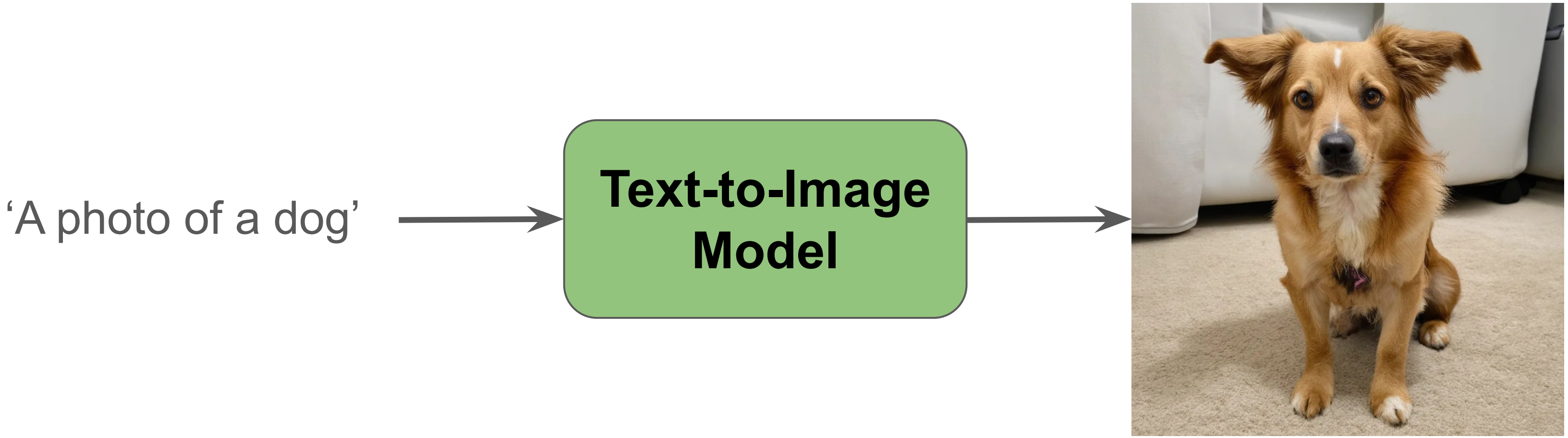
Instance Specific 2D and 3D Generation

Varun Jampani

Vice President of Research

Stability AI

Excellent Progress in Text-to-Image Models

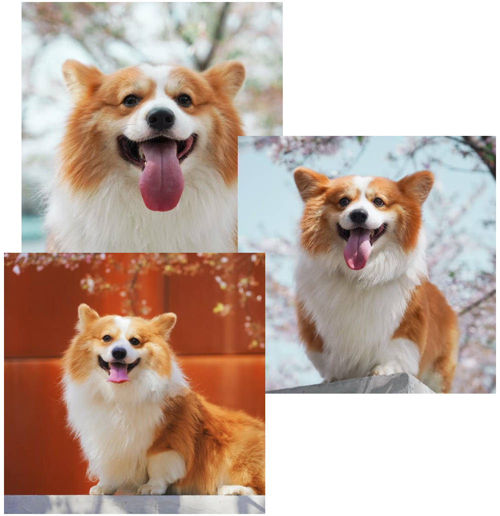


Instance-specific Generation and Control

Text control may not be enough → Often we like to generate *specific* instances

Instance images and image collections are quite common

- Tourist landmarks, Personal photo collections etc.



Research Question

How to adapt Text-to-Image models for

Instance-specific 2D and 3D generation?

Instance-specific 2D Generation

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

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

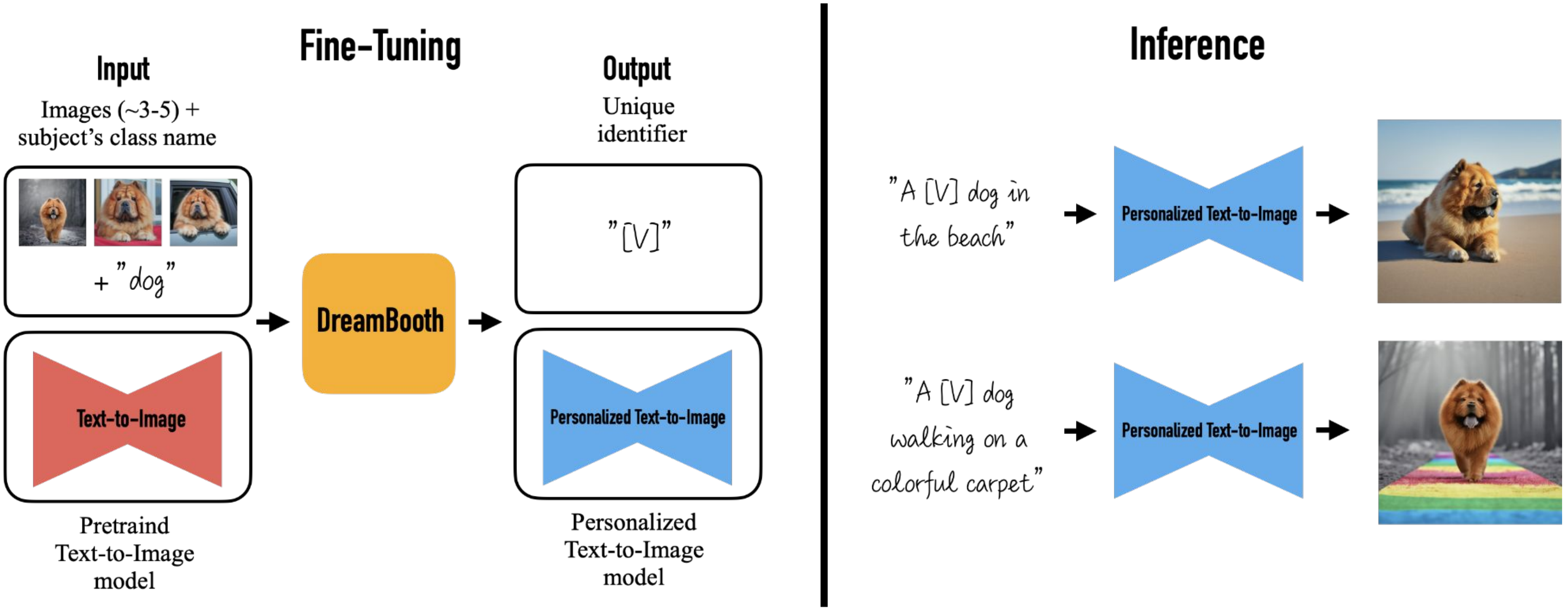
CVPR'23 Best Student Paper Honorable Mention

DreamBooth

Sample results of DreamBooth

1. Ruiz et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. CVPR 2023

DreamBooth - Simple Fine-tuning Approach



Sample DreamBooth Results

ZipLora: Any Subject in Any Style by Effectively Merging LoRAs

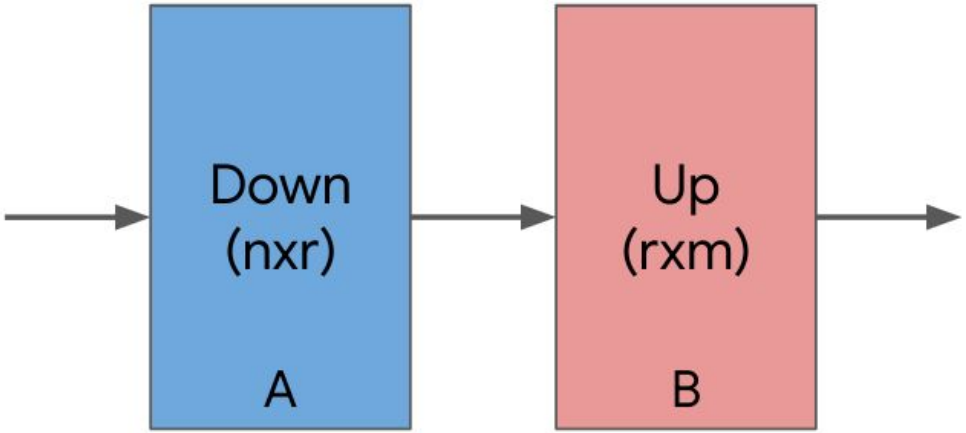
Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik,
Yuanzhen Li, Varun Jampani

ECCV'24

LoRA DreamBooth

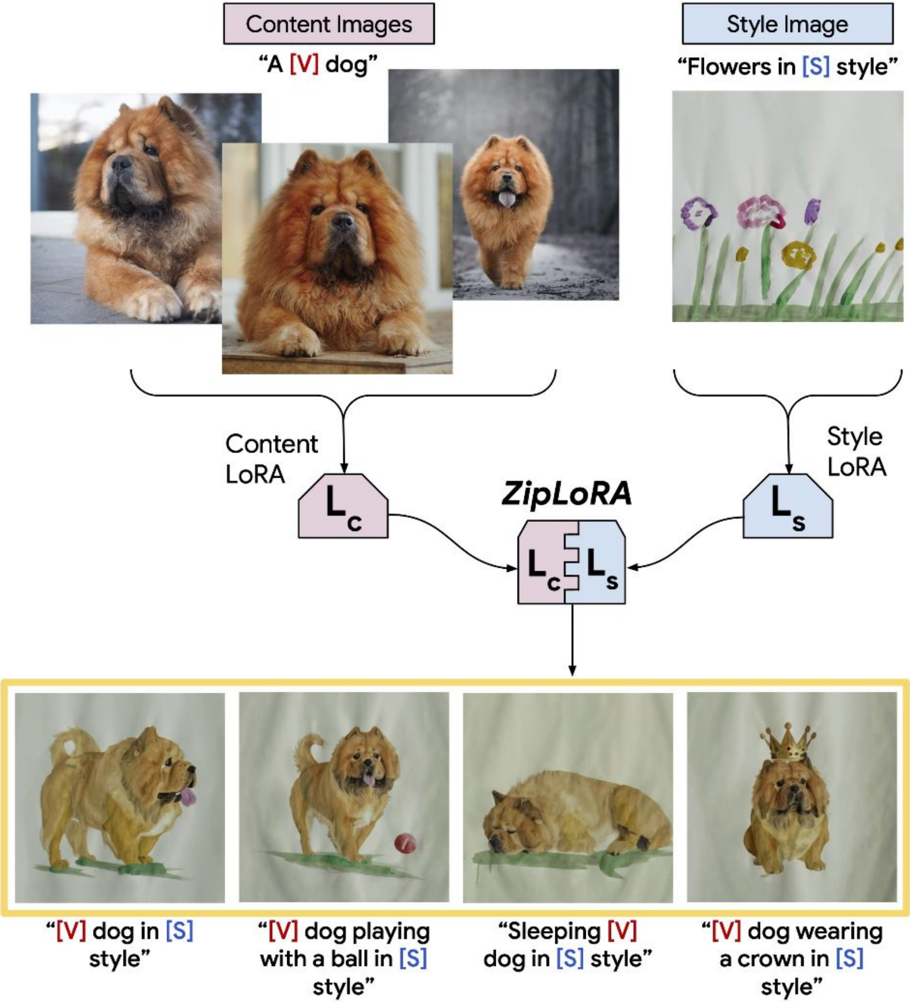
Decompose the network weight difference matrices into low-rank matrices

Finetune the network weights in this lower dimensional space



ZipLoRA

Generates any subject in any style by effectively combining the content and style LoRAs



