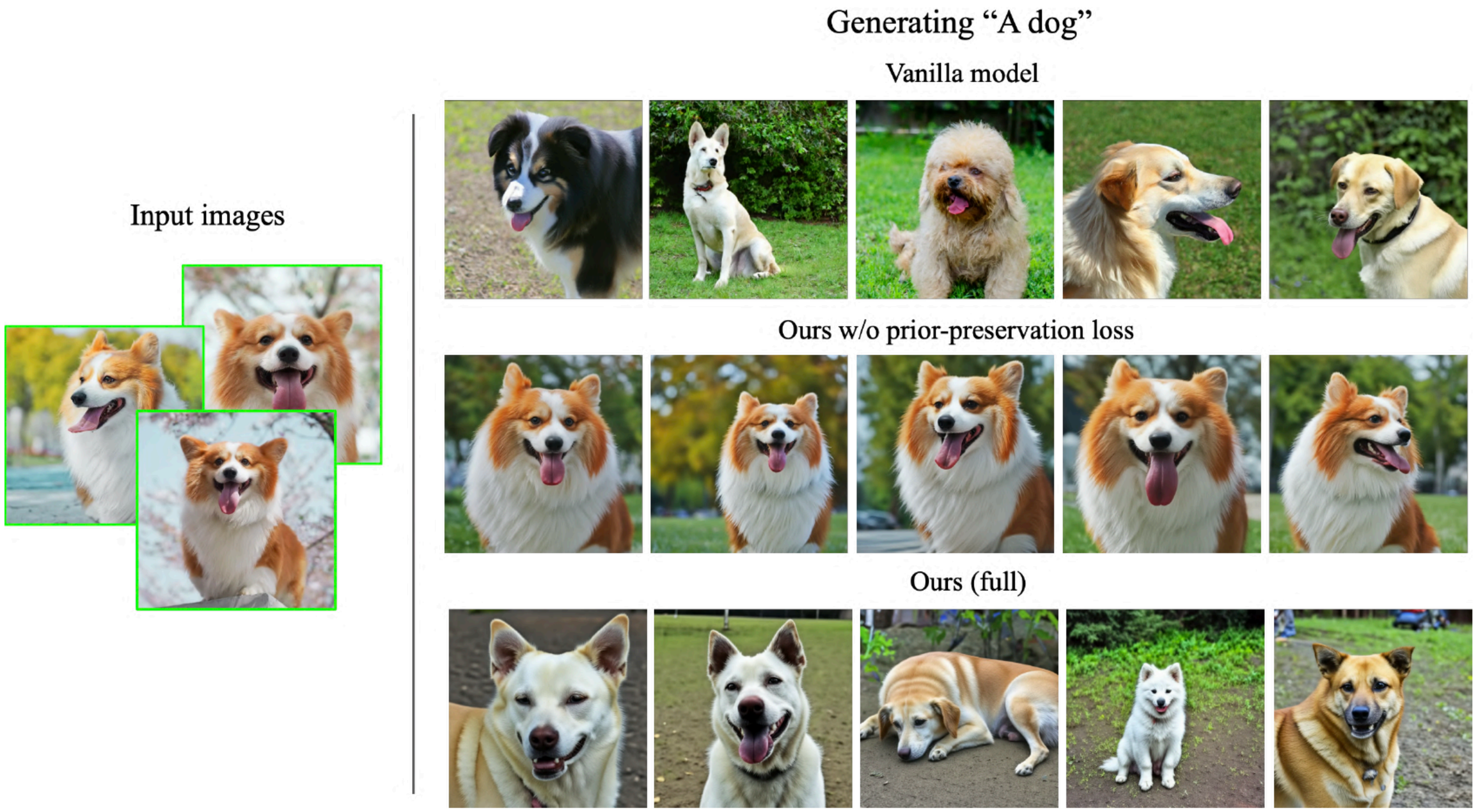
- Overfit to both the context and the appearance of the subject

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 drif: 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 $\mathbf{x}_{\text{pr}} = \hat{\mathbf{x}}(\mathbf{z}_{t_1}, \mathbf{c}_{\text{pr}})$ $\mathbf{z}_{t_1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$\mathbf{c}_{\text{pr}} := \Gamma(f(\text{"a [class noun]"}))$

- New loss function:

