Diffusion Models for Visual Content Creation

Niloy Mitra, Duygu Ceylan, Paul Guerrero,

Daniel Cohen-Or, Or Patashnik, Chun-Hao Huang, Minhyuk Sung

Part 4: Personalization & Editing

https://geometry.cs.ucl.ac.uk/courses/diffusion4ContentCreation_sigg24/

Presentation Schedule

Introduction to Diffusion Models

Guidance and Conditioning Sampling

Attention

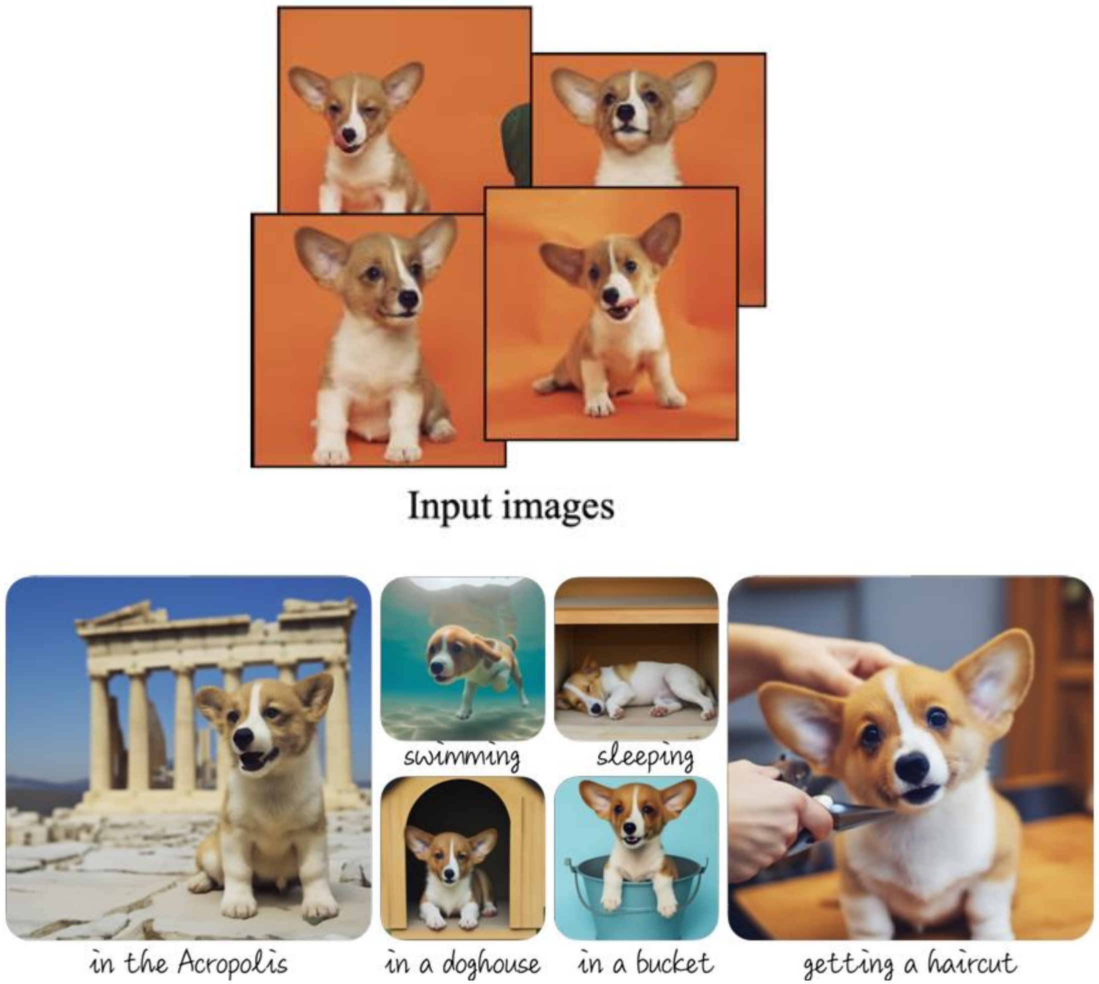
Break

Personalization and Editing

Beyond Single Images

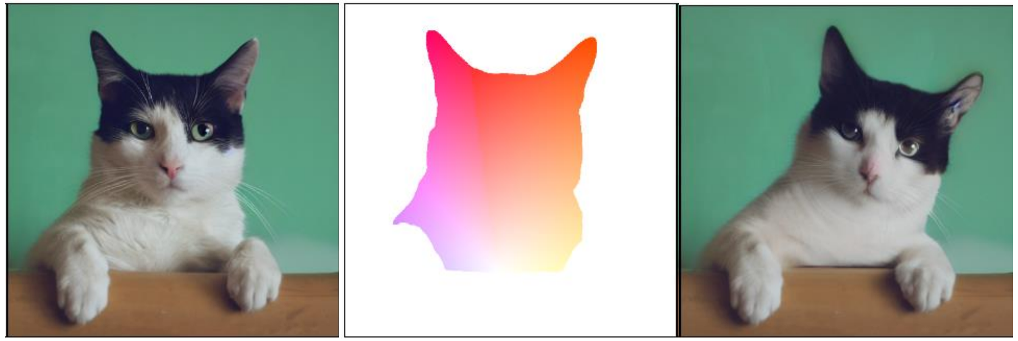
Diffusion Models for 3D Generation

Personalization

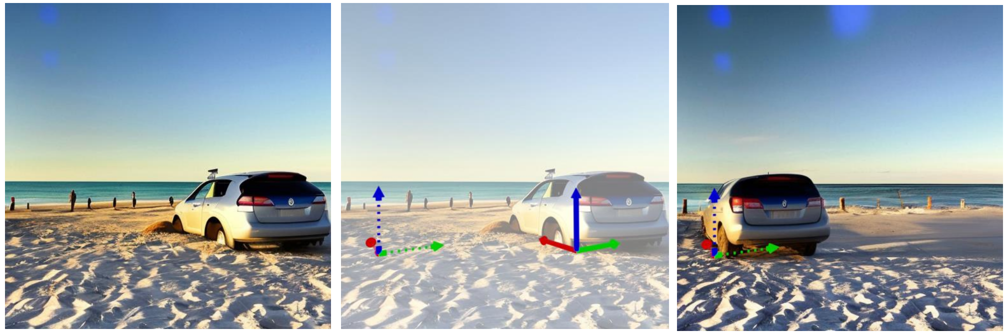


DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation,
Ruiz et al., CVPR 2023

Image Editing



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

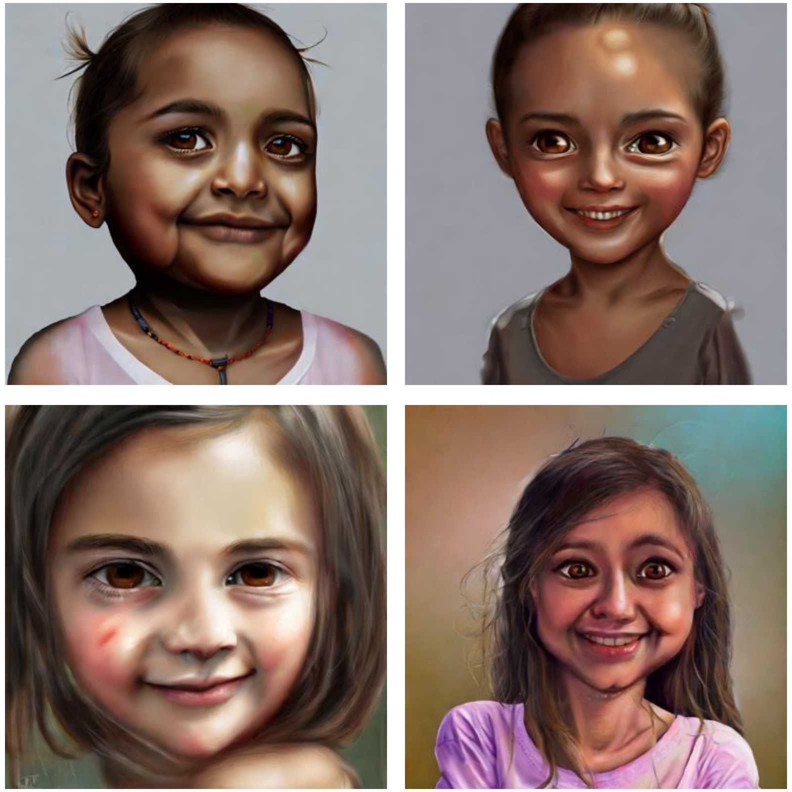


Diffusion Handles: Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D
Pandey et al., CVPR 2024

Personalization

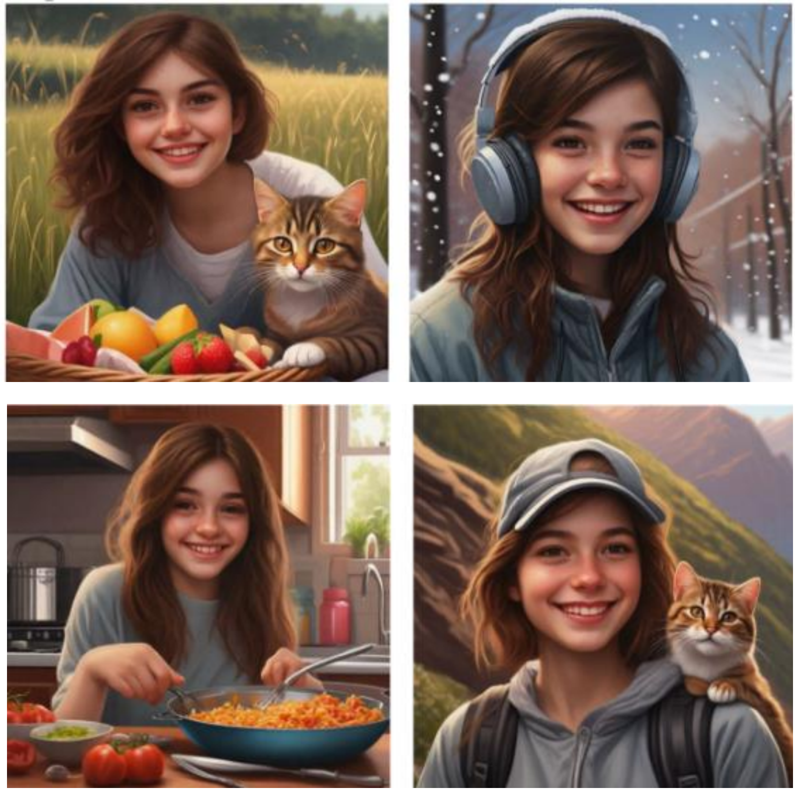
“a hyper-realistic digital painting of a happy girl, with brown eyes”

Without Personalization



Generated with StabelDiffusion 2.1

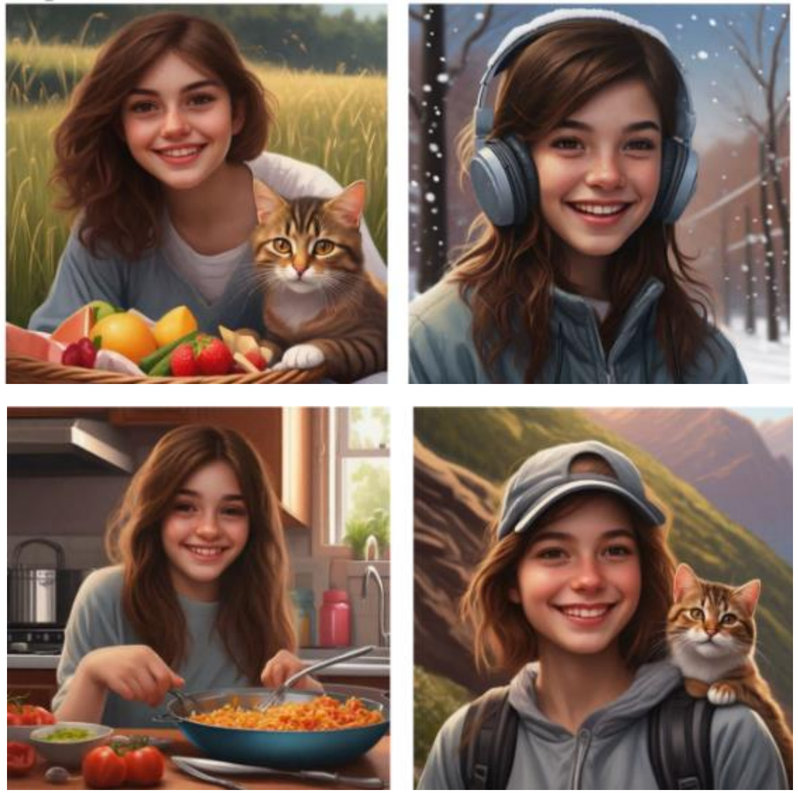
With Personalization



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

Personalization

With Personalization



Same subject in different settings.