Observation-1: StyleLoRA on SDXL works well

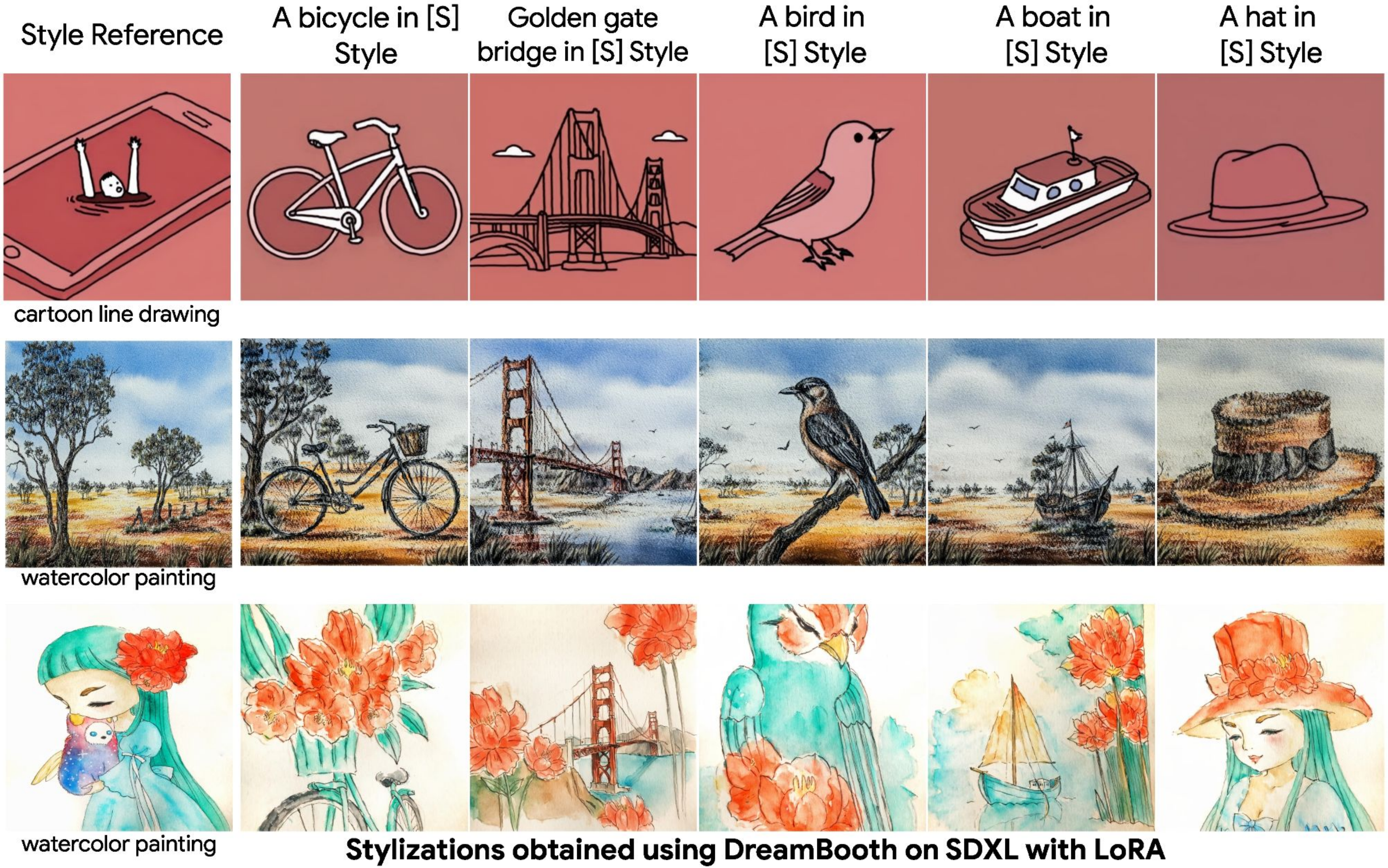
LoRA on SD [1]

LoRA on SDXL [2]

A statue in [s] style
[s]: ‘matte black sculpture’

A cat in [s] style
A man in [s] style
Old lay in [s] style

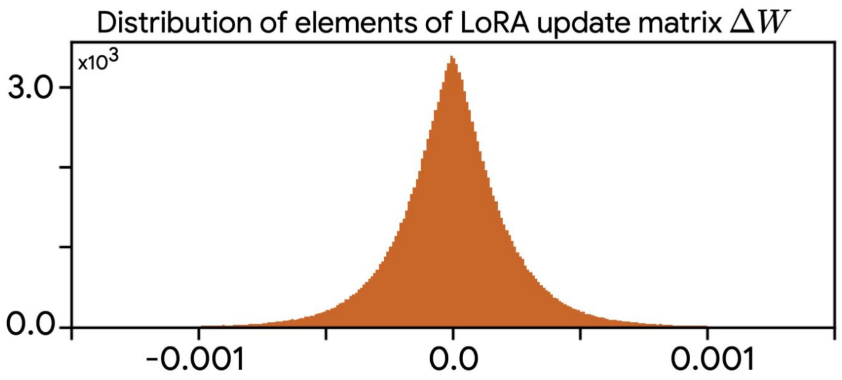
1. Rombach et al. High-resolution image synthesis with latent diffusion models. *CVPR 2022*
2. Podell et al. SDXL: Improving latent diffusion models for high-resolution image synthesis. *2023*

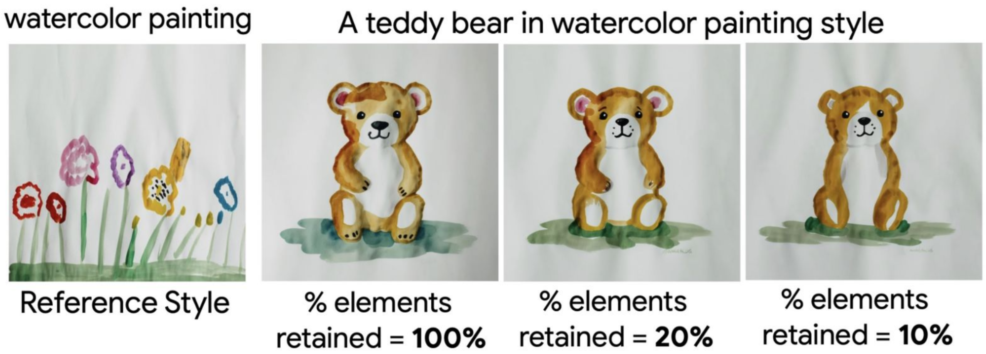


Stylizations obtained using DreamBooth on SDXL with LoRA

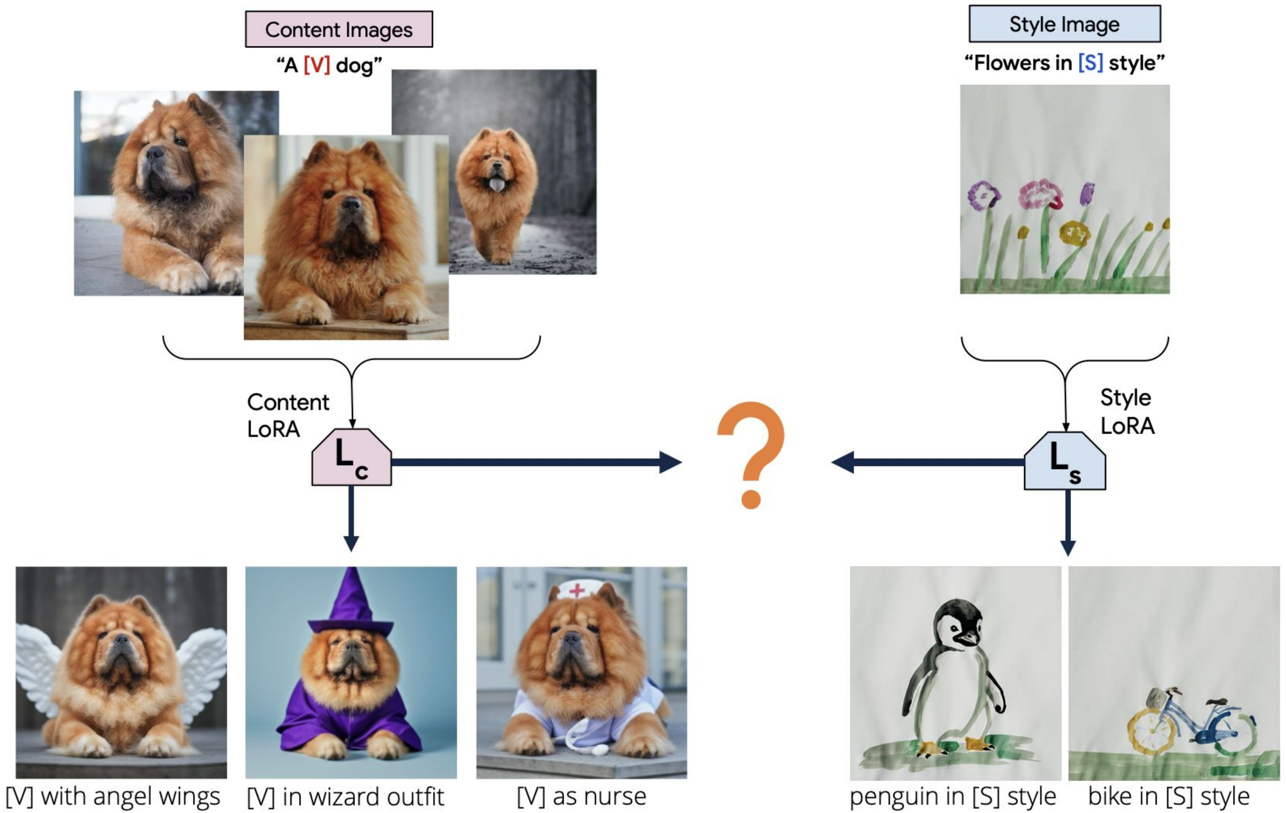
Observation-2: LoRA weights matrices are sparse