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

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

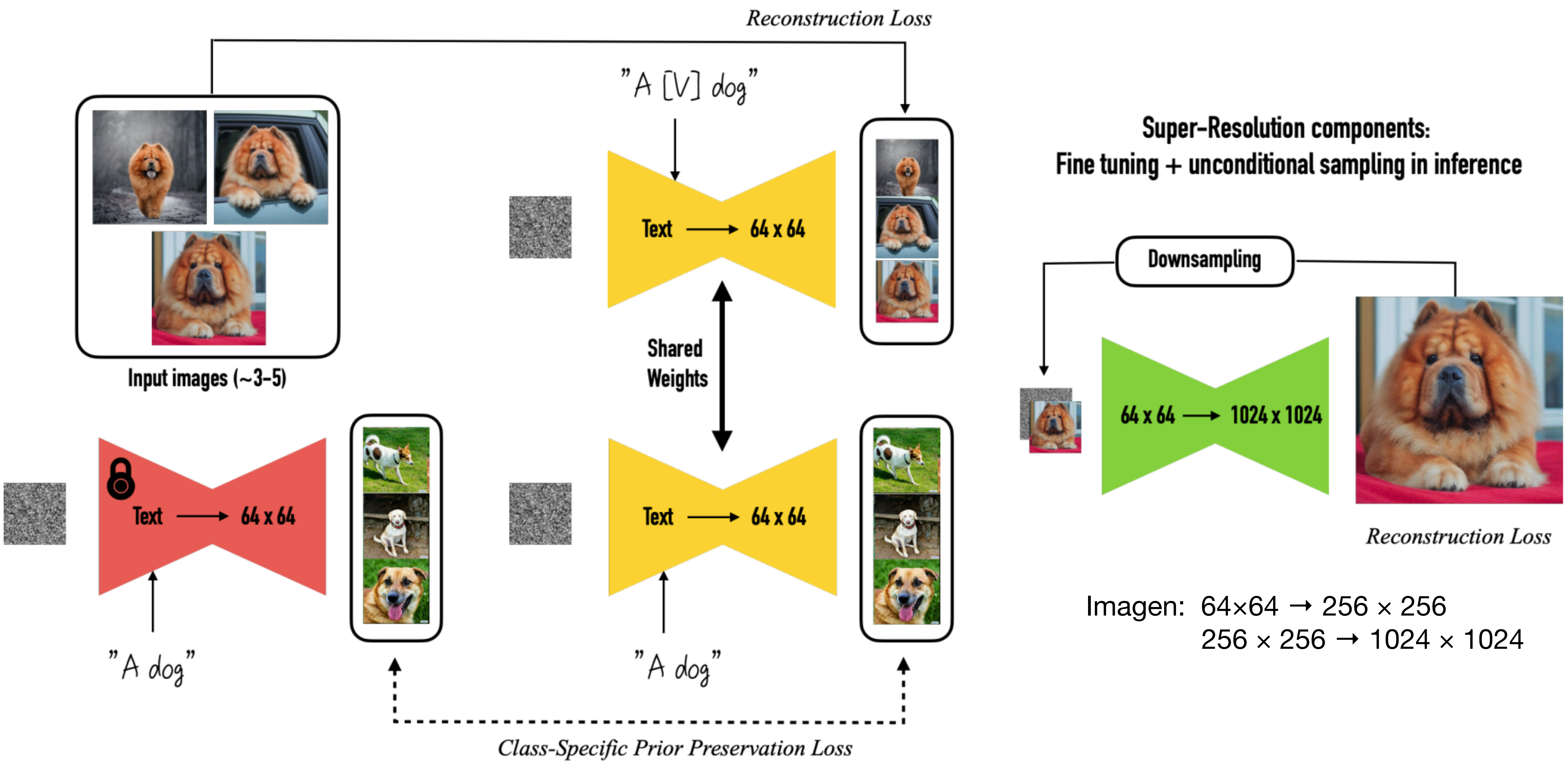
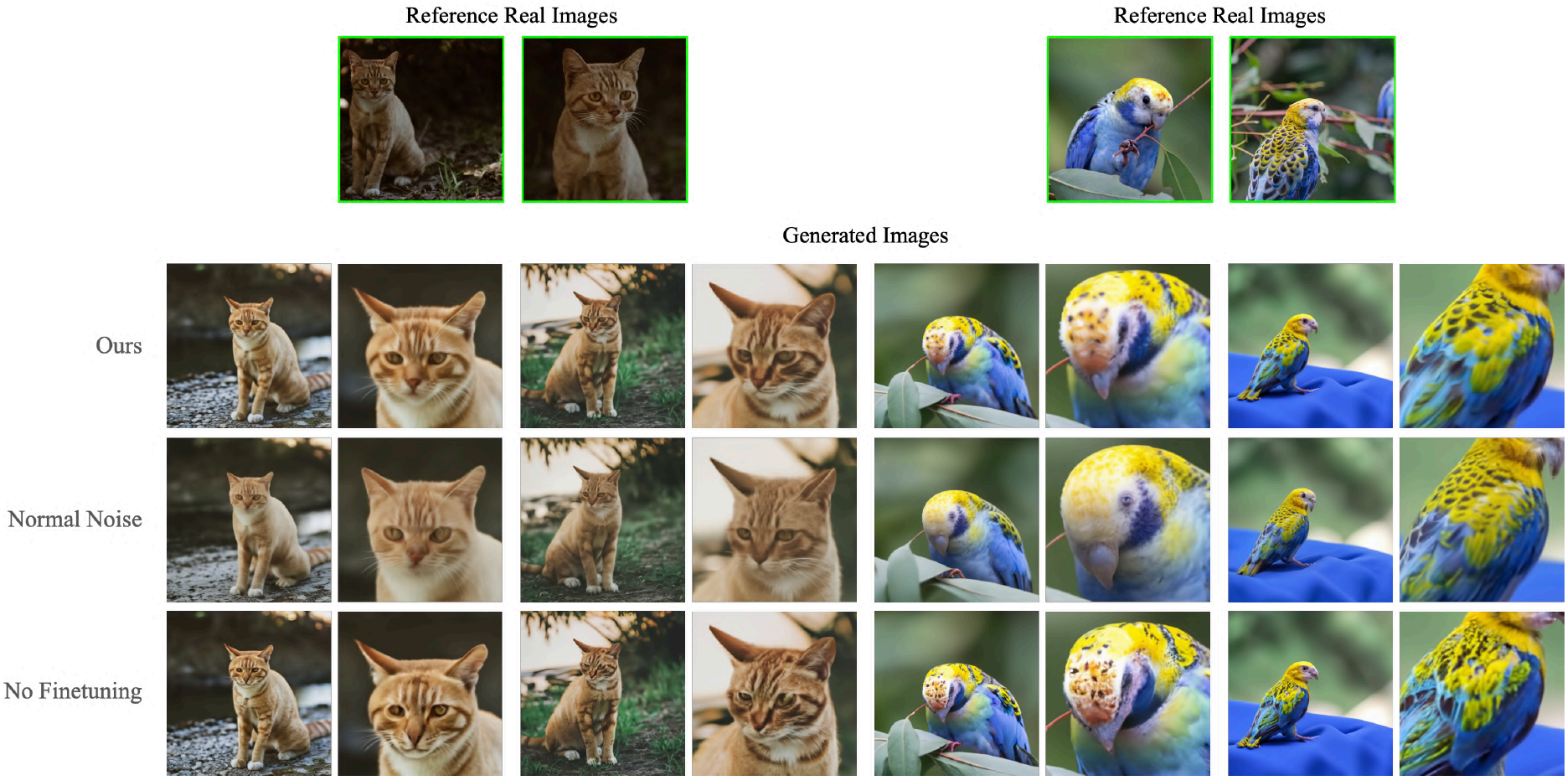
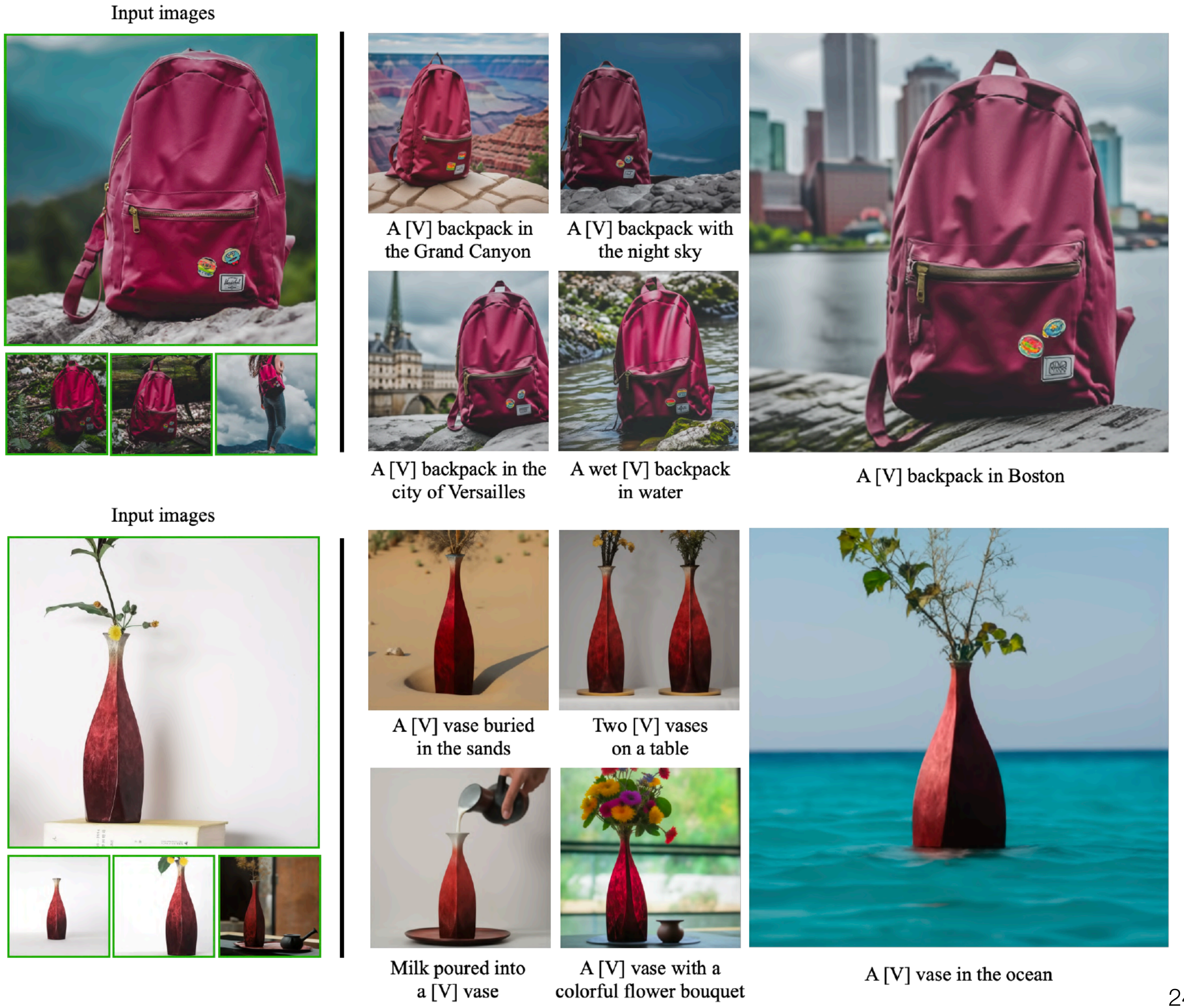