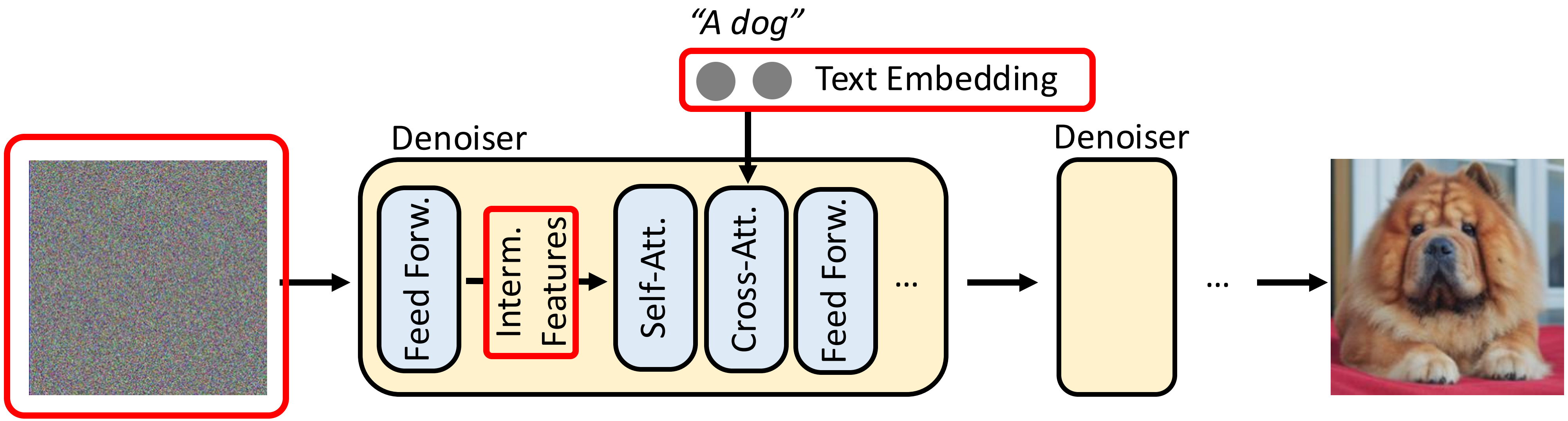
ConsiStory: Training-Free Consistent Text-to-Image Generation
Tewel et al., ArXiv Feb. 2024

Personalization:

Generative Model
+ **Identity Preservation**

Identity Preservation

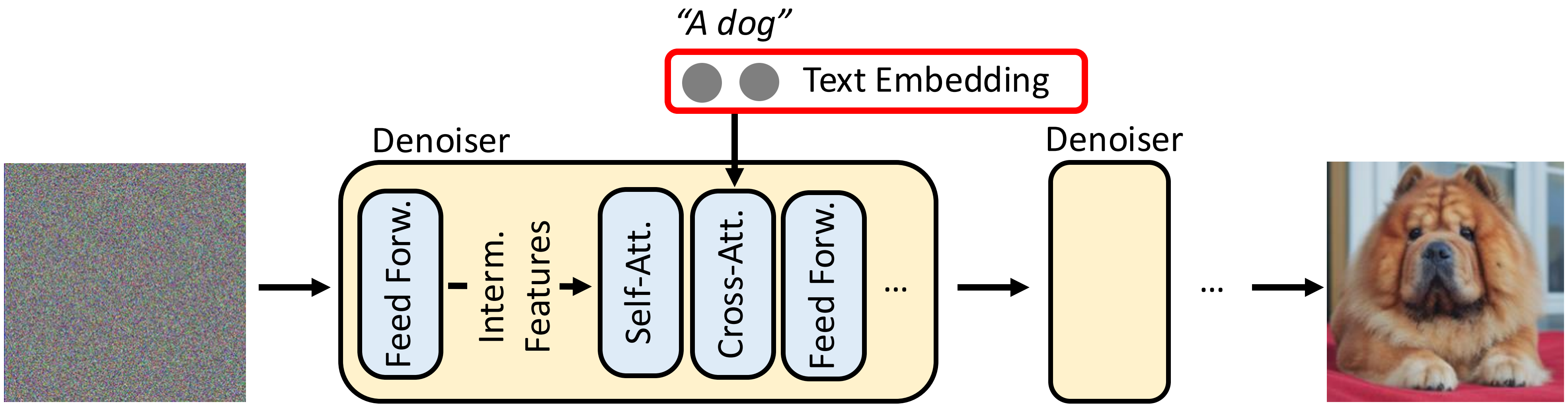
What can we use to control the identity of a generated subject?



Each of these have been used.

Identity Preservation

What can we use to control the identity of a generated subject?



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

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

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

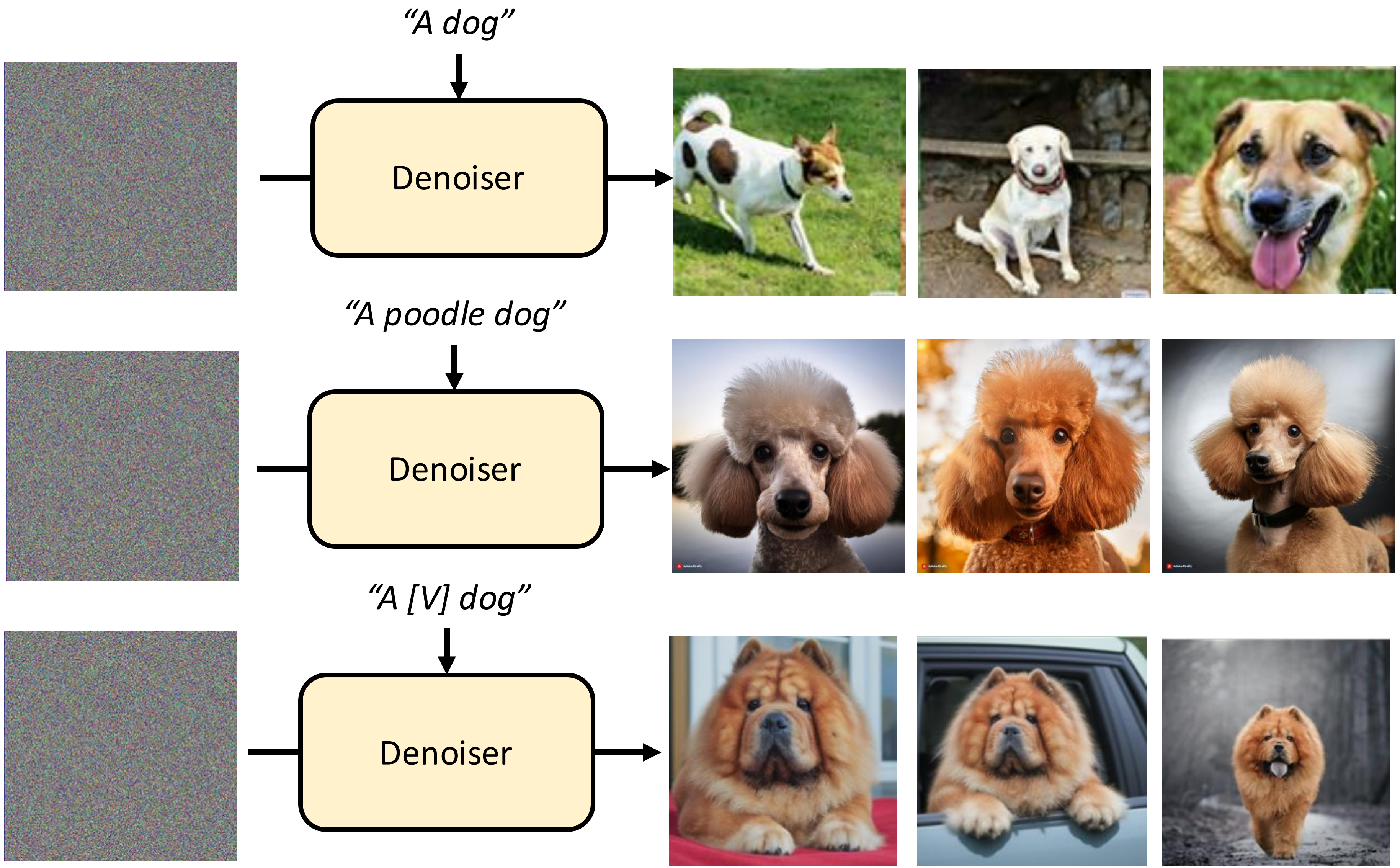
FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention, Gal et al., ArXiv May 2023

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

IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models, Ye et al., ArXiv August 2023

IMPRINT: Generative Object Compositing by Learning Identity-Preserving Representation, Song et al., CVPR 202

ID Preservation With Text Embeddings



"A dog"

Denoiser

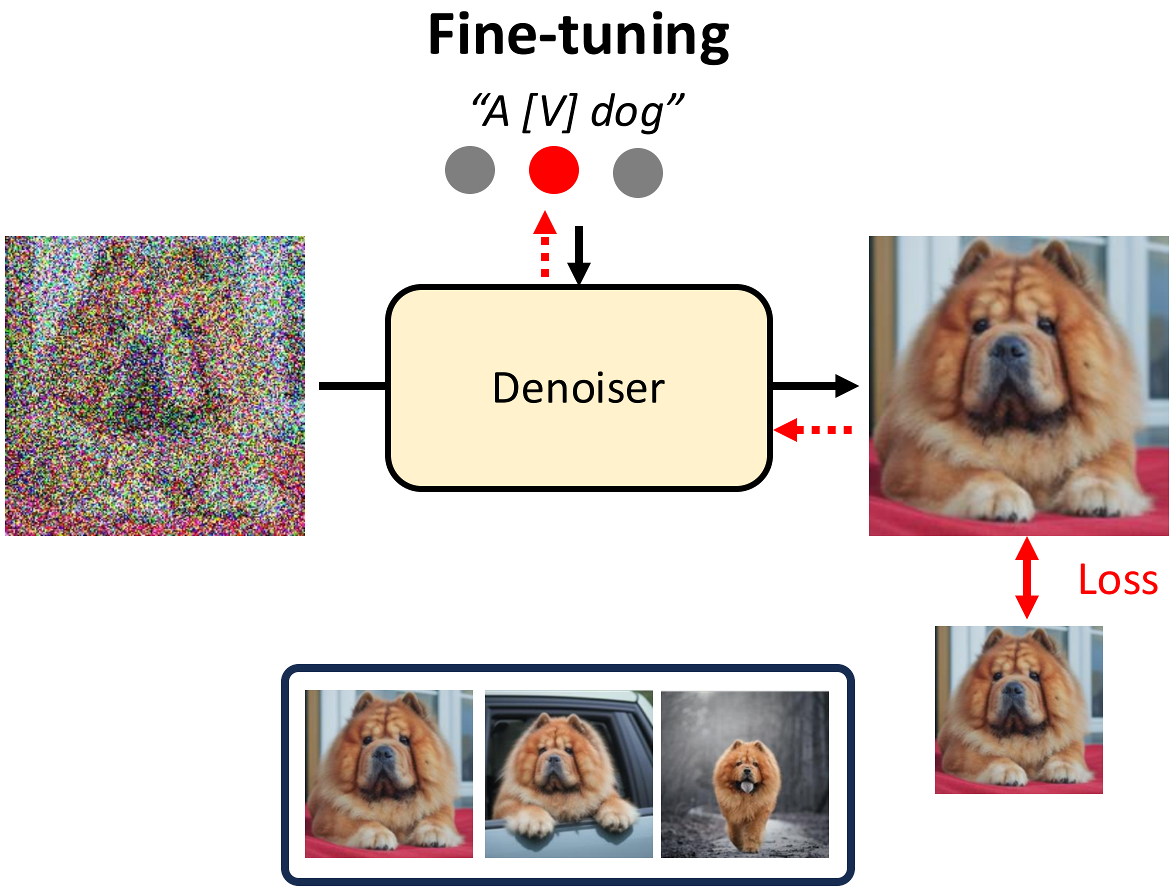
"A poodle dog"

Denoiser

"A [V] dog"

Denoiser

ID Preservation With Text Embeddings – Fine-Tune Tokens



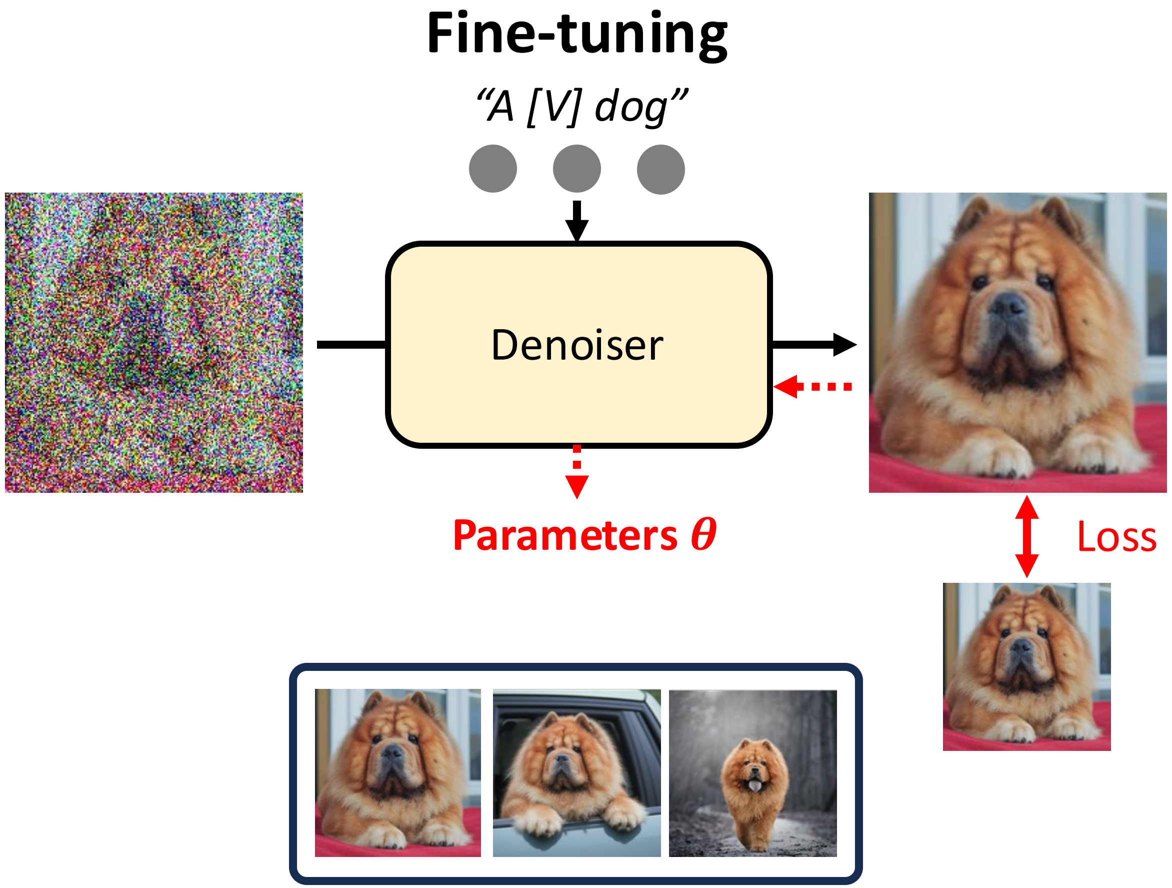
Fine-tuning

“A [V] dog”

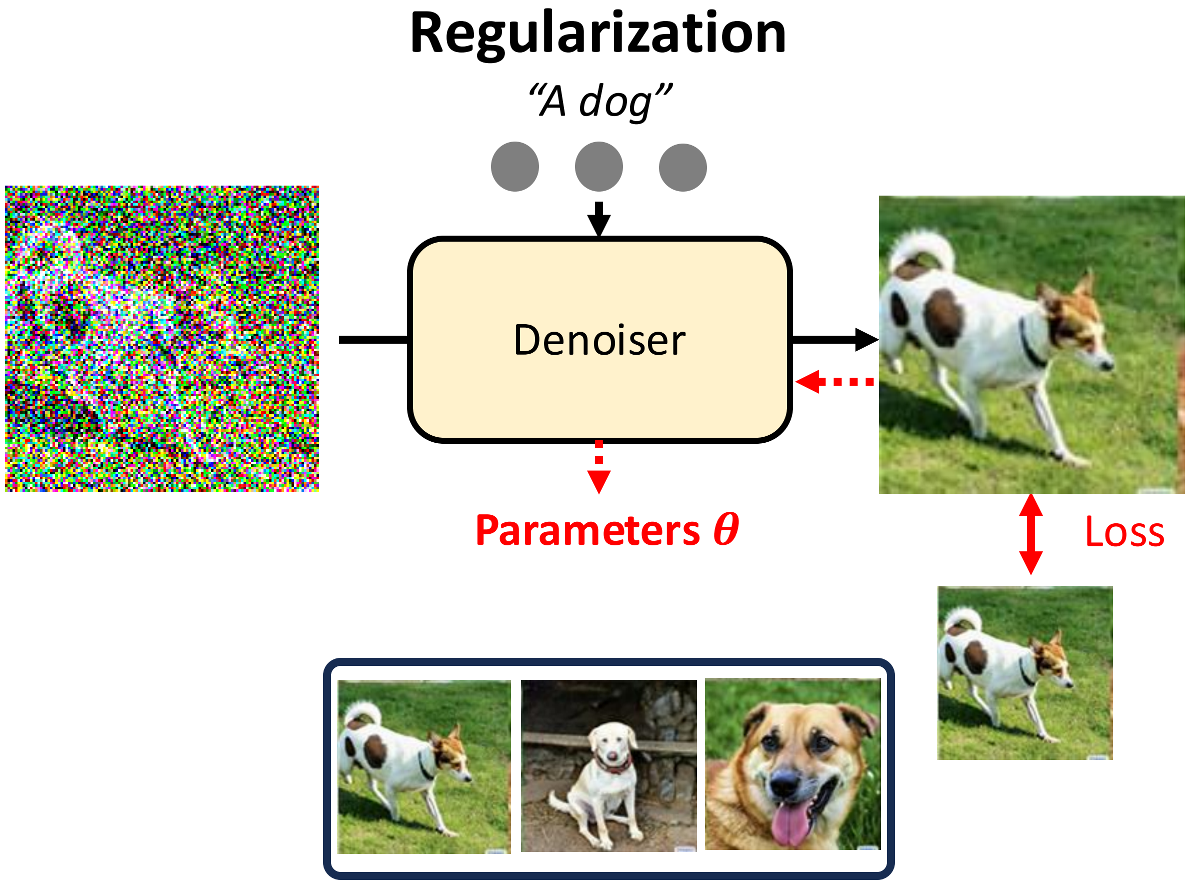
3-5 example images showing the identity.

