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

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, Kfir Aberman

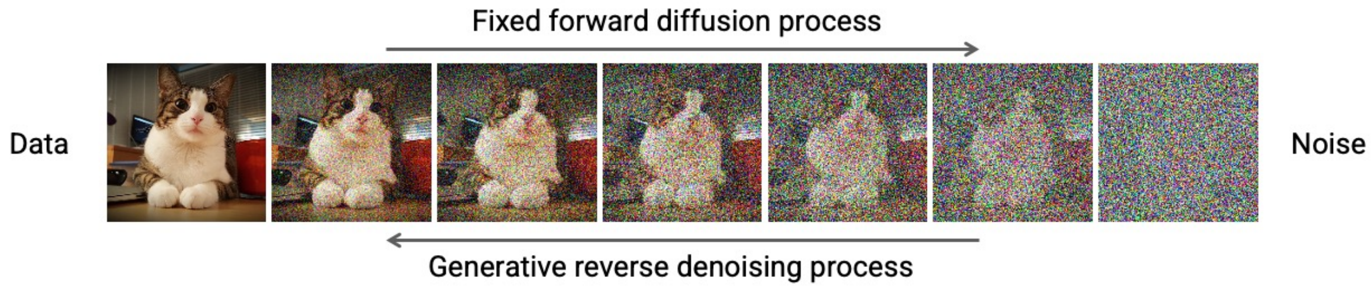
Google Research – Boston University -- CVPR 2022
Presented by: Ehsan Latif and Chetan Dhamane

Outline

- Introduction
- Background
- Main Contribution
- Method
- Experiments
- Results
- Limitations
- Conclusion

Diffusion Models

- Diffusion models are probabilistic generative models that are trained to learn a data distribution by the gradual denoising of a variable sampled for gaussian distribution.



Introduction

- Paper aim to solve the problem of generating realistic images from textual descriptions.
- The challenge of this problem lies in generating images that are both visually appealing and semantically consistent with the given textual description.
- Previous techniques for text-to-image generation have faced limitations such as mode collapse, limited diversity, and semantic consistency.
- The authors believe that there is a need for a more advanced and effective technique for text-to-image generation, which is why they propose the DreamBooth framework.
- The authors aim to answer the research question: "Given a textual description of an object and its attributes, how can we generate an image of the described object?"
- This problem is important to solve as it has various applications, including creative tools, video games, and robotics.

Overview

- DreamBooth—an AI-powered photo booth—can generate a myriad of images of the subject in different contexts (right), using the guidance of a text prompt. The results exhibit natural interactions with the environment, as well as novel articulations and variation in lighting conditions, all while maintaining high fidelity to the key visual features of the subject.

Background

- Recent text-to-image generation, such as GAN-based approaches, conditional GANs, and attention-based models.
- Limitations of these techniques, including:
  - **Mode collapse** (where the model generates similar images repeatedly),
  - **Limited diversity** (where the generated images are too similar to each other),
  - **Poor semantic consistency** between the generated images and textual descriptions (where the images generated do not accurately reflect the content described in the text).
- The subject-driven diffusion model is designed to generate a diverse set of images from textual descriptions which improves image quality.
- The fine-tuning module is designed to fine-tune the subject-driven diffusion model for a specific subject (the object or scene described in the text) which generate images that are semantically consistent.


Fidelity
New contexts
Fidelity
New contexts
Fidelity
New contexts
Imagen
Imagen
Imagen
Input Images
Image-guided, DALL-E2
Text-guided, Imagen
Ours

Comparison

Main Contribution

- Proposed DreamBooth framework, a novel text-to-image generation technique.
- Aim to address the limitations of previous techniques, such as mode collapse, limited diversity, and poor semantic consistency, by proposing the DreamBooth framework.
- Highlight the advantages of the DreamBooth framework, including its ability to generate high-quality images with high semantic consistency and improved diversity compared to previous techniques.
- Emphasize the efficiency and effectiveness of the DreamBooth framework, as it generates images faster and with higher quality compared to previous techniques.
- Extensive evaluation of the DreamBooth framework, including both qualitative and quantitative evaluations.