Imagen: 64×64 → 256 × 256
256 × 256 → 1024 × 1024

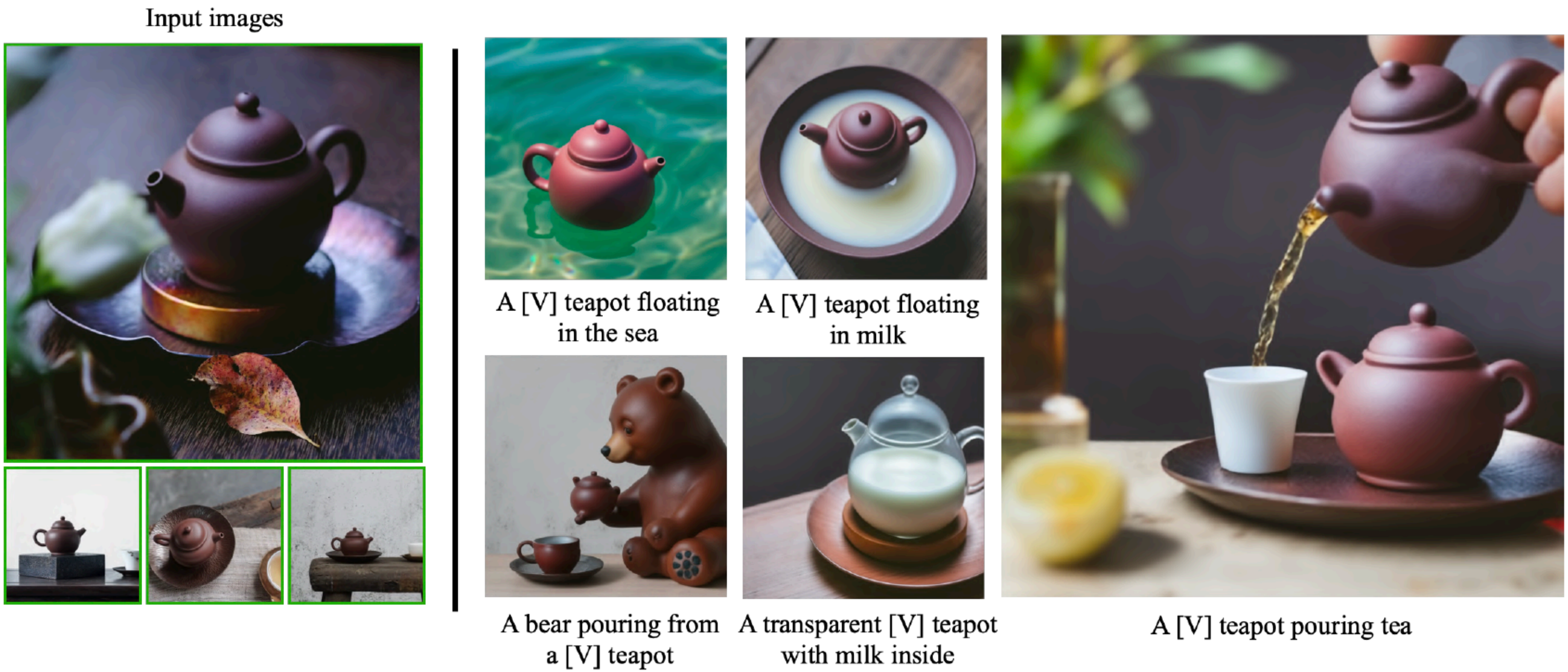
Reduce the level of noise augmentation from 10-3 to 10-5 during fine-tuning of the 256×256 SR model.



