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

CVPR 2023 (Award Candidate)

[paper] [project]

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

(Google Research)

2024.01.17

Mijin Koo

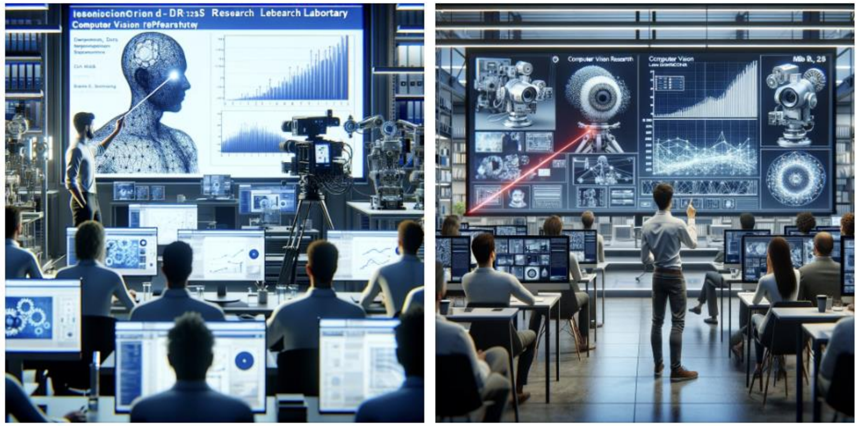
Contents

Text-to-Image models

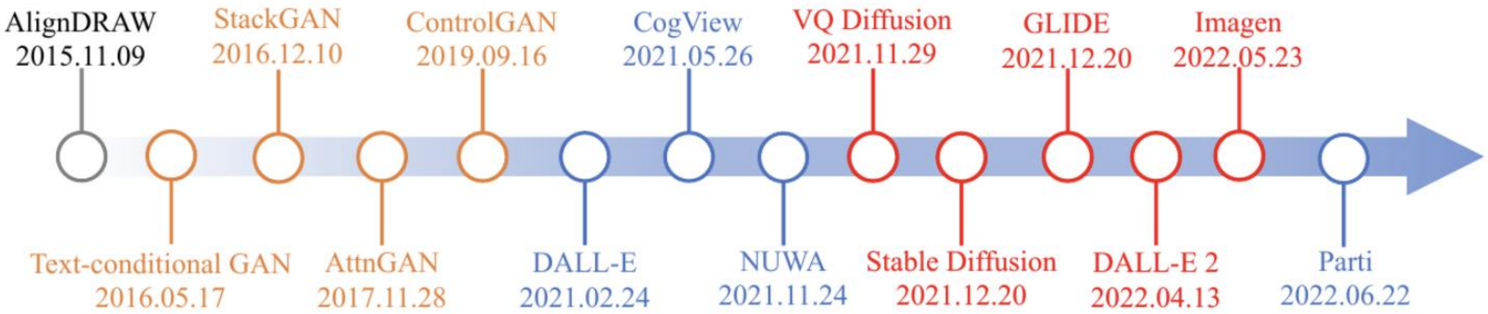
- **Paradigm of text-to-image task**
  - GAN-based
  - DALLE: autoregressive methods (not diffusion)
  - GLIDE, Imagen: image synthesis with diffusion models
  - Stable diffusion: image synthesis in latent space with diffusion models (LDM)

You
컴퓨터비전을 연구하는 MIPAL 연구실에서 논문 세미나 발표하는 인턴의 모습

DALL·E



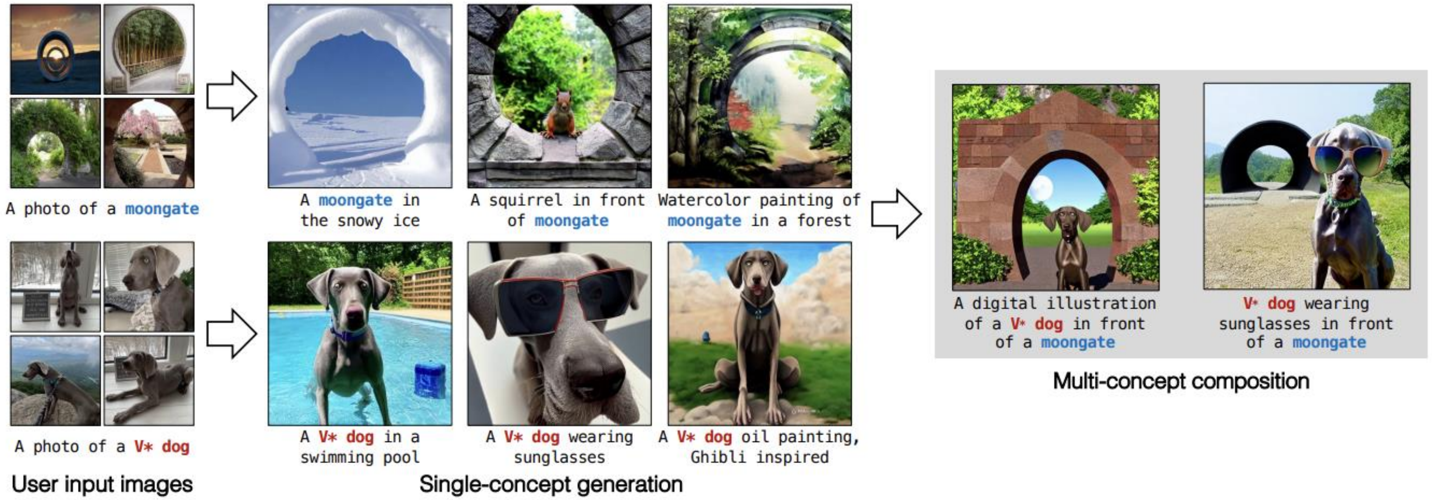
Here are two images showing an intern presenting a paper at a seminar in the MIPAL research laboratory, focused on computer vision research.



Text-to-image Diffusion Models in Generative AI: A Survey (arXiv, 2023)

Personalization with Diffusion Models

- **Introduction to the concept of Personalization**



Multi-concept customization of text-to-image diffusion (CVPR, 2023)

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation (CVPR, 2023)

Diffusion Models | DDPM

- **Unconditional generation**

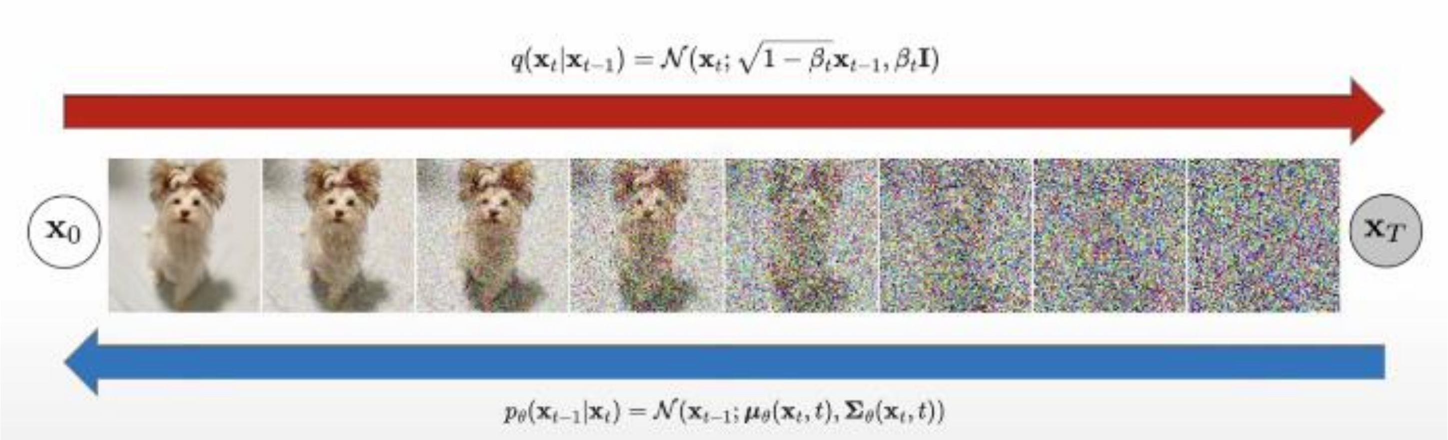
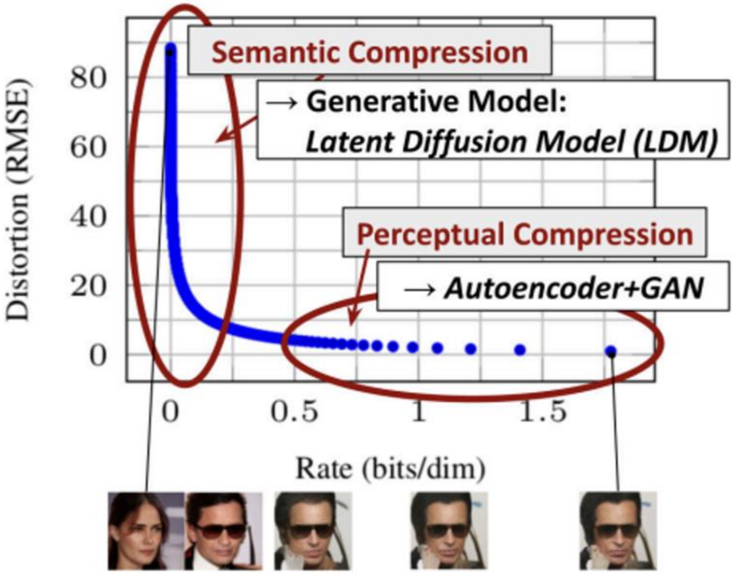


