Diffusion Models for Visual Computing

Paul Guerrero

Niloy Mitra, Daniel Cohen-Or, Minhyuk Sung, Chun-Hao Huang, Duygu Ceylan

Part 4: Personalization & Editing

https://geometry.cs.ucl.ac.uk/courses/diffusion4VC_eg24/

Presentation Schedule

Introduction to Diffusion Models

Guidance and Conditioning Sampling

Attention

Break

Personalization and Editing

Beyond Single Images

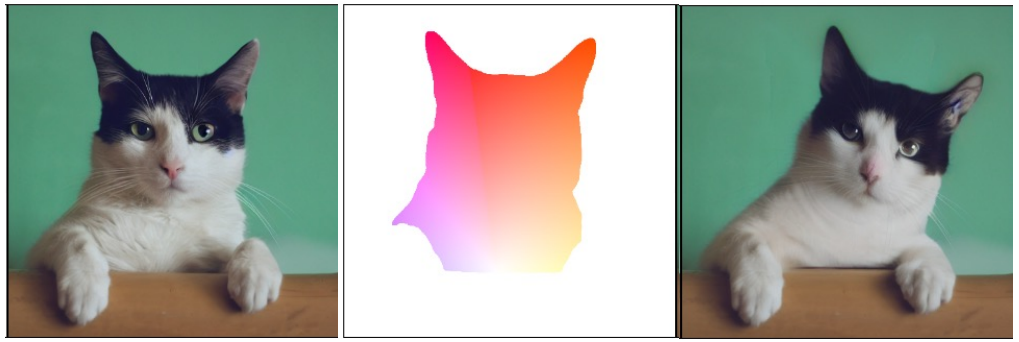
Diffusion Models for 3D Generation

Personalization

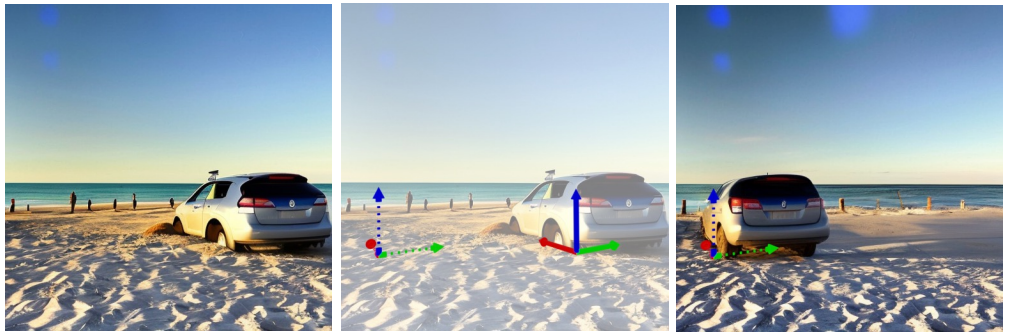
Input images

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, Ruiz et al., CVPR 2023

Editing



Motion Guidance: Diffusion-Based Image Editing with Differentiable Motion Estimators, Geng and Owens, ICLR 2024



Diffusion Handles: Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D Pandey et al., CVPR 2024

Personalization

"a hyper-realistic digital painting of a happy girl, with brown eyes"

Without Personalization

Generated with StabelDiffusion 2.1

With Personalization

ConsiStory: Training-Free Consistent Text-to-Image Generation
Tewel et al., ArXiv Feb. 2024

Personalization

With Personalization

Same subject in different settings.

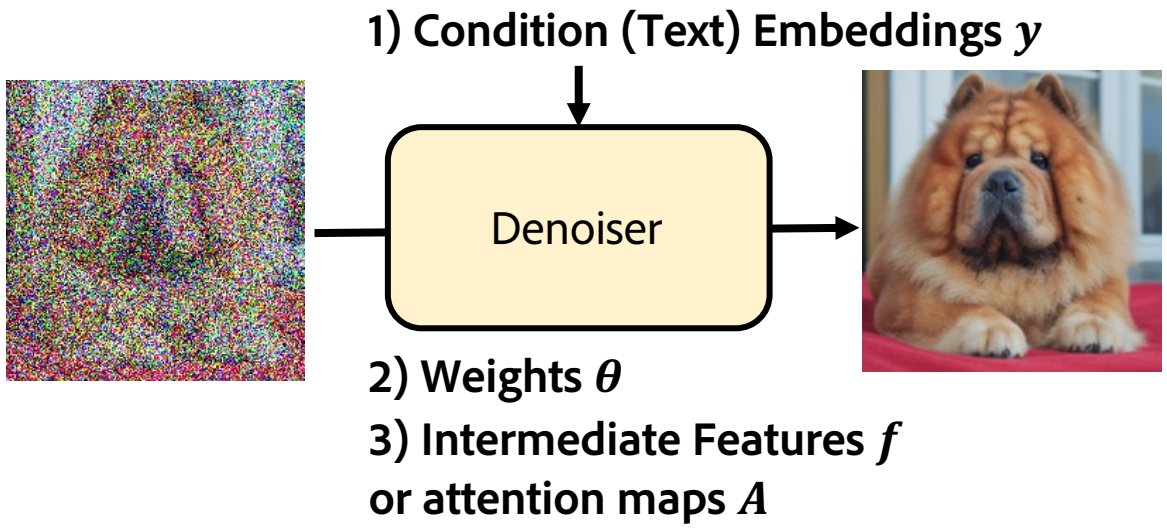
Personalization:

Generative Model
+ Identity Preservation

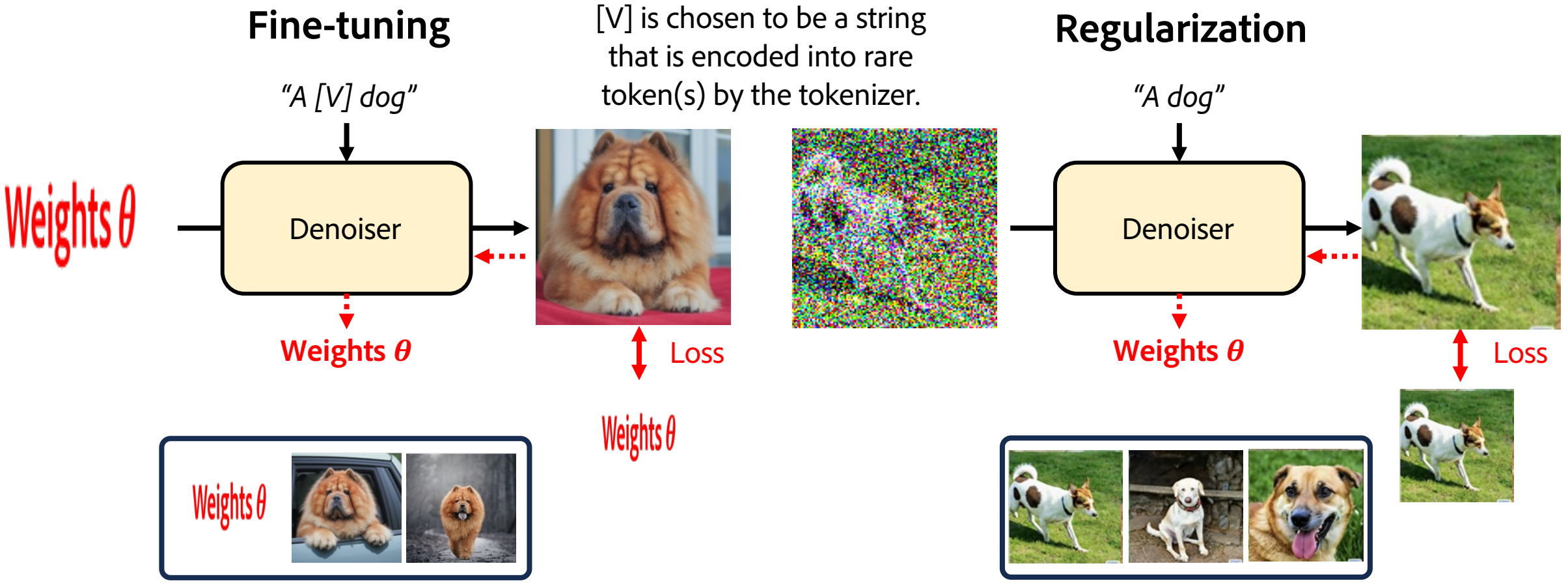
ConsiStory: Training-Free Consistent Text-to-Image Generation
Tewel et al., ArXiv Feb. 2024

Identity Preservation

How can we represent the identity of a subject?



ID Preservation by Fine-Tuning Denoiser Weights



3-5 example images showing the identity.

Regularization data can be generated by the original, non-finetuned model, or can come from a large dataset.