Most elements have small magnitude and have little effect on generation

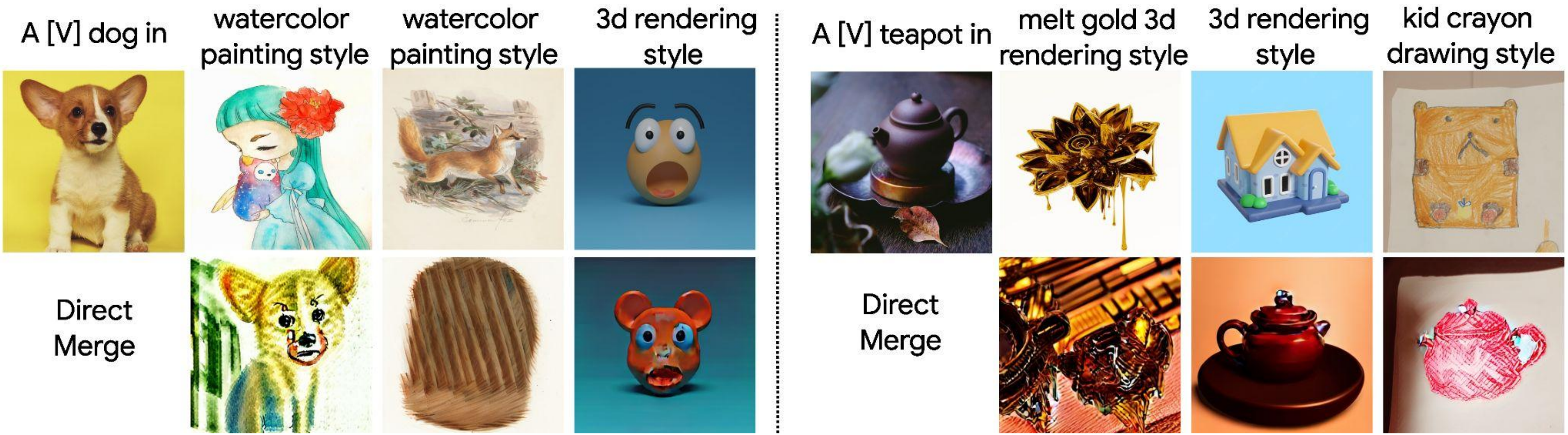




Merging *Subject* and *Style* LoRAs



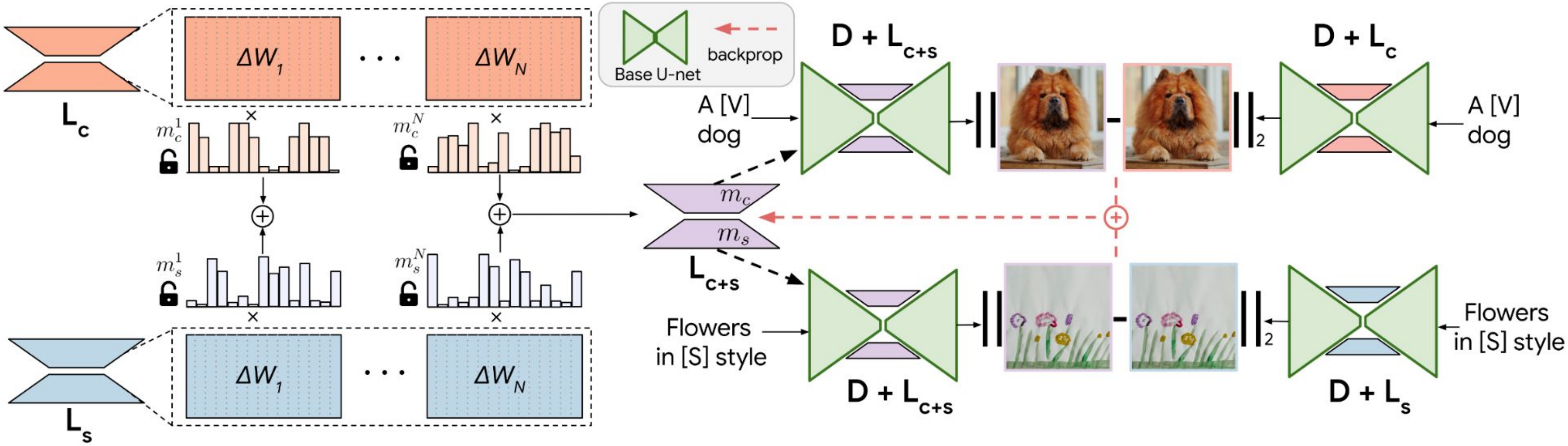
Simple LoRA addition does not work



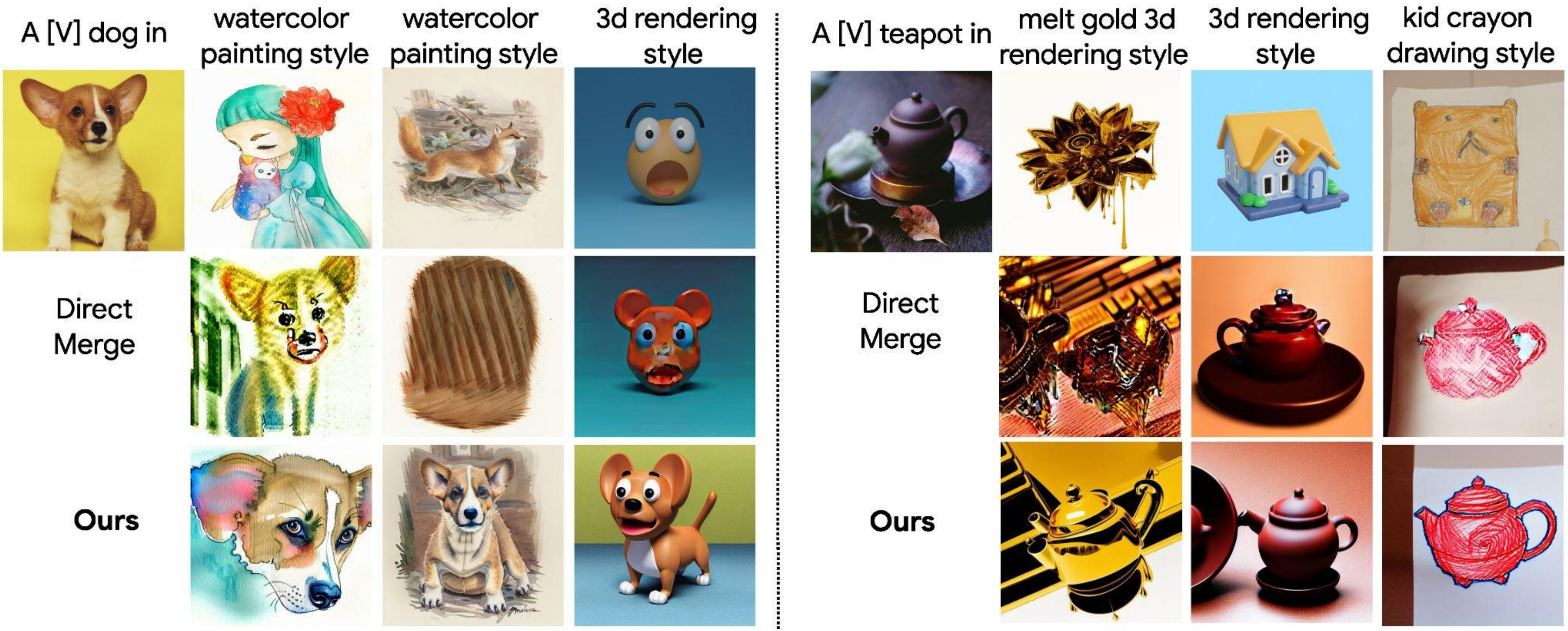
ZipLoRA Approach

Optimize layer-wise coefficients to merge

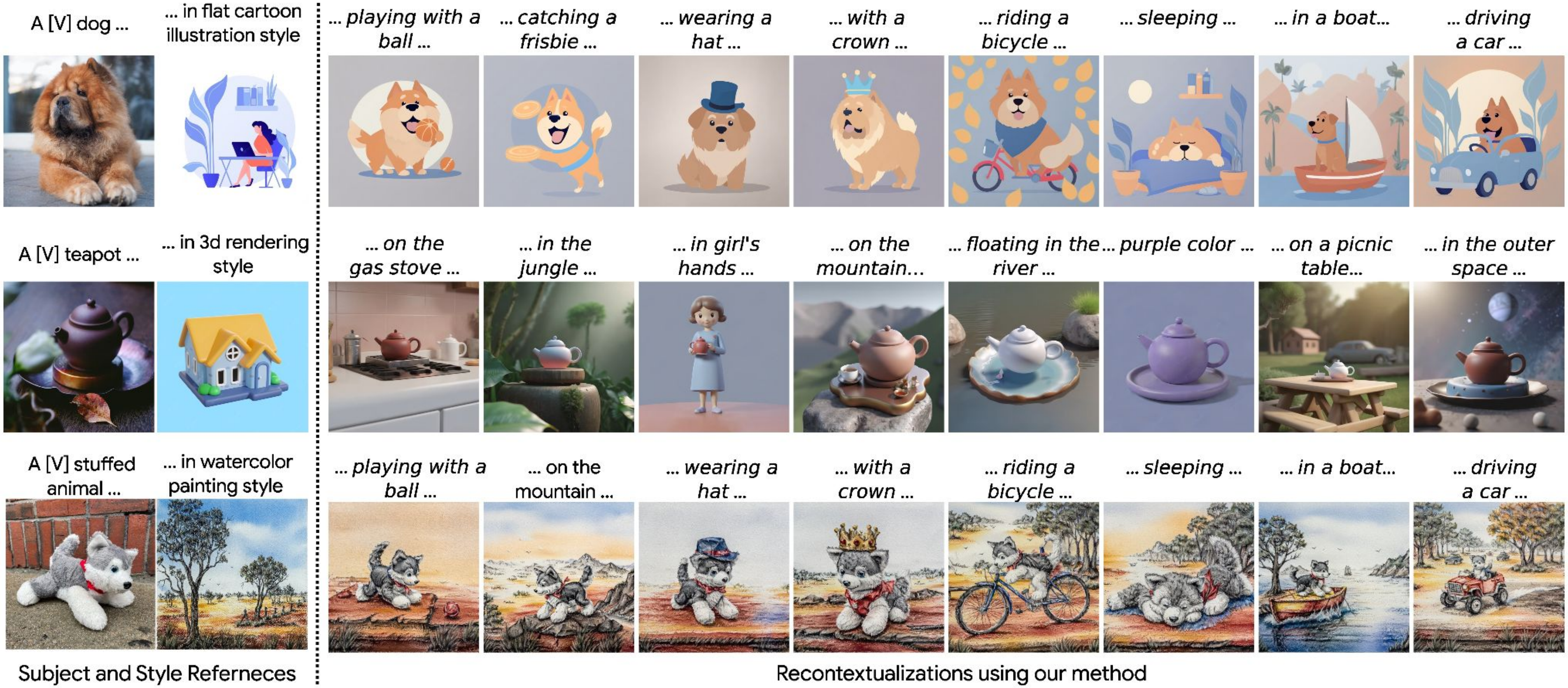
Preserve as much as original subject and style LoRAs



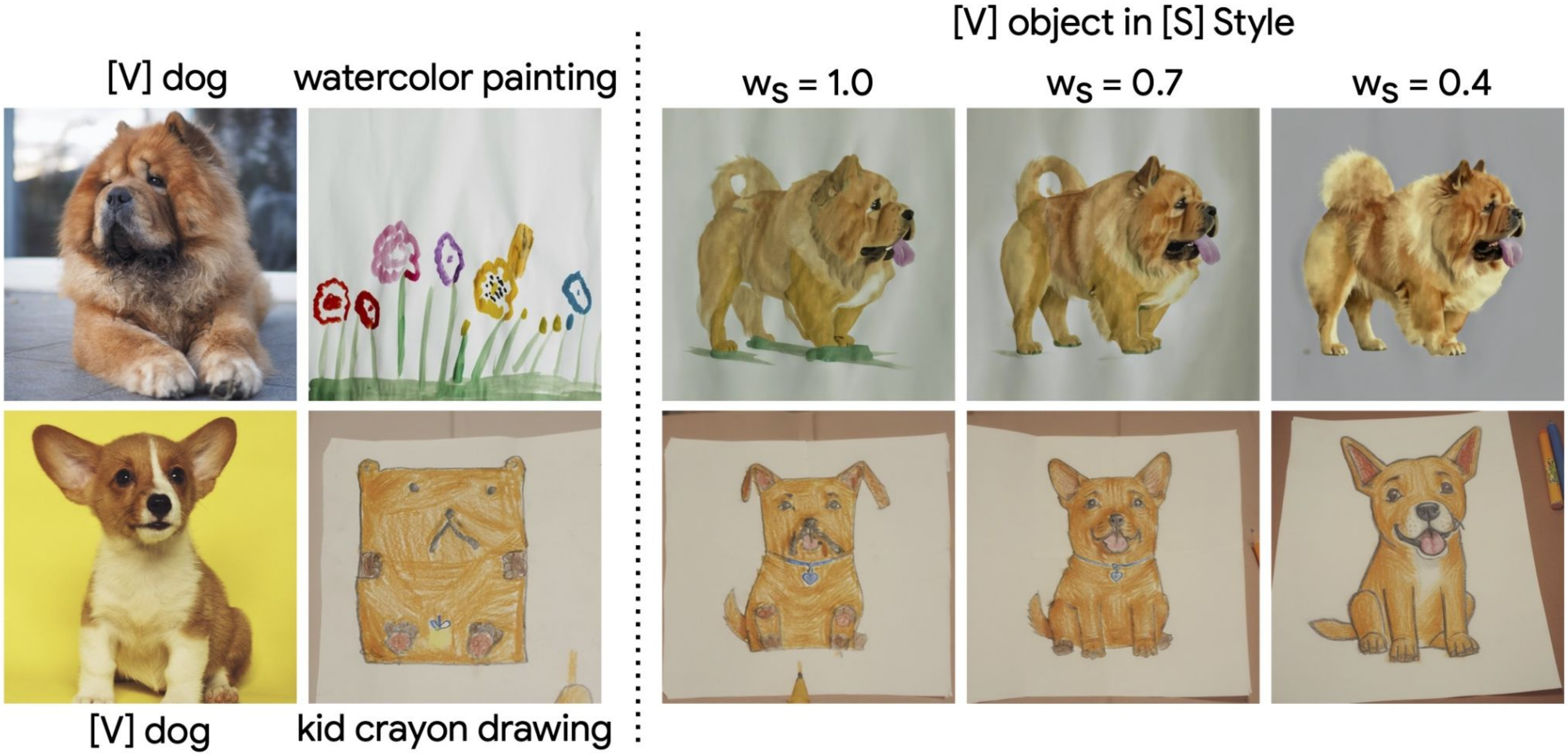
Comparison to Direct Merge



Sample Results



Control the extent of stylization



ZipLoRA: Summary

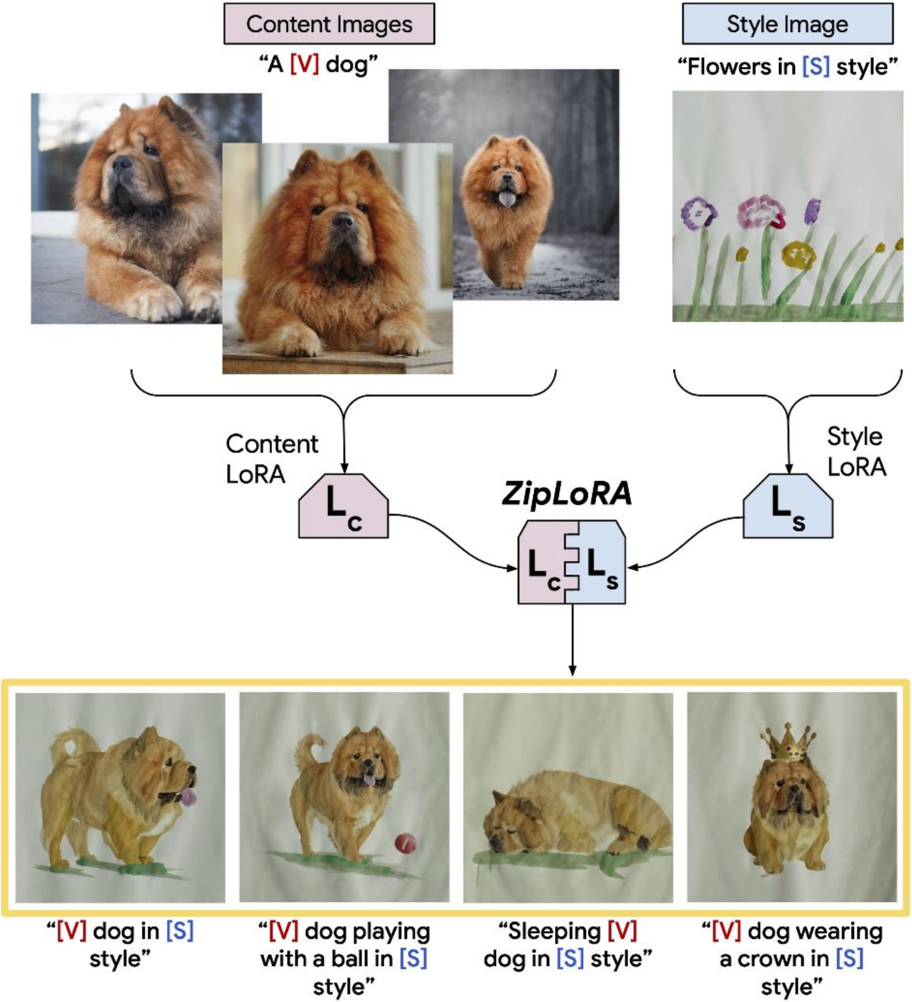
Style LoRA on SDXL works well

LoRA weight matrices are sparse

ZipLoRA proposes a simple yet effective LoRA merging strategy to generate any subject in any style