Preliminaries

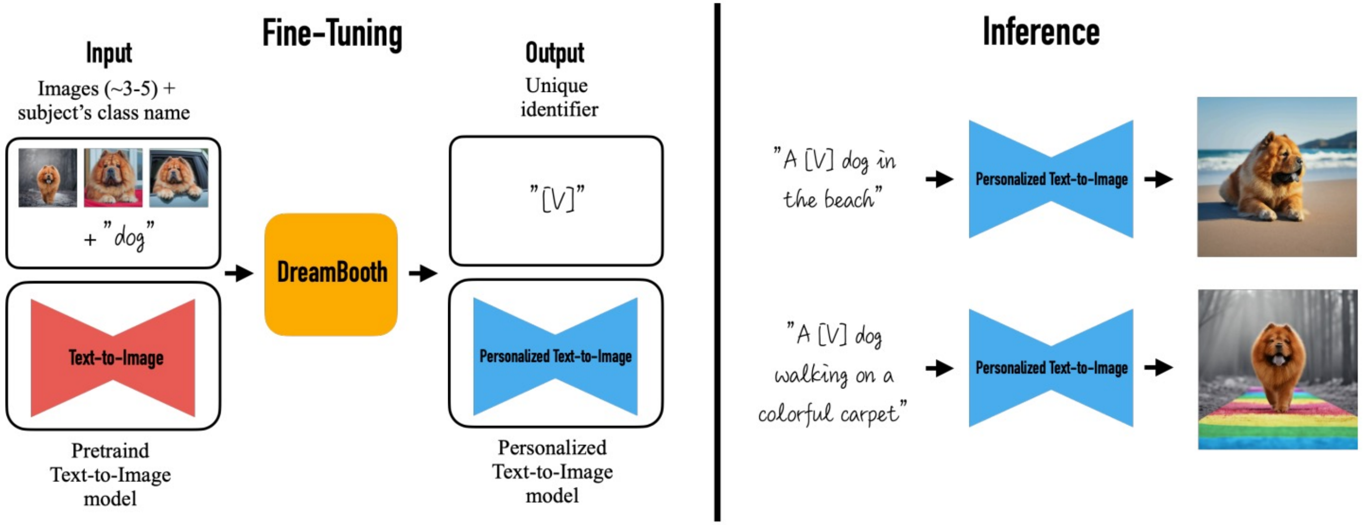
- Cascaded Text-to-Image Diffusion Models:
  - Learning the reverse process of fixed-length Markovian forward process.
  - $\hat{\mathbf{x}}_\theta$ A conditional diffusion model
  - $\mathbf{z}_t := \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$ Noised Image: $\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},t}\left[w_t \|\hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x}\|_2^2\right]$
  - x: Ground truth image, c: conditional factor, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$: A noise term,
  - α_t, σ_t, w_t : control the noise scheduled and sample quality for super resolution

Preliminaries (Cont'd)

- Vocabulary Encoding:
  - Use of latter model.
  - Text prompt P, conditioning embedding c, tokenizer f, language model Γ
  - To produce embedding: $\mathbf{c} := \Gamma(f(\mathbf{P}))$
  - The language model Γ is conditional on the token identifier vector to produce embedding c.
  - The text to image diffusion model is directly conditioned to c.