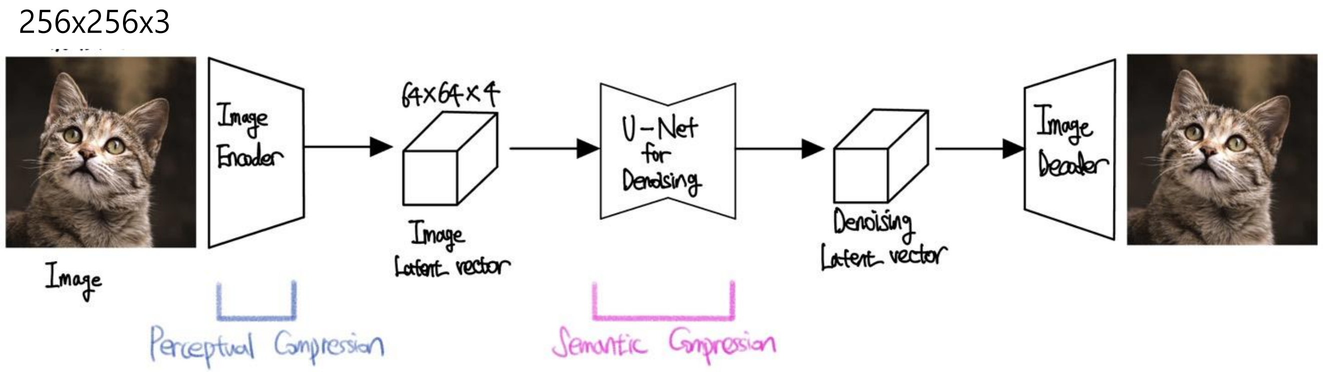
Figure 6: Unconditional CIFAR10 progressive generation ($\hat{\mathbf{x}}_0$ over time, from left to right). Extended samples and sample quality metrics over time in the appendix (Figs. 10 and 14).

Denoising diffusion probabilistic models (NIPS, 2020)

Diffusion Models | LDM

- **Motivation**
  - Diffusion models typically **operate directly in pixel space**, optimization of powerful DMs often consumes hundreds of GPU days and inference is expensive due to sequential evaluations.
- **Proposed Method**
  - We apply them in the **latent space** of powerful pretrained autoencoders.





High-Resolution Image Synthesis with Latent Diffusion Models (CVPR, 2022)

Diffusion Models | LDM

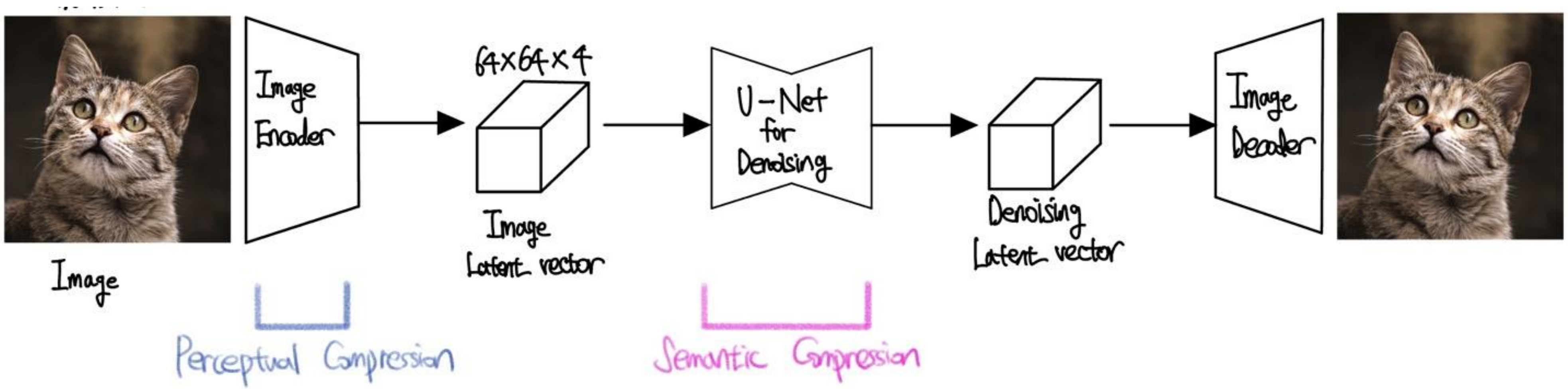
- **Method**
  1. Train autoencoder (Perceptual compression)
     - Train an autoencoder which provides a lower-dimensional (and thereby efficient) representational space which is perceptually equivalent to the data space.
  2. Train DM in latent space
     - We do not need to rely on excessive spatial compression, as we train DMs in the learned latent space
     - The reduced complexity also provides efficient image generation from the latent space with a single network pass.