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

ID Preservation With Text Embeddings – Fine-Tune Params.



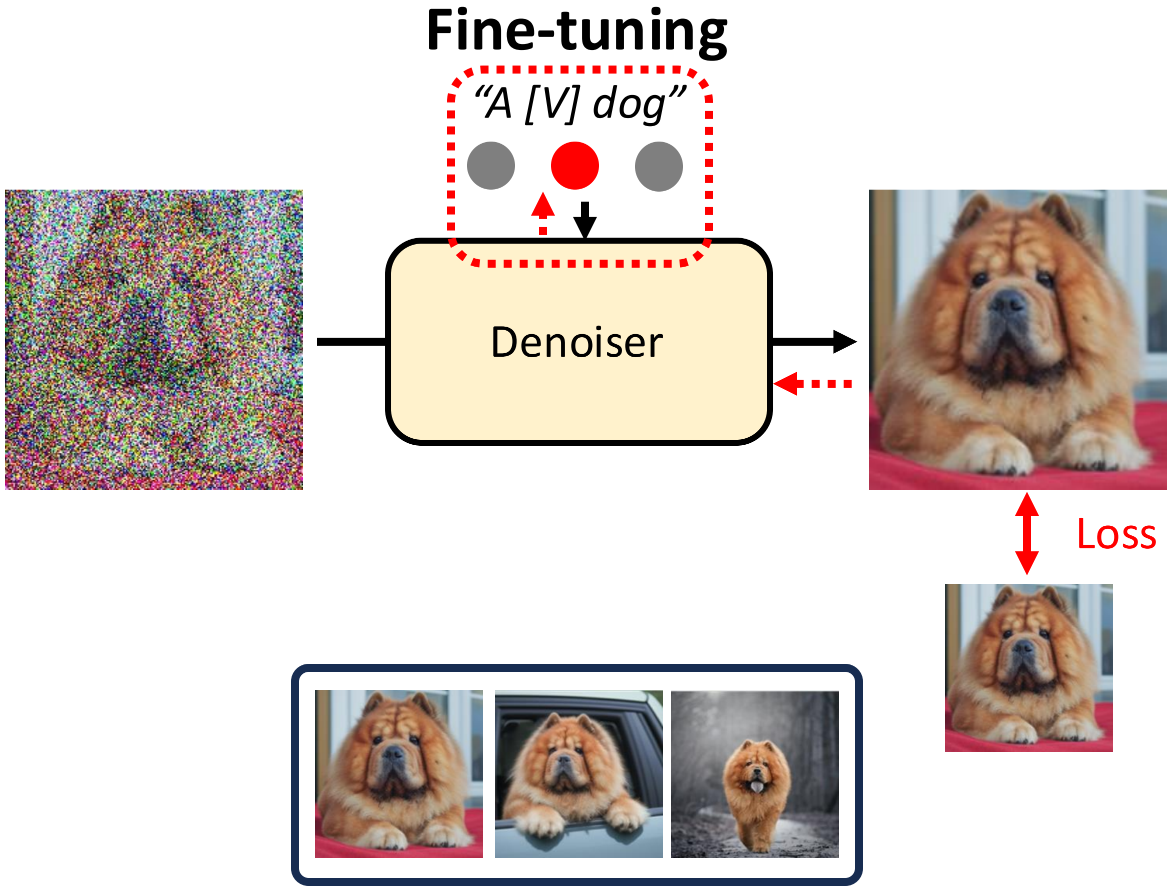
3-5 example images showing the identity.

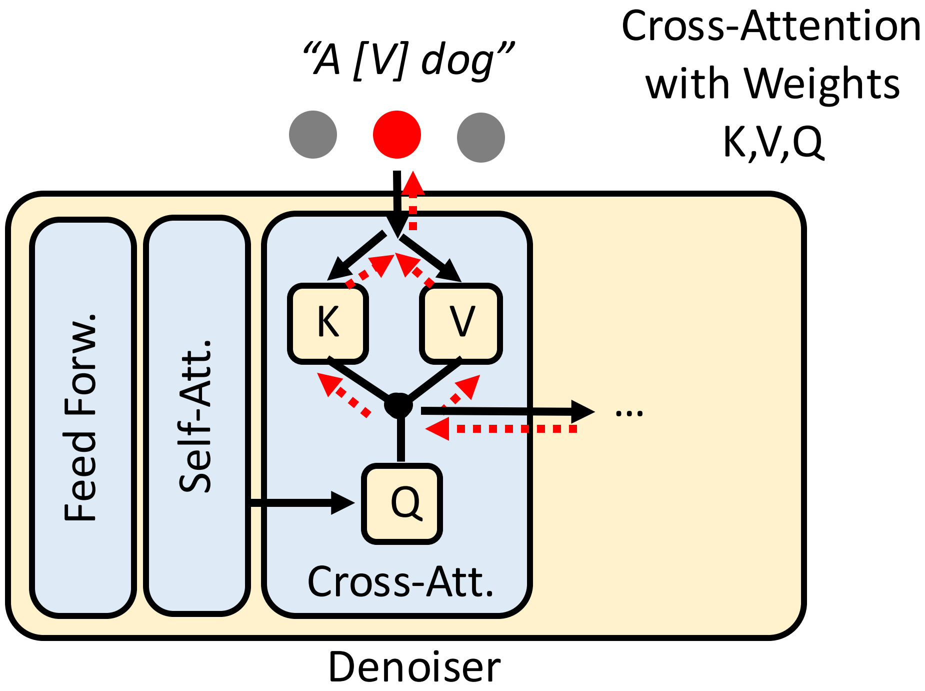


Regularization data can be generated by the original, non-finetuned model, or can come from a large dataset.

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

ID Preservation With Text Embeddings – Fine-Tune Params.





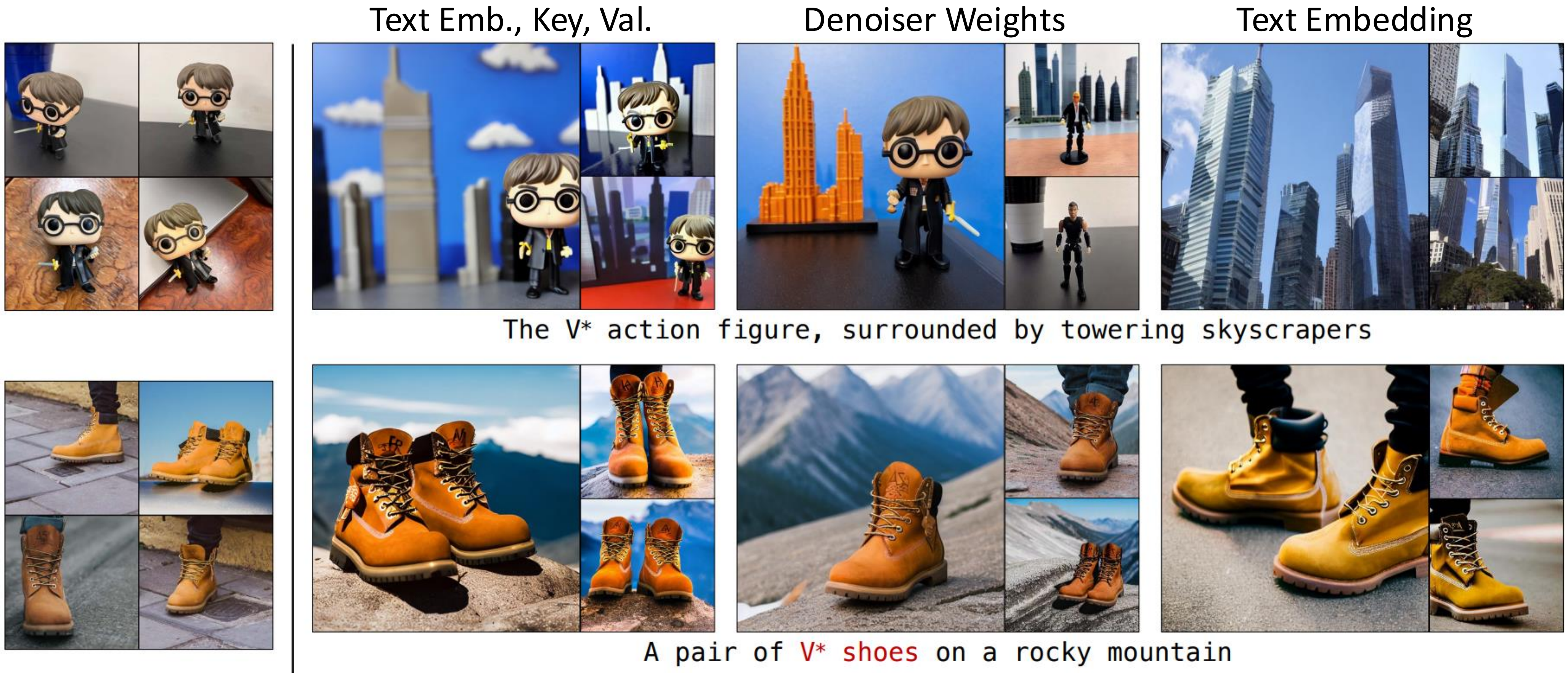
Multi-Concept Customization of Text-to-Image Diffusion, Kumari et al., CVPR 2023

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

ID Preservation With Text Embeddings

Fine-tuning text embeddings, keys and values.

ID preservation is close to tuning denoiser weights, while requiring less storage.

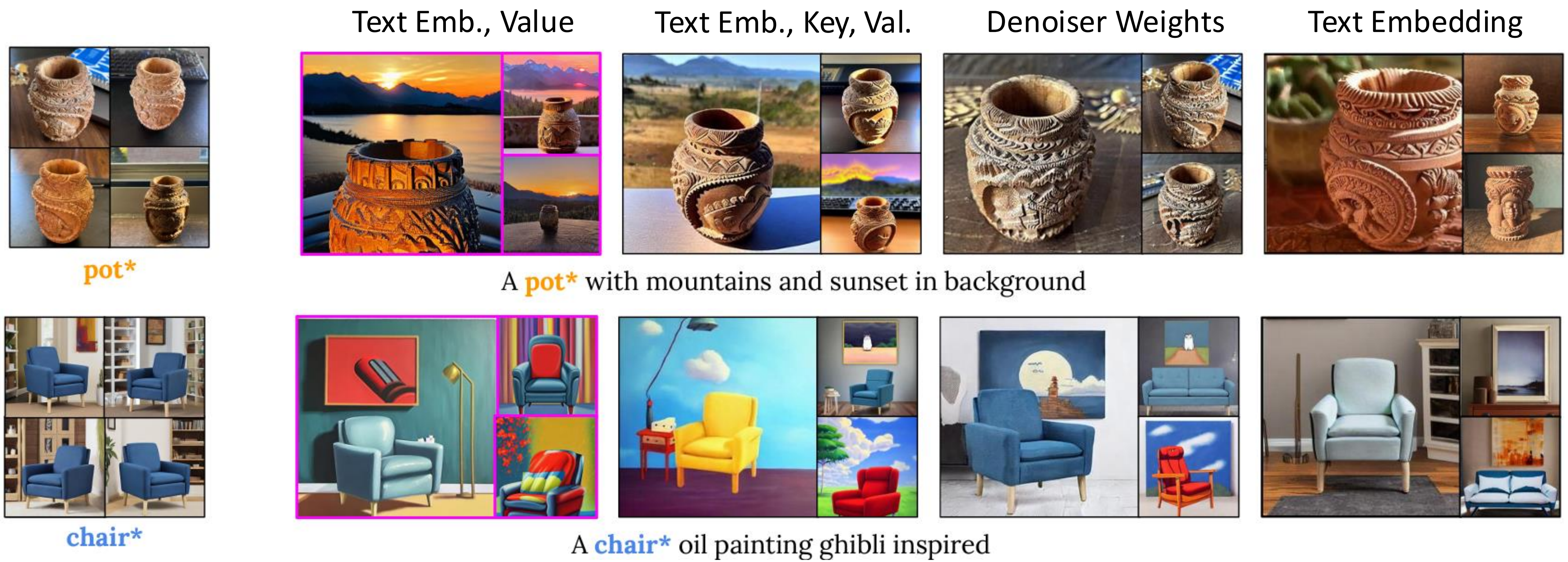

Text Emb., Key, Val.
Denoiser Weights
Text Embedding
The V* action figure, surrounded by towering skyscrapers
A pair of V* shoes on a rocky mountain

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

ID Preservation With Text Embeddings

Fine-tuning text embeddings and values only.