Method

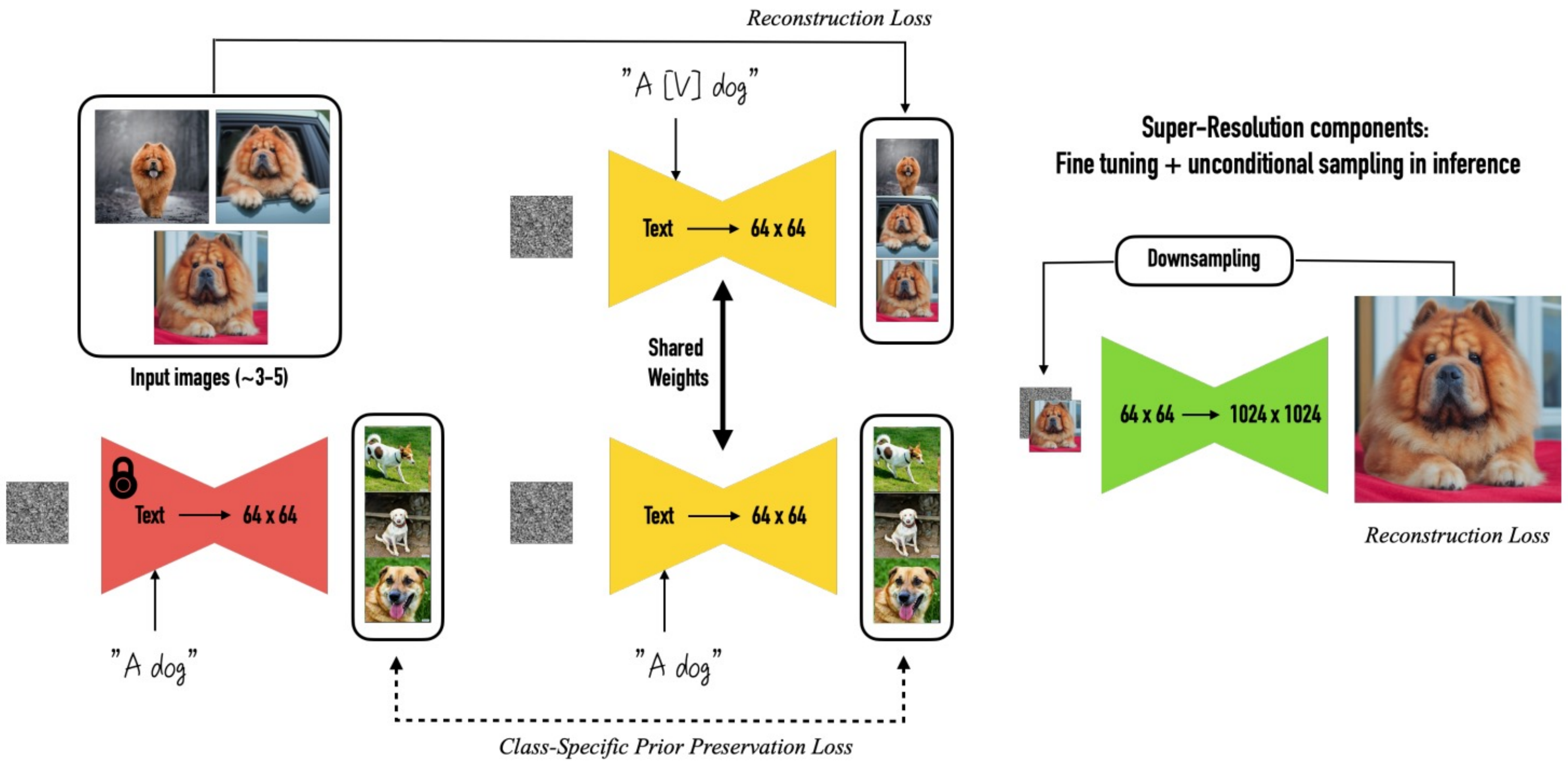
- Take few images (~3-5) as input of a subject (e.g., dog)
- The corresponding class name "dog"
- Return a fine-tuned text to image model that encodes a unique identifier that refers to the image.
- Used pre-trained image model as a base.



Method (cont'd)

- Label all images of the subject "a <identifier> <class noun>".
- Unique identifier without a class noun yields increased time, decreased performance.
- Rare token identifier to create the unique identifier.
- Fine-tune model using the classic denoising loss can cause overfitting and language drift.
- *Autogenous class-specific prior preserving loss* to prevent above mentioned issues.
- Finally super resolution applied.

Method (Cont'd)