- DM

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\Big]$$

- LDM

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\Big]$$

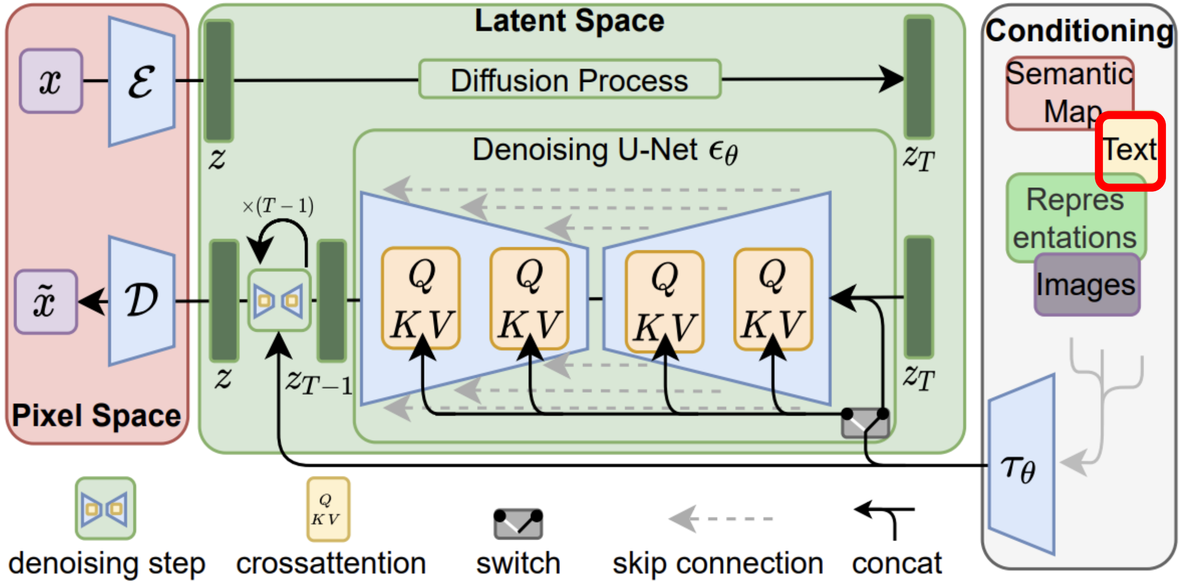


Diffusion Models | LDM

- **For conditional generation**
  - Conditioning mechanism
    - Modeling conditional distributions of the form $p(z|y)$
    - This can be implemented with a conditional denoising autoencoder $\epsilon_\theta(z_t, t, y)$
    - y can be text, semantic maps, or other image for image-to-image translation tasks
    - Integrates transformers with the DM's UNet backbone, allowing for various types of token-based conditioning mechanisms
  - Cross-attention mechanism
    - Effective for learning attention-based models of various input modalities
    - Domain specific encoder $\tau_\theta(y)$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \text{ with}$$

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t),\ K = W_K^{(i)} \cdot \tau_\theta(y),\ V = W_V^{(i)} \cdot \tau_\theta(y).$$



- Unconditional

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), \epsilon \sim \mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right]$$

- Conditional

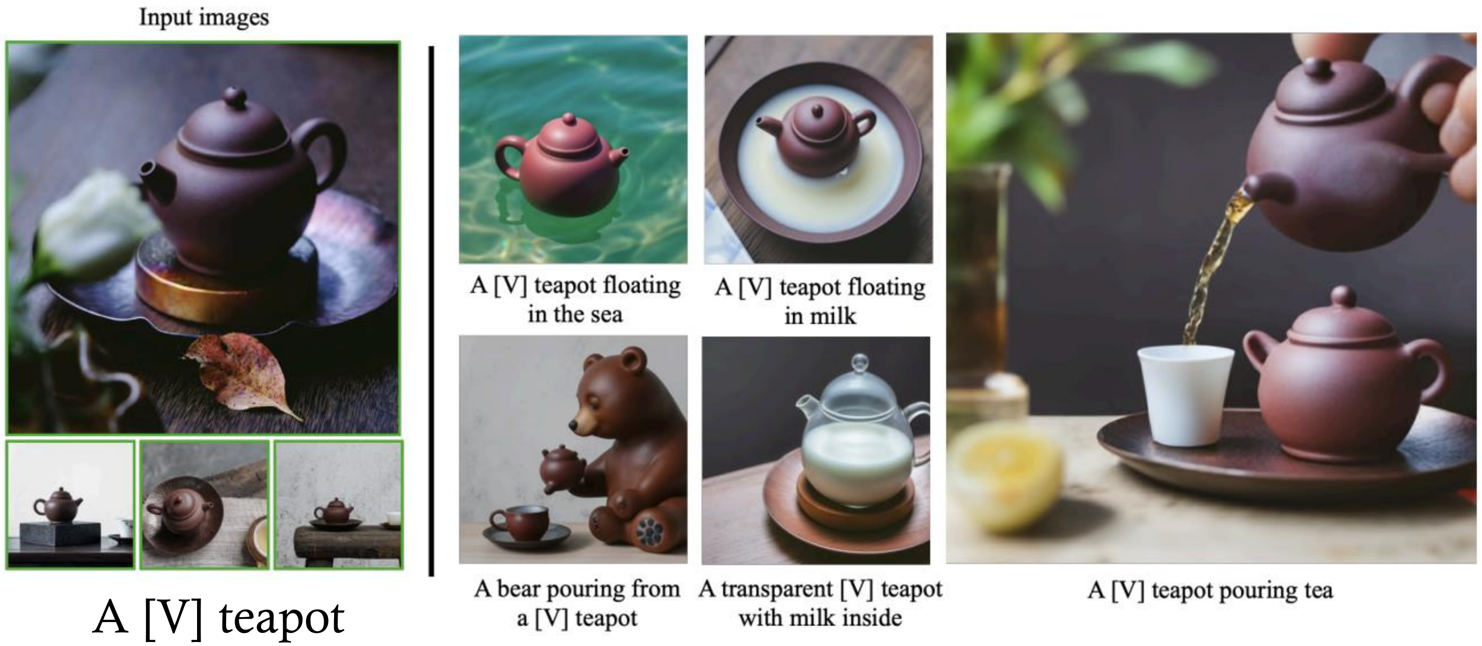
$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right],$$

DreamBooth | Overview

- **Motivation**
  - Large text-to-image models enable high-quality and diverse synthesis of images form a given text prompt. However, they lack the ability mimic the appearance of subjects in a given reference set and synthesize novel renditions of them in different contexts.
- **Proposed method**
  - So we present a new approach for "personalization" of text-to-image diffusion models.
  - Given as input just a few images of a subject, we fine-tune a pretrained text-to-image model such that it learns to bind a unique identifier with that specific subject.