Outlook:

- Learn multiple subjects as well as multiple styles in a single network



Instance 2D Editing

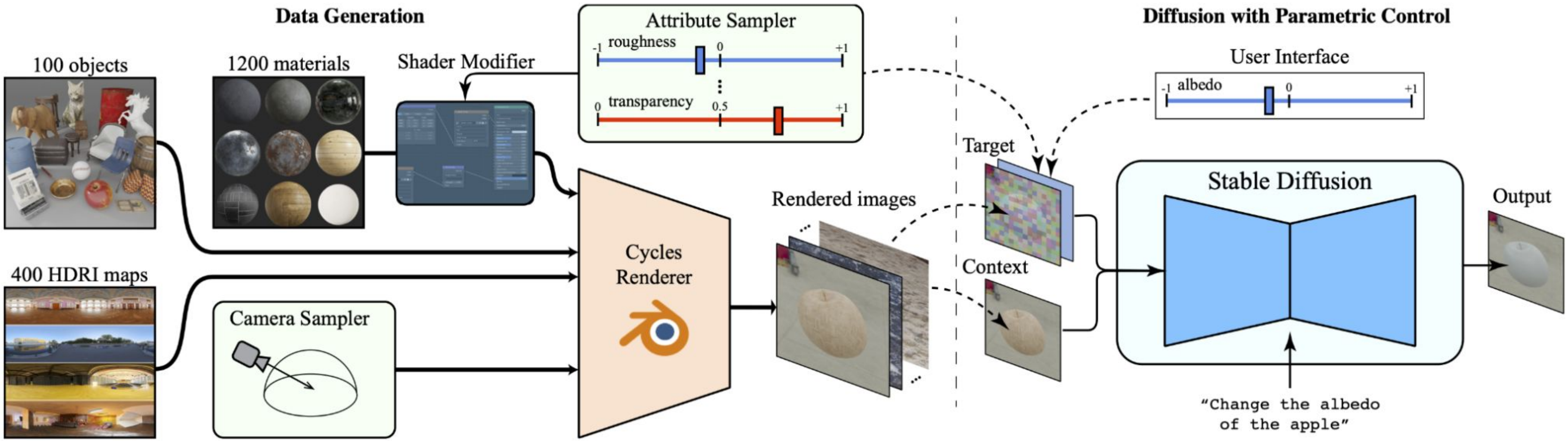
Alchemist: Parametric Control of Material Properties with Diffusion Models

Prafull Sharma, Varun Jampani, Yuanzhen Li, Xuhui Jia, Dmitry Lagun, Fredo Durand, William T. Freeman, Mark Matthews
[CVPR’24 oral]

Stable diffusion can be fine-tuned for precise material control in images

Alchemist: Parametric Control of Material Properties with Diffusion Models

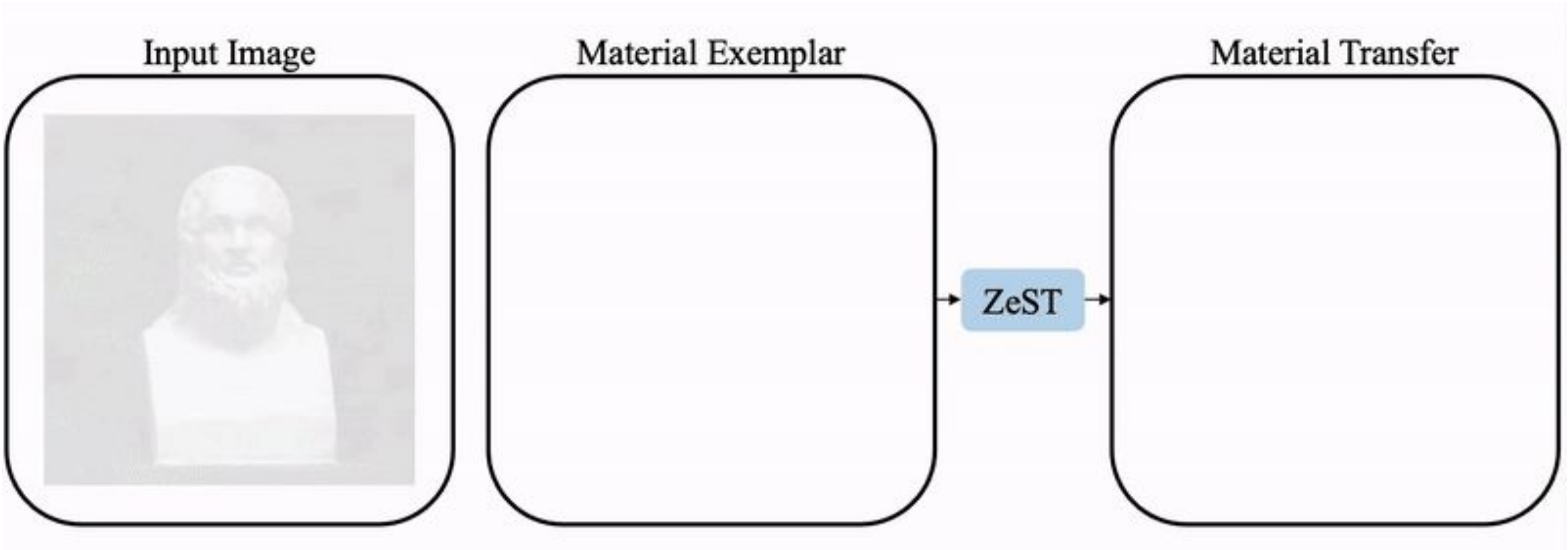
Prafull Sharma, Varun Jampani, Yuanzhen Li, Xuhui Jia, Dmitry Lagun, Fredo Durand, William T. Freeman, Mark Matthews
[CVPR'24 oral]



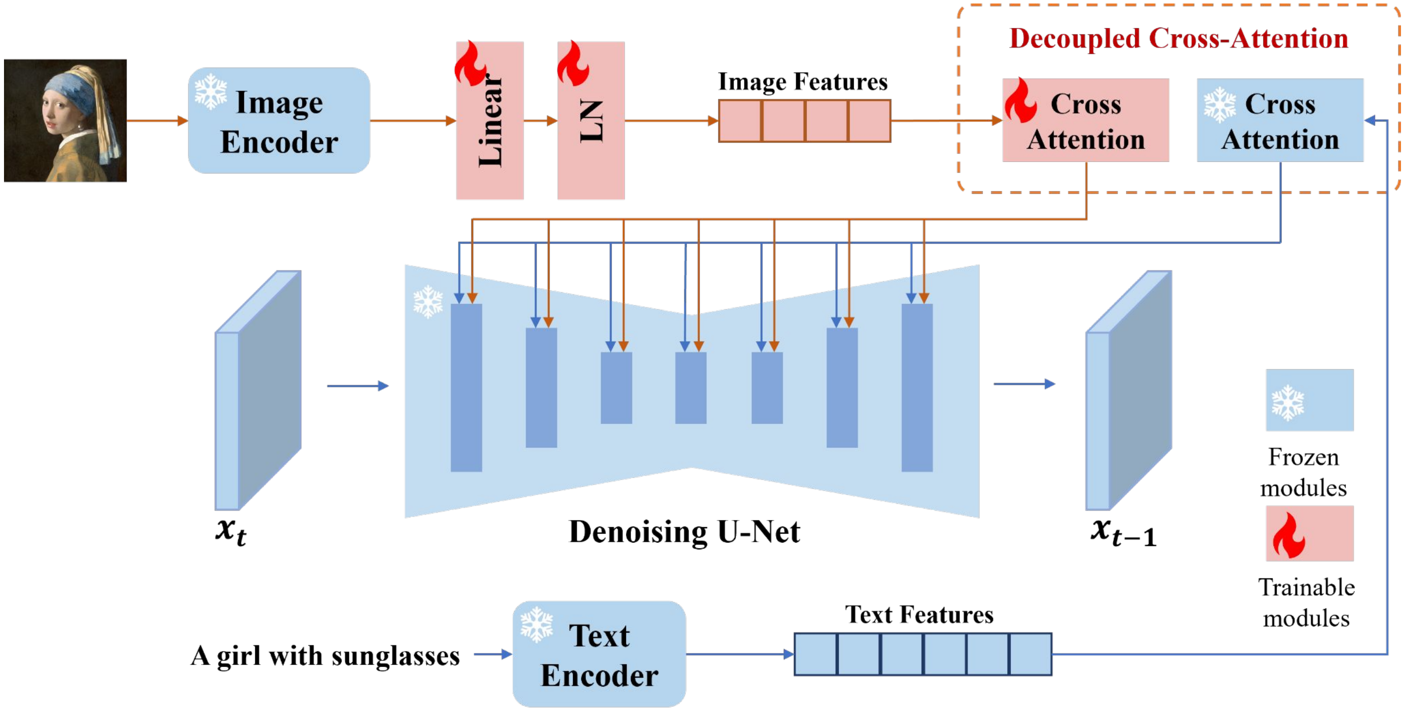
ZeST: Zero-shot Material Transfer from a Single Image

Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik,
Yuanzhen Li, Varun Jampani