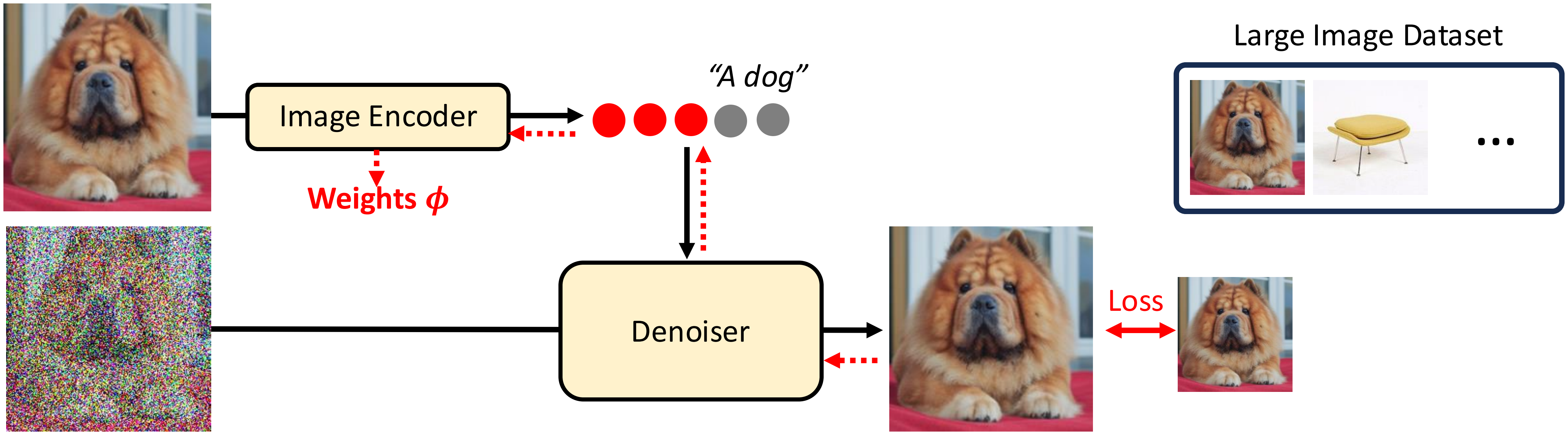
ID preservation is slightly worse than tuning denoiser weights but follows the prompt better.


Text Emb., Value
Text Emb., Key, Val.
Denoiser Weights
Text Embedding
pot*
A pot* with mountains and sunset in background
chair*
A chair* oil painting ghibli inspired

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

ID Preservation With Text Embeddings – Train Encoder

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



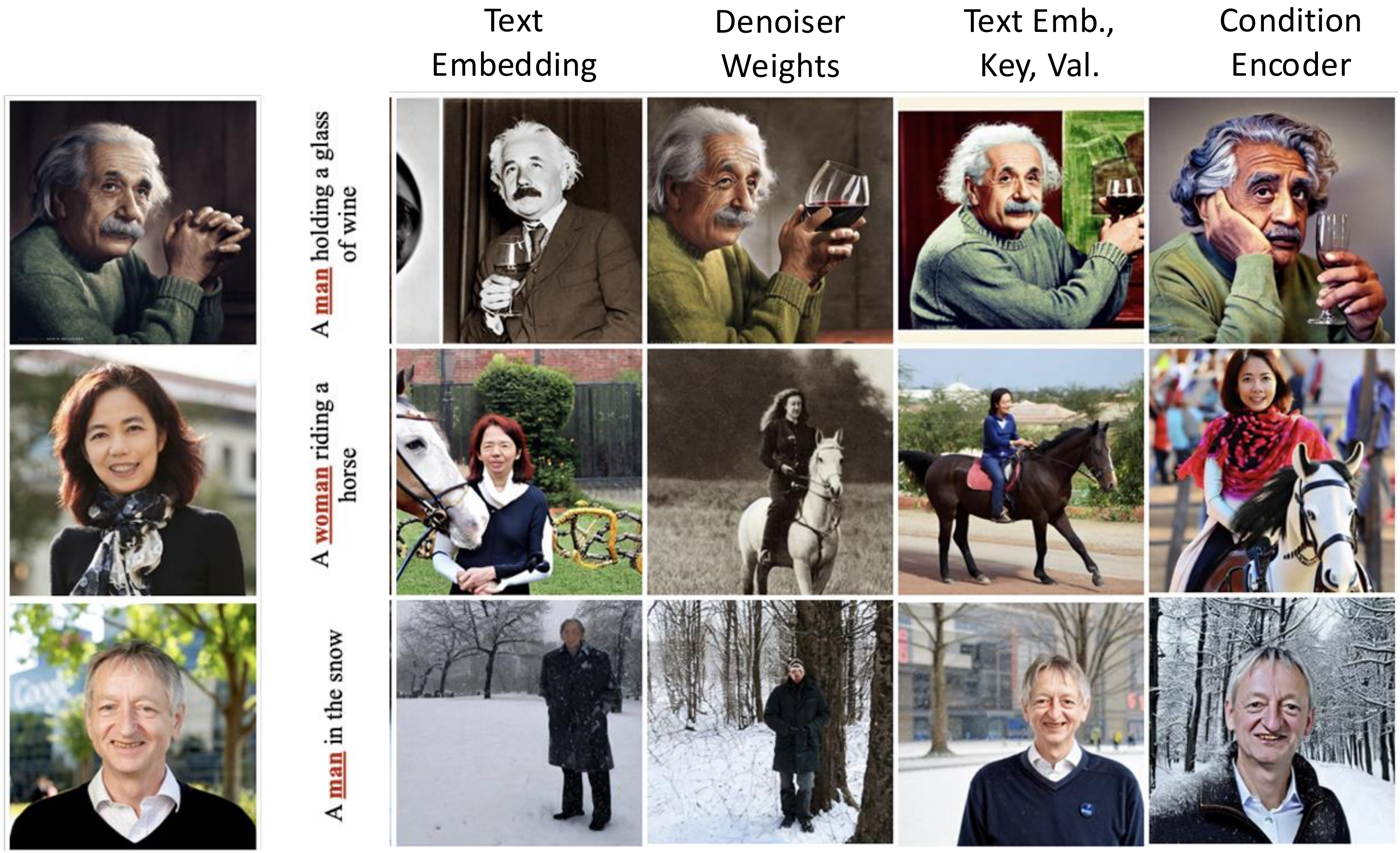
FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention, Gal et al., ArXiv May 2023

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

IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models, Ye et al., ArXiv August 2023

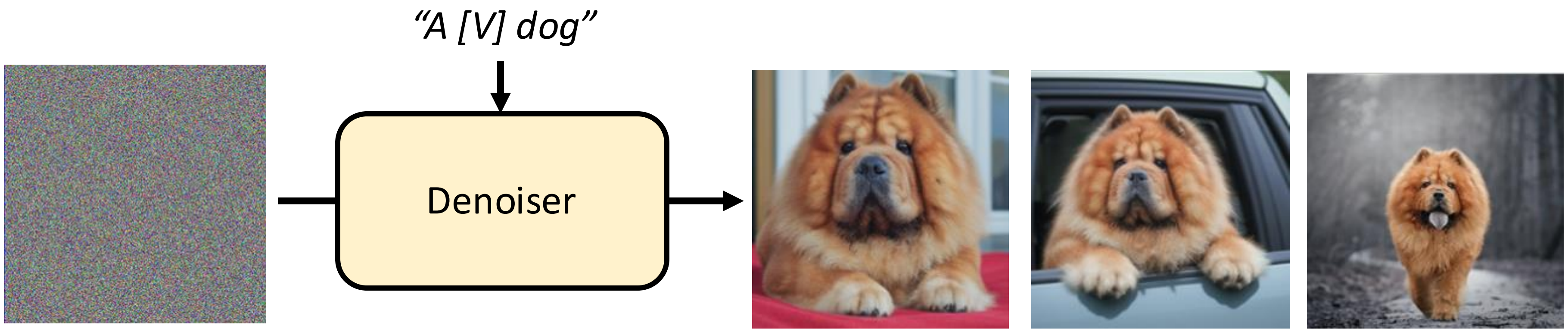
IMPRINT: Generative Object Compositing by Learning Identity-Preserving Representation, Song et al., CVPR 2024

ID Preservation with Text Embeddings



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

ID Preservation With Text Embeddings - Summary

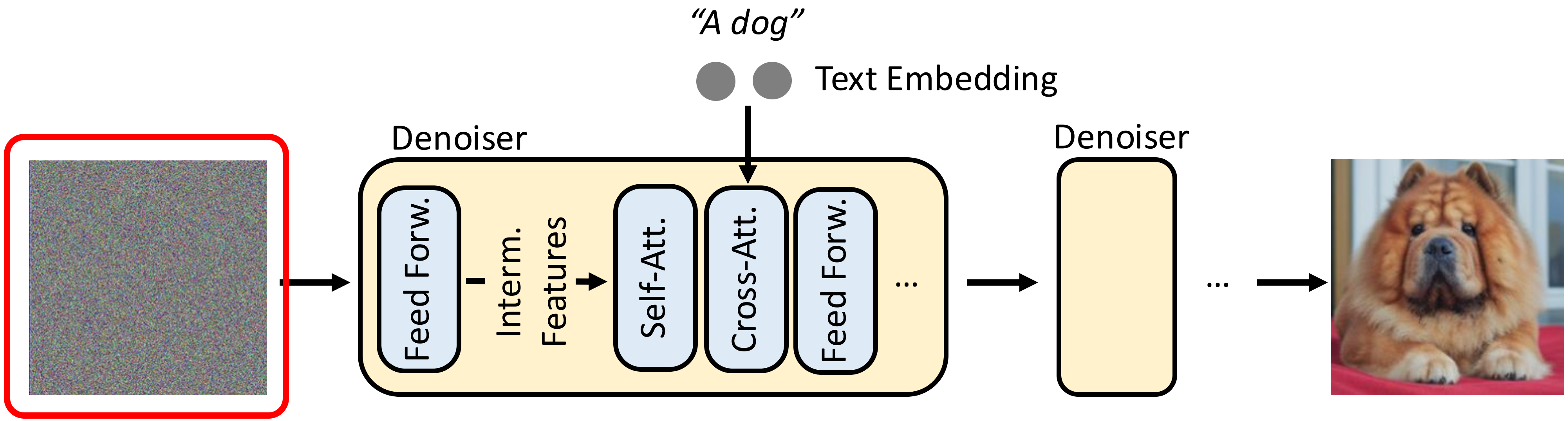


How do we associate [V] with the subject?

Strategy	Inference Speed	Memory Per Identity	Preservation
Fine-tune text embedding token	Medium	Low	Low
Fine-tune network parameters	Slow to Medium	Medium to High	High
Train image encoder & fine-tune parameters	Fast	None	Medium

Identity Preservation

What can we use to control the identity of a generated subject?

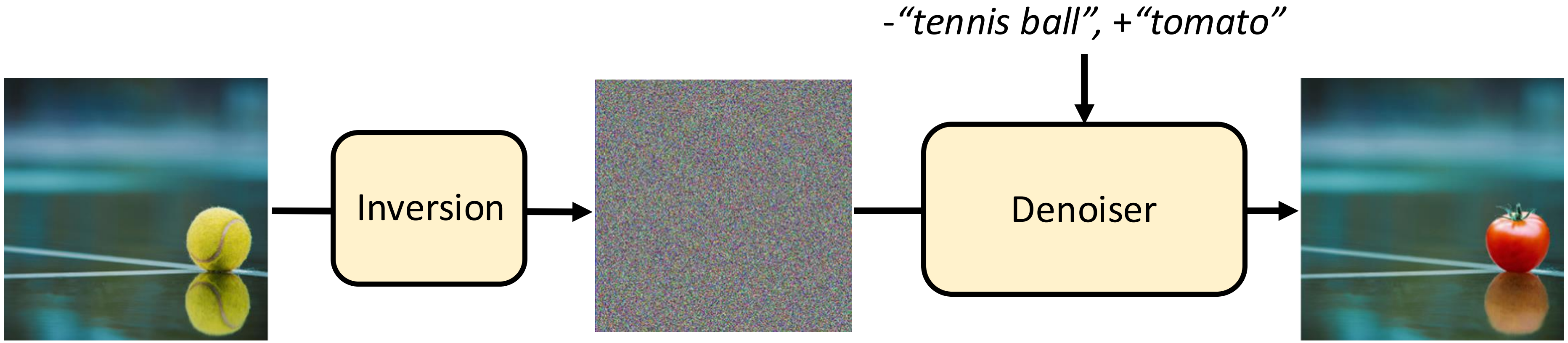


Null-text Inversion for Editing Real Images using Guided Diffusion Models, Mokady and Hertz et al., CVPR 2023

An Edit Friendly DDPM Noise Space: Inversion and Manipulations, Huberman-Spiegelglas et al., CVPR 2024

LEDITS++: Limitless Image Editing using Text-to-Image Models, Brack et al., CVPR 2024

ID Preservation Through the Input Noise



-*“tennis ball”*, +*“tomato”*

Inversion

Denoiser

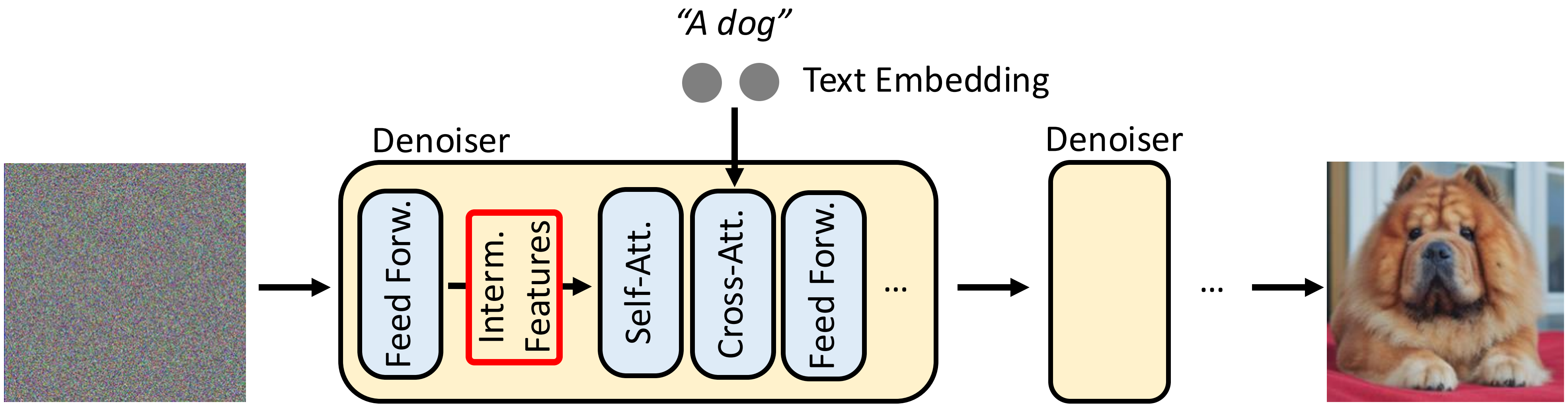
Null-text Inversion for Editing Real Images using Guided Diffusion Models, Mokady and Hertz et al., CVPR 2023

An Edit Friendly DDPM Noise Space: Inversion and Manipulations, Huberman-Spiegelglas et al., CVPR 2024

LEDITS++: Limitless Image Editing using Text-to-Image Models, Brack et al., CVPR 2024