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, Ruiz et al., CVPR 2023

ID Preservation by Fine-Tuning Denoiser Weights

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, Ruiz et al., CVPR 2023

ID Preservation by Fine-Tuning Denoiser Weights

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation,
Ruiz et al., CVPR 2023

Code

https://github.com/google/dreambooth

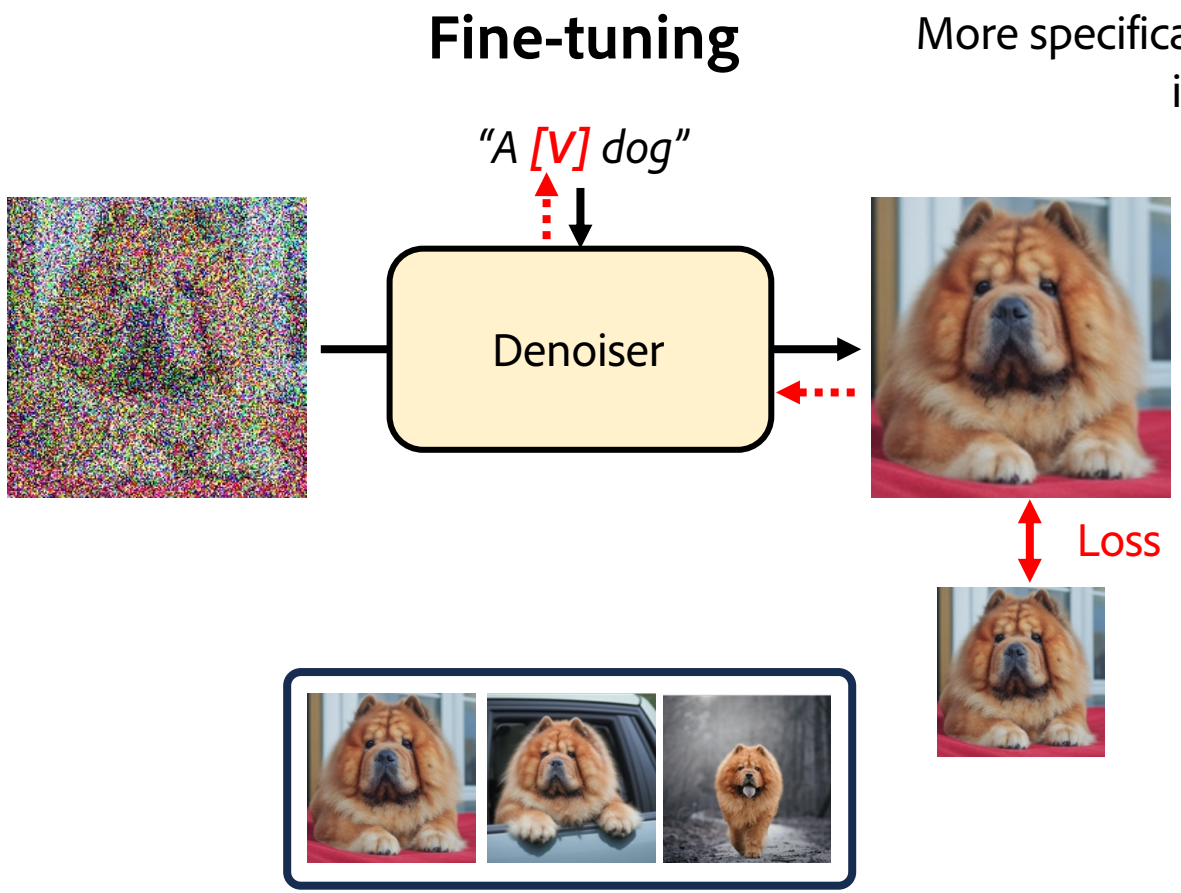
Code

https://huggingface.co/docs/diffusers/en/training/dreambooth

Demo

https://huggingface.co/spaces/multimodalart/dreambooth-training

ID Preservation by Fine-Tuning Text Embeddings



An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, Gal et al., ICLR 2023

ID Preservation by Fine-Tuning Text Embeddings

Compared to fine-tuning denoiser weights:

- Only the optimized embedding needs to be stored, not the full denoiser weights.
- ID preservation is a bit weaker.

Input Images

"A photo of [V]"

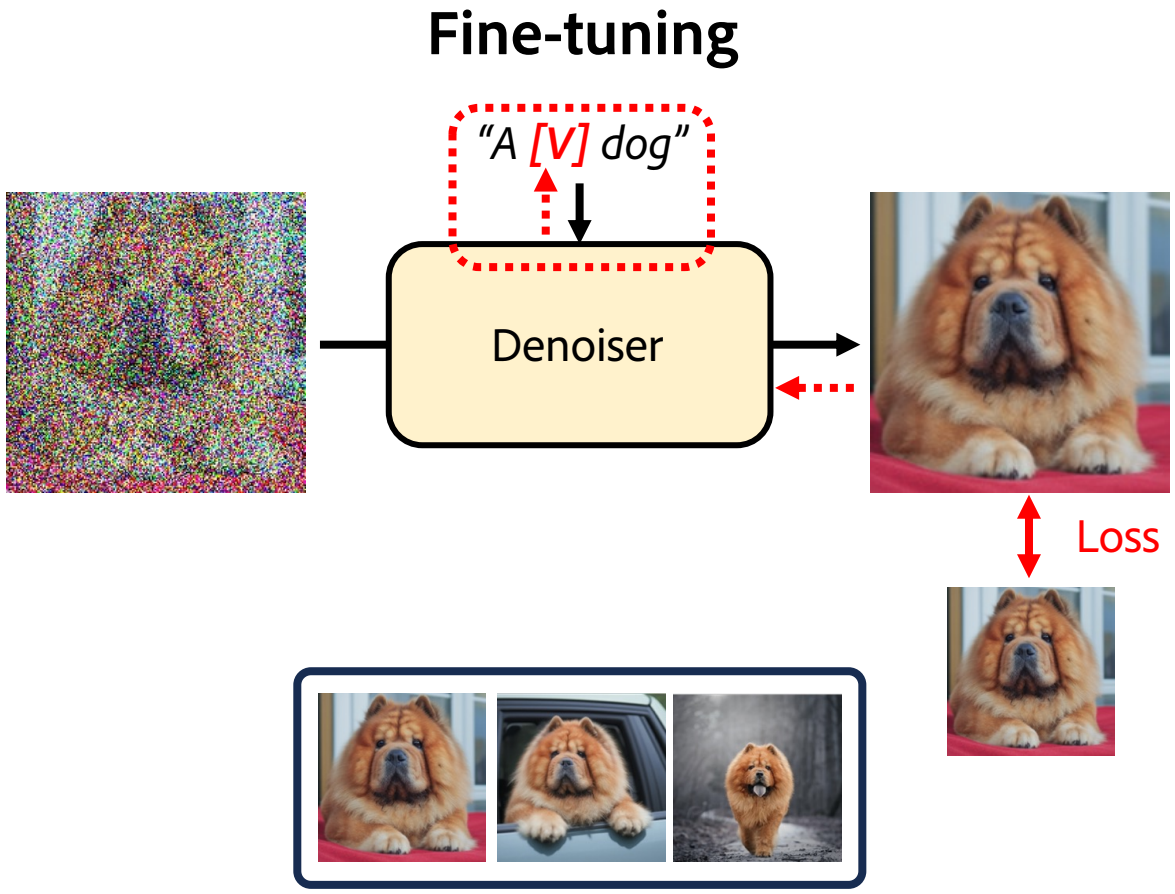
"A photo of S_*" "A photo of S_* on the beach" "A photo of S_* on the moon" "Elmo holding a S_*"

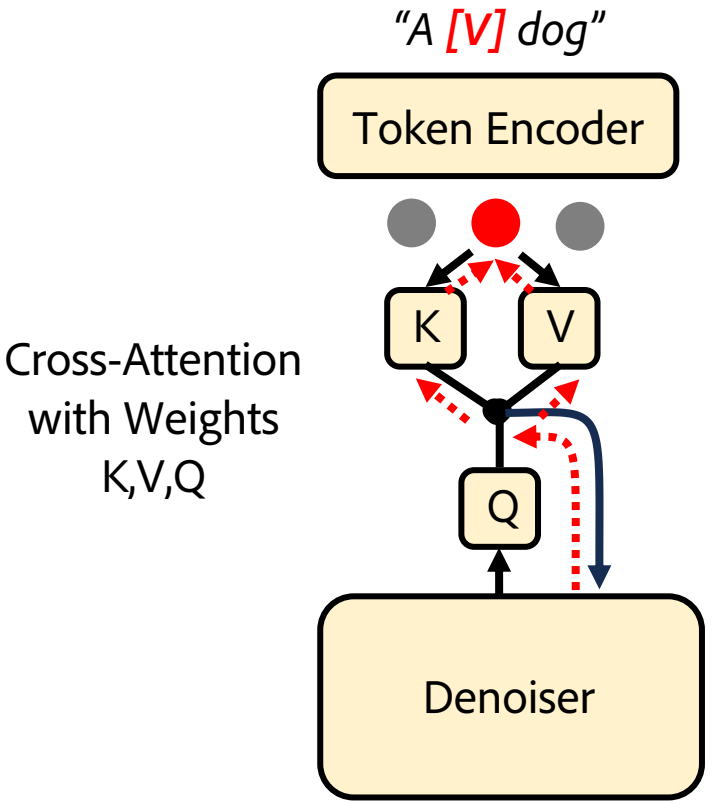
An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, Gal et al., ICLR 2023