ECCV'24

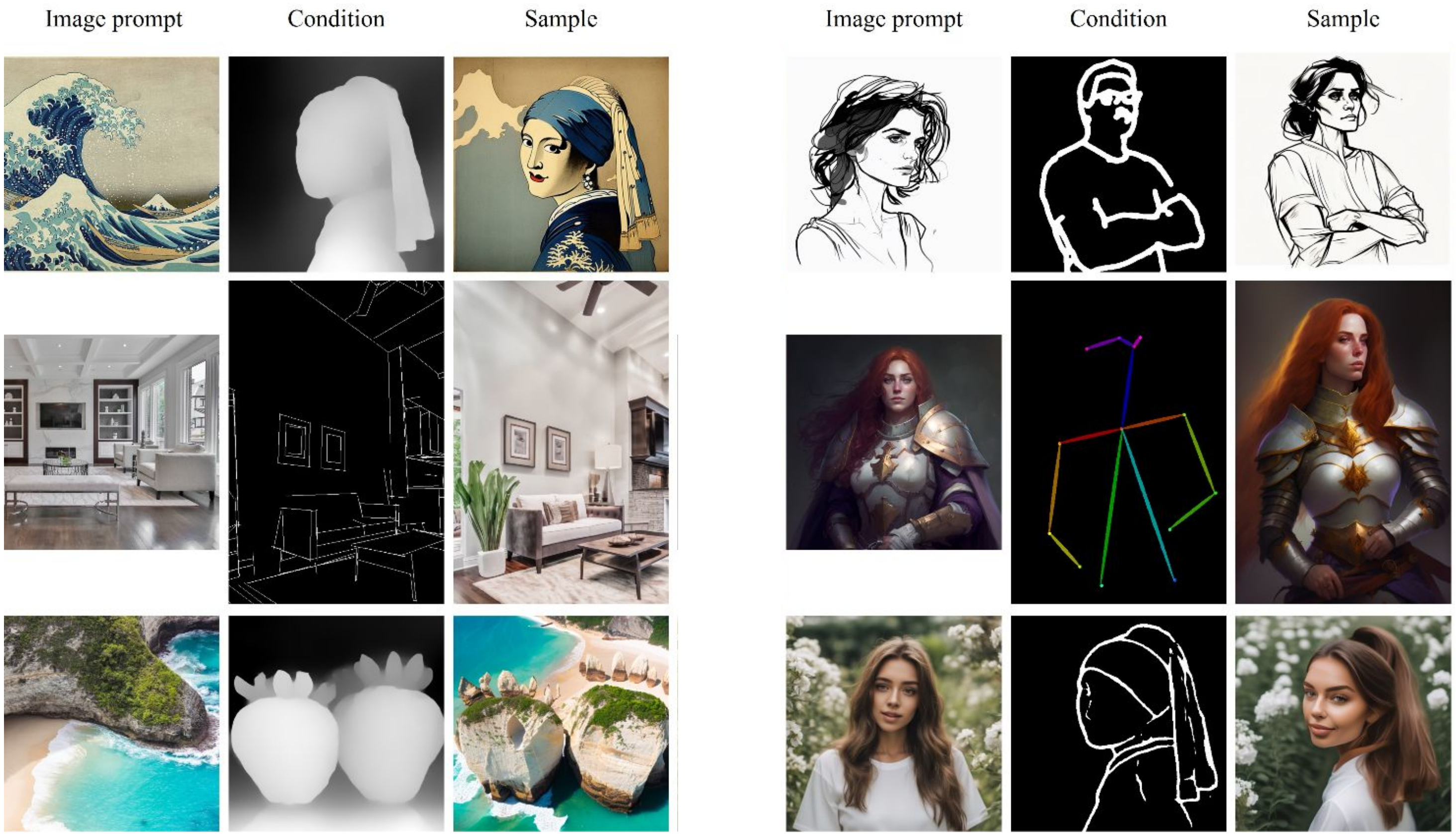

Input Image
Material Exemplar
Material Transfer
ZeST

IPAdapter - Using Images to Prompt Diffusion Models

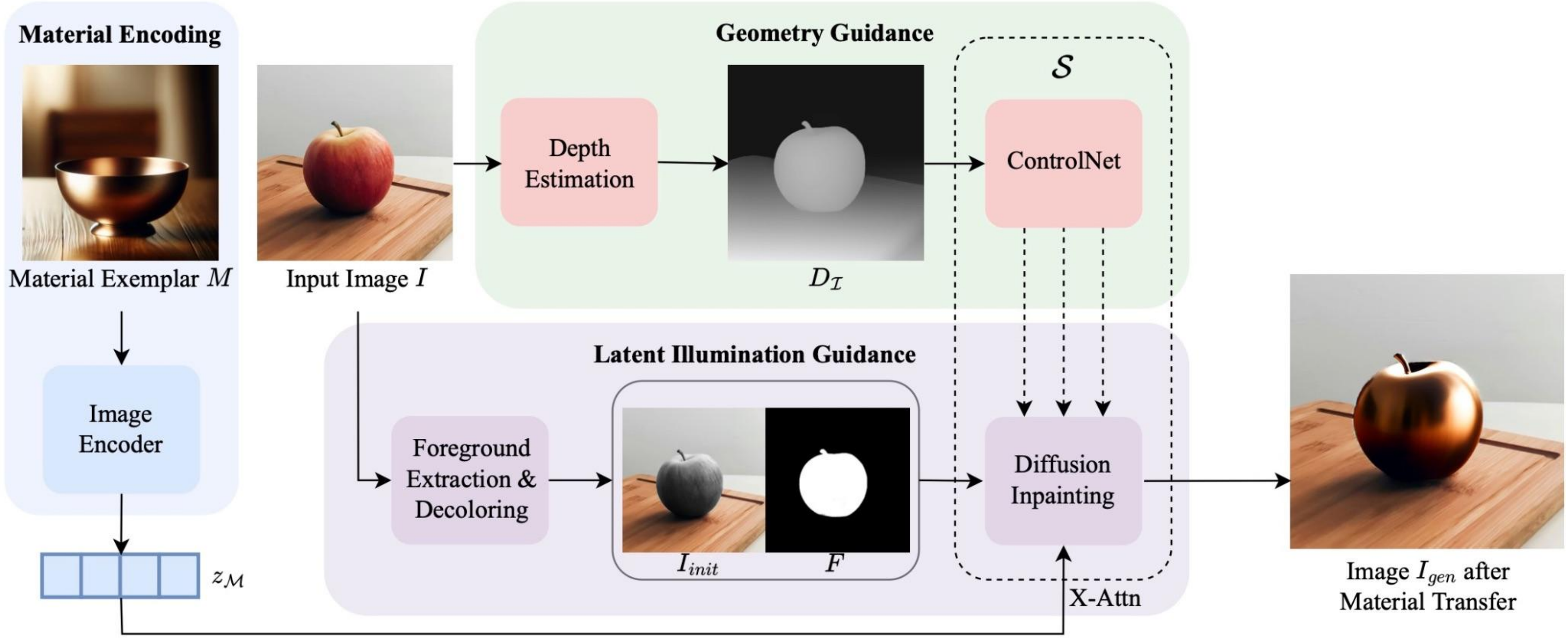


1. Ye et al. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv 2023.*

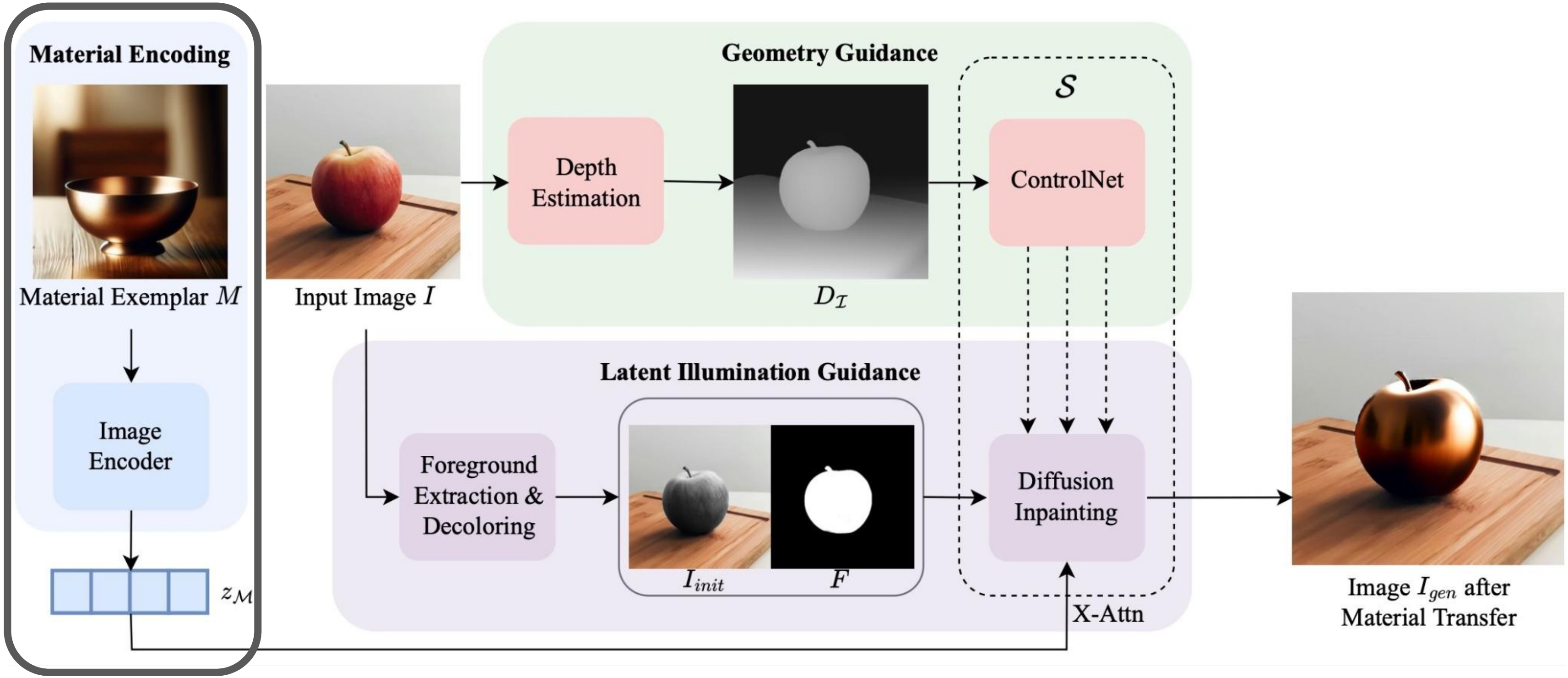
IP-Adapter Sample Results



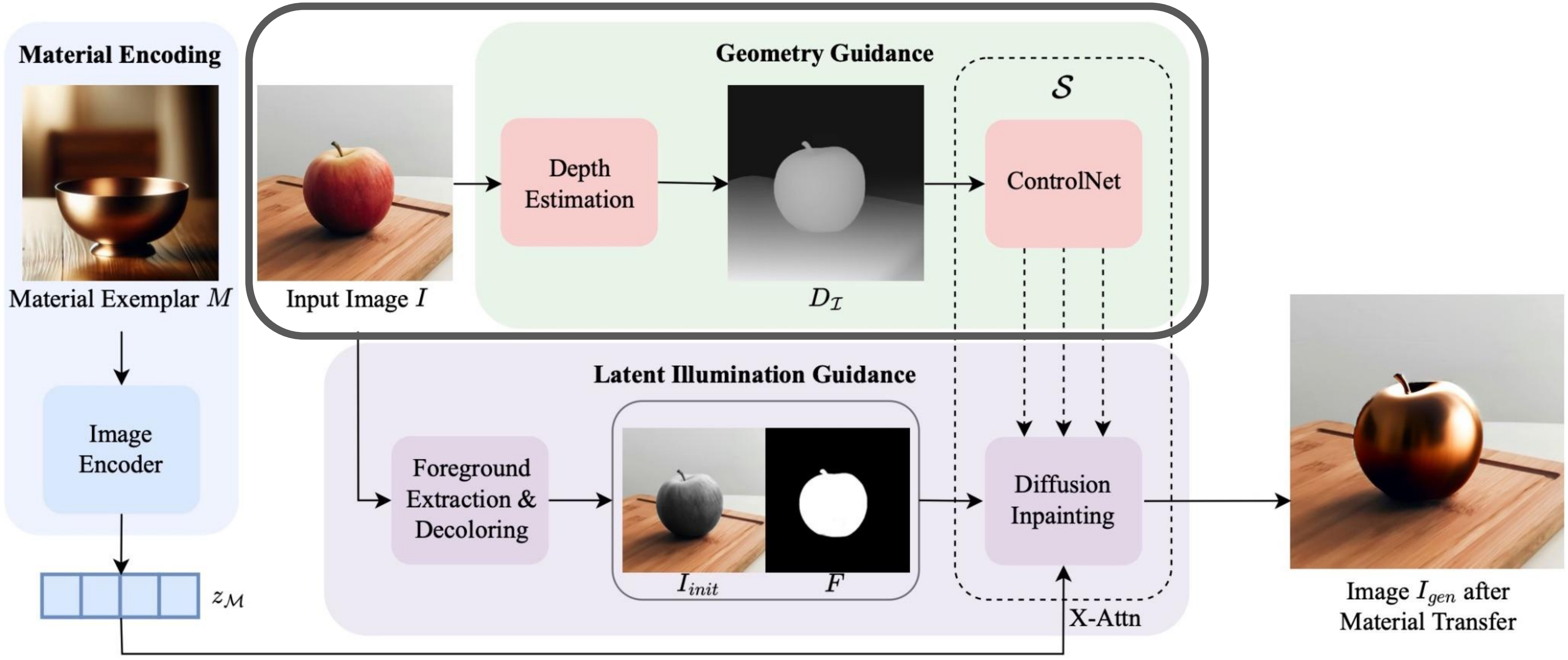
ZeST Architecture



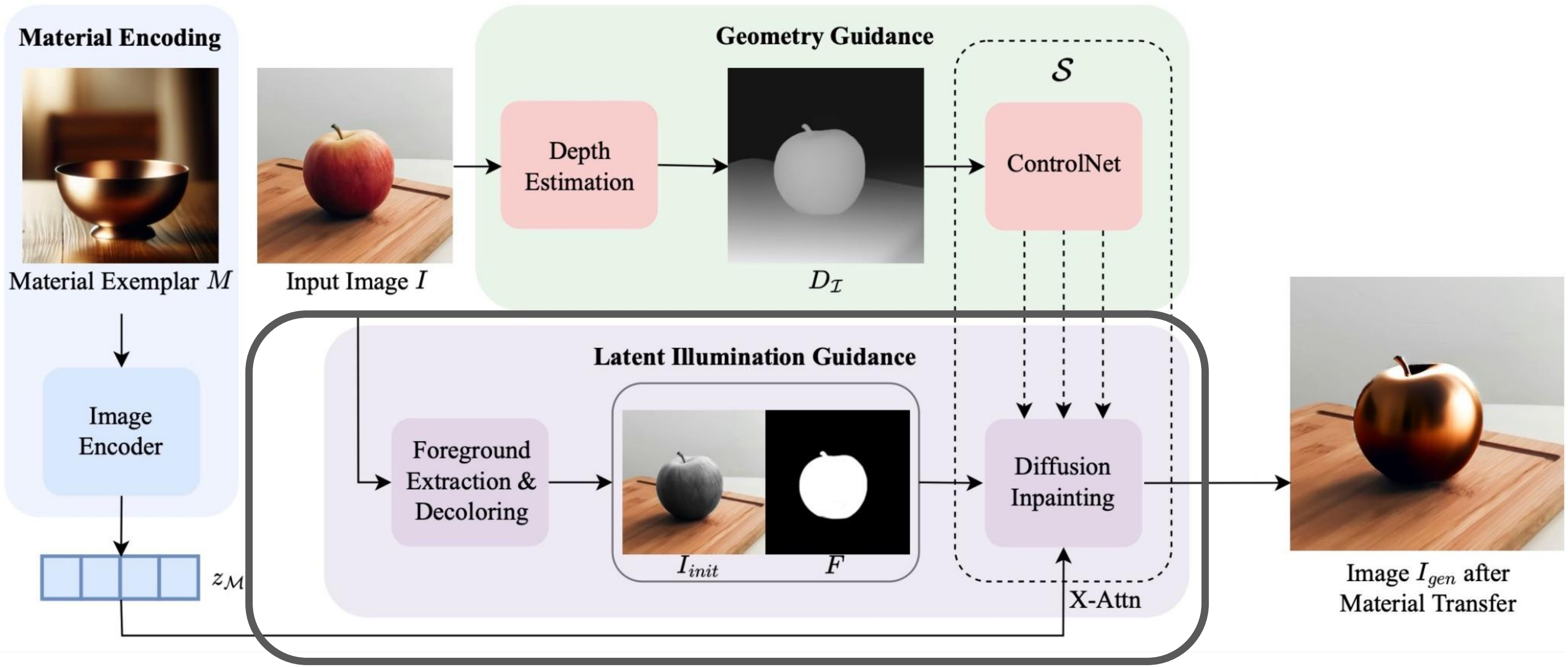
ZeST Architecture



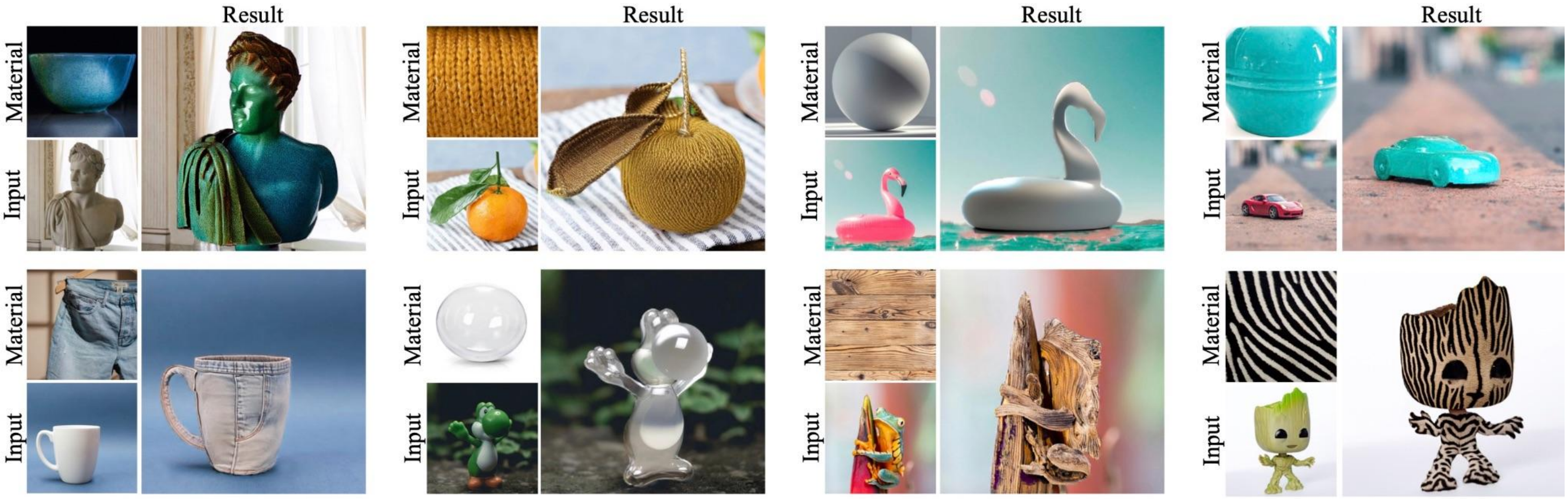
ZeST Architecture



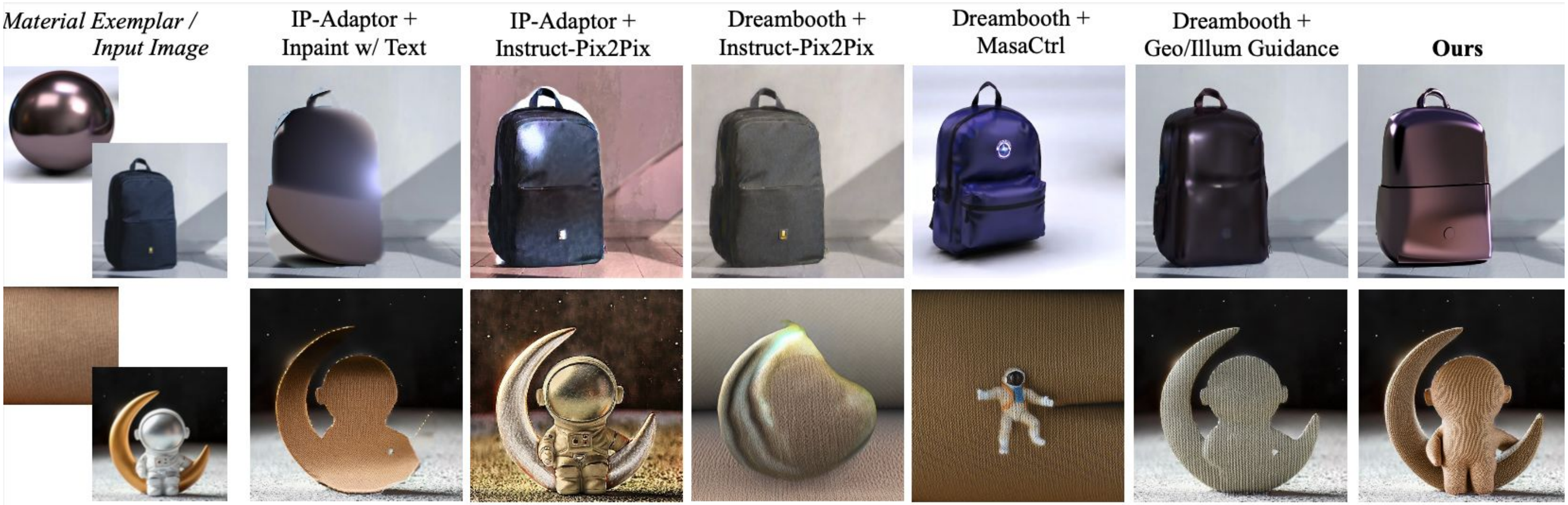
ZeST Architecture



ZeST Results



Comparison with Baselines



Multi-object Material Transfer

ZeST - Summary

IP-Adapter based diffusion inpainting works well for material transfer