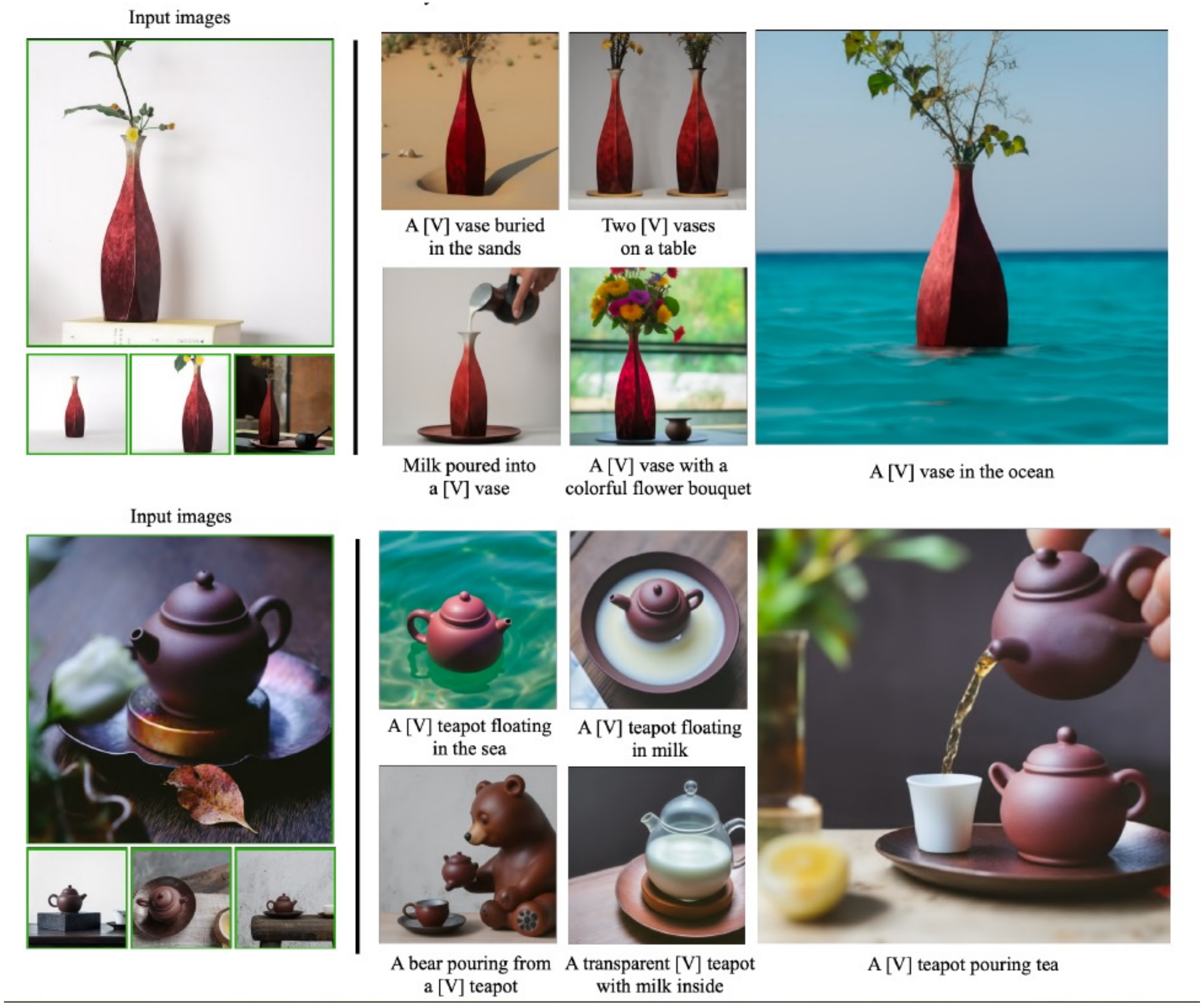


Experiment

- The authors find a large expanse of potential text-guided semantic modifications of the subject instance, which includes,
  - Recontextualization
  - Artistic Renditions
  - Expression manipulation
  - Accessorizing
  - Property Modification
- Across all these varied semantic modifications, the model can preserve unique visual features that give the subject its identity.

Results

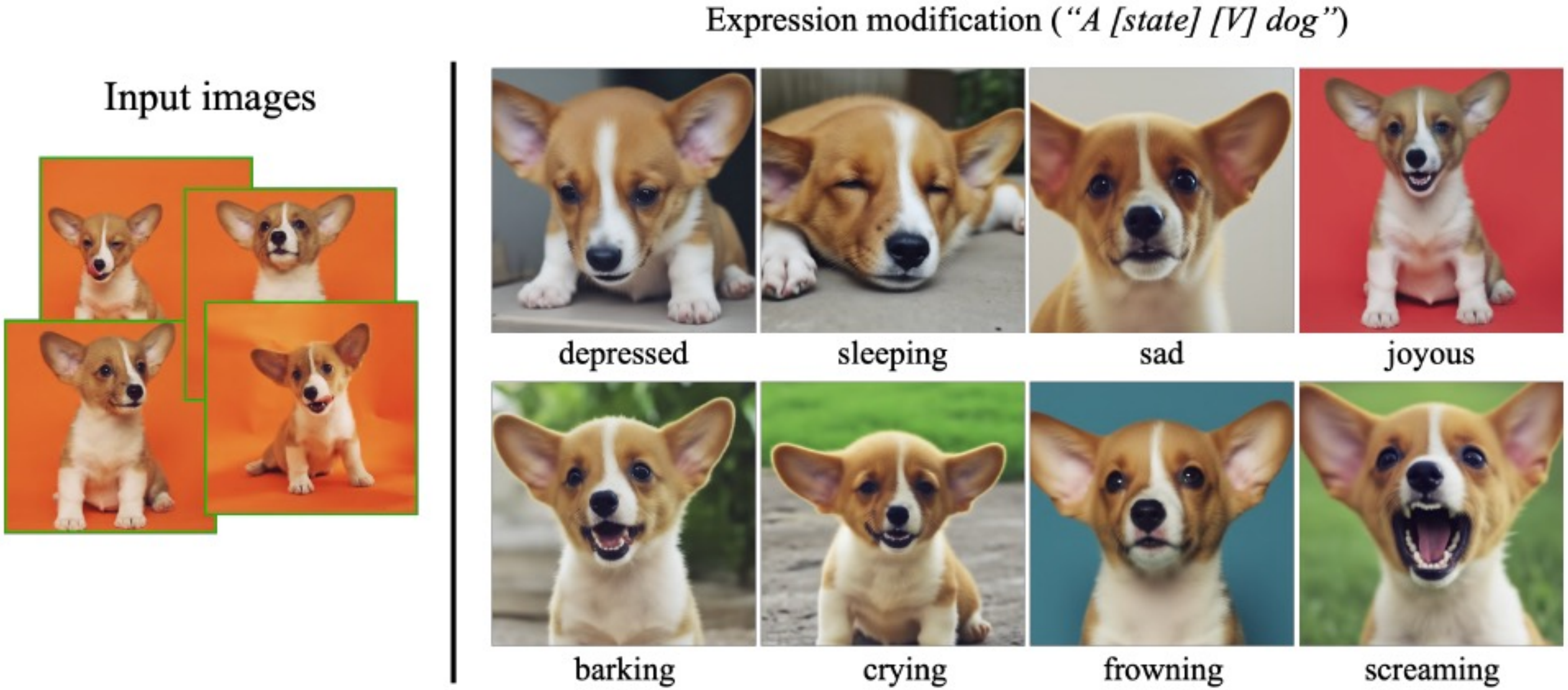
- Recontextualization.
- “a [V] [class noun] [context description]”
- Ex:“a [V] clock with the Eiffel Tower in the background”