Code

https://github.com/rinongal/textual_inversion

Fine-Tuning Text Embeddings & Cross-Att. Weights





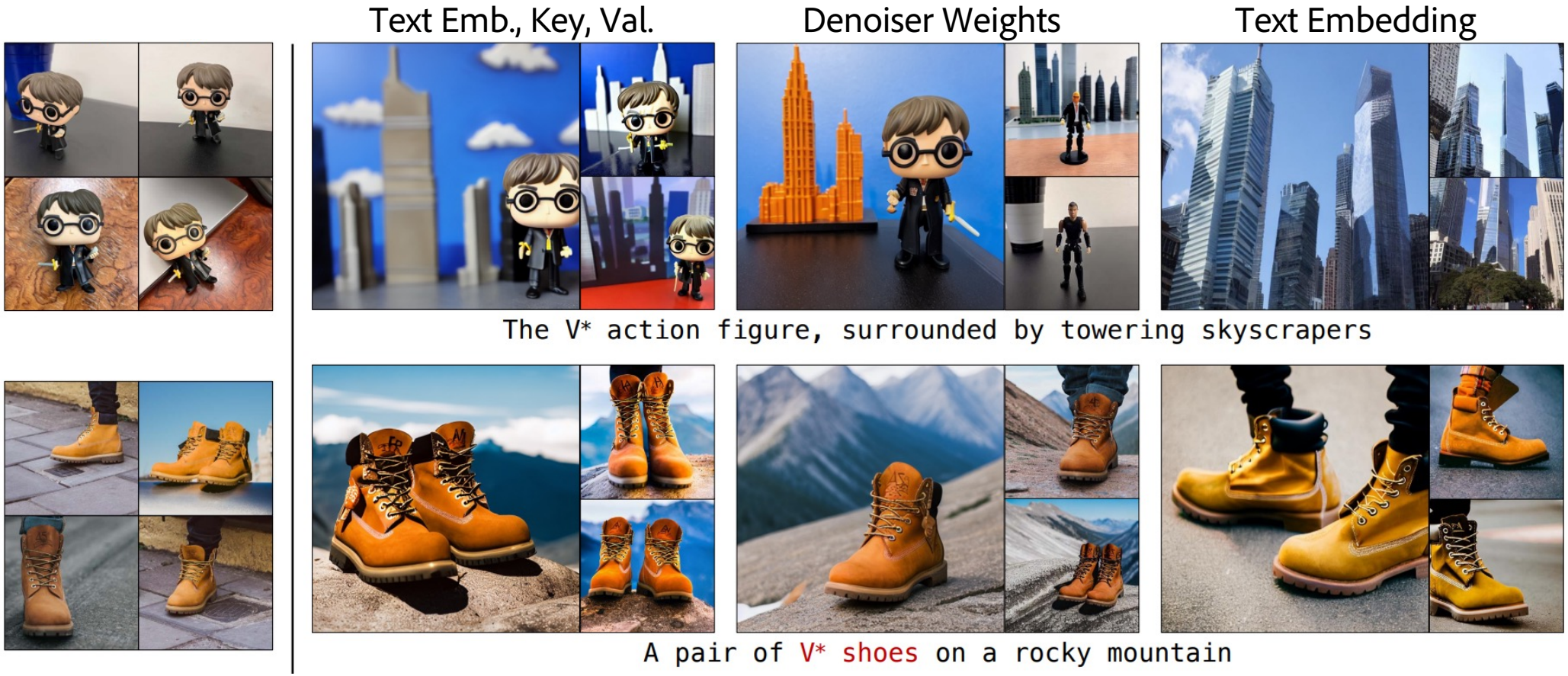
Cross-Attention with Weights K,V,Q

Multi-Concept Customization of Text-to-Image Diffusion, Kumari et al., CVPR 2023

Key-Locked Rank One Editing for Text-to-Image Personalization, Tewel et al., SIGGRAPH 2023

Fine-Tuning Text Embeddings & Cross-Att. Weights

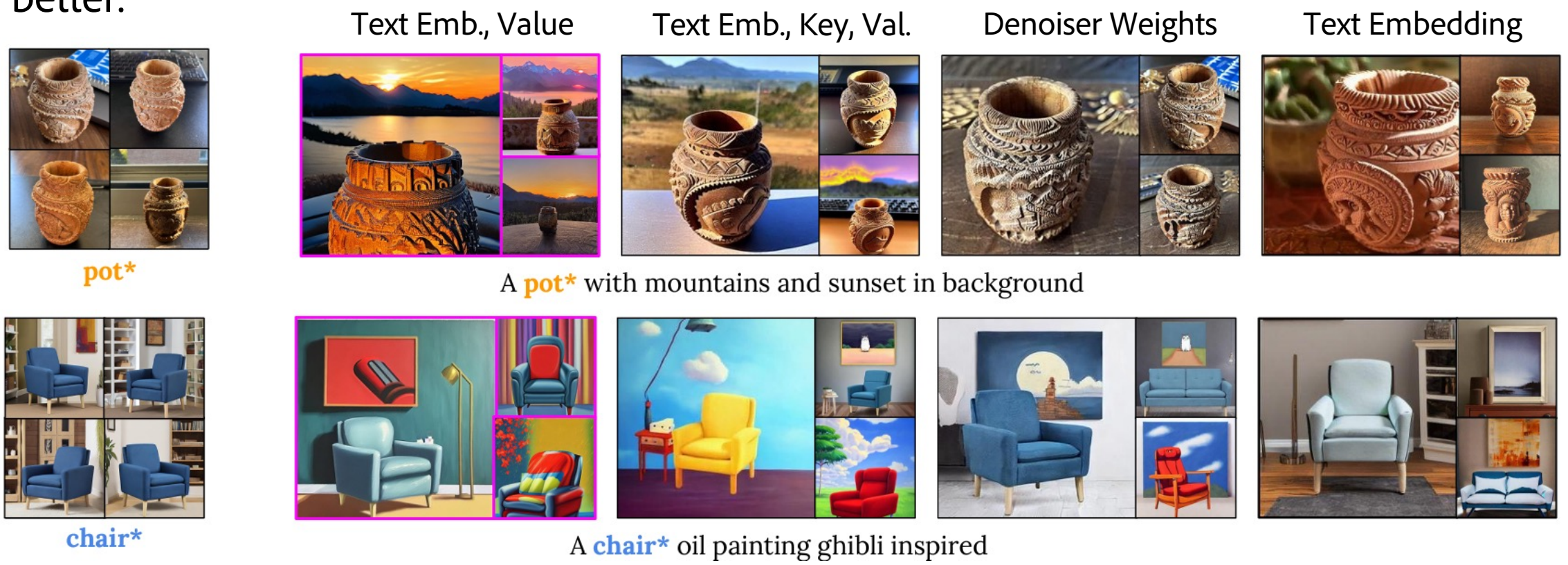
Fine-tuning text embeddings, keys and values.
ID preservation is close to tuning denoiser weights, while requiring less storage.



Multi-Concept Customization of Text-to-Image Diffusion, Kumari et al., CVPR 2023

Fine-Tuning Text Embeddings & Cross-Att. Weights

Fine-tuning text embeddings and values only.
ID preservation is slightly worse than tuning denoiser weights but follows the prompt better.



Key-Locked Rank One Editing for Text-to-Image Personalization, Tewel et al., SIGGRAPH 2023

Custom Diffusion

Multi-Concept Customization of Text-to-Image Diffusion

(fine-tuning keys, values and text embeddings)

Code

https://github.com/adobe-research/custom-diffusion

Code

https://huggingface.co/docs/diffusers/en/training/custom_diffusion

Perfusion

Key-Locked Rank One Editing for Text-to-Image Personalization

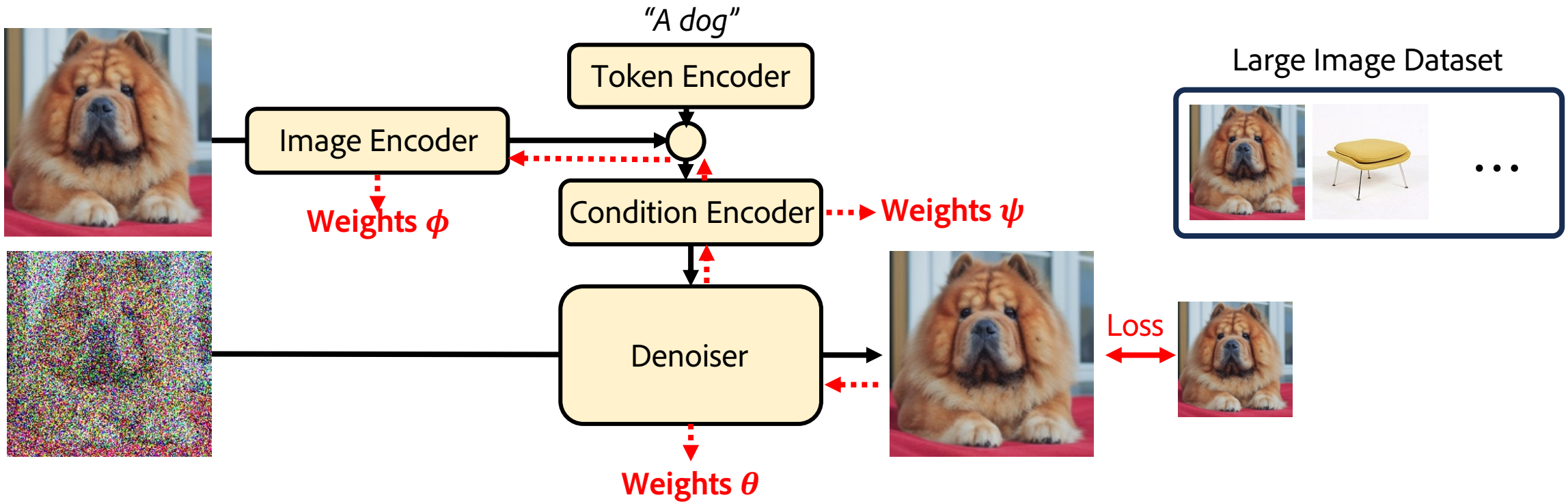
(fine-tuning values and text embeddings only)

Code

https://github.com/ChenDarYen/Key-Locked-Rank-One-Editing-for-Text-to-Image-Personalization

