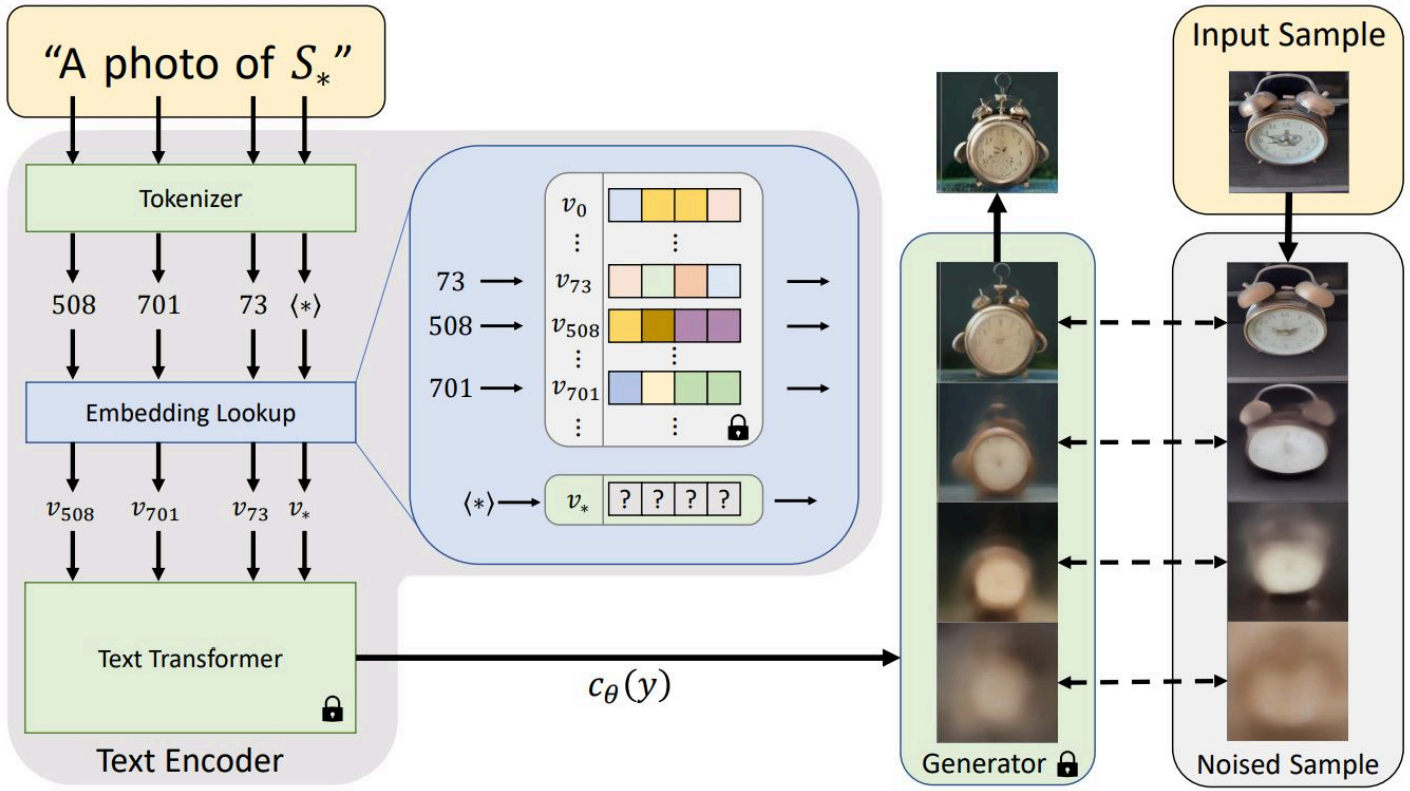
Subject-driven Generation: Textual Inversion & Dreambooth

[1] "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion" (Gal et al, 2022)

[2] "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation" (Ruiz et al, CVPR 2023)