Bypasses the complex task of explicit 3D shape and material estimation

Outlook:

- Control over different materials aspects (Base color, Specularity, roughness etc.)
- Going beyond images - Video and 3D material transfer

Instance-specific 3D Generation

Problem Setting

Input (3-6 casual captures)

Text Prompt: A photo of a dog

Output: 3D NeRF model

Subject-Driven Image Generation

- DreamBooth [1], Textual Inversion [2] etc.

Sample results of DreamBooth

1. Ruiz et al. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. CVPR 2023
2. Gal et al. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. ICLR 2023

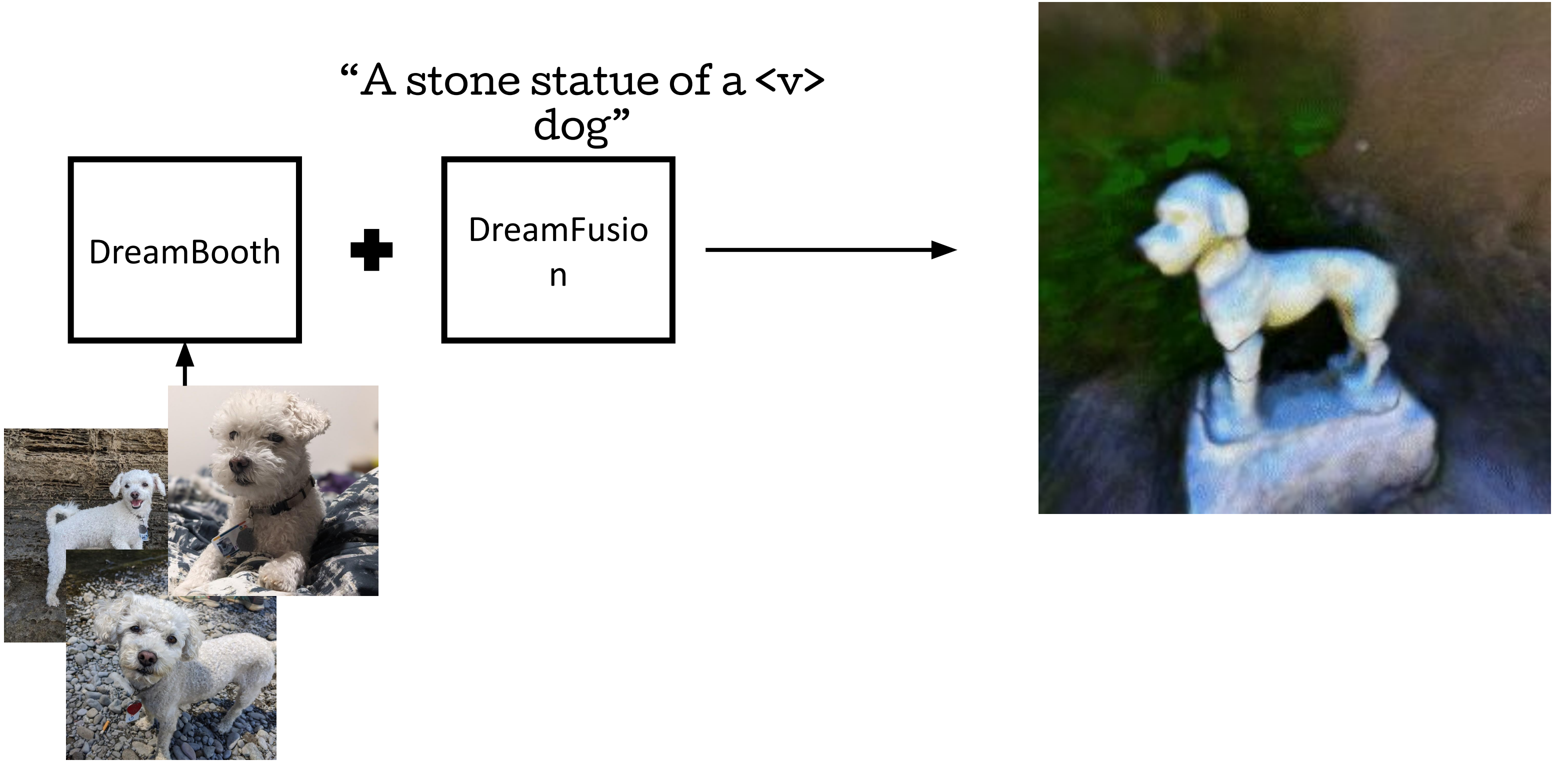
Text-to-3D Generation with T2I Models

- DreamFusion [1], Latent-NeRF [2], Magic3D [3]