ID Preservation with Learned Condition Encoders

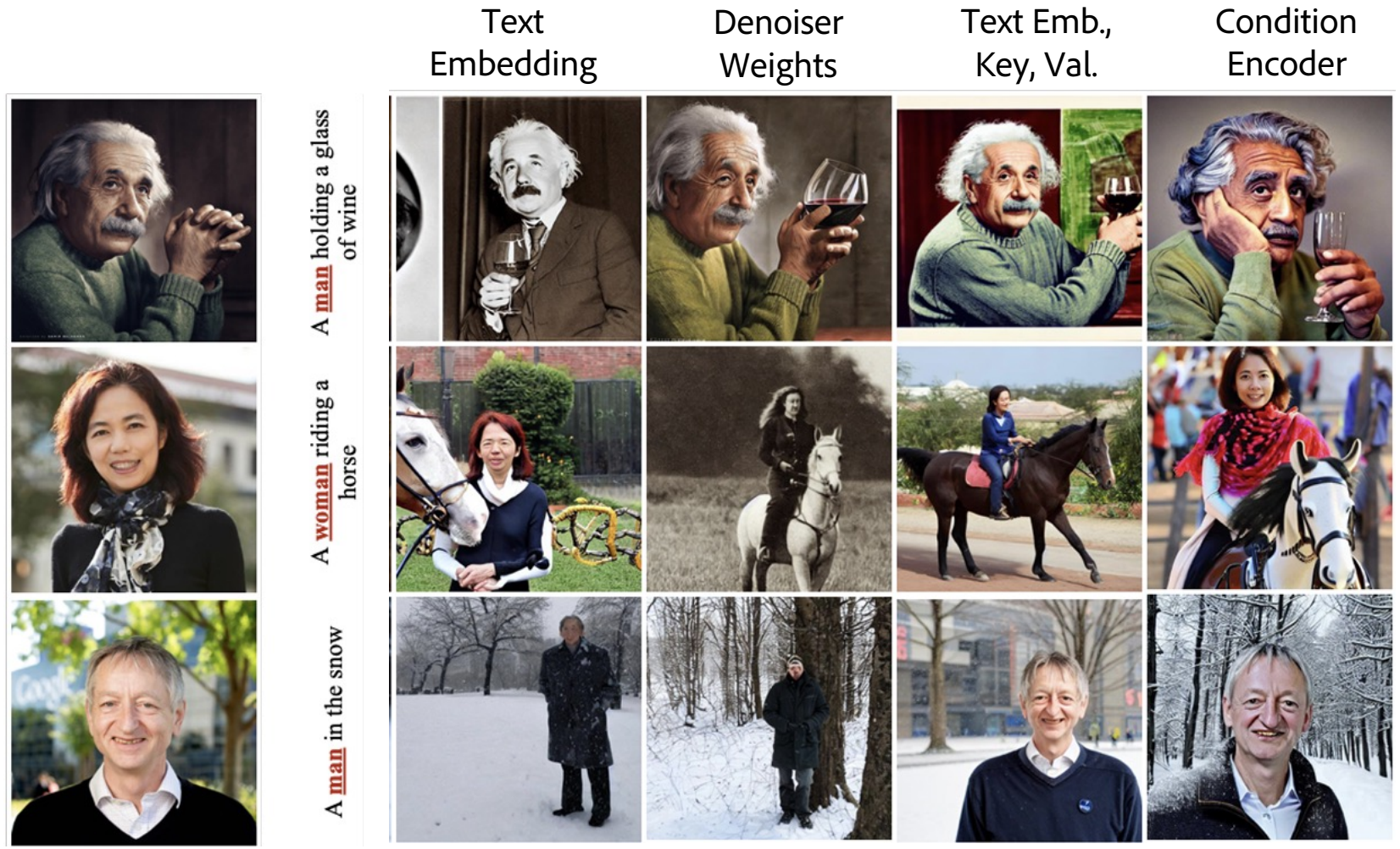
Motivation: avoid the need to fine-tune for each object identity.



FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention, Gal et al., ArXiv May 2023

BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing, Li et al., NeurIPS 2024

ID Preservation with Learned Condition Encoders



FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention, Gal et al., ArXiv May 2023

ID Preservation with Learned Condition Encoders

BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing,
Li et al., NeurIPS 2024

FastComposer

uning-Free Multi-Subject Image Generation with Localized Attention

(CLIP-based image encoder)

Code

https://github.com/mit-han-lab/fastcomposer/tree/main

Demo

https://replicate.com/cjwbw/fastcomposer

BLIP-Diffusion

Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing

(BLIP-based image encoder)

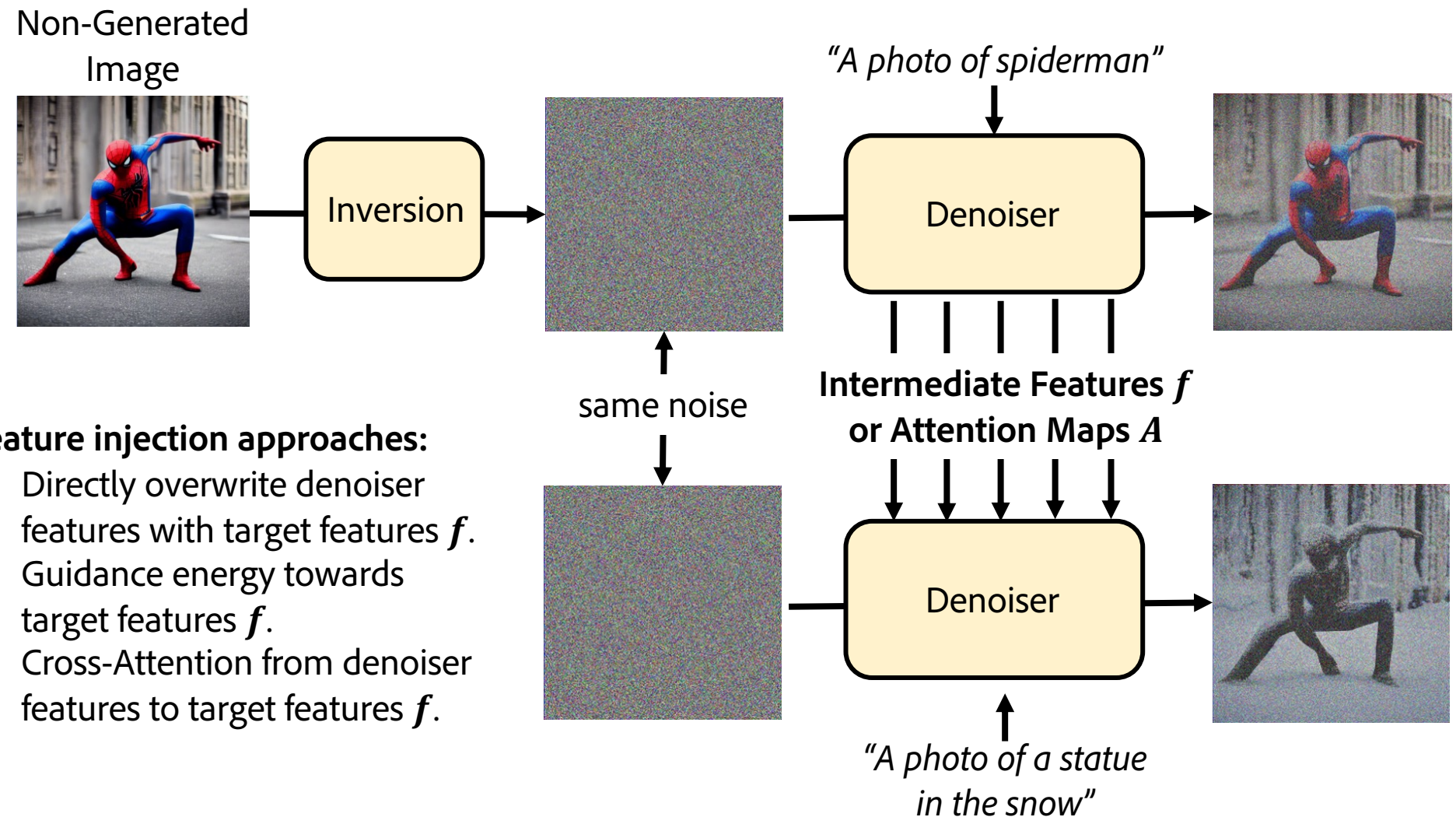
Code

https://huggingface.co/docs/diffusers/en/api/pipelines/blip_diffusion

Code

https://github.com/salesforce/LAVIS/tree/main/projects/blip-diffusion

ID Preservation through Intermediate Features



Feature injection approaches:

1) Directly overwrite denoiser features with target features f.
2) Guidance energy towards target features f.
3) Cross-Attention from denoiser features to target features f.

Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation, Tumanyan et al., CVPR 2023

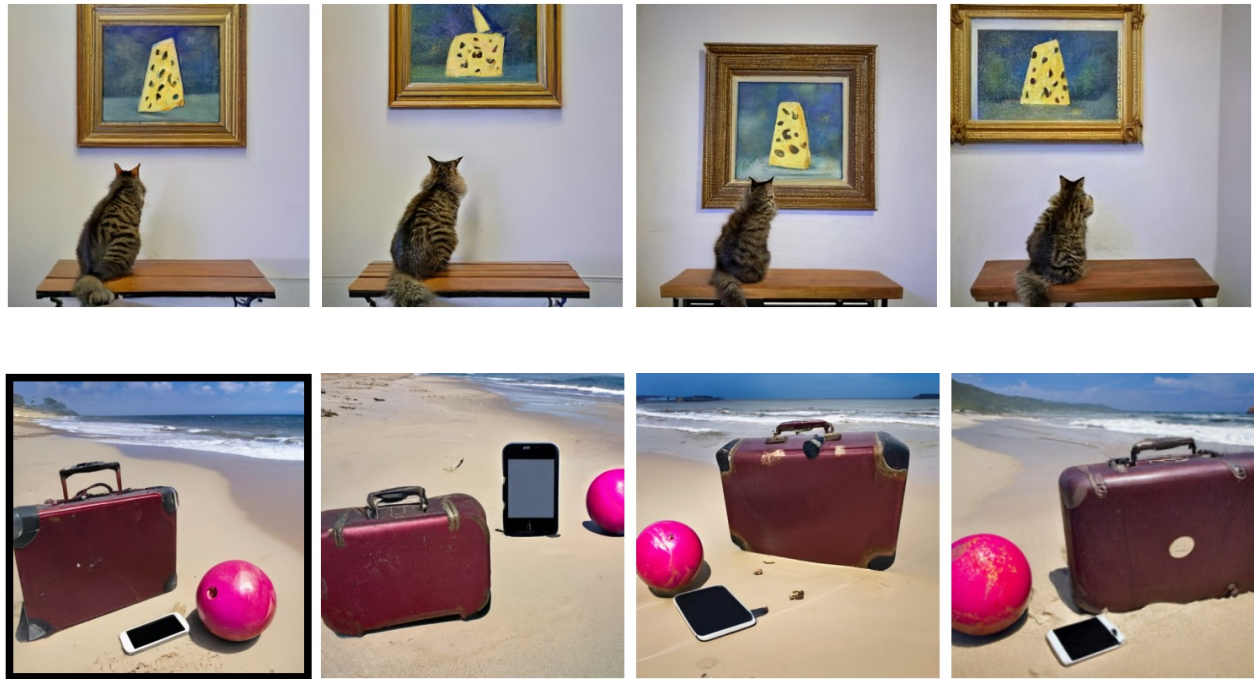
Diffusion Self-Guidance for Controllable Image Generation, Epstein et al., NeurIPS 2023

MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing, Cao et al., ICCV 2023

ID Preservation through Intermediate Features

- No training or fine-tuning needed.
- Intermediate features entangle identity with location and context, thus it requires additional work to allow for large changes in locations & contexts.