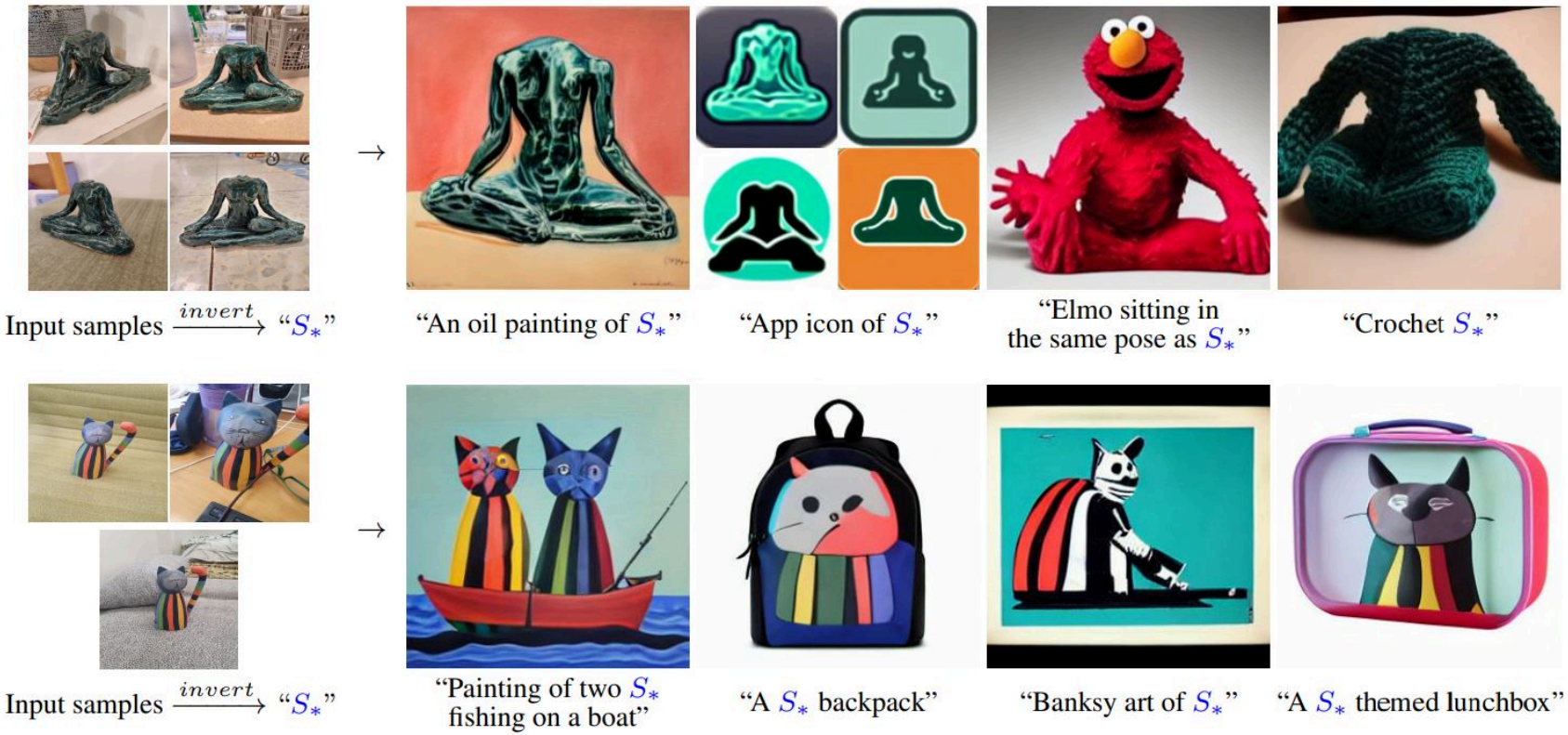
Textual Inversion



Here, the only trainable thing is the S_* token.