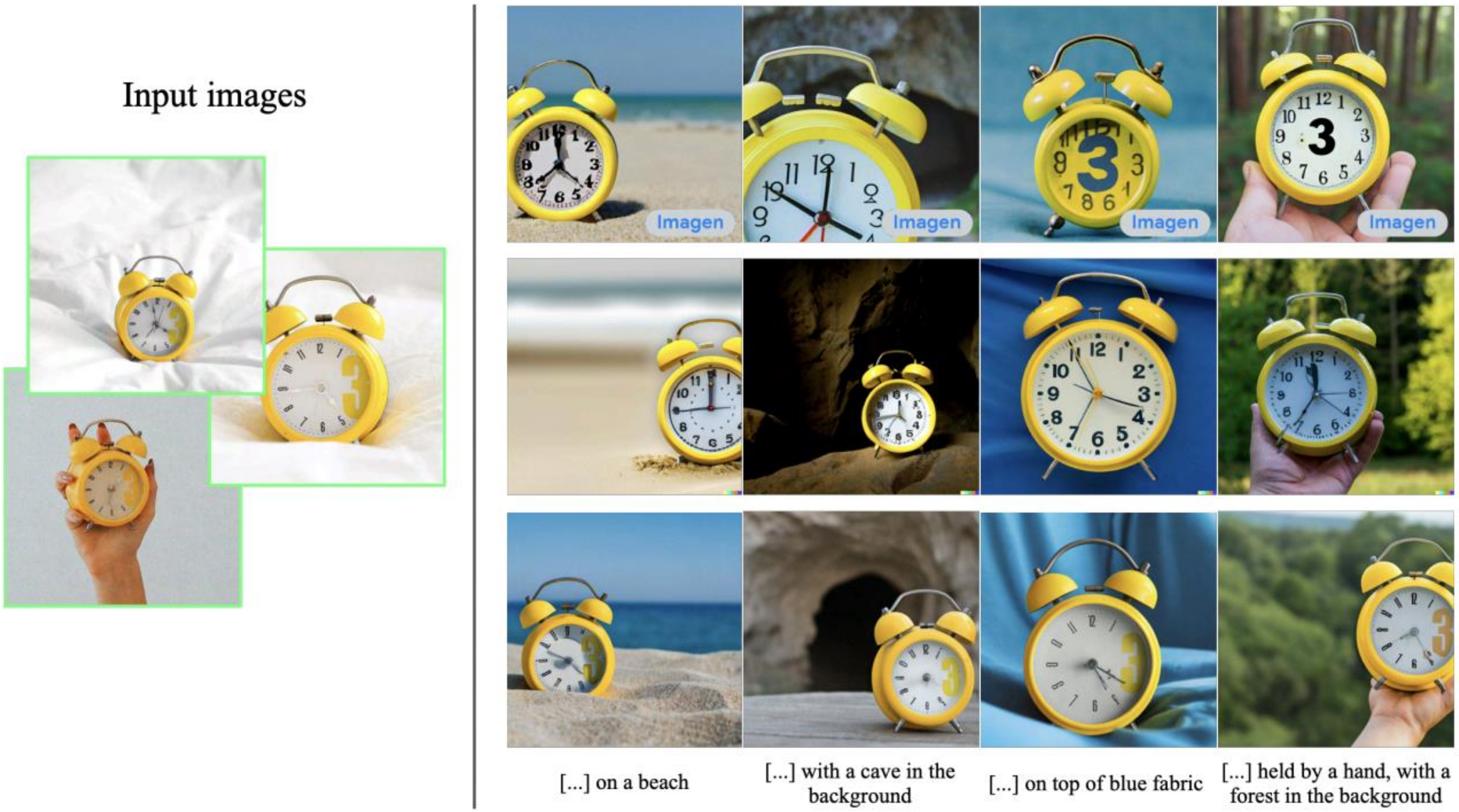
DreamBooth | Method

- **Personalizatoin of Text-to-image models**
  - Designing prompts for few-shot personalization
    - Input images paired with a text prompt containing a **unique identifier** and the **name of the class** the subject belongs to (e.g., “A [V] dog”- “a [identifier] [class noun]”)
      - ➢ Unique identifier for implanting specific subject
      - ➢ Class descriptor for providing prior of the class
    - Leverage the model’s prior of the specific class and entangle it with the embedding of our subject’s unique identifier



A [V] teapot

DreamBooth | Method



Detailed prompt, Imagen

Detailed prompt, DALLE-2

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

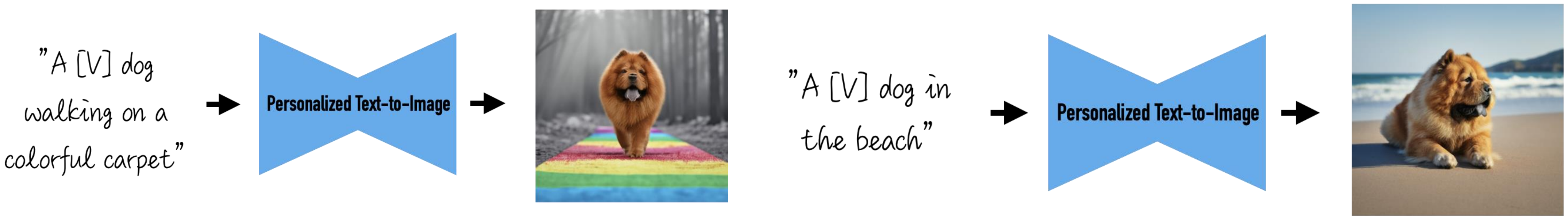
Ours

“a [V] clock in the ~”

DreamBooth | Method

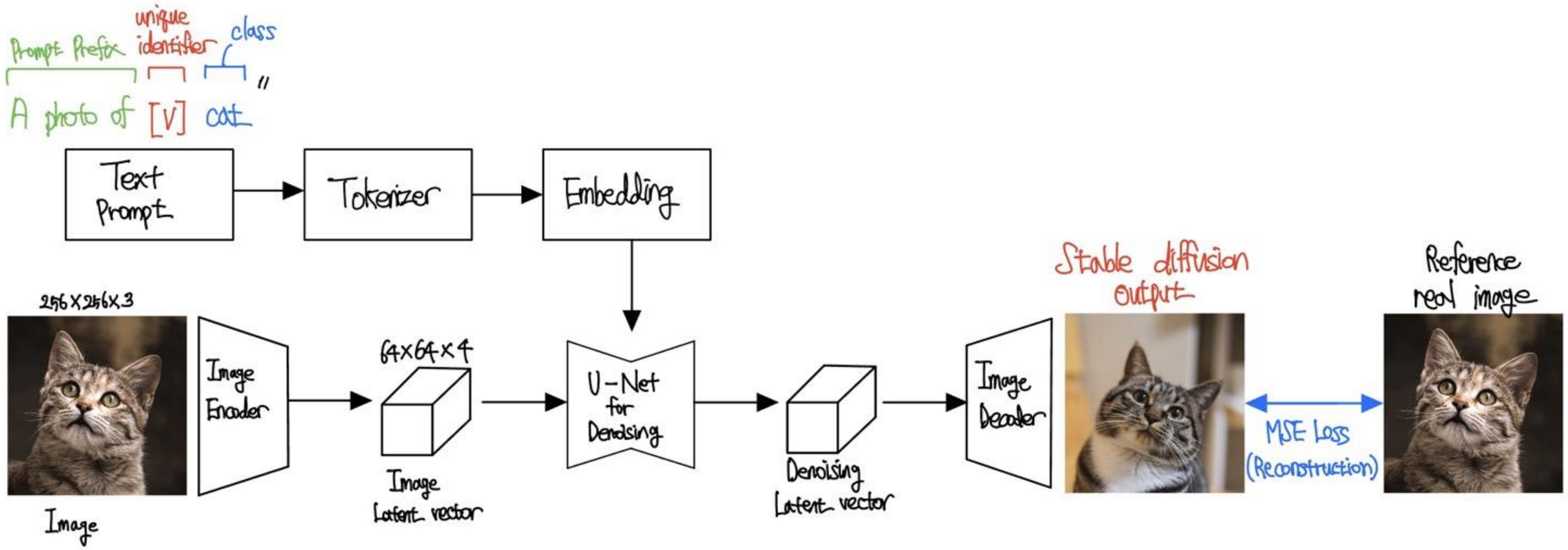
- **Personalizatoin of Text-to-image models**
  - Rare-token identifier (unique identifier)
    - Problem with existing English words
      - Existing English words(e.g. “unique”, “special”) are suboptimal for identifiers, since model has to learn to disentangle them for their original meaning and re-entangle them to reference the subject
    - Random character approach
      - Hazardous method: Select random English characters for an identifier (e.g., "xxy5syt00").
      - Tokenizer might treat each letter separately, leading to similar weaknesses as common English words.
    - **Rare token approach**
      - Unique identifier needs to be a weak prior in both the language model and diffusion model

      1. Find rare tokens in vocabulary
      2. Invert these tokens into text space to minimize strong priors ([V]=“sks”)



DreamBooth | Method

- **Class-specific Prior Preservation Loss**
  - [Goal] Achieve maximum subject fidelity by fine-tuning all model layers.
  - Challenge 1 – Language drift
    - Fine-tuning layers conditioned on text embeddings may lead to language drift.
    - Language drift observed in both language and diffusion models.
    - Diffusion model forgets how to generate subjects of the same class during fine-tuning.
  - Challenge 2 – Reduced output diversity
    - Fine-tuning on a small image set may limit output diversity in viewpoints, poses, and articulations.
    - Model training only with reconstruction loss, it can lead to reduce variability in output poses and views.