Results: Art Renditions

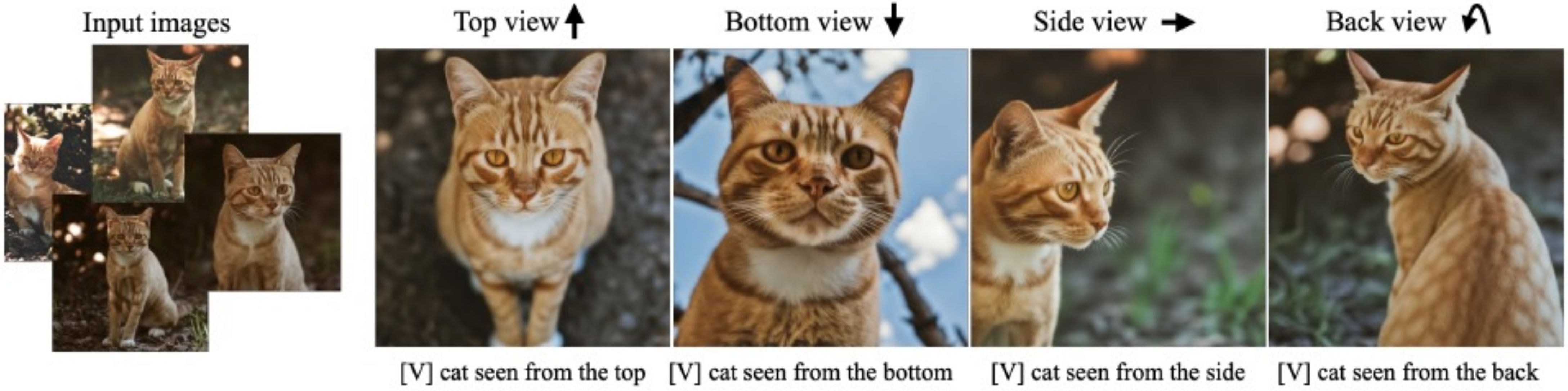
- "a painting of a [V] [class noun] in the style of [famous painter]".
- "a statue of a [V] [class noun] in the style of [famous sculptor]"