Guidance energy towards target features f.



Diffusion Self-Guidance for Controllable Image Generation, Epstein et al., NeurIPS 2023

Cross-Attention from denoiser features to target features f.

MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing, Cao et al., ICCV 2023

ID Preservation through Intermediate Features

- No training or fine-tuning needed.
- Intermediate features entangle identity with location and context, thus it requires additional work to allow for large changes in locations & contexts.

Directly overwrite denoiser features with target features f.

Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation, Tumanyan et al., CVPR 2023

MasaCtrl

Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing
(Cross-Attention-Based Feature Injection)

Code

https://github.com/TencentARC/MasaCtrl

PnP-Diffusers

Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation
(Overwrite-Based Feature Injection)

Demo

https://huggingface.co/spaces/hysts/PnP-diffusion-features

Code

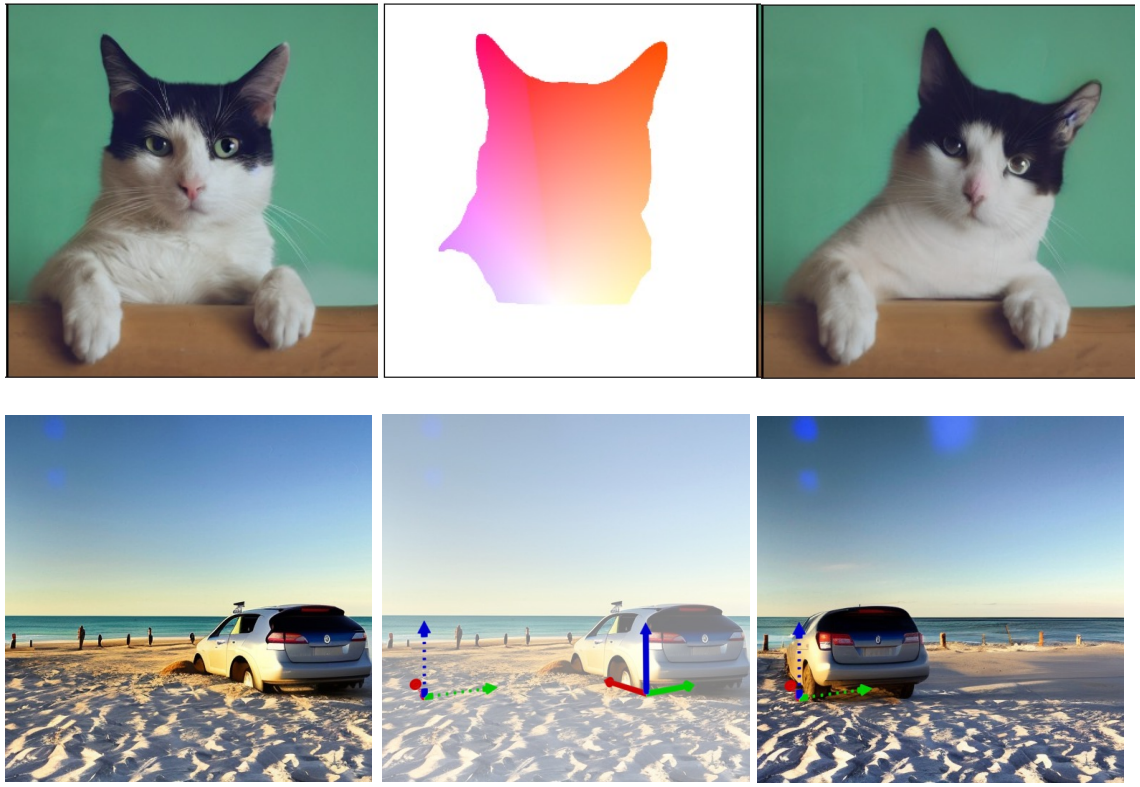
https://github.com/MichalGeyer/pnp-diffusers

Editing with Generative Models

Personalization

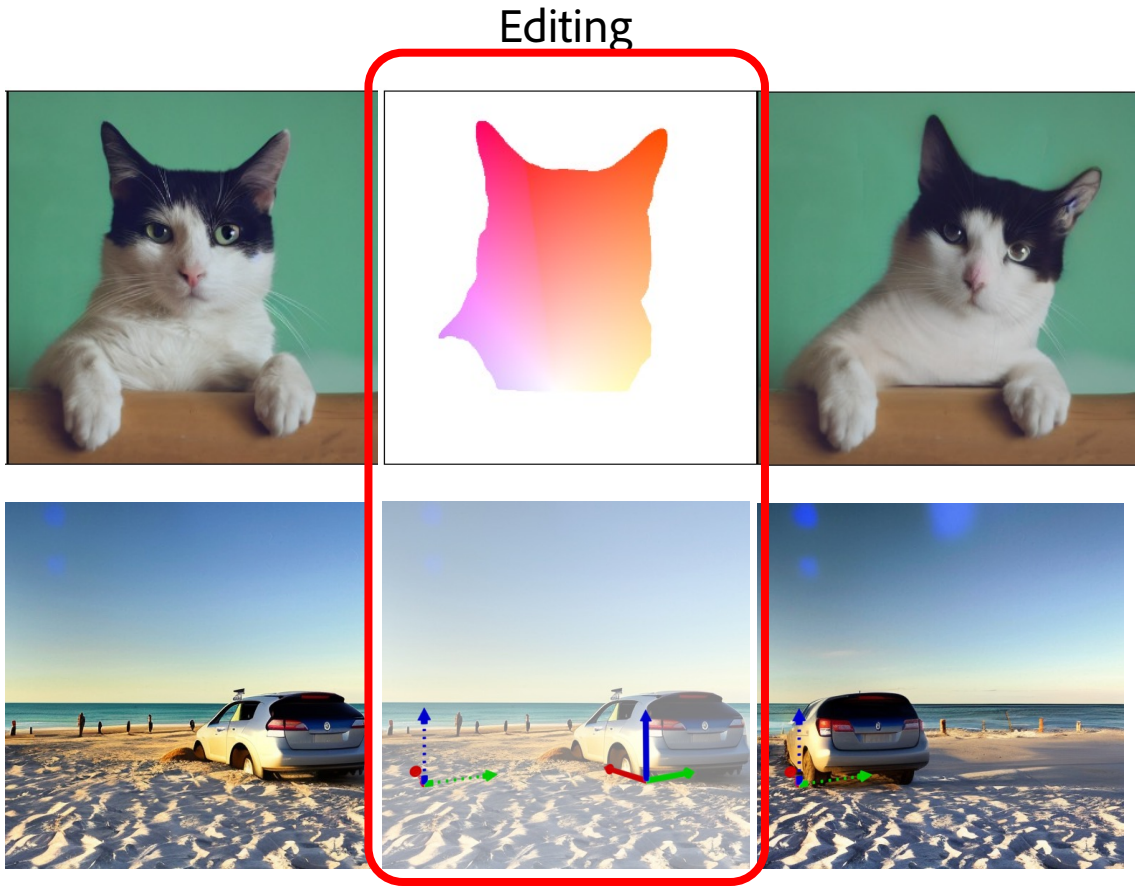
ConsiStory: Training-Free Consistent Text-to-Image Generation
Tewel et al., ArXiv Feb. 2024

Editing



Diffusion Handles: Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D Pandey et al., CVPR 2024
Motion Guidance: Diffusion-Based Image Editing with Differentiable Motion Estimators, Geng and Owens, ICLR 2024

Editing with Generative Models



Same subject, same scene.
Subject property changed by user **edit**.
(Property such as position, pose, etc.)