Experimental Results On Recontextualization



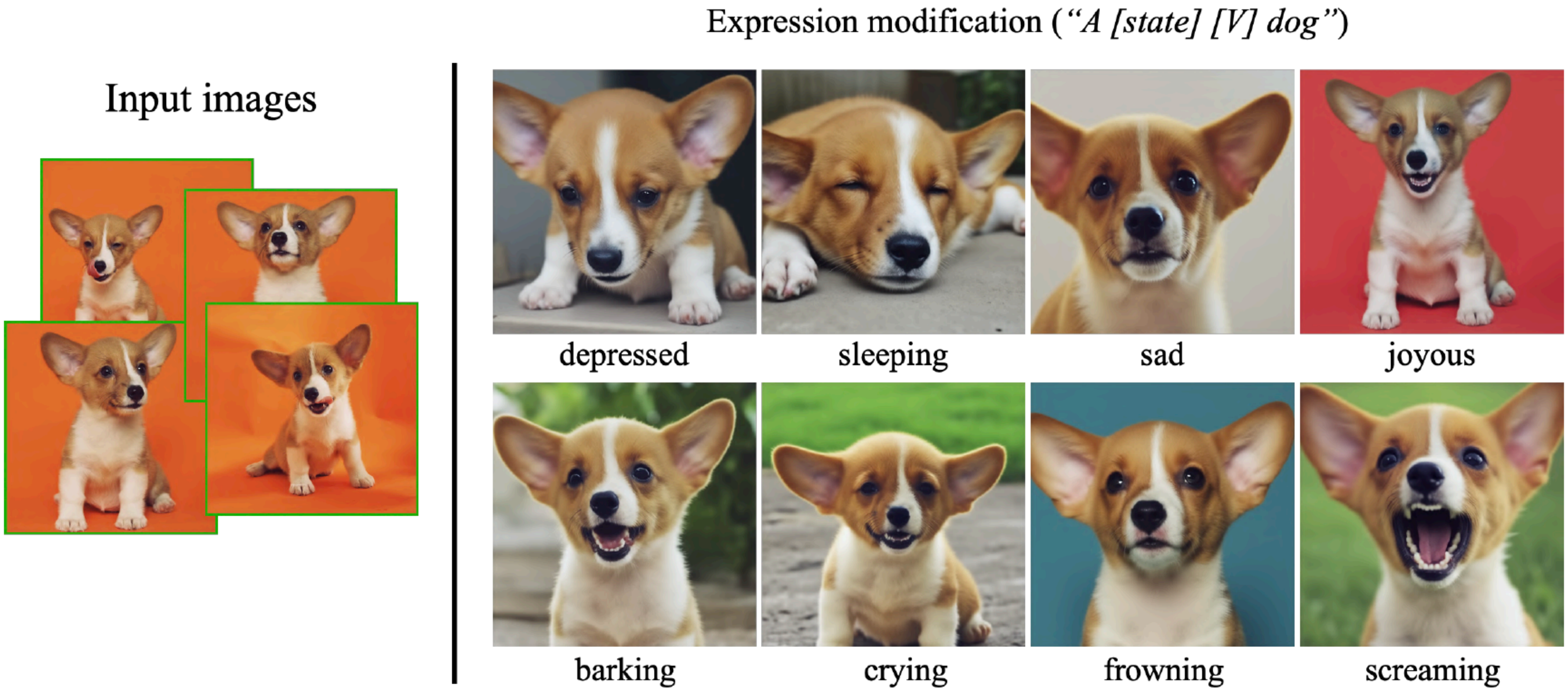
Experimental Results On Recontextualization