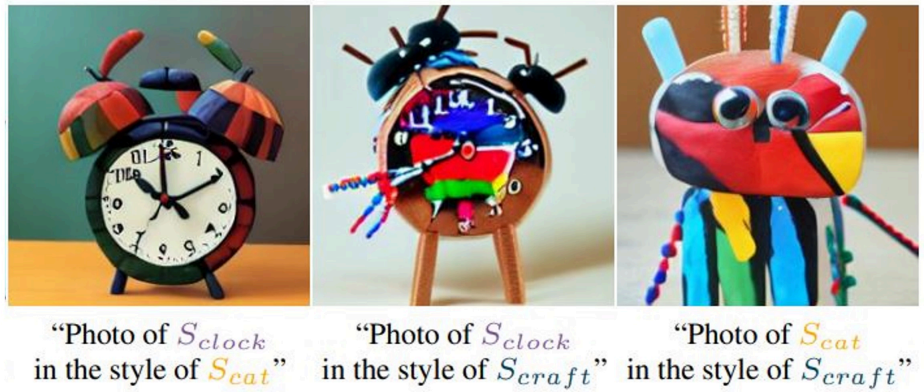
All other parameters are frozen.

Textual Inversion



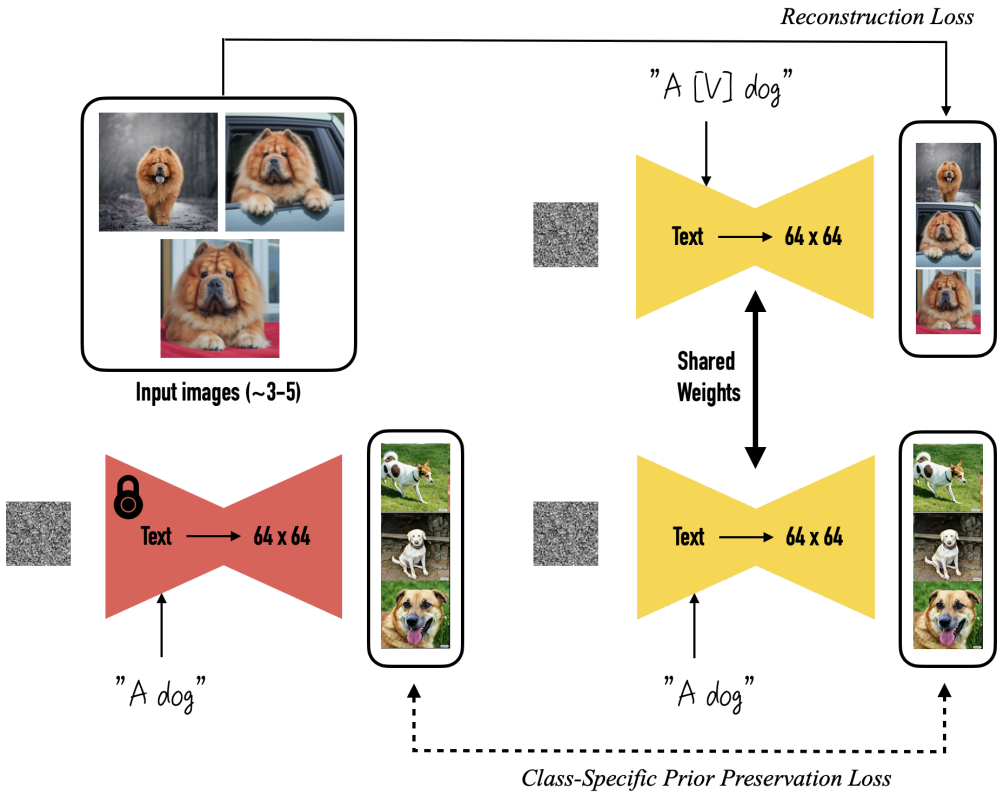
In this way, you can add a new token that represents a specific concept.