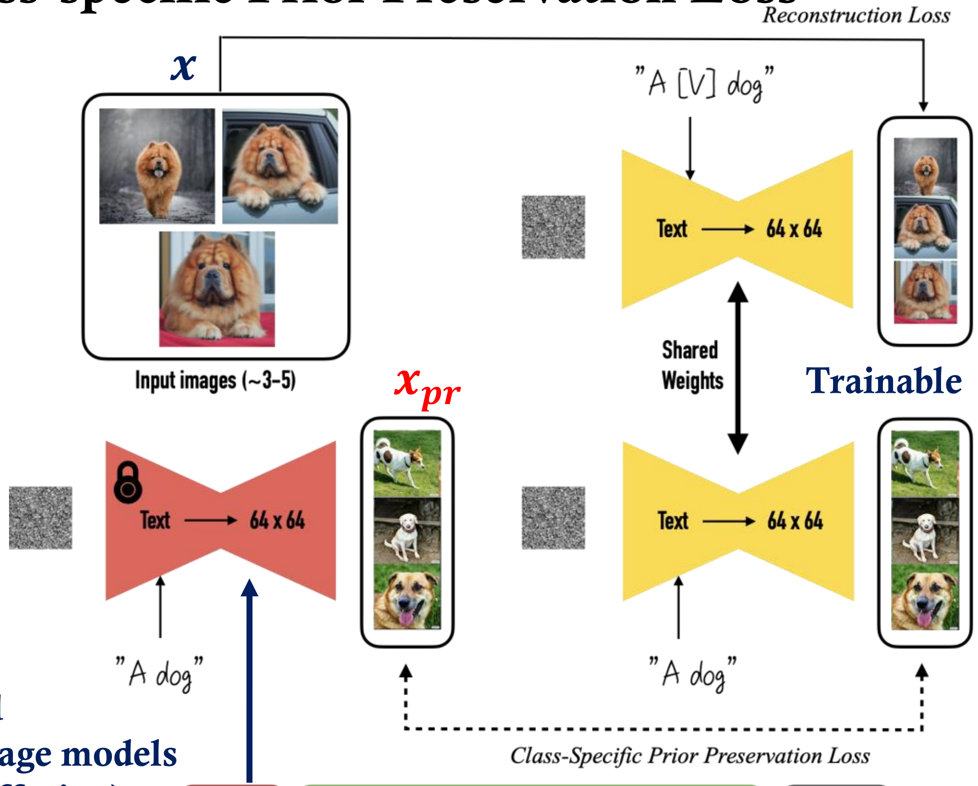
DreamBooth | Method

- **Class-specific Prior Preservation Loss**
  - Proposed solution: Autogenous Class-specific Prior Preservation Loss (PPL)
    - [Goal] Mitigate diversity reduction and language drift.
    - Supervise the model with its own generated samples during few-shot fine-tuning.
    - Generate data using ancestral sampler on frozen pre-trained diffusion model.
    - Loss Function
      - Combines reconstruction loss and prior-preservation term.
      - Prior-preservation term: Supervises the model with its own generated images.
      - λ controls the relative weight of the prior-preservation term.

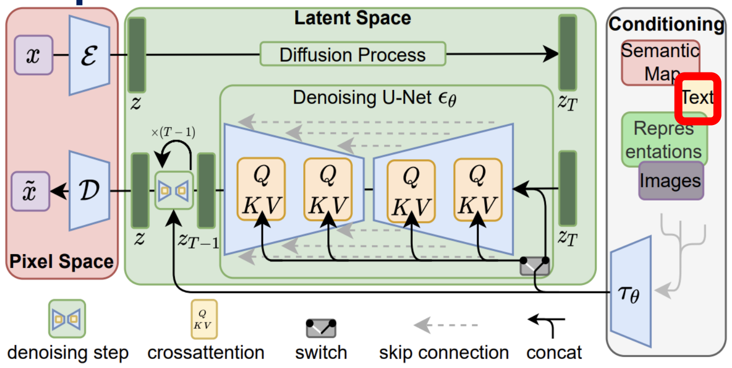
$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}\Big[\overbrace{\boxed{w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2}}^{\text{Reconstruction loss}}+\lambda\underbrace{\boxed{w_{t'}\|\hat{\mathbf{x}}_\theta(\alpha_{t'}\mathbf{x}_{\text{pr}}+\sigma_{t'}\boldsymbol{\epsilon}',\mathbf{c}_{\text{pr}})-\mathbf{x}_{\text{pr}}\|_2^2}}_{\text{Prior-preservation loss}}\Big]$$

DreamBooth | Method

- **Class-specific Prior Preservation Loss**



Pretrained text-to-image models (Stable Diffusion)

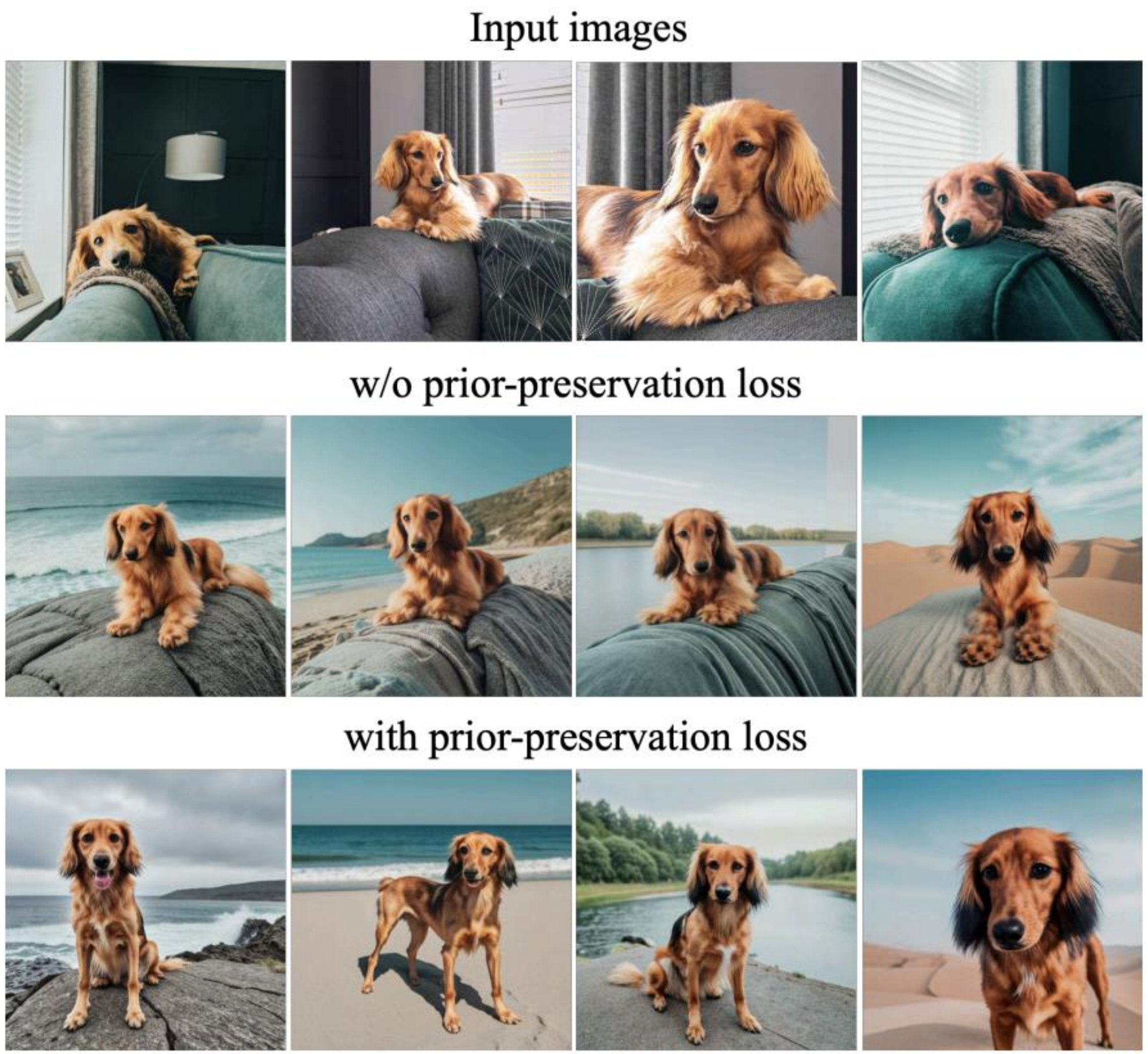


$$\mathbb{E}_{\mathbf{x},\mathbf{c},\epsilon,\epsilon',t}[\overbrace{\boxed{w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\epsilon,\mathbf{c})-\mathbf{x}\|_2^2}}^{\text{Reconstruction loss}}+\lambda\underbrace{\boxed{w_{t'}\|\hat{\mathbf{x}}_\theta(\alpha_{t'}\mathbf{x}_{\text{pr}}+\sigma_{t'}\epsilon',\mathbf{c}_{\text{pr}})-\mathbf{x}_{\text{pr}}\|_2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

- Generated data $x_{pr} = \hat{x}(z_{t_1}, c_{pr})$
- Random initial noise $z_{t_1} \sim N(0, I)$
- Conditioning vector $c_{pr} = \tau(f(\text{"}a\ [class]\text{"}))$
  - Tokenizer f