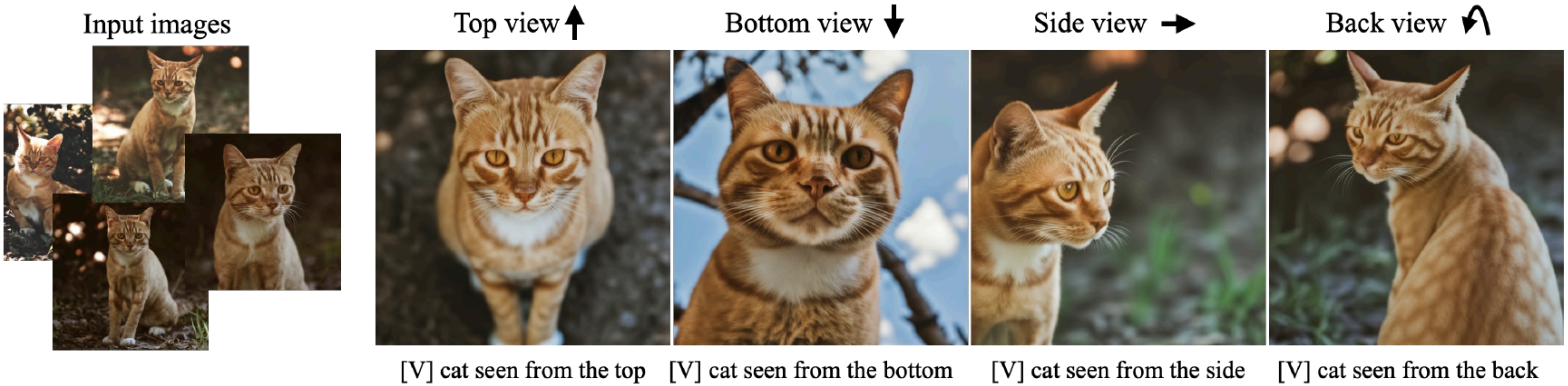
Experimental Results
On Art Renditions

“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



Experimental Results
On Novel View Synthesis