Identity Preservation

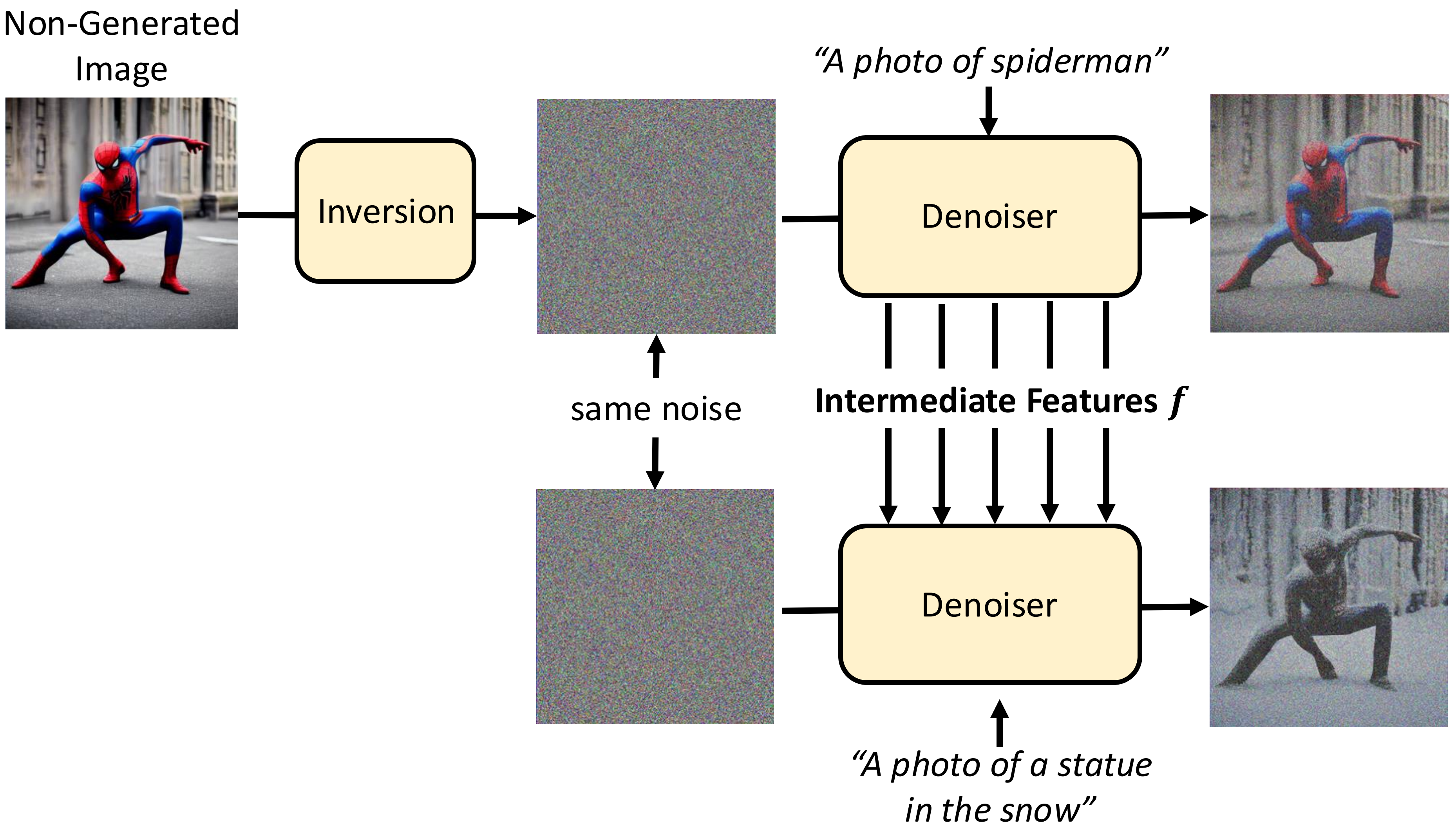
What can we use to control the identity of a generated subject?



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

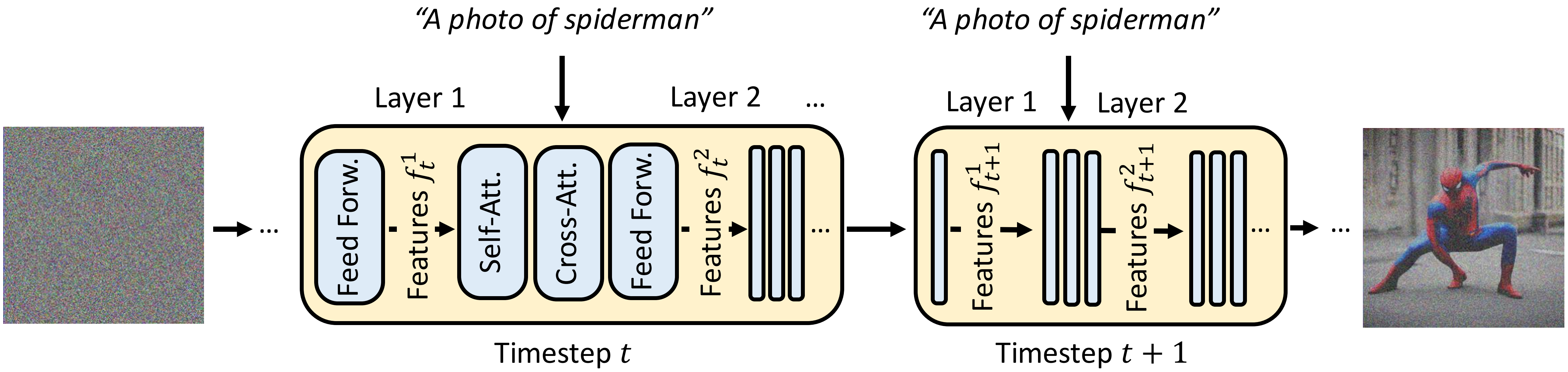
Diffusion Handles Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D, Pandey et al. CVPR 2024
Magic Fixup: Streamlining Photo Editing by Watching Dynamic Videos, Alzayer et al., ArXiv March 2024
ConsiStory: Training-Free Consistent Text-to-Image Generation, Tewel et al., Siggraph 2024

ID Preservation through Intermediate Features



ID Preservation through Intermediate Features

Which Features?



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

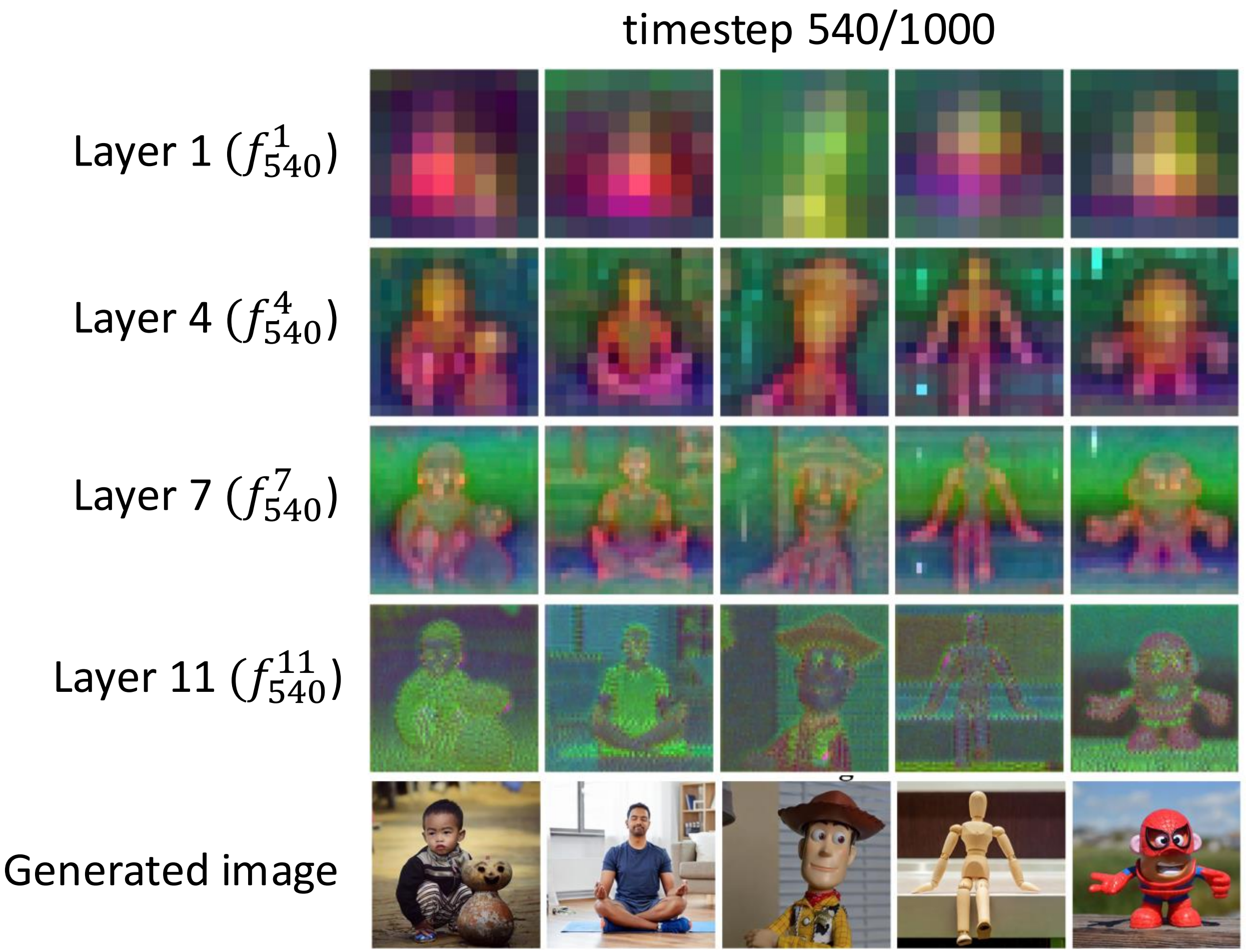
Diffusion Handles Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D, Pandey et al. CVPR 2024

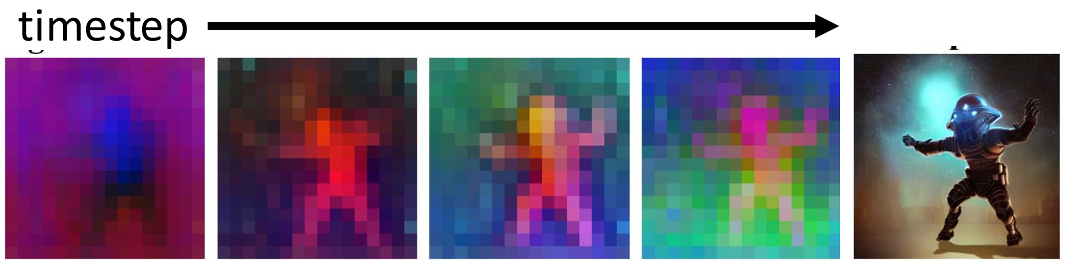
Magic Fixup: Streamlining Photo Editing by Watching Dynamic Videos, Alzayer et al., ArXiv March 2024

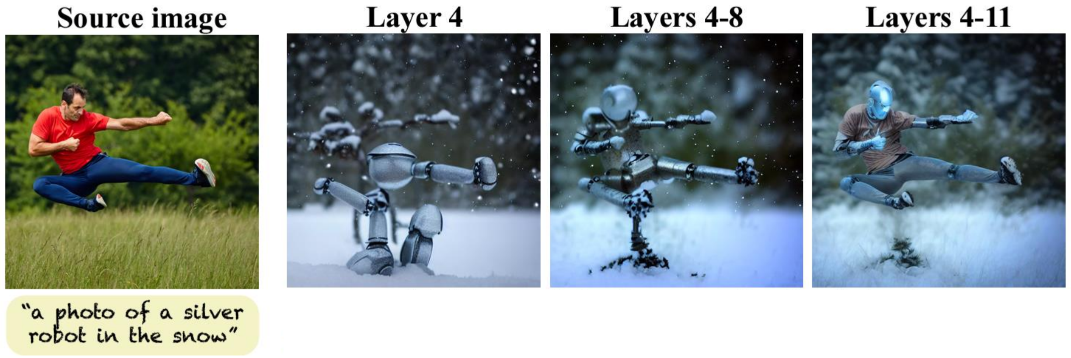
ConsiStory: Training-Free Consistent Text-to-Image Generation, Tewel et al., Siggraph 2024

ID Preservation through Intermediate Features

Earlier layers & timesteps typically contain more semantic concepts, later layers & timesteps also describe details





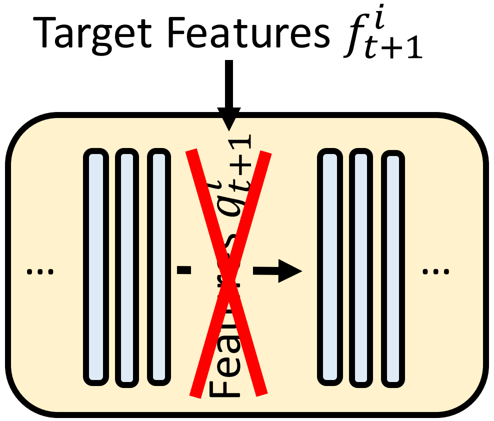


Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation, Tumanyan et al., CVPR 2023

ID Preservation through Intermediate Features

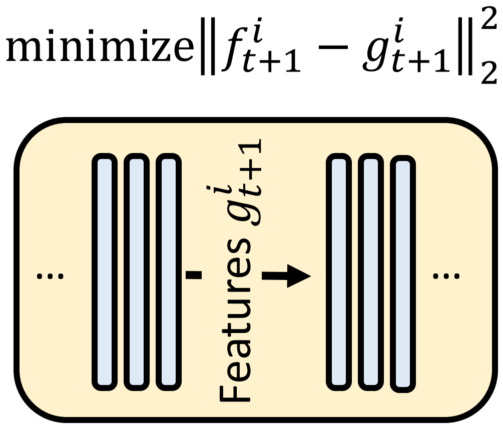
How are Target Features injected?

Overwrite denoiser features $\boldsymbol{g}$ with target features $\boldsymbol{f}$.



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

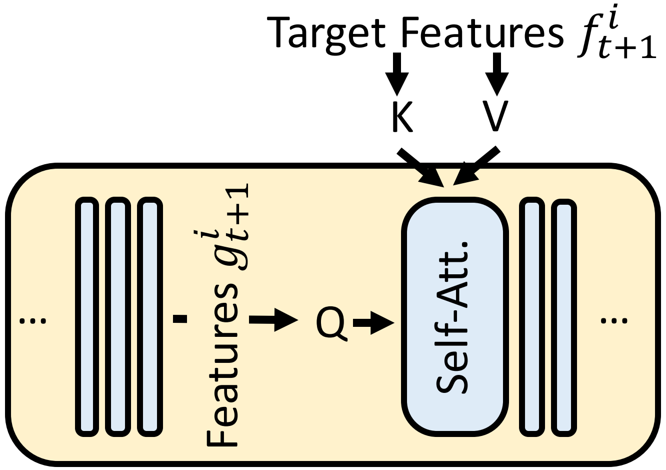
Guidance energy towards target features $\boldsymbol{f}$.