Textual Inversion



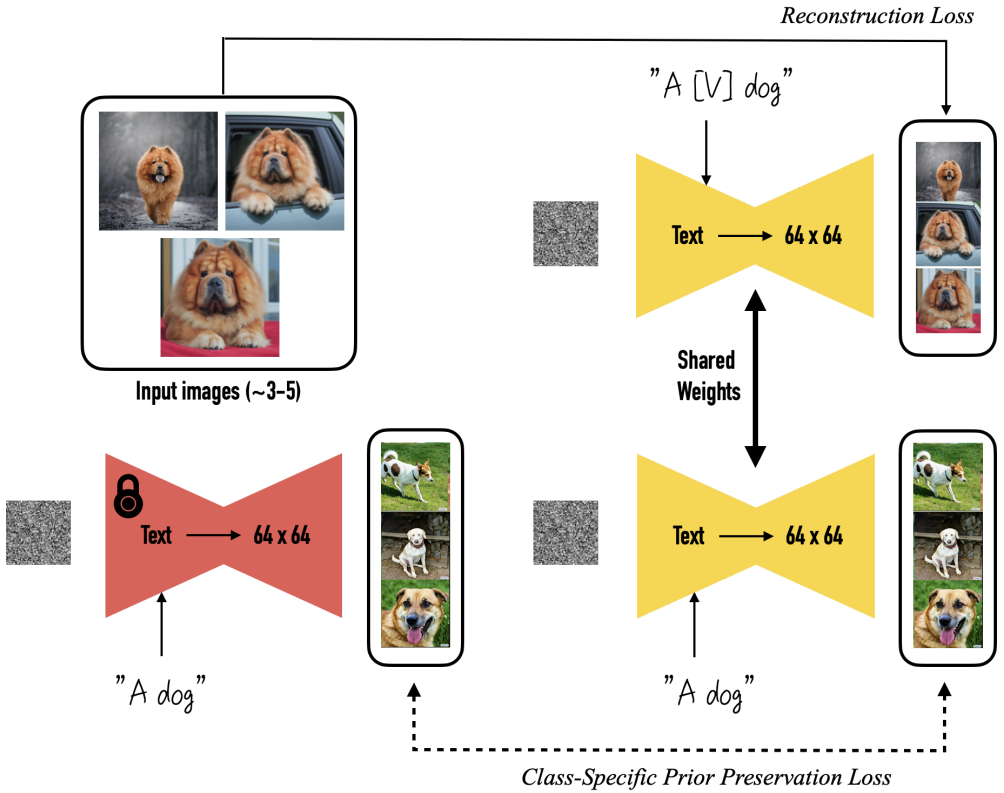
You can combine these textual-inversion-tokens like the figure above.

Dreambooth



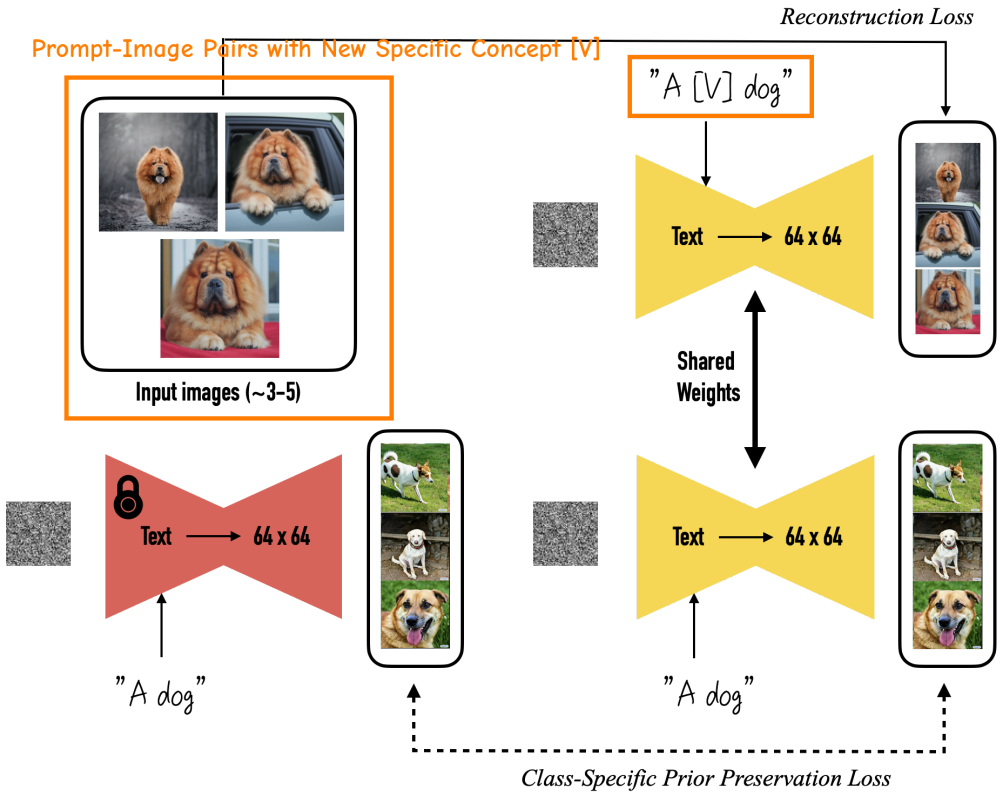
Dreambooth learns new concept (i.e., `[V]` token in the figure above) like Textual Inversion does.