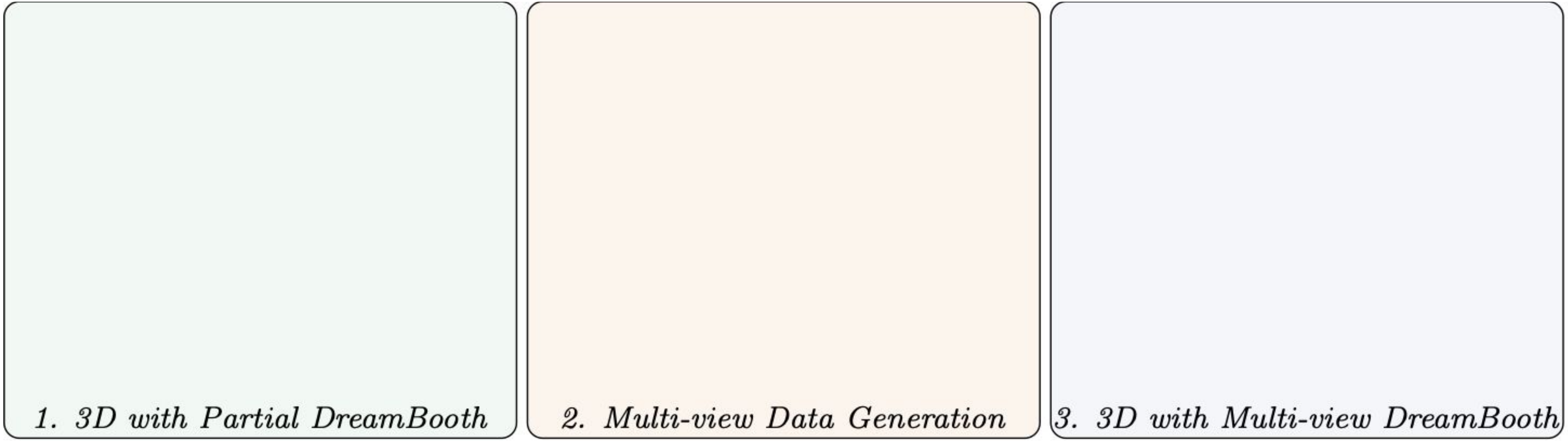
Sample results of DreamFusion

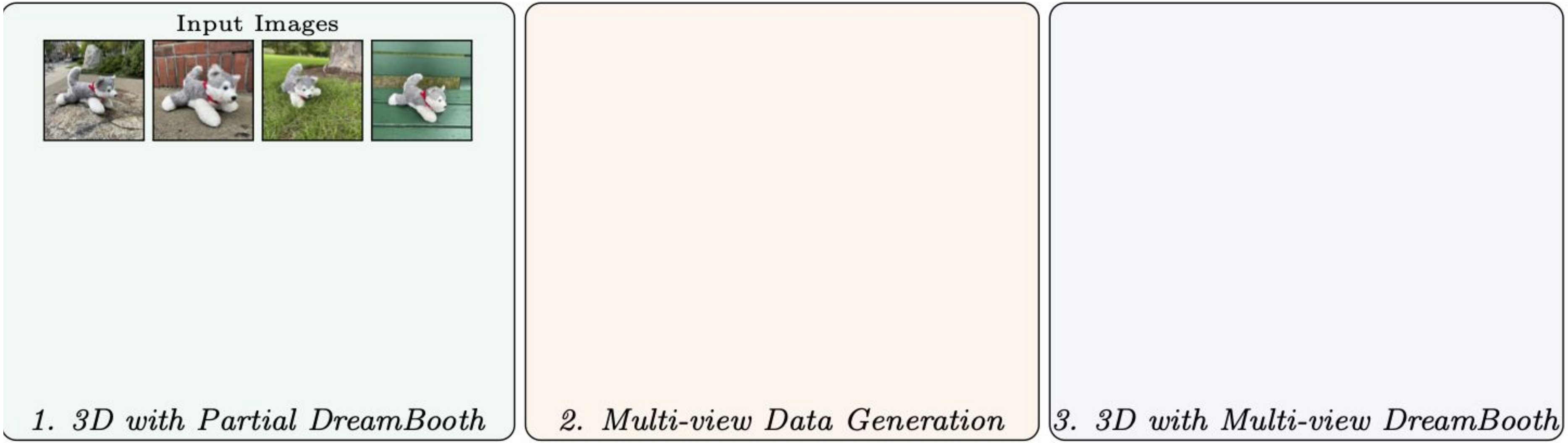
1. Poole et al. DreamFusion: Text-to-3D using 2D diffusion. ICLR 2023
2. Metzer et al. Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures. arXiv 2022
3. Lin et al. Magic3D: High-Resolution Text-to-3D Content Creation. arXiv 2022

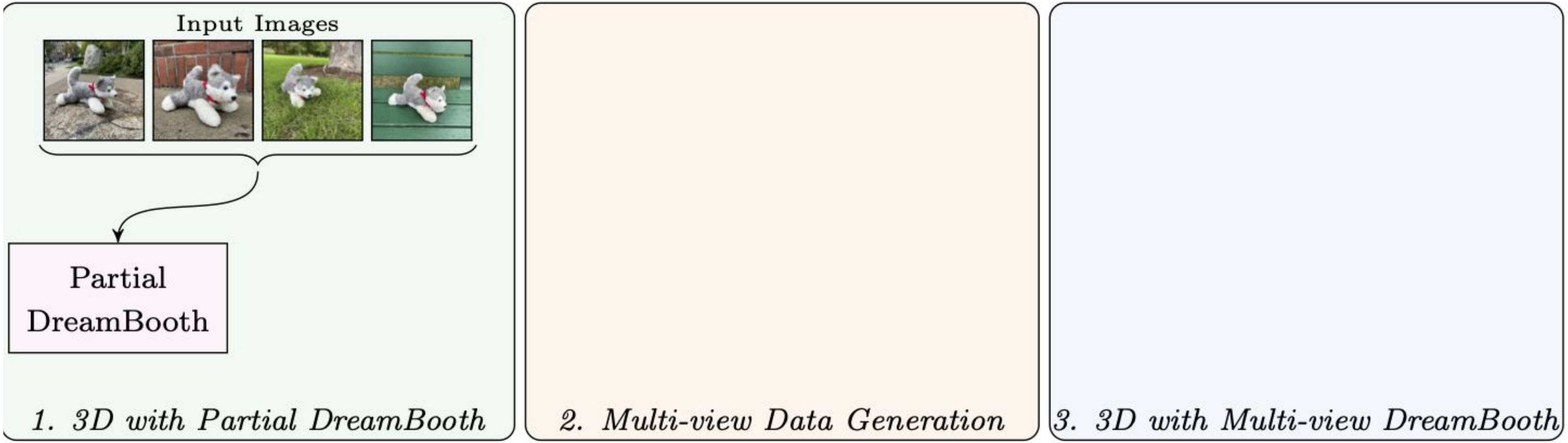
DreamBooth + DreamFusion

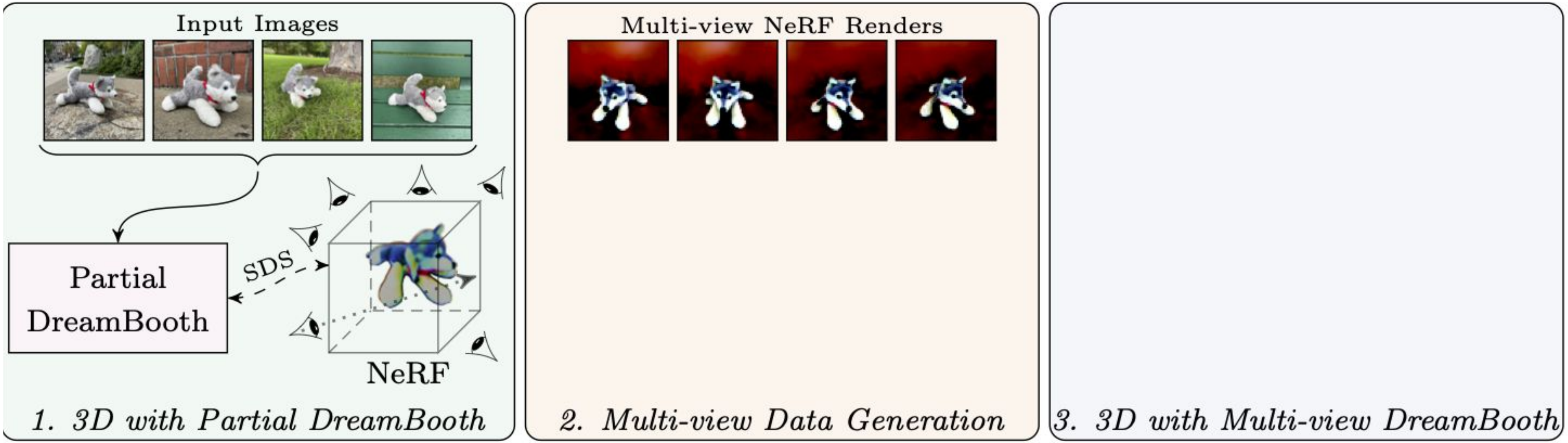


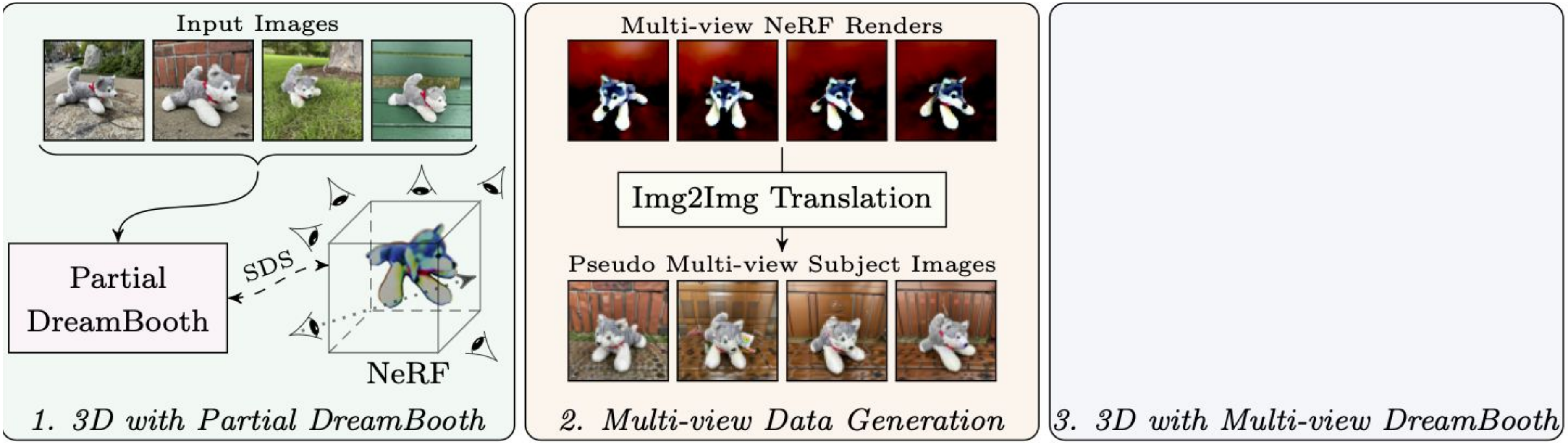
Approach

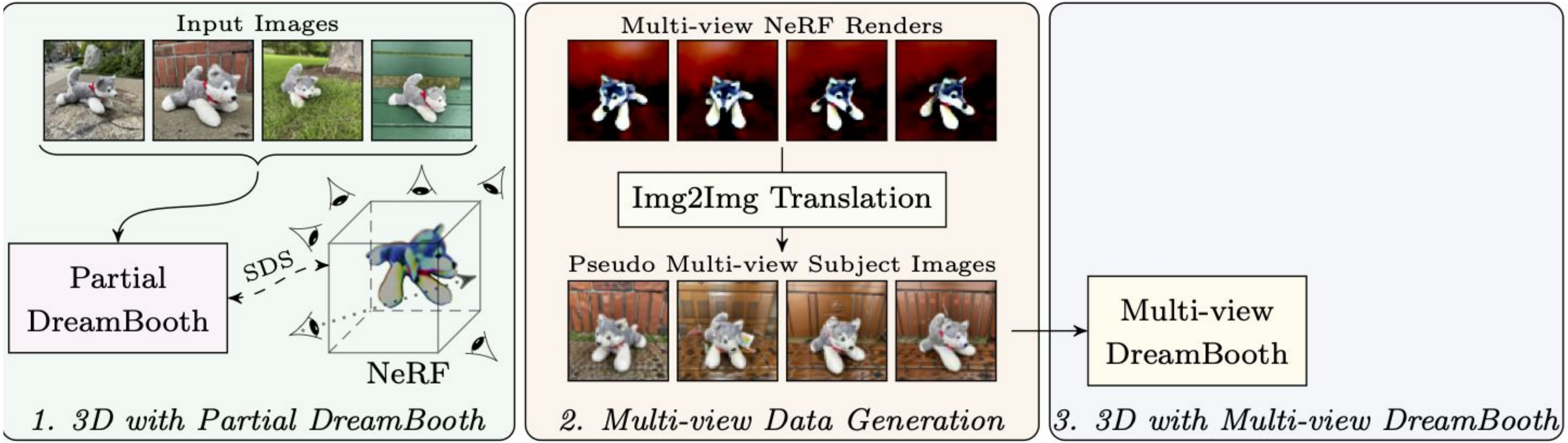

1. 3D with Partial DreamBooth
2. Multi-view Data Generation
3. 3D with Multi-view DreamBooth