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, Pandey et al. CVPR 2024

Cross-Attention from denoiser features $\boldsymbol{g}$ to target features $\boldsymbol{f}$.



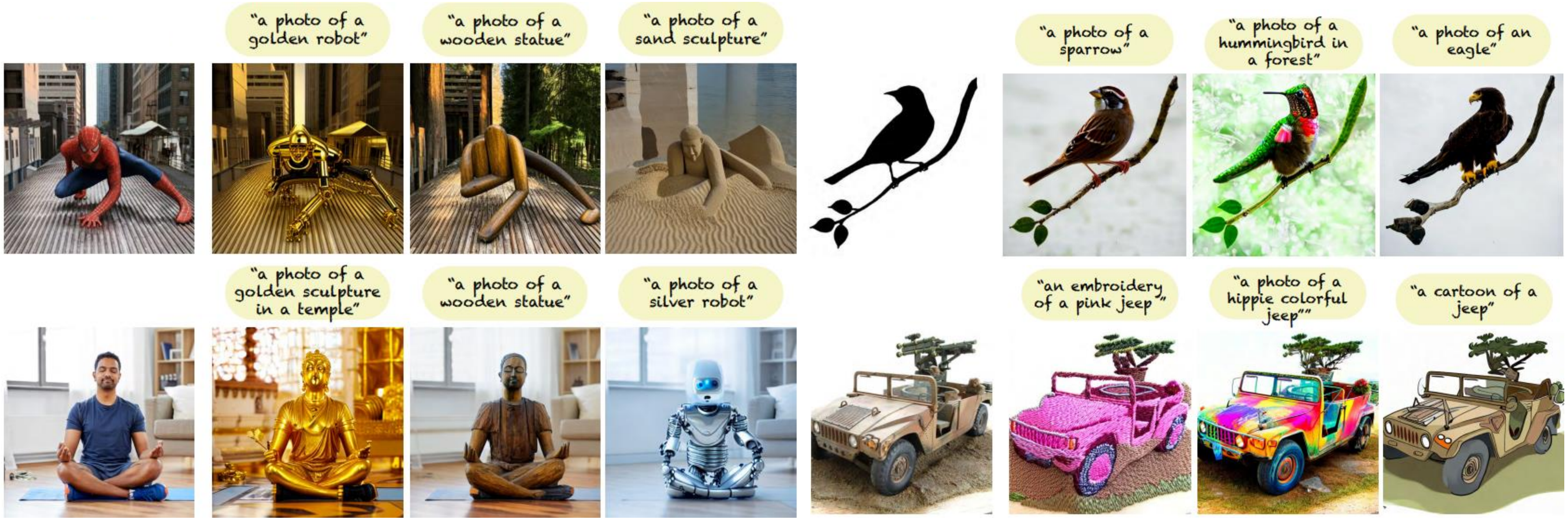
Magic Fixup: Streamlining Photo Editing by Watching Dynamic Videos, Alzayer et al., ArXiv March 2024

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

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

ID Preservation through Intermediate Features

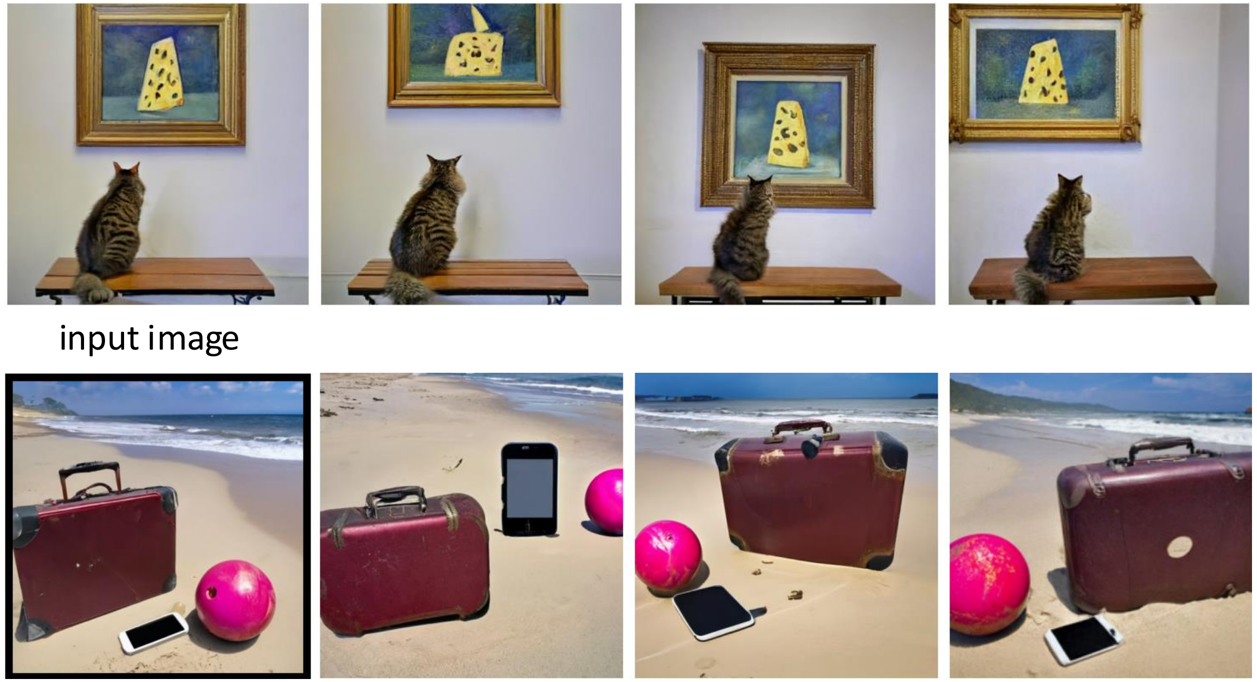
Overwrite denoiser features with target features f.


"a photo of a golden robot"
"a photo of a wooden statue"
"a photo of a sand sculpture"
"a photo of a sparrow"
"a photo of a hummingbird in a forest"
"a photo of an eagle"
"a photo of a golden sculpture in a temple"
"a photo of a wooden statue"
"a photo of a silver robot"
"an embroidery of a pink jeep"
"a photo of a hippie colorful jeep""
"a cartoon of a jeep"

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

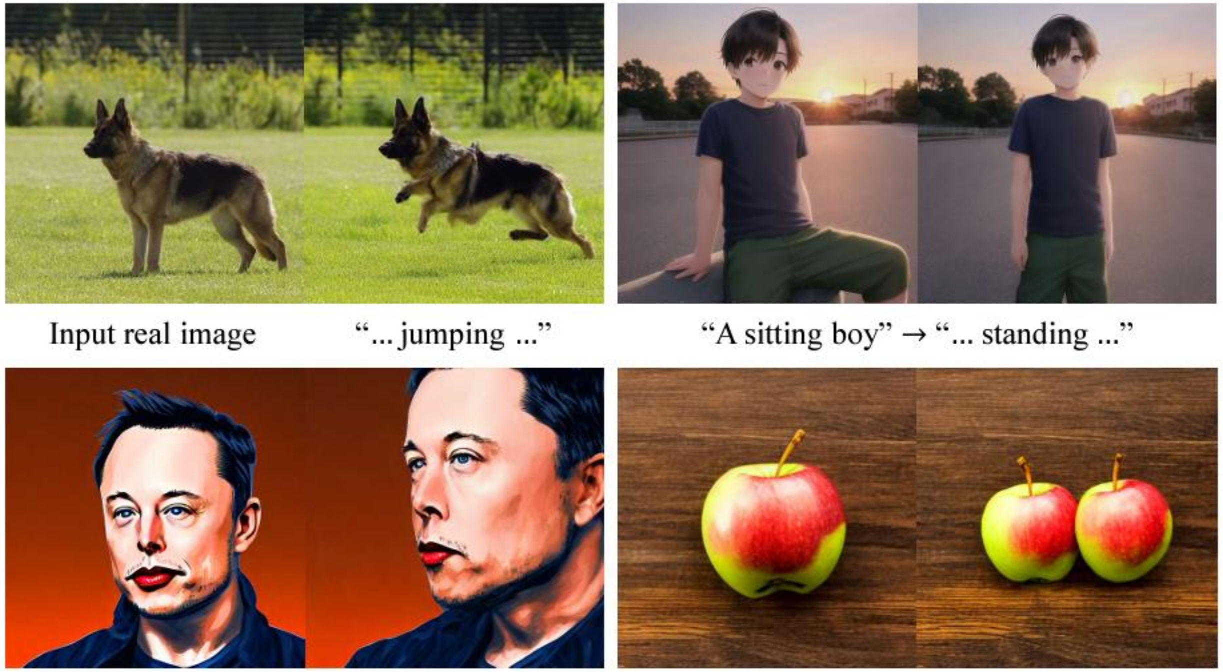
ID Preservation through Intermediate Features

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.

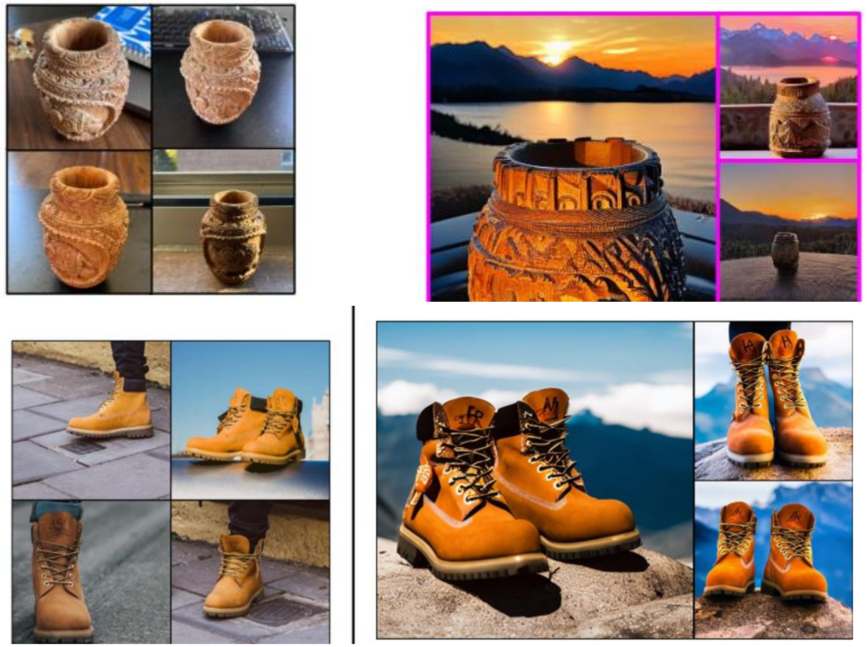


MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing, Cao et al., ICCV 2023

ID Preservation – Summary

Text Embeddings

- Specific Subject(s)
- Includes less details
- Does not include Layout
- Requires Training/Fine-Tuning



Intermediate Features (and Noise)

- Entire Image
- Includes details
- Includes layout
- Many variants are training-free

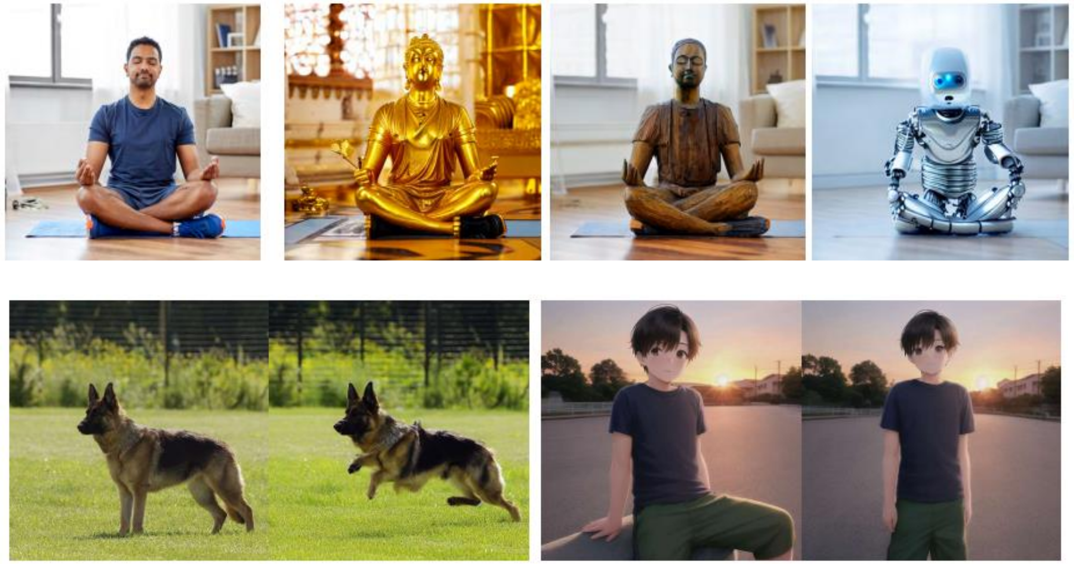
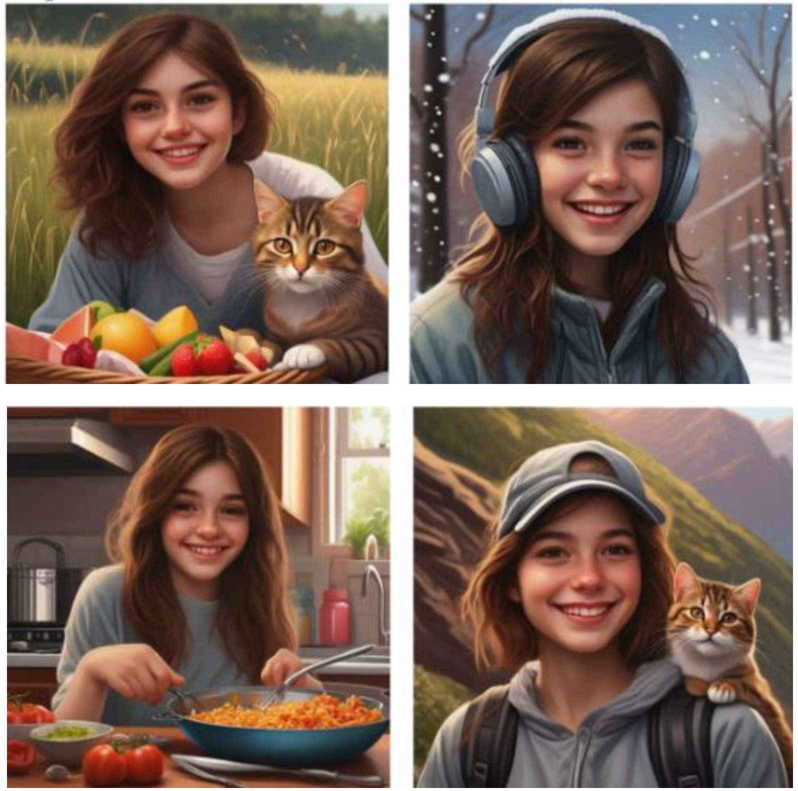