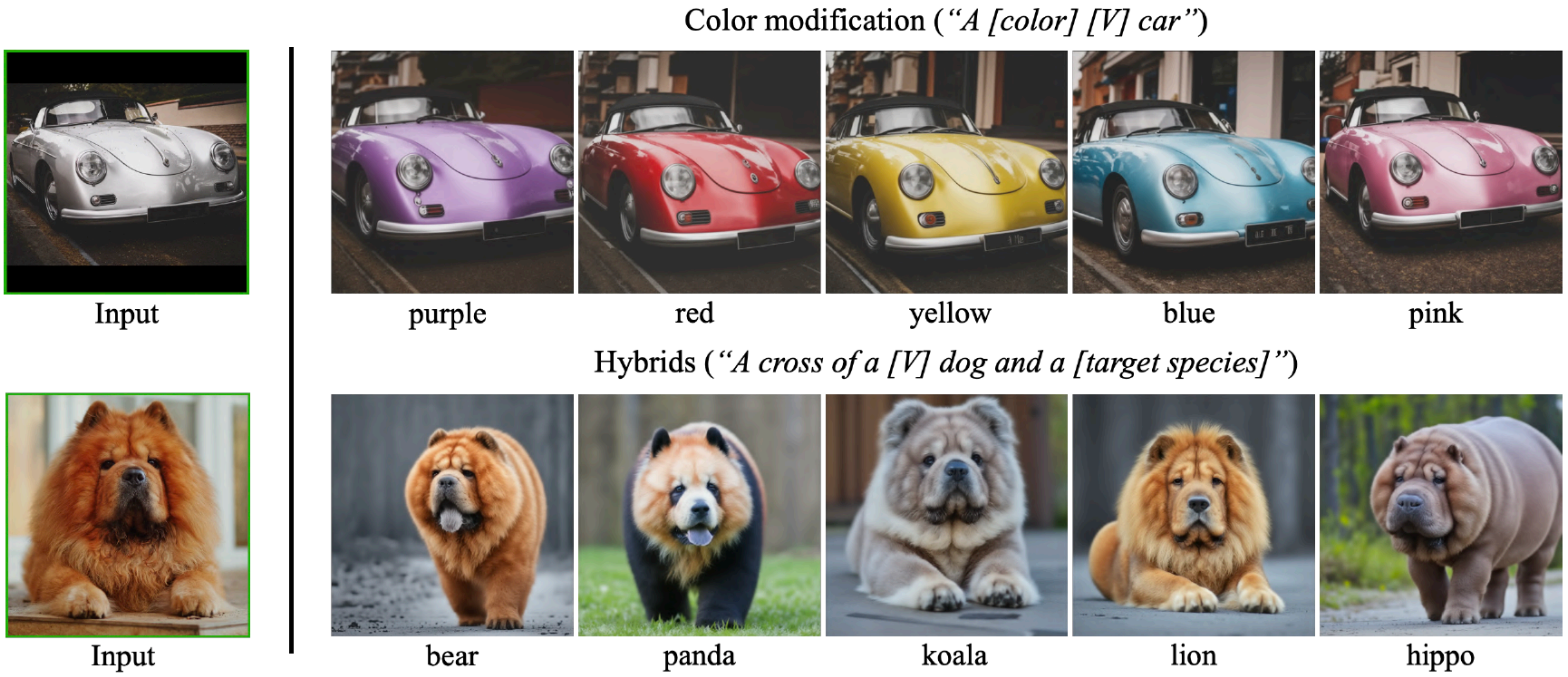


Experimental Results On Accessorization

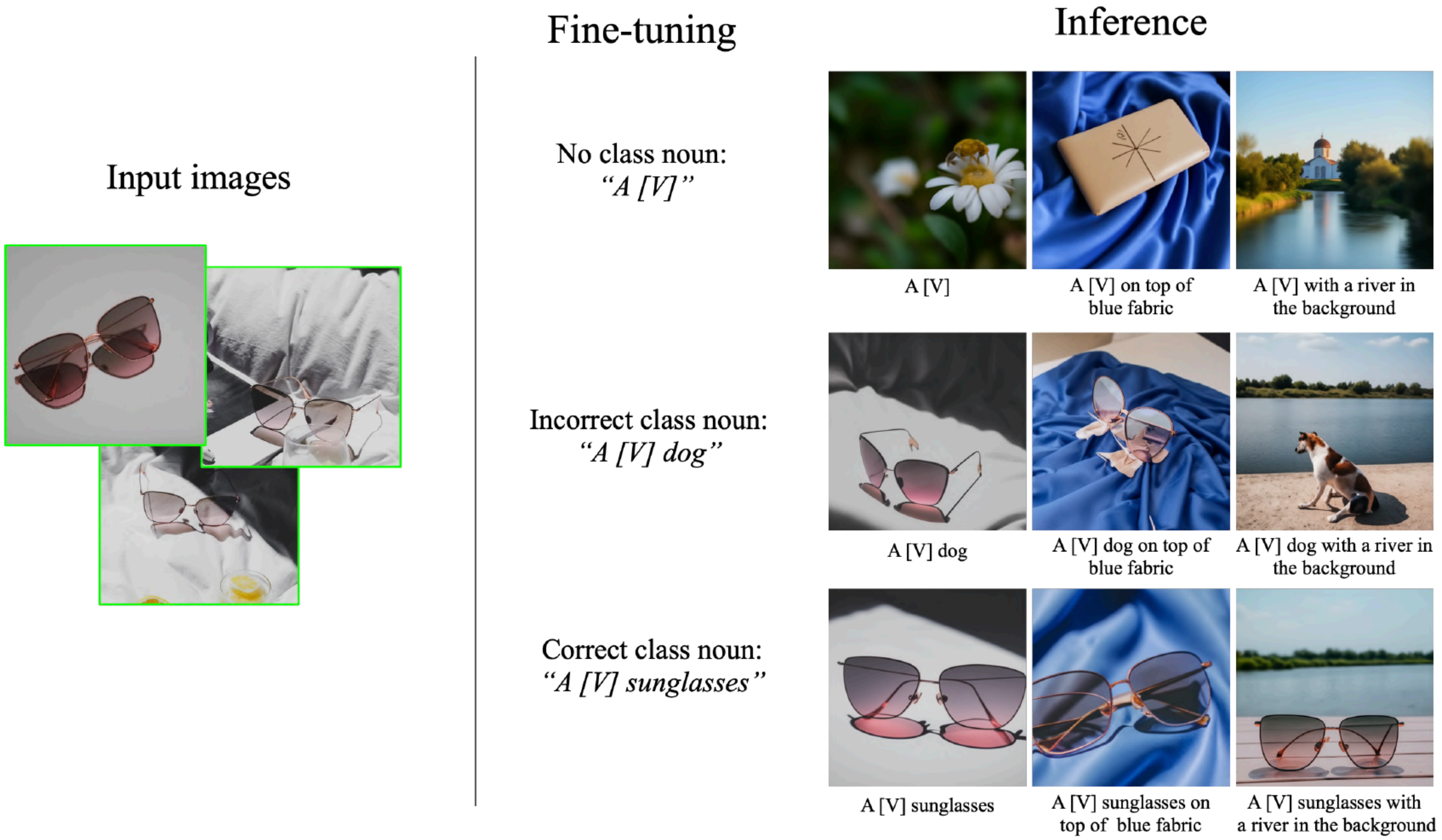
“a [V] [class noun] wearing [accessory]”