Editing:

Generative Model
+ Identity Preservation
+ Edit Control

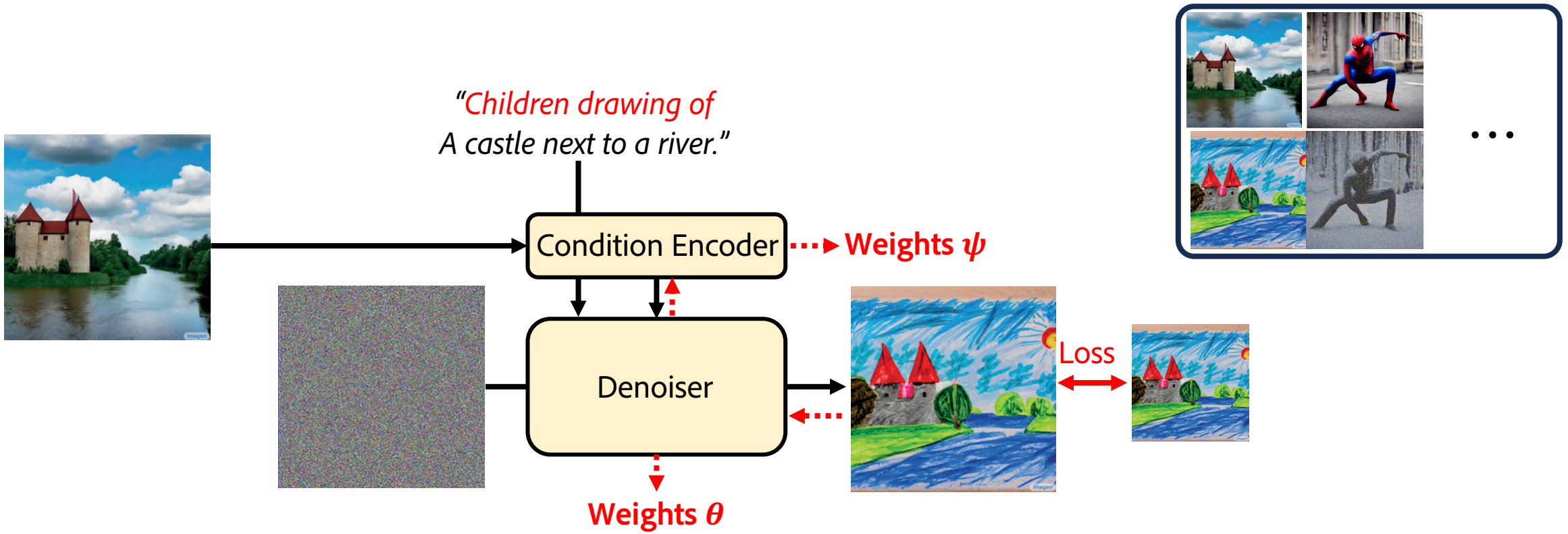
Diffusion Handles: Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D Pandey et al., CVPR 2024
Motion Guidance: Diffusion-Based Image Editing with Differentiable Motion Estimators, Geng and Owens, ICLR 2024

Image-to-Image Translation

Edit Control: Text Prompt

Identity Preservation: Condition Encoder

Large Paired Image Dataset

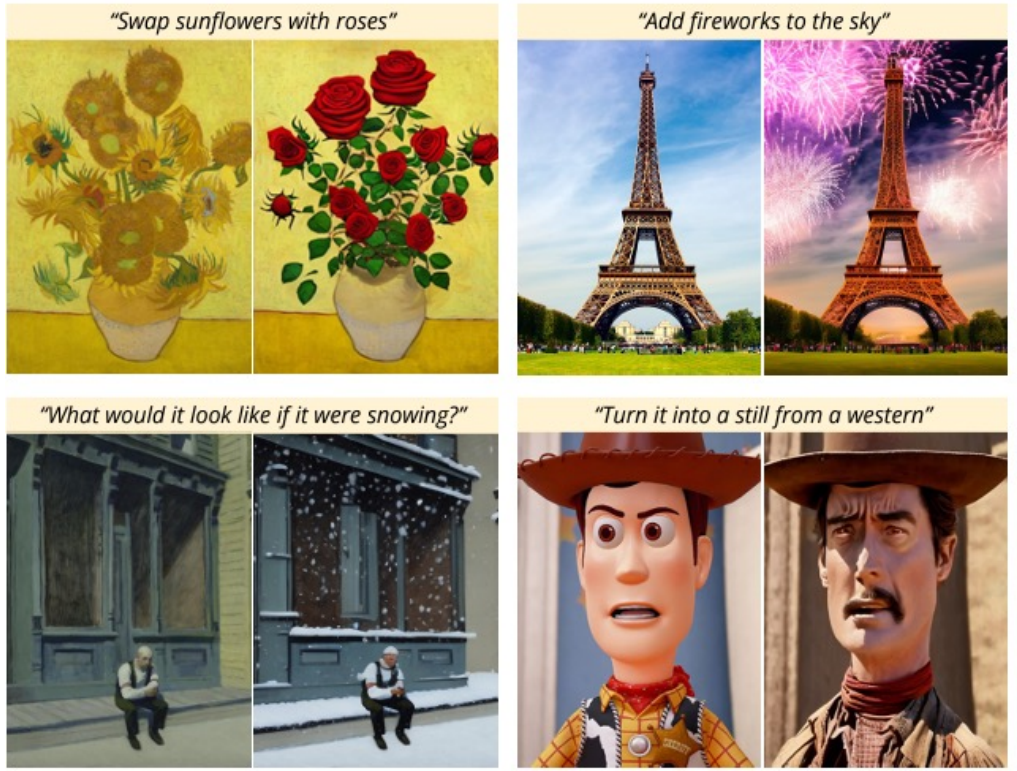


InstructPix2Pix: Learning to Follow Image Editing Instructions,
Brooks et al., CVPR 2023

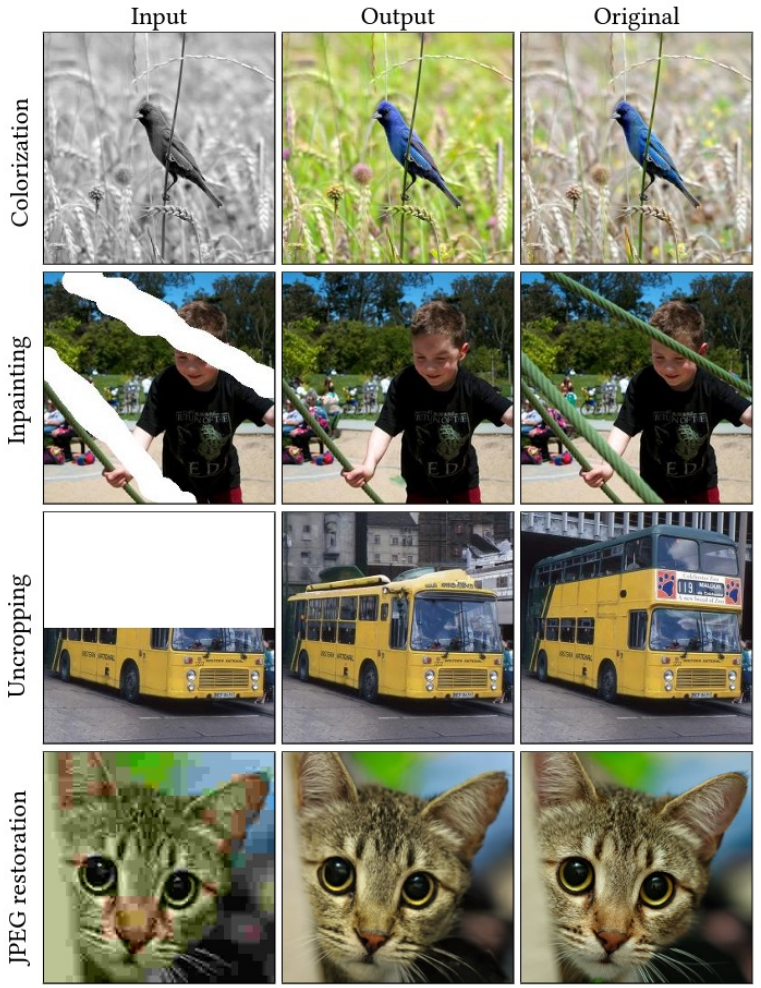
Palette: Image-to-Image Diffusion Models,
Saharia et al., SIGGRAPH 2022

Image-to-Image Translation

- Requires paired image dataset.
- Text prompt provides only coarse control.



InstructPix2Pix: Learning to Follow Image Editing Instructions,
Brooks et al., CVPR 2023



Palette: Image-to-Image Diffusion Models,
Saharia et al., SIGGRAPH 2022

InstructPix2Pix

Learning to Follow Image Editing Instructions

(Edit Instructions)