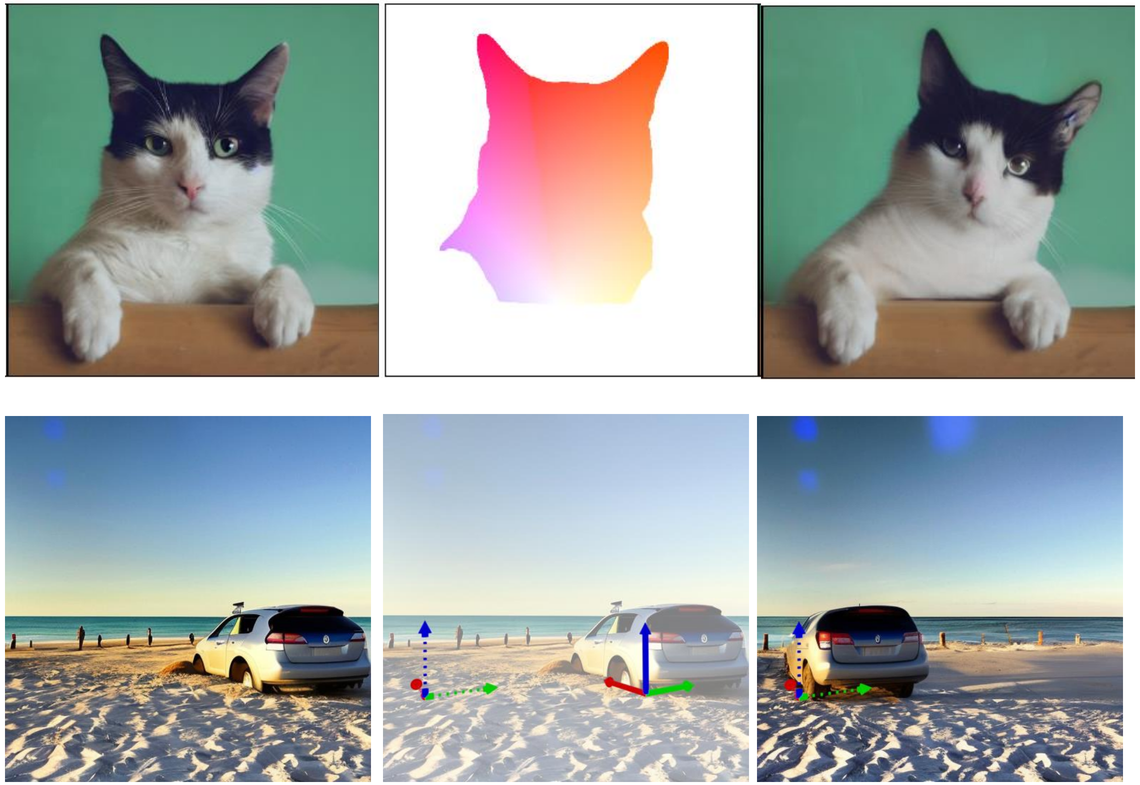
Image Editing with Generative Models

Personalization



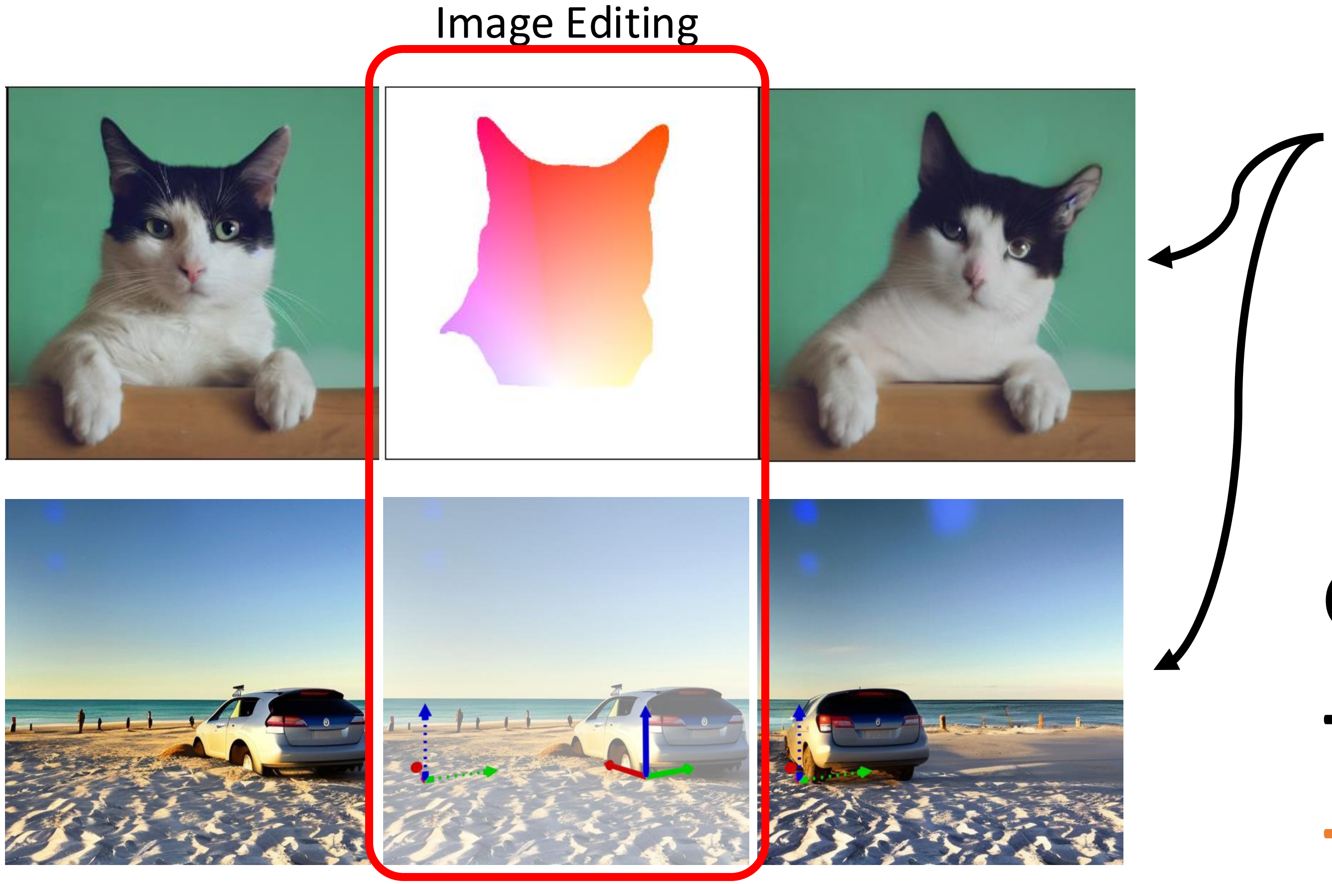
ConsiStory: Training-Free Consistent Text-to-Image Generation
Tewel et al., ArXiv Feb. 2024

Image 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

Image Editing with Generative Models



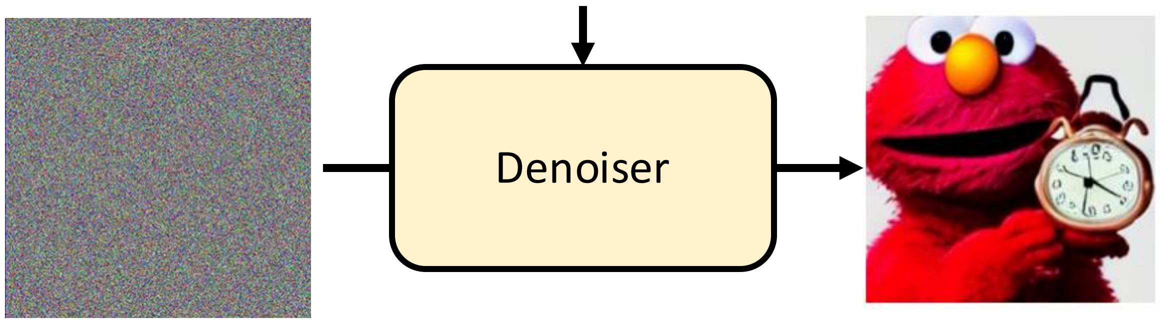
Same subject, same scene.
Subject property changed by user **edit**.
(Property such as position, pose, etc.)

Image 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

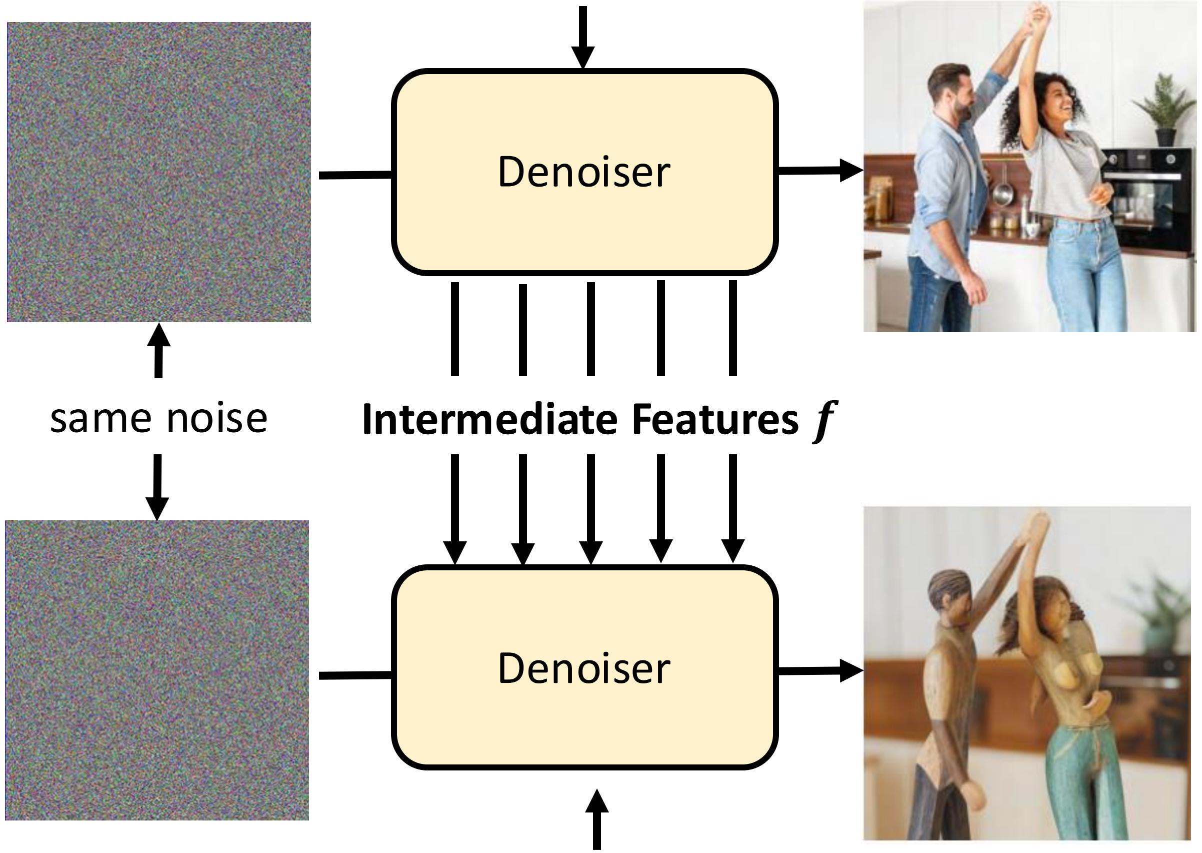
Edit Control Through Text



*“Elmo holding a **[V]**”*

Denoiser

Fine-tuned on



“A couple dancing.”

Denoiser

same noise

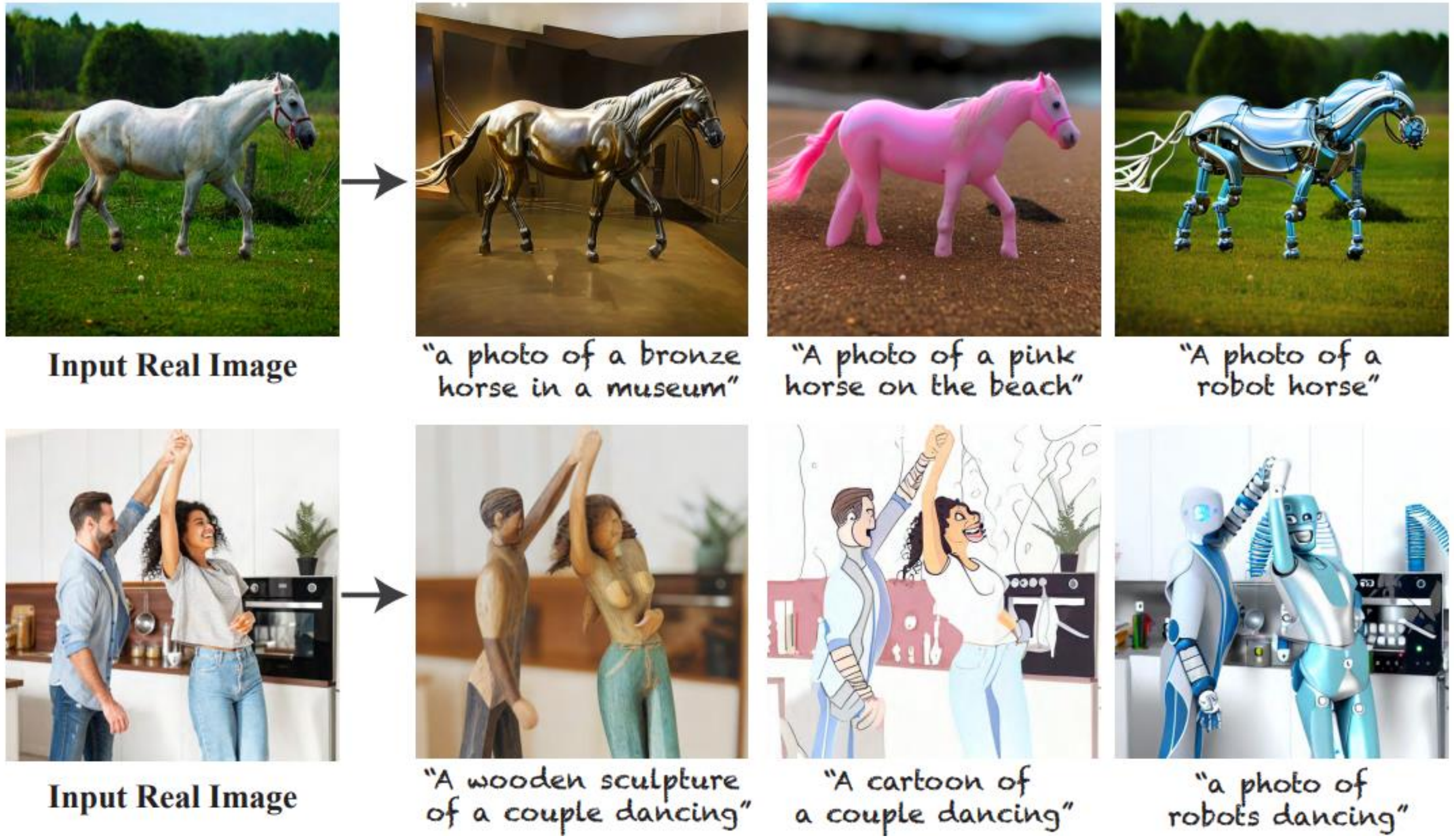
Intermediate Features f

Denoiser

“A wooden sculpture of A couple dancing.”