Dreambooth



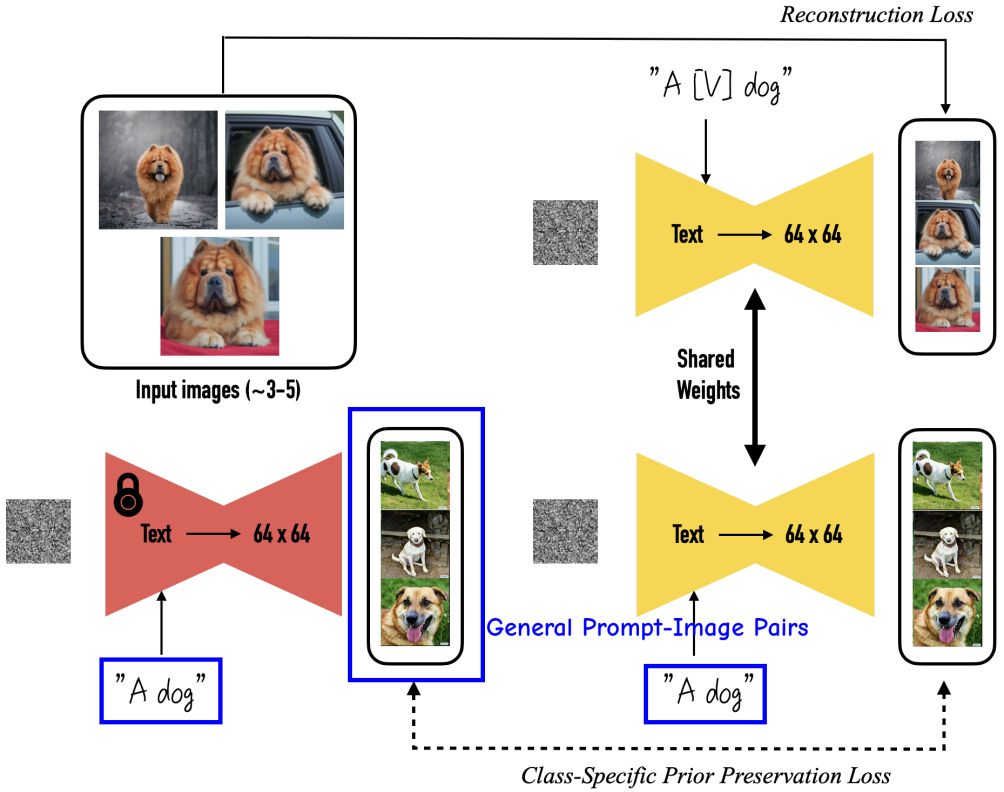
However, the UNet of Dreambooth is not frozen.

Dreambooth



The UNet is also trained to learn the new concept with the prompt-image pairs for the new concept (i.e., `[V]`).