Code

https://github.com/timothybrooks/instruct-pix2pix

Demo

https://huggingface.co/spaces/timbrooks/instruct-pix2pix

Palette

Image-to-Image Diffusion Models

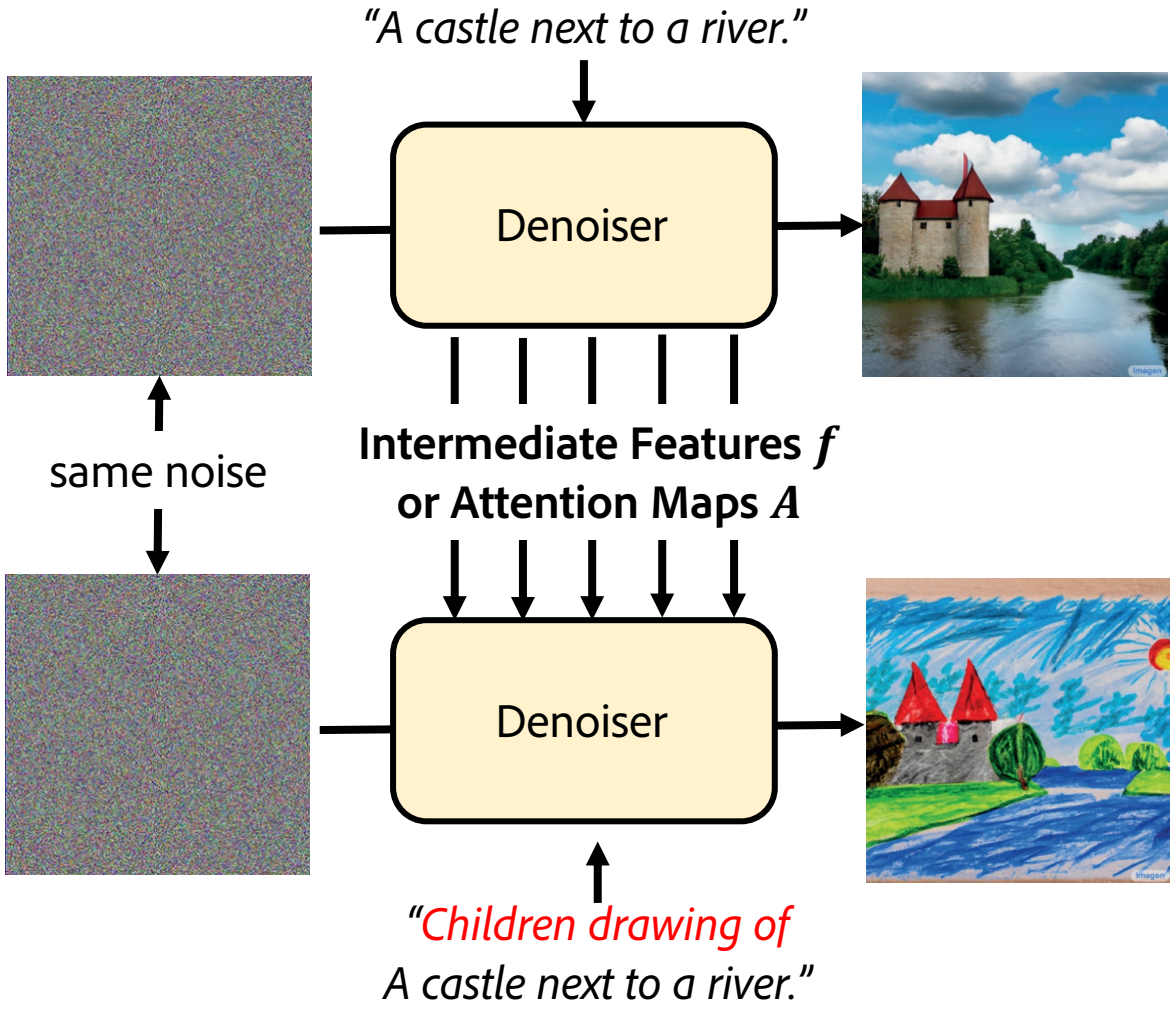
(Per-Task Fine-tuning)

Code

https://github.com/Janspiry/Palette-Image-to-Image-Diffusion-Models

Image-to-Image Translation

Edit Control:
Text Prompt
Identity Preservation:
Intermediate Features / Attention Maps



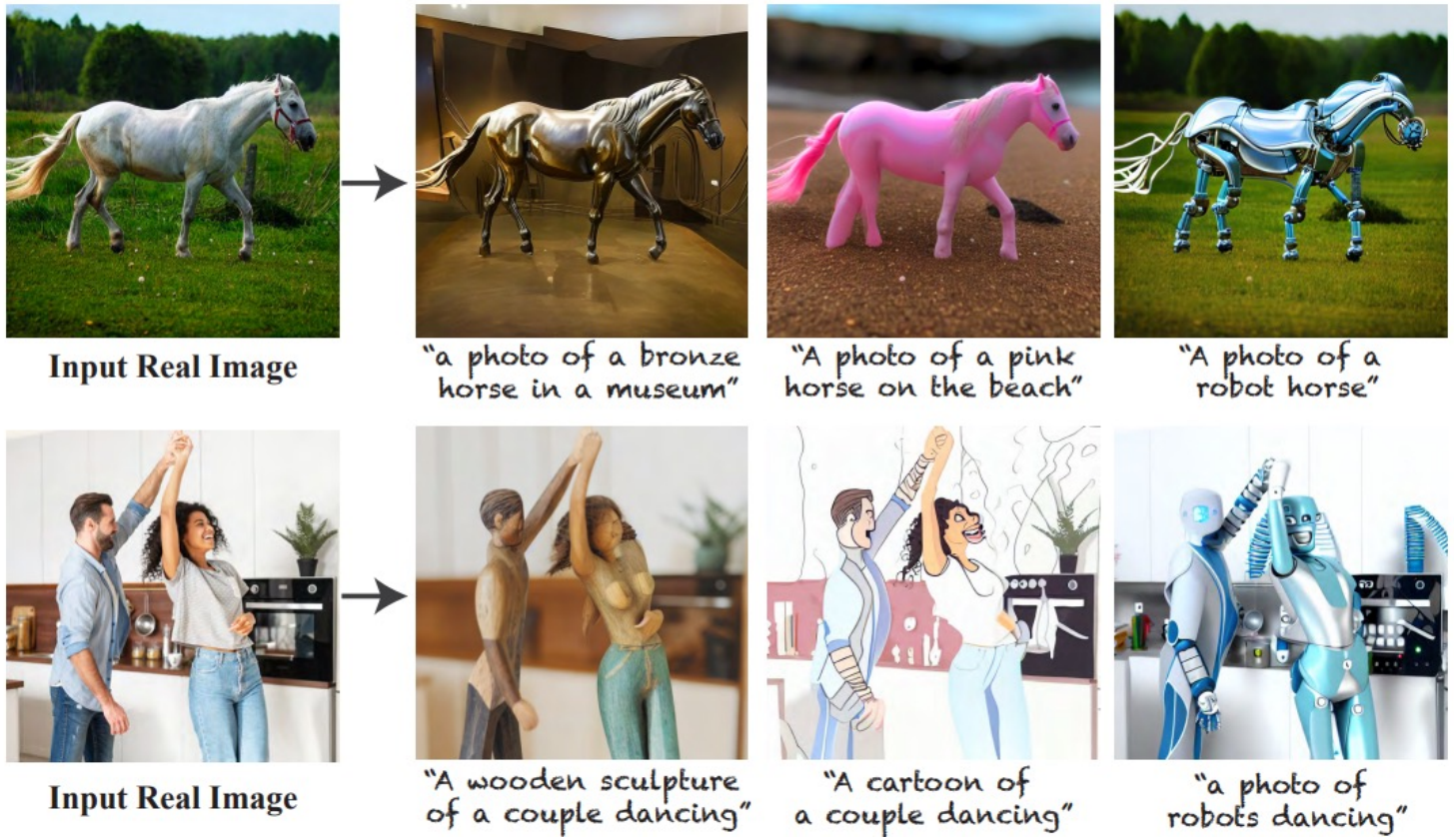
Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation, Tumanyan et al., CVPR 2023

Prompt-to-Prompt Image Editing with Cross Attention Control, Hertz et al., ArXiv Aug. 2022

MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing, Cao et al., ArXiv Aug. 2022

Image-to-Image Translation

- No training or fine-tuning needed.
- Cannot strongly change scene layout (object positions in the image remain roughly the same)
- Text prompt provides only coarse control.



Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation, Tumanyan et al., CVPR 2023

Prompt-to-Prompt Image Editing with Cross Attention Control, Hertz et al., ArXiv Aug. 2022

Image-to-Image Translation

- No training or fine-tuning needed.
- Cannot strongly change scene layout (object positions in the image remain roughly the same)
- Text prompt provides only coarse control.

MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing, Cao et al., ICCV 2023

PnP-Diffusers

Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation

(Overwrite-Based Feature Injection)

Demo

https://huggingface.co/spaces/hysts/PnP-diffusion-features

Code

https://github.com/MichalGeyer/pnp-diffusers

Prompt-to-Prompt

Image Editing with Cross Attention Control

(Overwrite-Based Feature Injection)

Code

https://github.com/google/prompt-to-prompt/

MasaCtrl

Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing

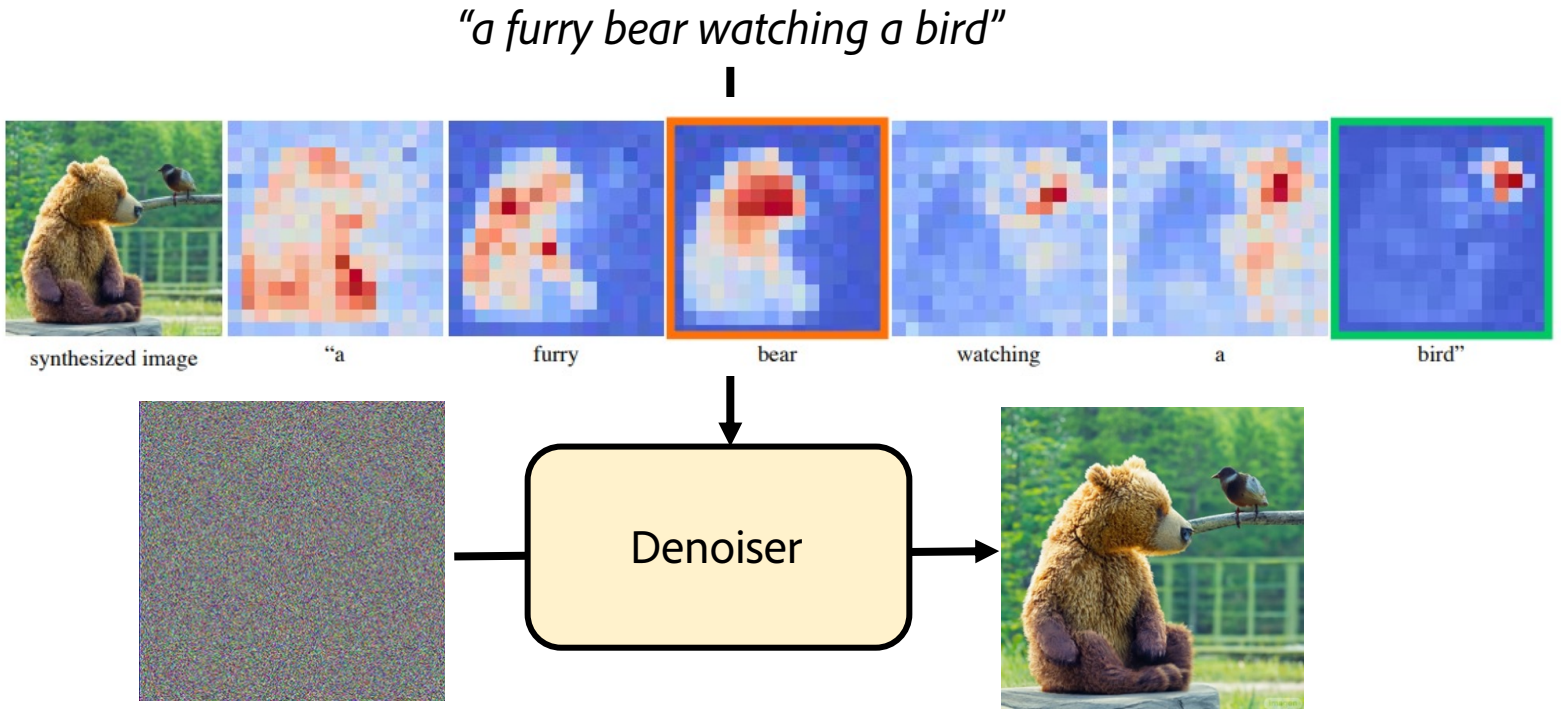
(Cross-Attention-Based Feature Injection)