Results: Expression Manipulation



Results: Text Guided View Synthesis

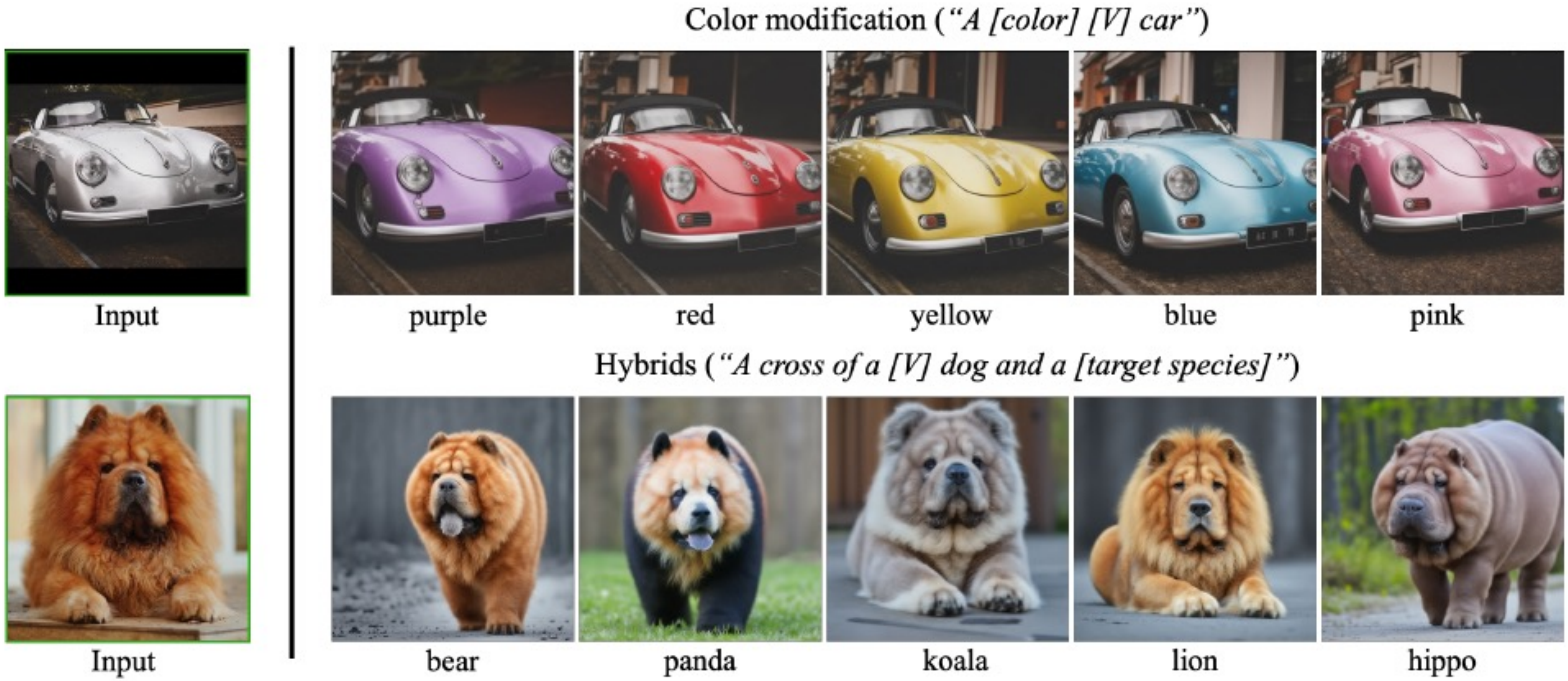
- The highlight is that the model has not seen this specific cat from behind, from below, or from above.



Results: Accessorization

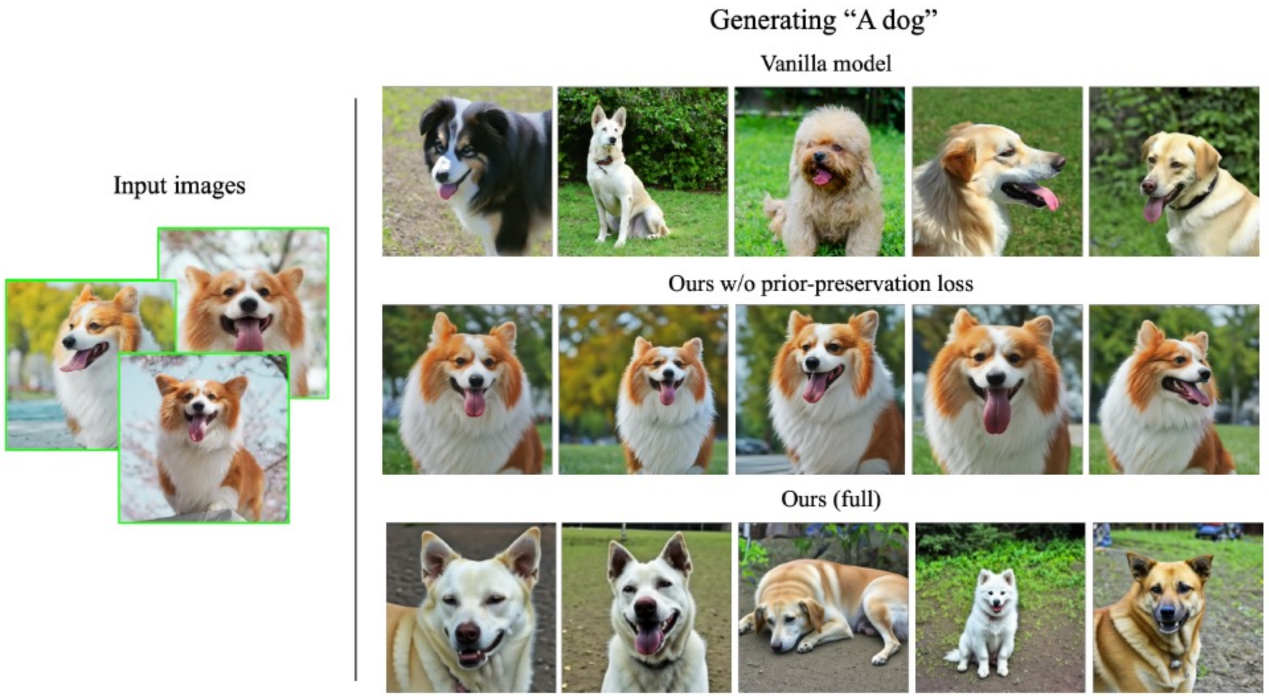
- “a [V] [class noun] wearing [accessory]”