Input images

Experimental Results
On Property Modification

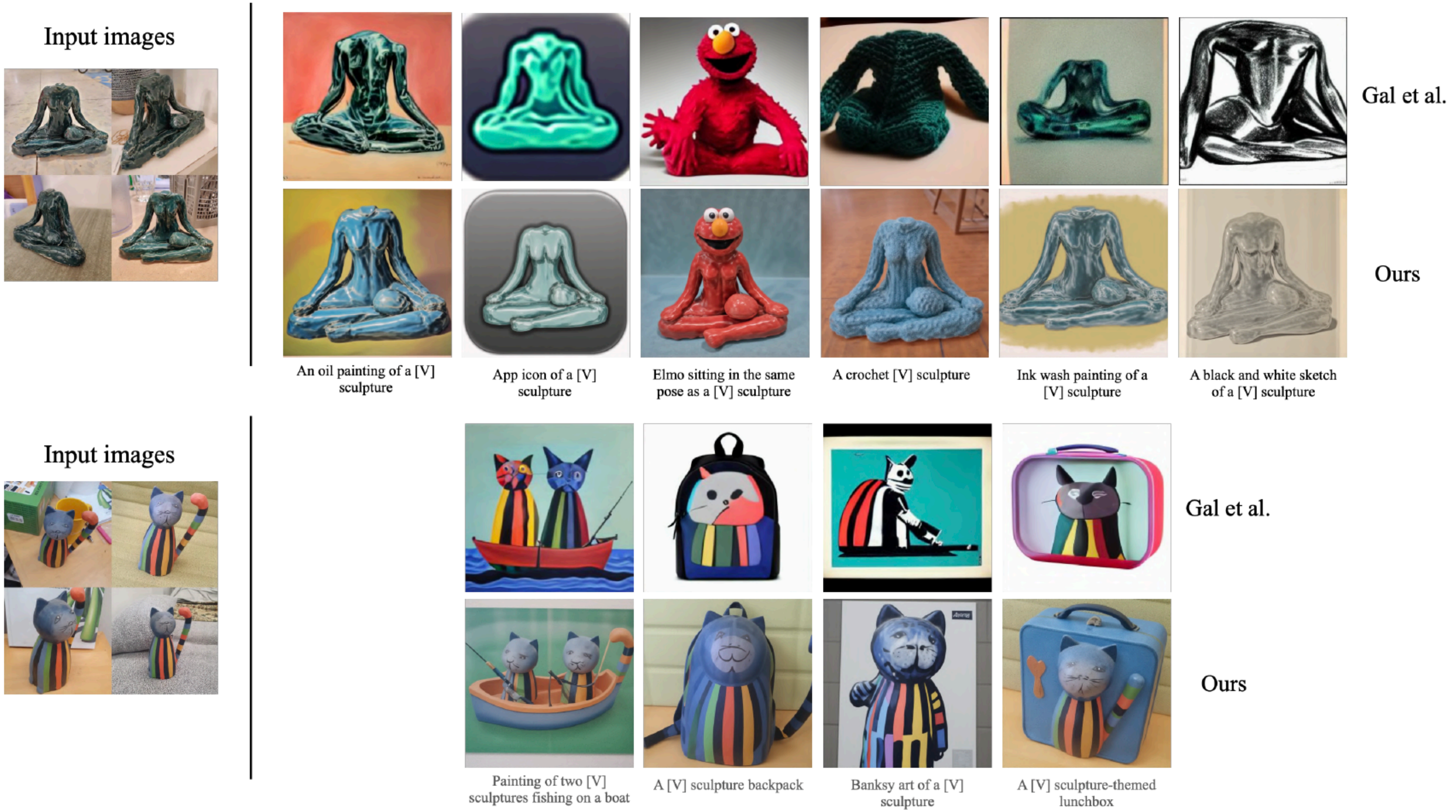


Ablation Studies
On Class-Prior Ablation



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.



Comparisons

Input images

[...] on a beach

[...] with a cave in the background

[...] on top of blue fabric

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

Detailed prompt, Imagen

Detailed prompt, DALLE-2

Ours

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

Limitation

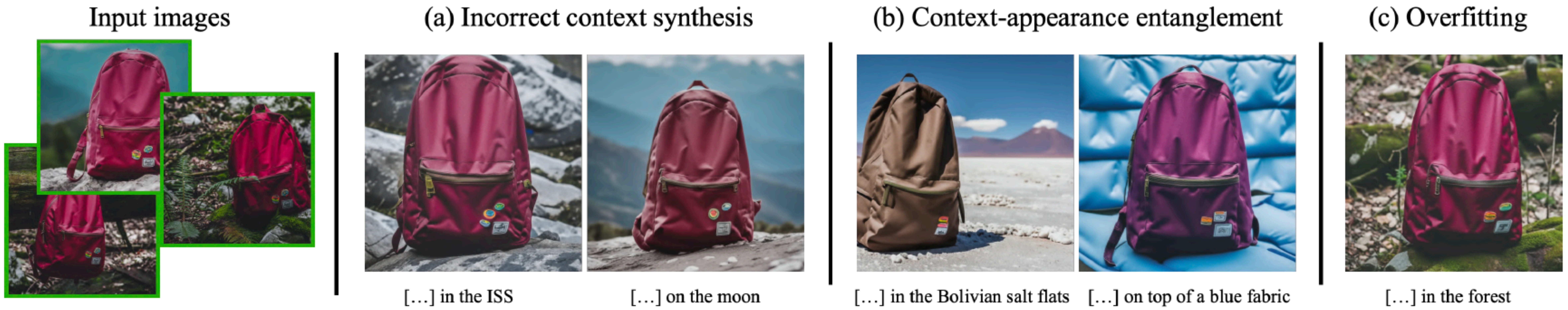


Figure 17: **Failure modes.** Given a rare prompted context the model might fail at generating the correct environment (a). It is possible for context and subject appearance to become entangled, with colors describing the context melding or changing the subject, or the model reverting to its prior with 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.