Code

https://github.com/TencentARC/MasaCtrl

Using Attention Maps & Intermediate Features

Edit Control: Transform Intermediate Features / Attention Maps
Identity Preservation: Intermediate Features / Attention Maps



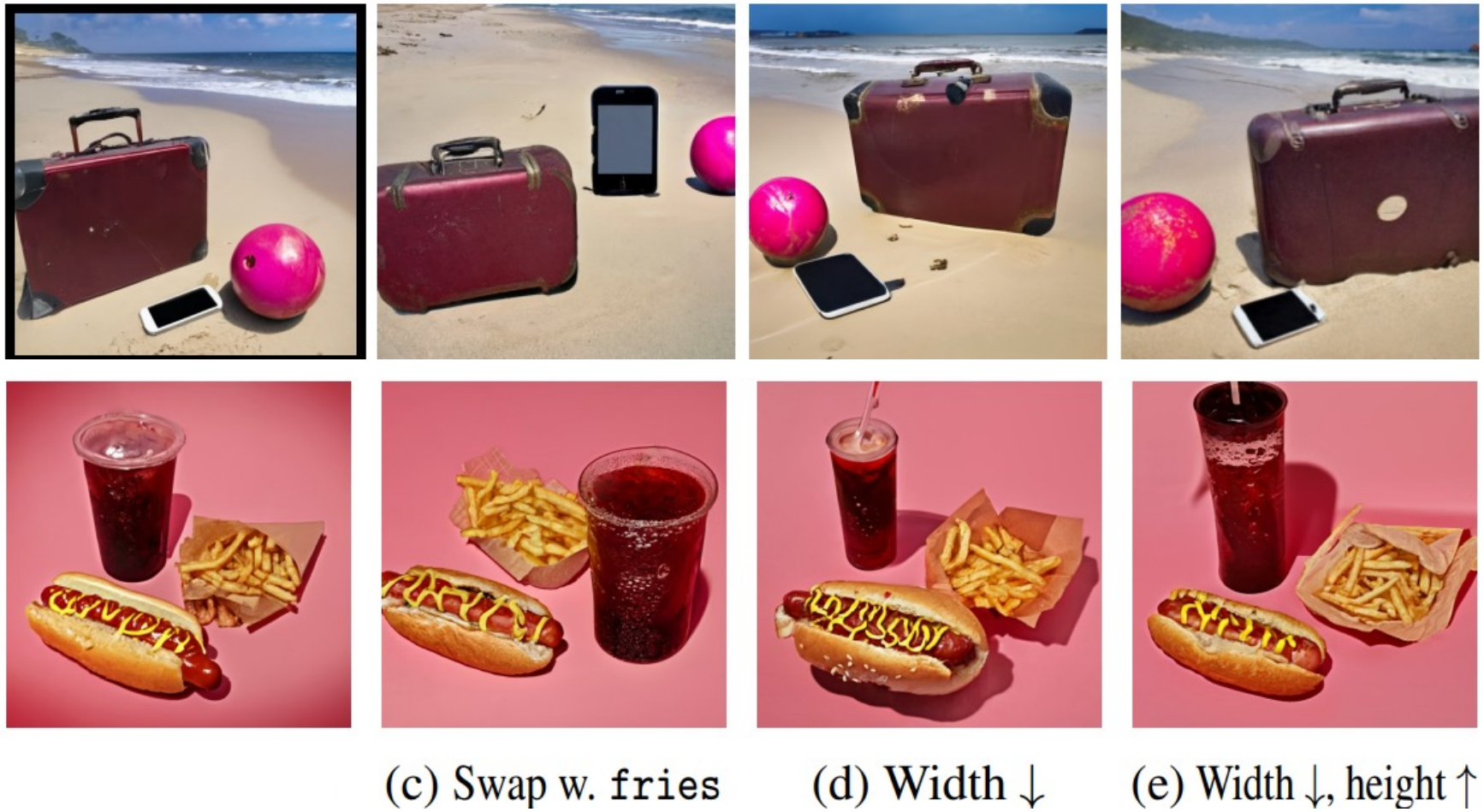
Attention maps from pixels to text tokens.

Prompt-to-Prompt Image Editing with Cross Attention Control, Hertz et al., ArXiv Aug. 2022

Diffusion Self-Guidance for Controllable Image Generation, Epstein et al., NeurIPS 2023

Diffusion Handles Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D, CVPR 2024

Using Attention Maps & Intermediate Features



Diffusion Self-Guidance for Controllable Image Generation,
Epstein et al., NeurIPS 2023

Prompt-to-Prompt

Image Editing with Cross Attention Control

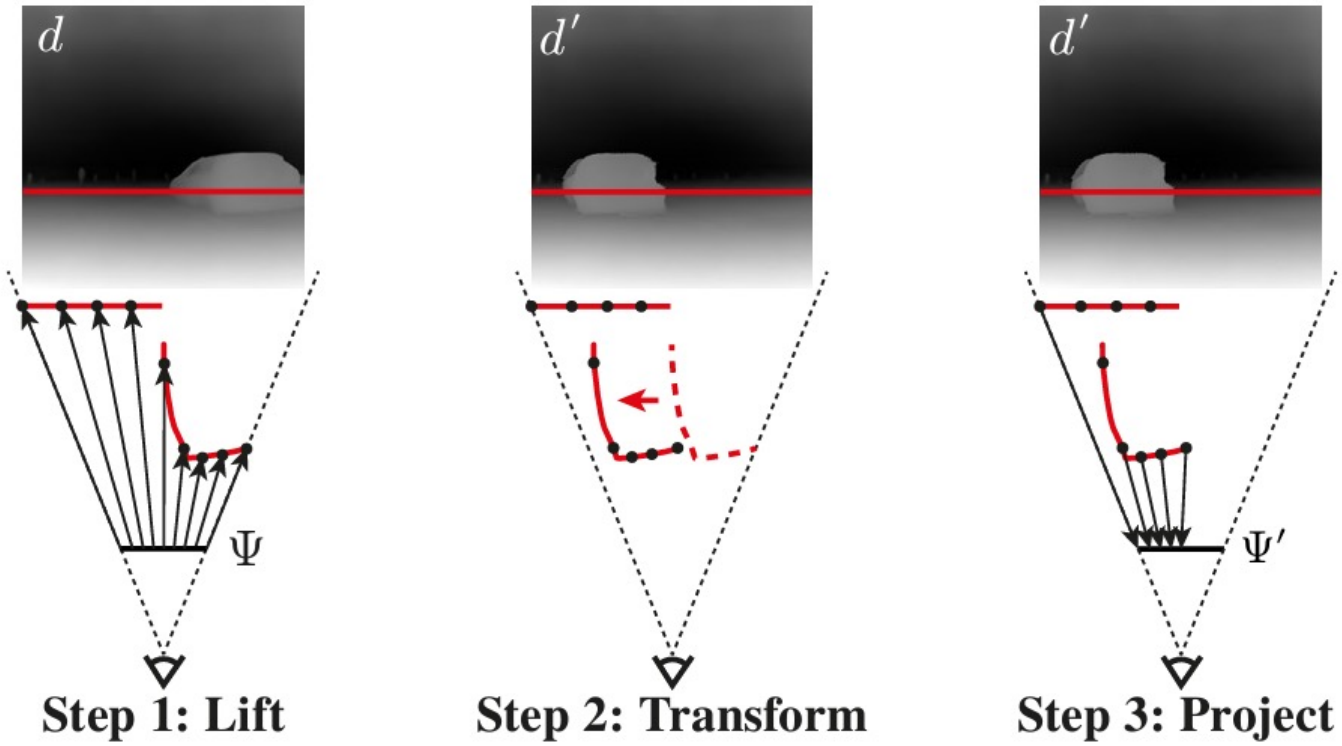
(Overwrite-Based Feature Injection)

Code

https://github.com/google/prompt-to-prompt/

Editing with Attention Maps and Intermediate Features

Attention maps / intermediate features can be 3D-transformed using monocular depth estimates.



Diffusion Handles Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D, CVPR 2024

Attention Maps & Intermediate Features

Attention maps / intermediate features can be 3D-transformed using monocular depth estimates.

Diffusion Handles Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D, CVPR 2024

Attention Maps & Intermediate Features

Attention maps / intermediate features can be 3D-transformed using monocular depth estimates.

Diffusion Handles Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D, CVPR 2024

Attention Maps & Intermediate Features

Attention maps / intermediate features can be 3D-transformed using monocular depth estimates.

Diffusion Handles Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D, CVPR 2024