Results: Property Modification



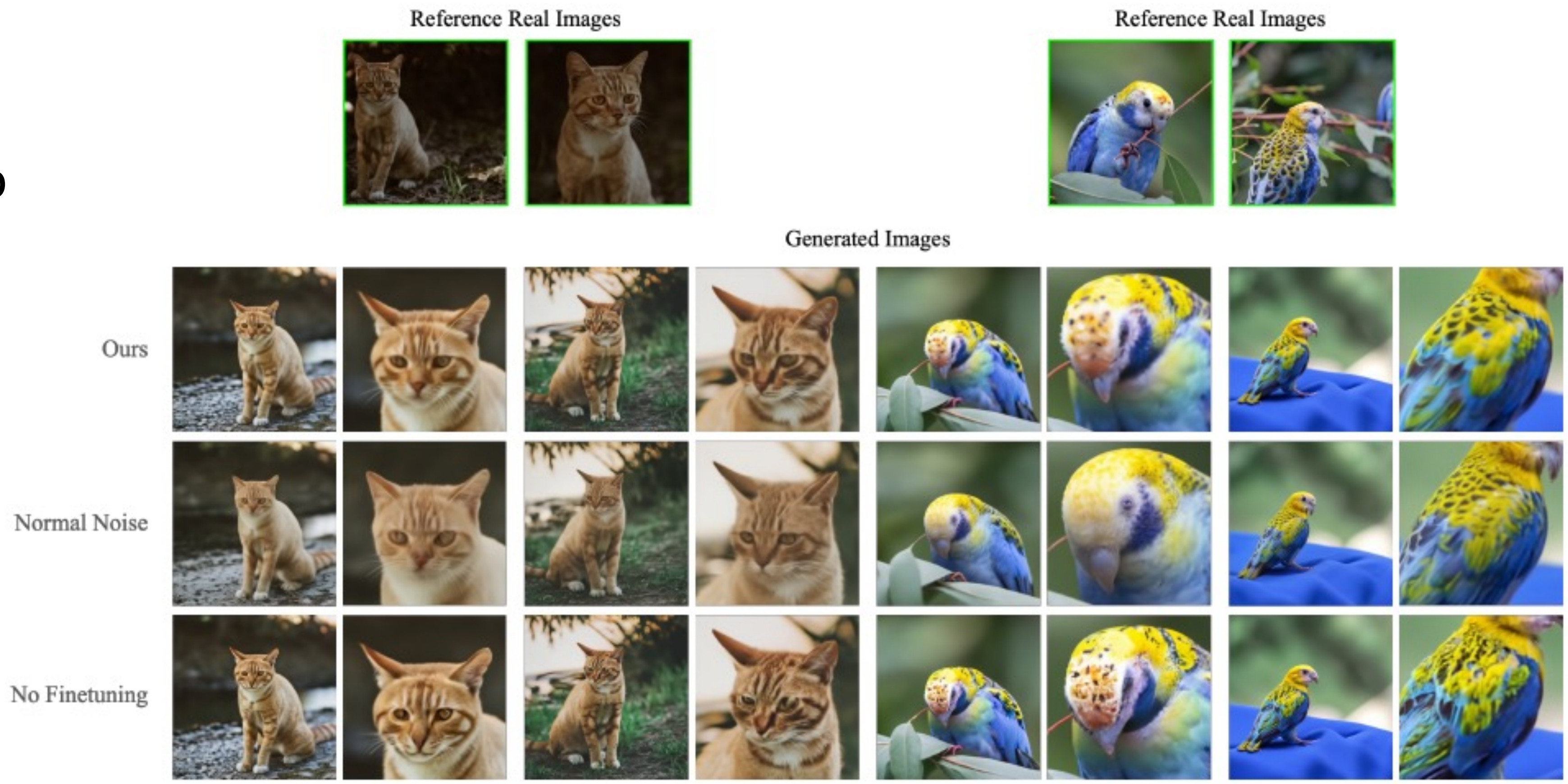
Comparison

- With the prior preservation loss, their results exhibit variation in the poses of the subject.
- Fine-tuning using images of our subject without prior-preservation loss results in language drift and the model loses the capability of generating other members of the subject's class.