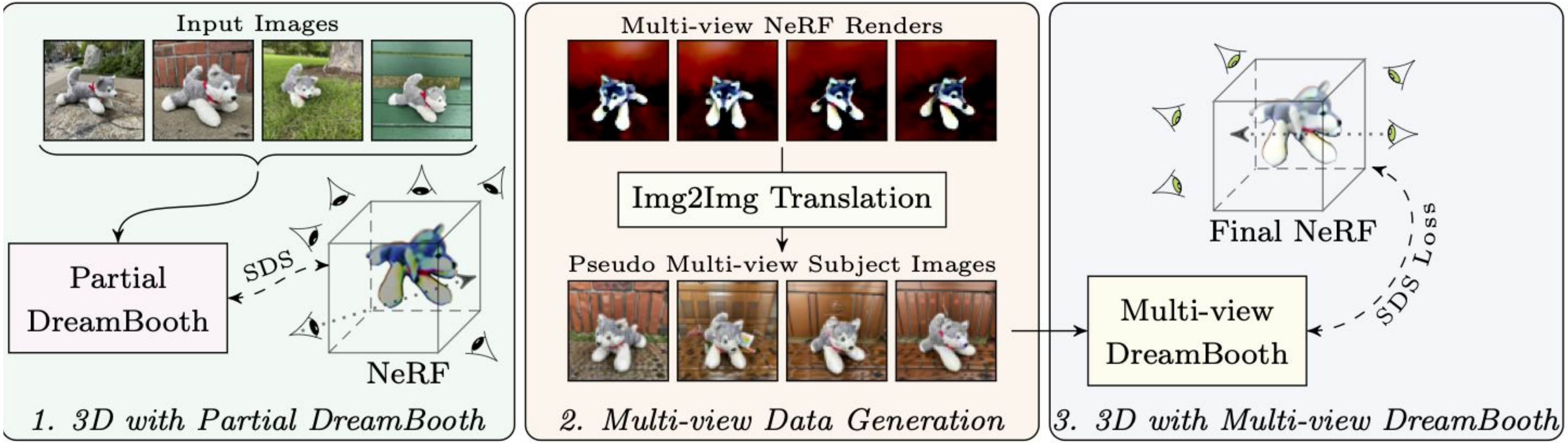

Input Images
1. 3D with Partial DreamBooth
2. Multi-view Data Generation
3. 3D with Multi-view DreamBooth


Input Images
Partial
DreamBooth
1. 3D with Partial DreamBooth
2. Multi-view Data Generation
3. 3D with Multi-view DreamBooth

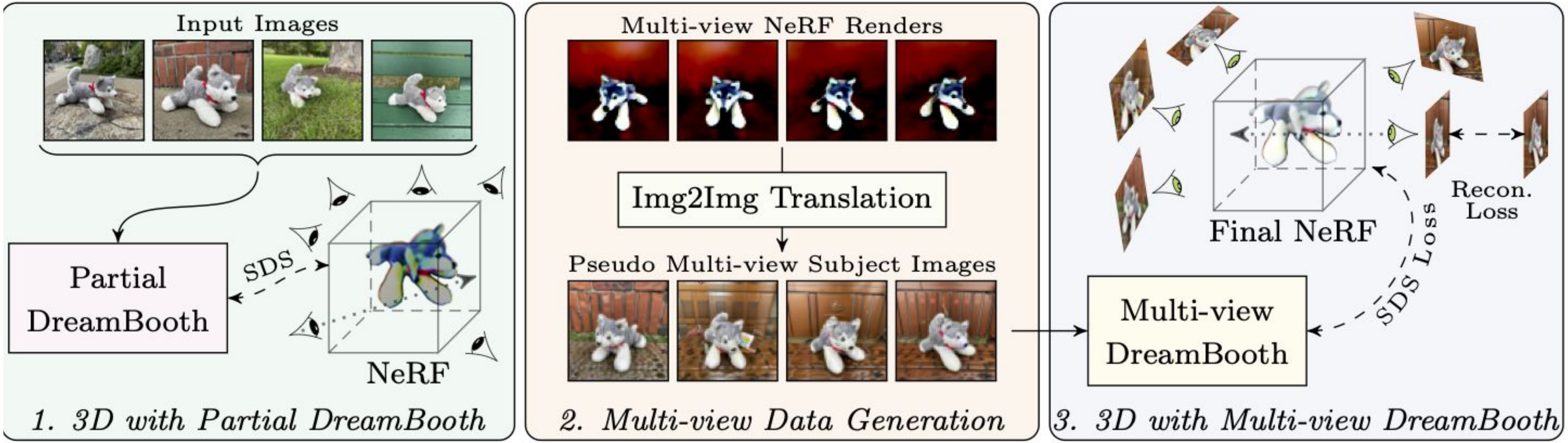

Input Images
Partial
DreamBooth
SDS
NeRF
1. 3D with Partial DreamBooth
Multi-view NeRF Renders
2. Multi-view Data Generation
3. 3D with Multi-view DreamBooth


Input Images
Partial DreamBooth
SDS
NeRF
1. 3D with Partial DreamBooth
Multi-view NeRF Renders
Img2Img Translation
Pseudo Multi-view Subject Images
2. Multi-view Data Generation
3. 3D with Multi-view DreamBooth


Input Images
Partial
DreamBooth
SDS
NeRF
1. 3D with Partial DreamBooth
Multi-view NeRF Renders
Img2Img Translation
Pseudo Multi-view Subject Images
2. Multi-view Data Generation
Multi-view
DreamBooth
3. 3D with Multi-view DreamBooth



Input Images
Partial DreamBooth
SDS
NeRF
1. 3D with Partial DreamBooth
Multi-view NeRF Renders
Img2Img Translation
Pseudo Multi-view Subject Images
2. Multi-view Data Generation
Final NeRF
SDS Loss
Multi-view DreamBooth
3. 3D with Multi-view DreamBooth



Input Images
Partial DreamBooth
SDS
NeRF
1. 3D with Partial DreamBooth
Multi-view NeRF Renders
Img2Img Translation
Pseudo Multi-view Subject Images
2. Multi-view Data Generation
Final NeRF
Recon. Loss
SDS Loss
Multi-view DreamBooth
3. 3D with Multi-view DreamBooth

Results

Material Change

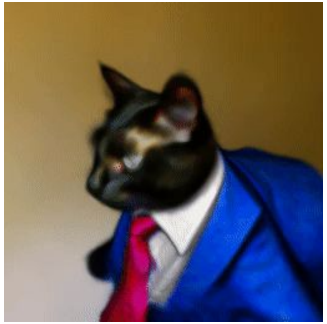
Input Exemplar | A photo of a <V> dog | *made of glass* | *A stone statue*

Accesorization

Input Exemplars

A photo of a <V> cat

with a tie



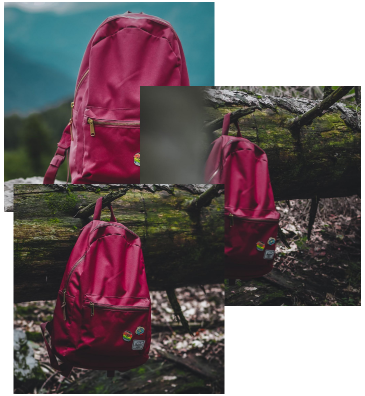
in a suit

A photo of a <V> dog

on a rainbow carpet

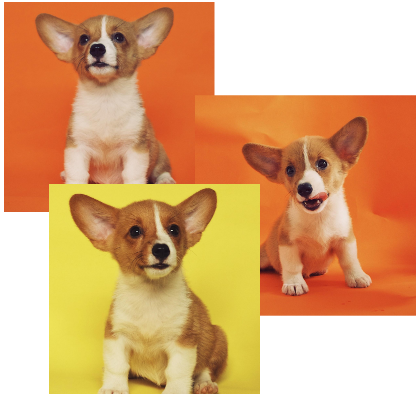
with a green umbrella

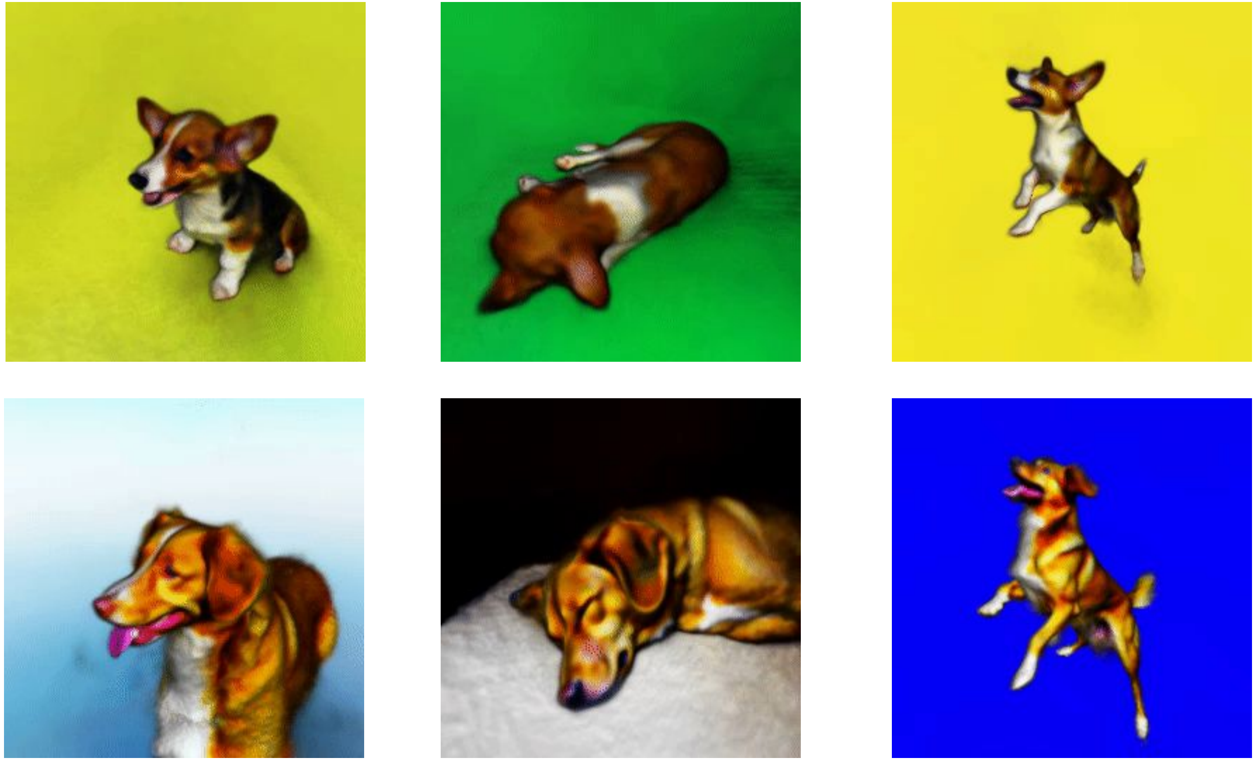
Color Change



A <V> pink backpack | **Green** | **Grey**

Pose change





Sitting Sleeping Jumping

DreamBooth3D - Summary

- Novel and effective 3-stage optimization scheme
- High subject-fidelity
- Versatile text based 3D editing

Concluding Remarks

Text-to-Image models can be effectively adapted for instance-specific 2D and 3D generation as well as manipulation

- Instance 2D generation (DreamBooth, ZipLoRA)
- Instance 2D editing (Alchemist, ZeST)
- Instance 3D generation (DreamBooth3D)

Outlook:

- Controllable instance specific 2D/3D generations where we can control different instance attributes.
- Using instance-specific generations for robust instance discrimination
  - Features, recognition, retrieval etc.

Thank You

Comments and suggestions are most welcome

varunjampani@gmail.com
varunjampani.github.io