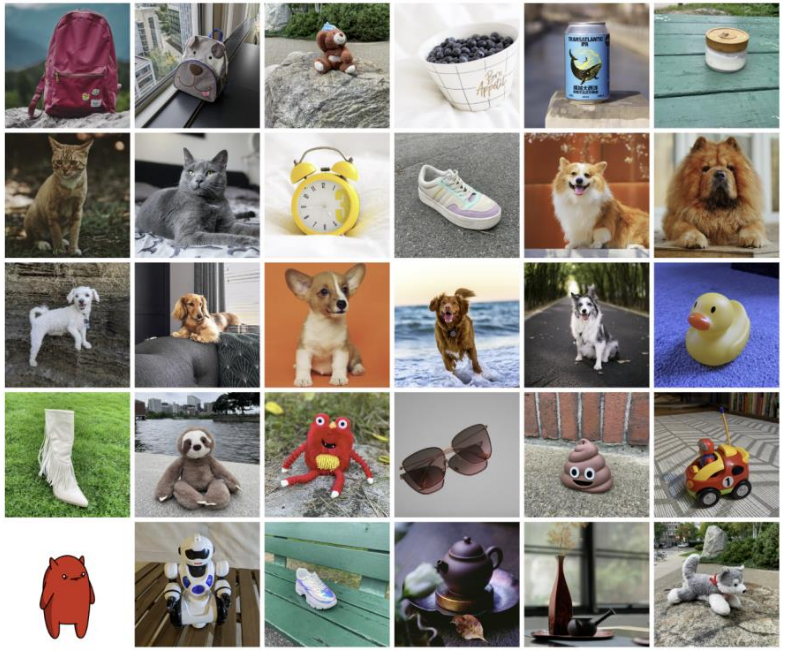
DreamBooth | Method

- **Class-specific Prior Preservation Loss**
  - Prior-preservation loss encourages output diversity and overcomes language drift
  - Naive fine-tuning can result in overfitting to input image context and subject appearance (e.g. pose).
  - PPL acts as a regularizer that alleviates overfitting and encourages diversity, allowing for more pose variability and appearance diversity.



DreamBooth | Experiments

- **Experiments Details**
  - Dreambooth capabilities
    - Enables text-guided semantic modifications of subject instances.
    - Modifications include recontextualization, subject property changes (material and species), art rendition, and viewpoint modifications.
    - Preserves unique visual features, maintaining subject identity and essence.
  - Datasets
    - Collection of 30 subjects, including objects and live subjects/pets
    - Data sources: Authors' collection and Unsplash

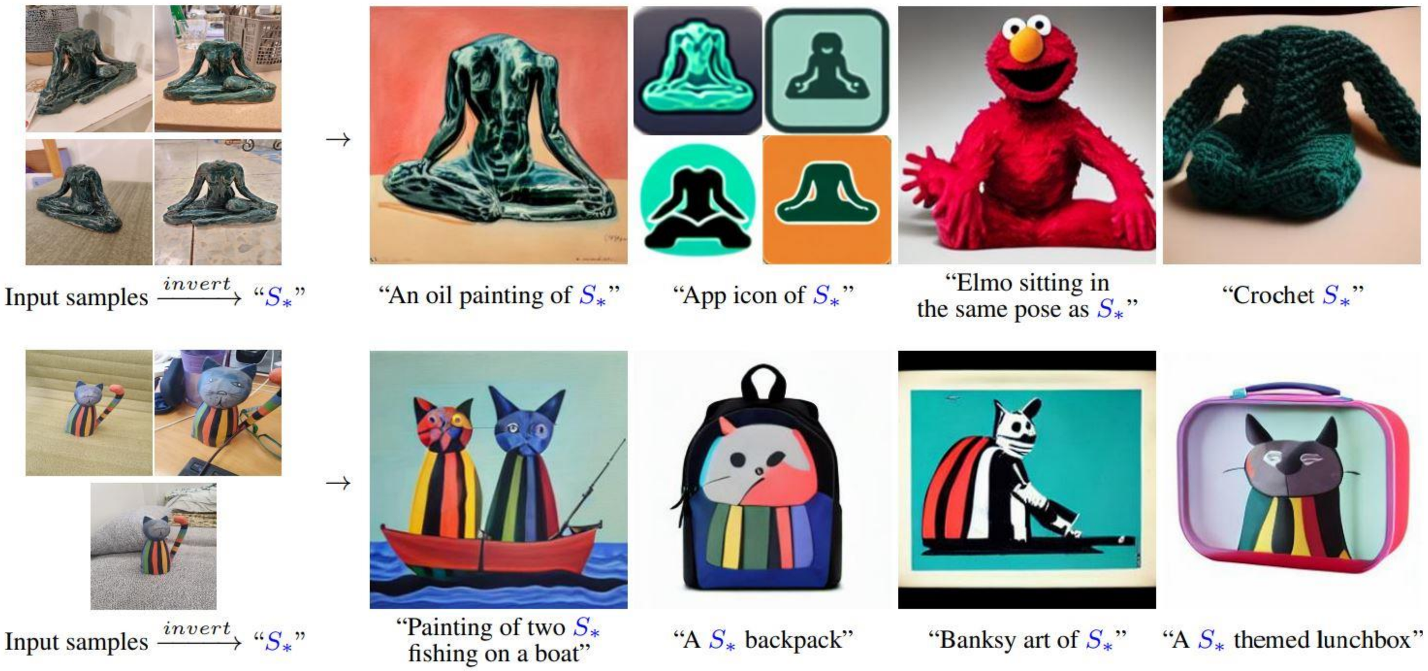


DreamBooth | Experiments

- **Evaluation metrics**
  1. Subject fidelity evaluation
     - CLIP-I (Cosine Similarity)
       - Average pairwise cosine similarity between CLIP embeddings of generated and real images.
       - CLIP-I commonly used but may not distinguish between different subjects with highly similar text descriptions.
     - DINO (ViTS/16 DINO Embeddings)
       - Average pairwise cosine similarity between ViTS/16 DINO embeddings of generated and real images.
       - DINO is our preferred metric, since it measures similarity considering unique features rather than ignoreing differences between subjects of the same class
  2. Prompt fidelity evaluation
     - CLIP-T (Cosine Similarity)
       - Average cosine similarity between prompt and image CLIP embeddings.
       - Measure the similarity between the prompt and the corresponding image embeddings.

DreamBooth | Experiments

- **Comparison with Textual Inversion**
  - Evaluation setup
    - Comparative analysis with Textual Inversion, recent concurrent work by Gal et al. [20].
    - Utilization of hyperparameters provided in the Textual Inversion work.
    - Image generation
      - DreamBooth: Imagen, Stable Diffusion
      - Textual Inversion: Stable Diffusion



An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion (ICLR, 2023)

DreamBooth | Experiments

- **Comparison with Textual Inversion**
  - Results

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
Real Images	0.774	0.885	N/A
DreamBooth (Imagen)	**0.696**	**0.812**	**0.306**
DreamBooth (Stable Diffusion)	0.668	0.803	0.305
Textual Inversion (Stable Diffusion)	0.569	0.780	0.255

Table 1. Subject fidelity (DINO, CLIP-I) and prompt fidelity (CLIP-T, CLIP-T-L) quantitative metric comparison.