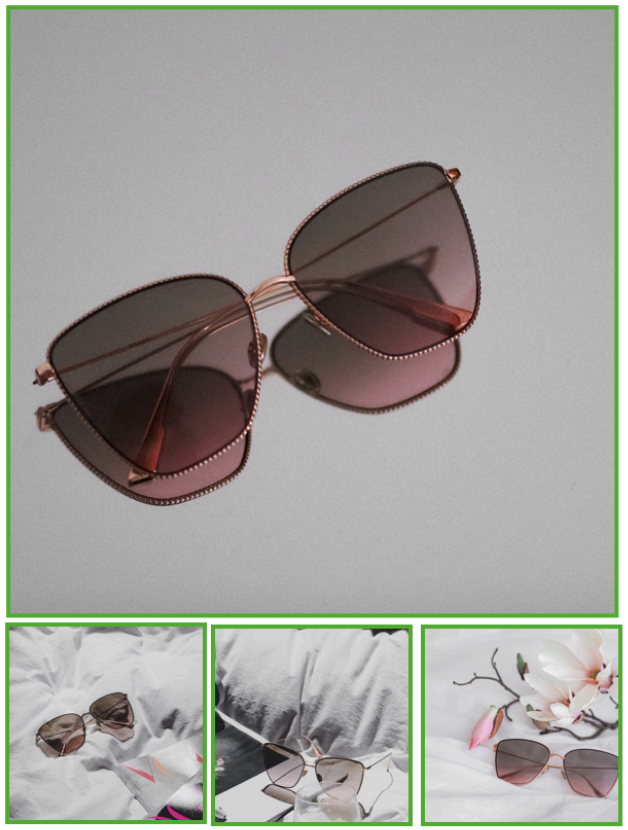
Dreambooth



In addition, **to preserve the existing pre-trained diffusion model's text-to-image generation mechanism**, the general prompt-image pairs without the specific concept are also used in the learning.

Dreambooth

Input images

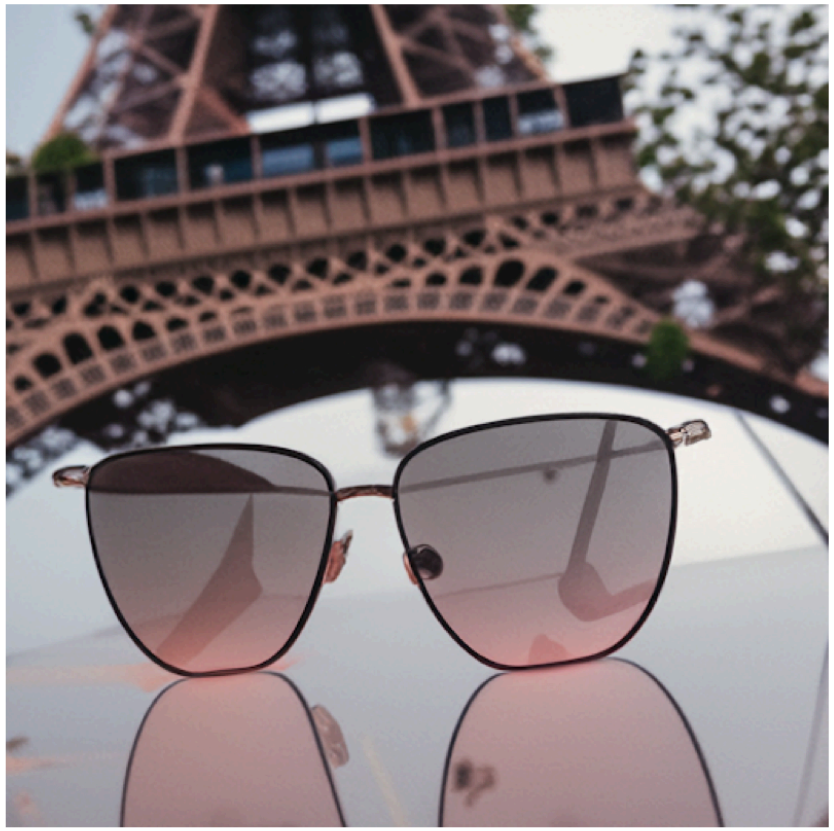


A [V] sunglasses in the jungle

A [V] sunglasses worn by a bear

A [V] sunglasses at Mt. Fuji

A [V] sunglasses on top of snow



A [V] sunglasses with Eiffel Tower in the background

Dreambooth

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

Dreambooth

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