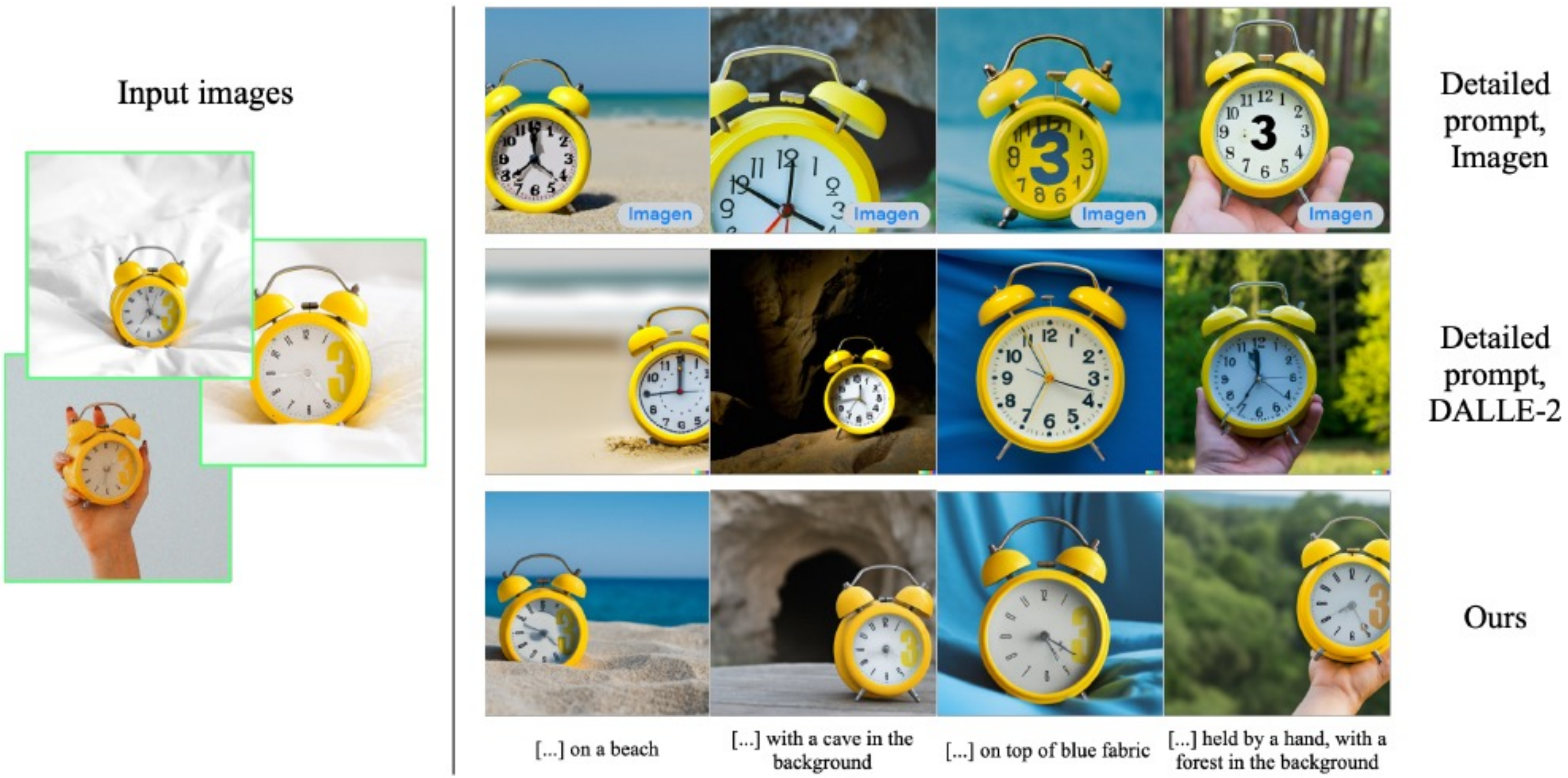


Comparison

- Using the normal level of noise augmentation of to train the models results in blurred high-frequency patterns,
- No fine-tuning results in hallucinated high-frequency patterns.



Comparison


Input images
Imagen
Imagen
Imagen
Imagen
Detailed prompt, Imagen
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

Conclusion

- Key idea: embed a given subject instance in the output domain of a text-to-image diffusion model by binding the subject to a unique identifier.
- Fine-tuning a pretrained text-to-image model without "forgetting" other visual concepts it had learned during training.
- this fine-tuning process can work given only 3-5 casually captured images of the subject ->accessible and easy to use
- fine-tuned model is able to reuse its learned knowledge of the visual world with holding the key features.

THANK YOU