Method	Subject Fidelity ↑	Prompt Fidelity ↑
DreamBooth (Stable Diffusion)	**68%**	**81%**
Textual Inversion (Stable Diffusion)	22%	12%
Undecided	10%	7%

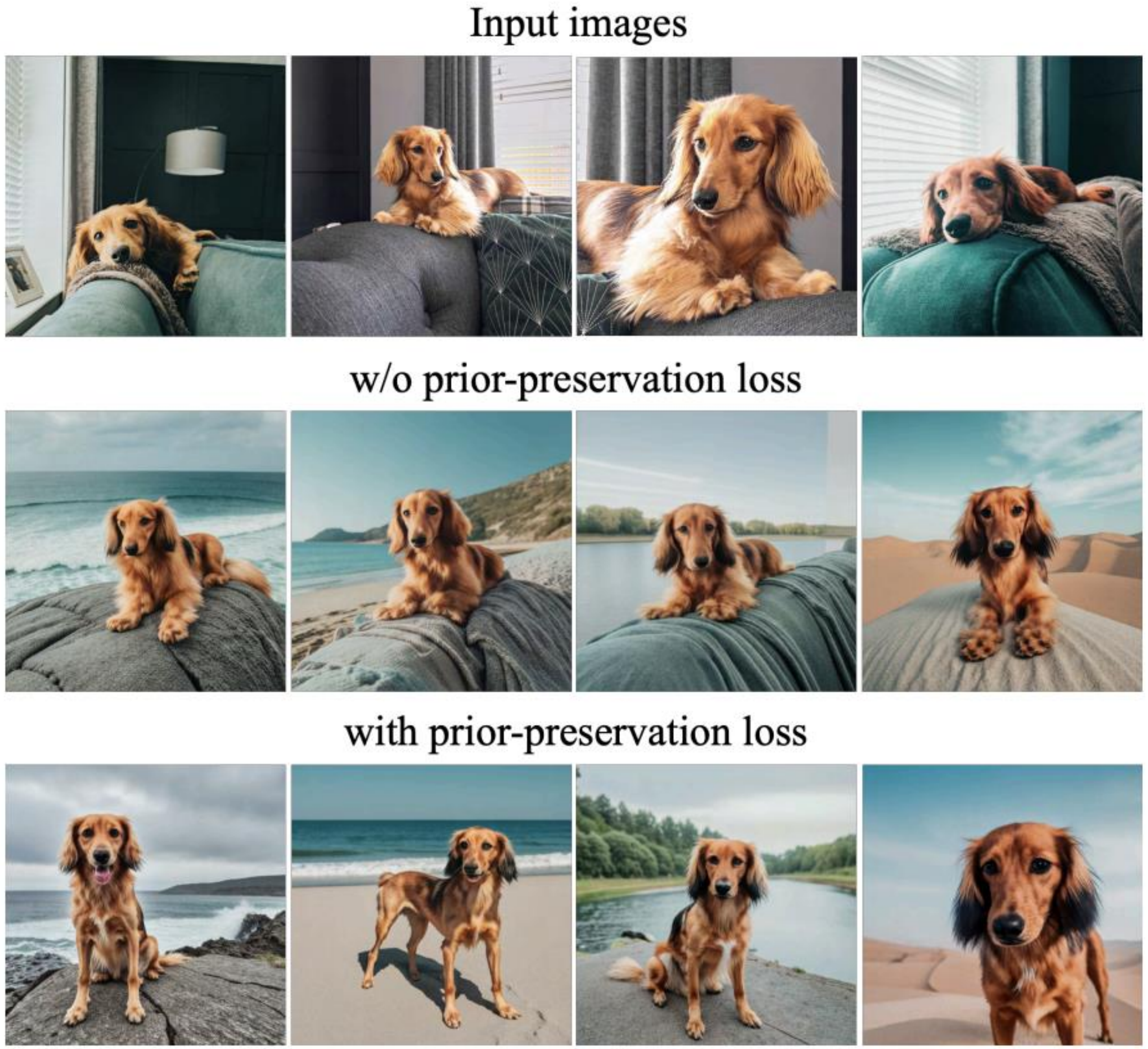
Table 2. Subject fidelity and prompt fidelity user preference.

DreamBooth | Ablation Study

- **Prior Preservation Loss Ablation**
  - PPL substantially counters language drift and helps retain the ability to generate diverse images of the prior class.
  - Higher diversity observed in the model trained with PPL, with slightly diminished subject fidelity.
  - Model trained with PPL overfits less to the reference images' environment.
  - Generates the dog in more diverse poses and articulations.

Method	PRES ↓	DIV ↑	DINO ↑	CLIP-I ↑	CLIP-T ↑
DreamBooth (Imagen) w/ PPL	**0.493**	**0.391**	0.684	0.815	**0.308**
DreamBooth (Imagen)	0.664	0.371	**0.712**	**0.828**	0.306

Table 3. Prior preservation loss (PPL) ablation displaying a prior preservation (PRES) metric, diversity metric (DIV) and subject and prompt fidelity metrics.



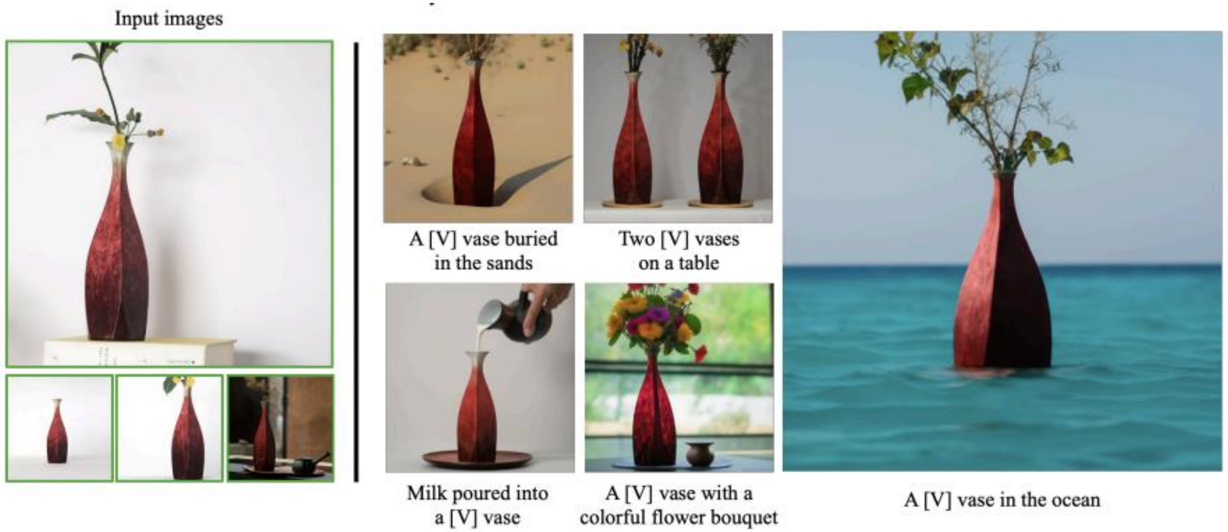
DreamBooth | Ablation Study

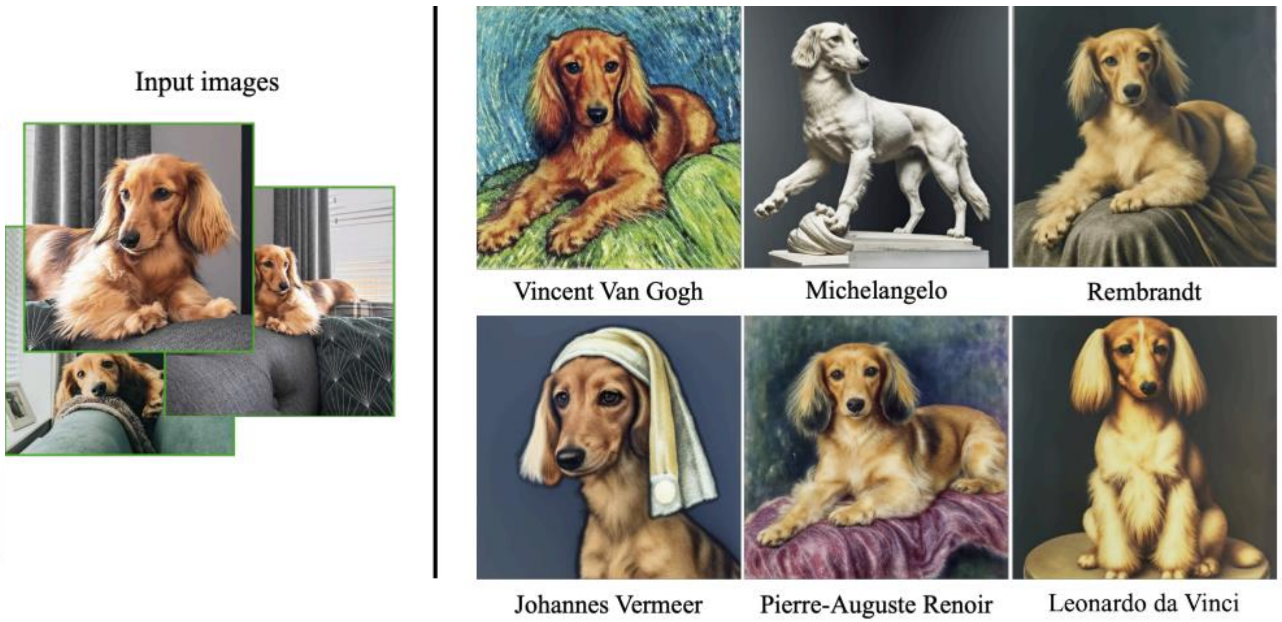
- **Class-Prior Ablation**
  - Correct class noun.
    - Allows faithful fitting to the subject and leverages the class prior.
    - Enables generation of the subject in various contexts.
  - No class noun.
    - Model struggles to learn the subject, has difficulty converging, and can generate erroneous samples.
  - Randomly sampled incorrect class noun.
    - Contentions observed, resulting in misshapen or erroneous subjects.

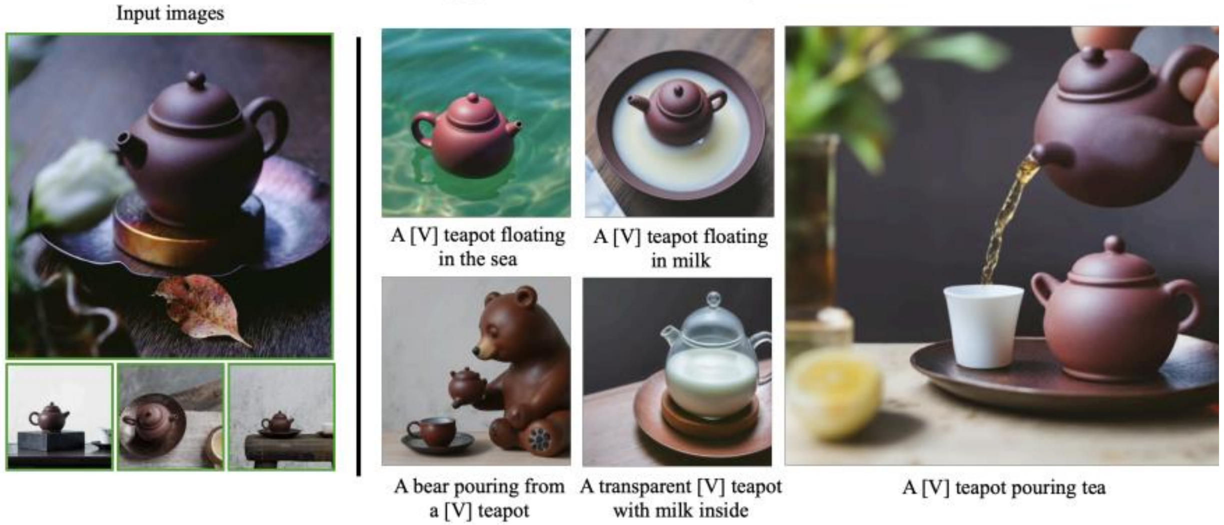
Method	DINO ↑	CLIP-I ↑
Correct Class	**0.744**	**0.853**
No Class	0.303	0.607
Wrong Class	0.454	0.728

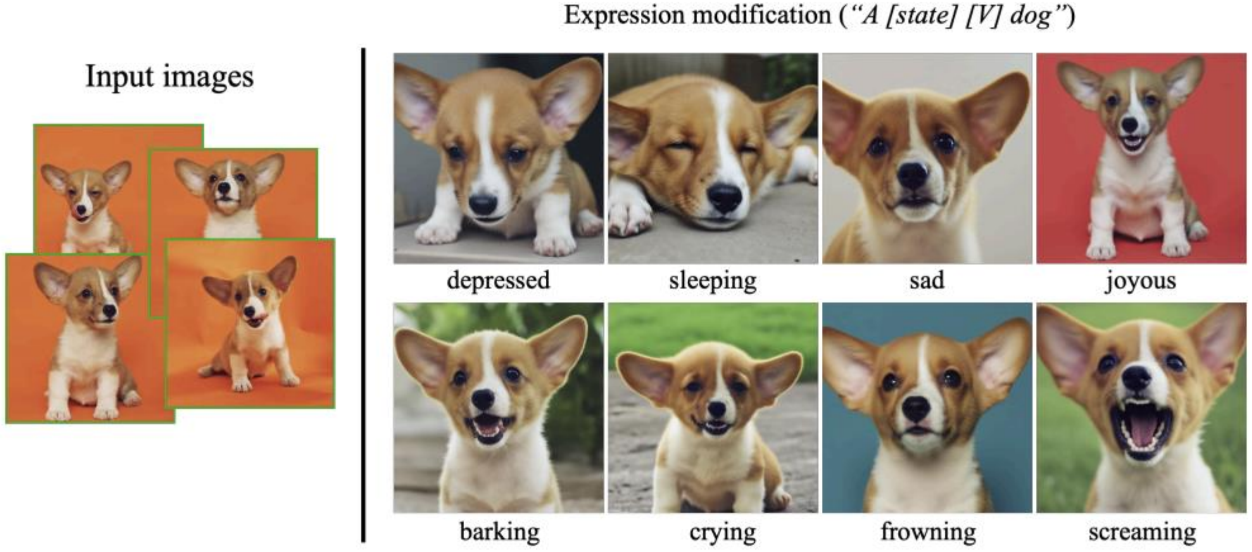
Table 4. Class name ablation with subject fidelity metrics.

DreamBooth | Results

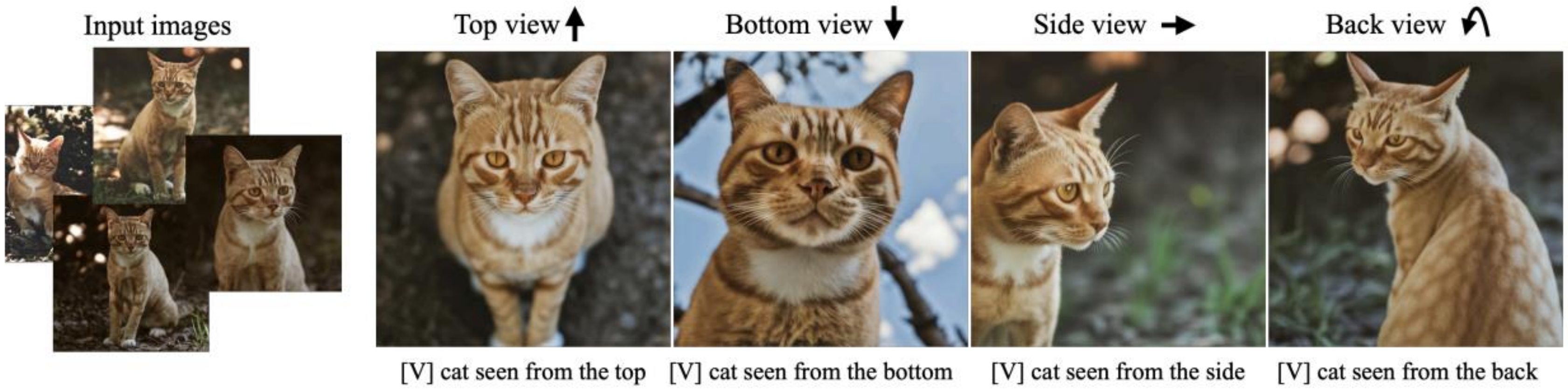








DreamBooth | Results



Thank you!