Edit Control Through Text

- Most widely-used form of control
- Very general in what it can control.
- Only coarse control. (No detailed control over locations/layouts/amounts/degrees.)



Input Real Image

"a photo of a bronze horse in a museum"

"A photo of a pink horse on the beach"

"A photo of a robot horse"

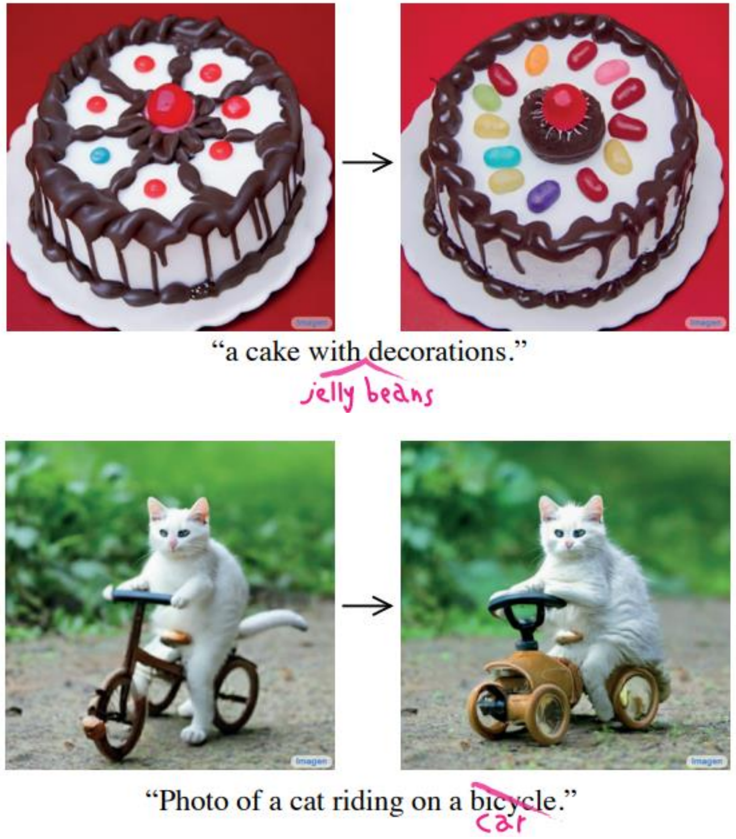
Input Real Image

"A wooden sculpture of a couple dancing"

"A cartoon of a couple dancing"

"a photo of robots dancing"

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



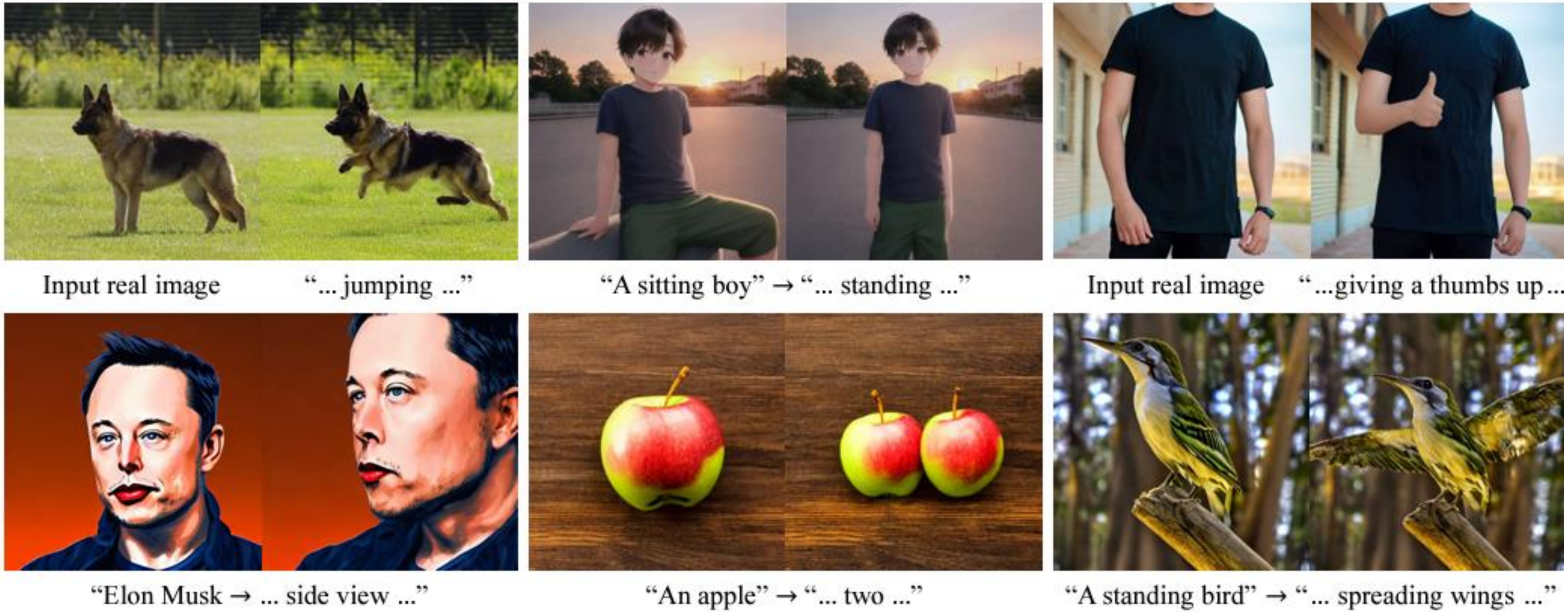
"a cake with decorations." jelly beans

"Photo of a cat riding on a ~~bicycle~~ car."

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

Edit Control Through Text

- Most widely-used form of control
- Very general in what it can control.
- Only coarse control. (No detailed control over locations/layouts/amounts/degrees.)



Input real image “... jumping ...”

“A sitting boy” → “... standing ...”

Input real image “...giving a thumbs up...

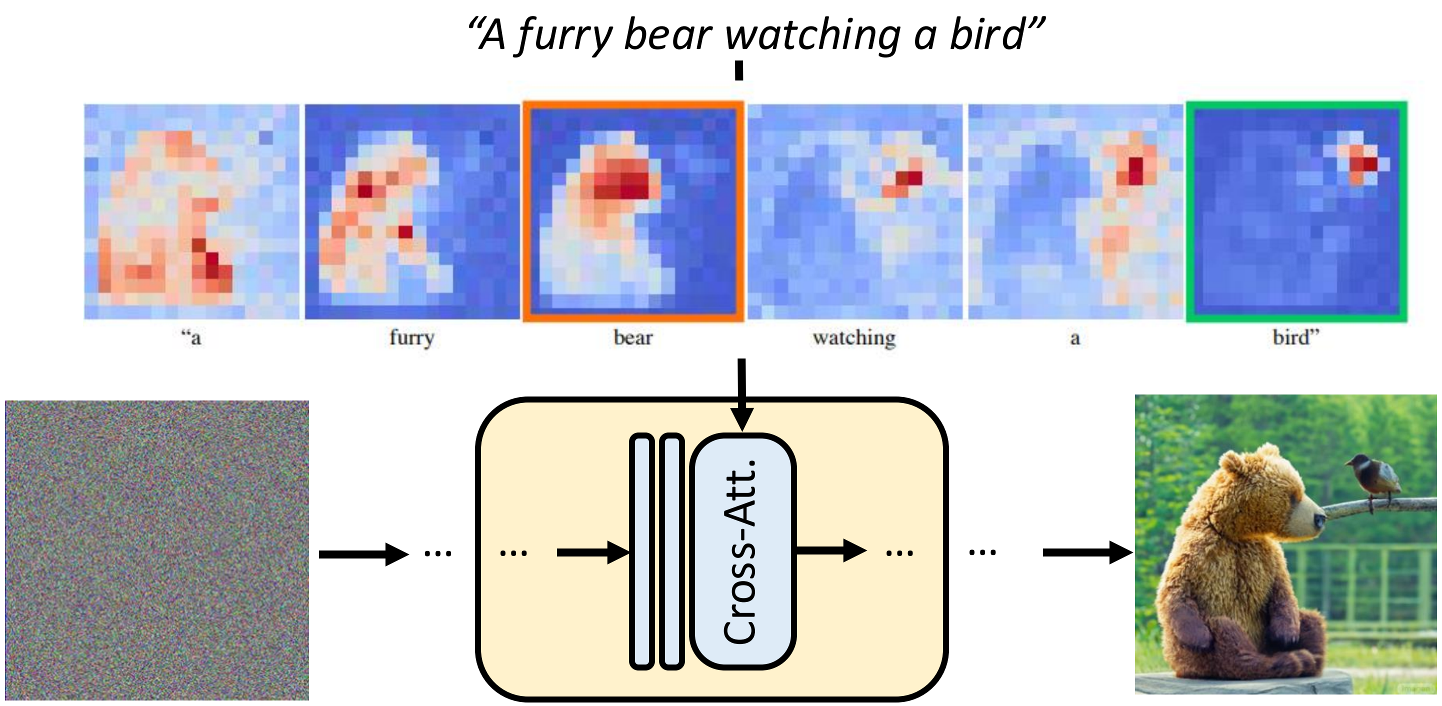
“Elon Musk → ... side view ...”

“An apple” → “... two ...”

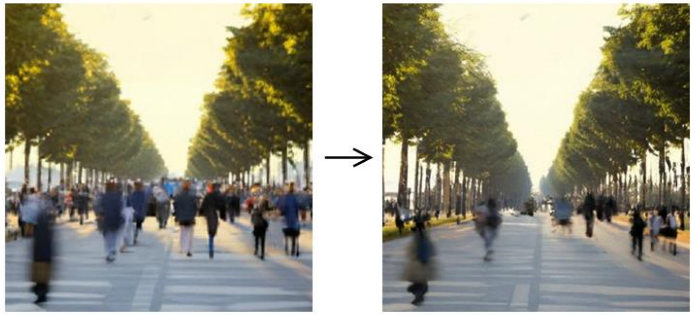
“A standing bird” → “... spreading wings ...”

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

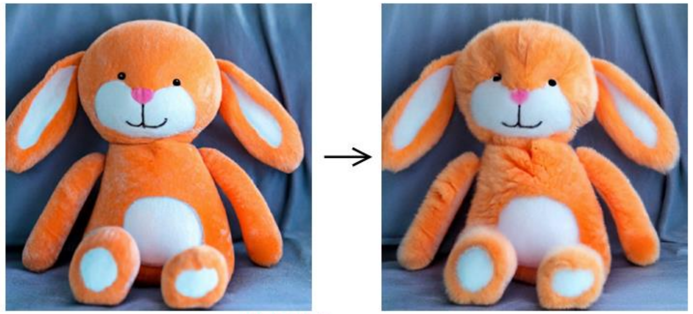
Edit Control Through Cross-Attention Maps



“A furry bear watching a bird”



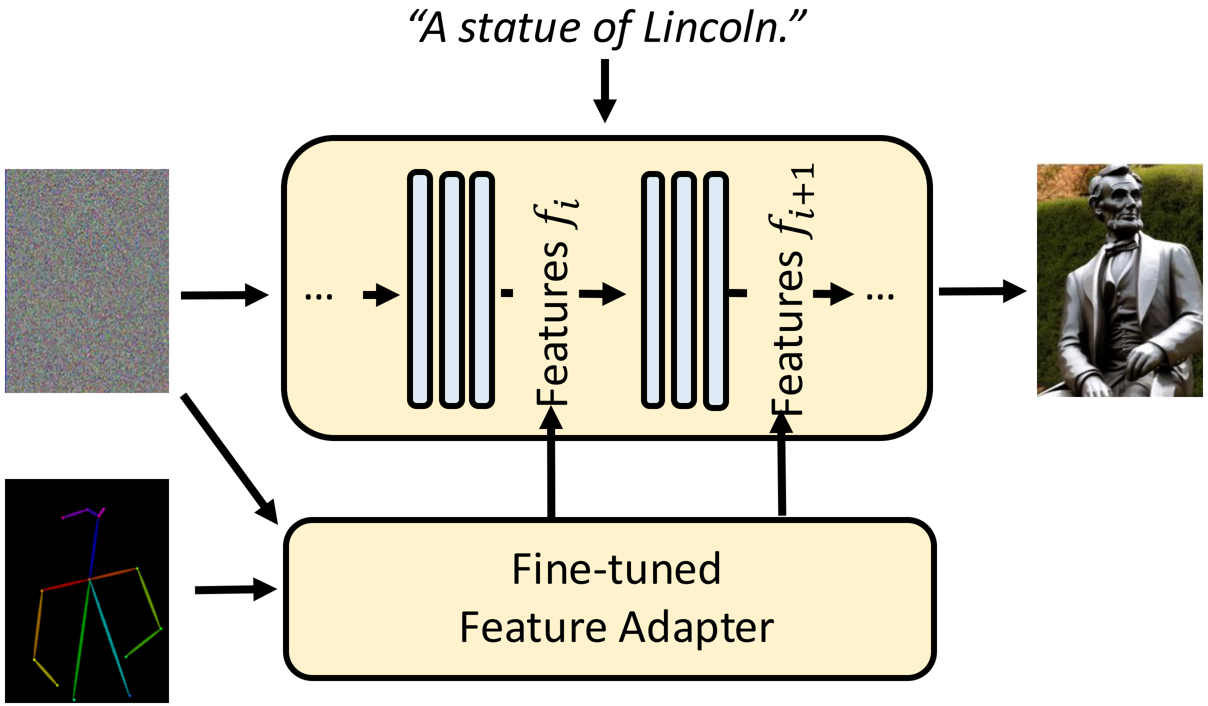
“The boulevards are crowded today.”



“My fluffy bunny doll.”

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

Edit Control Through Learned Modifications of Intermediate Features



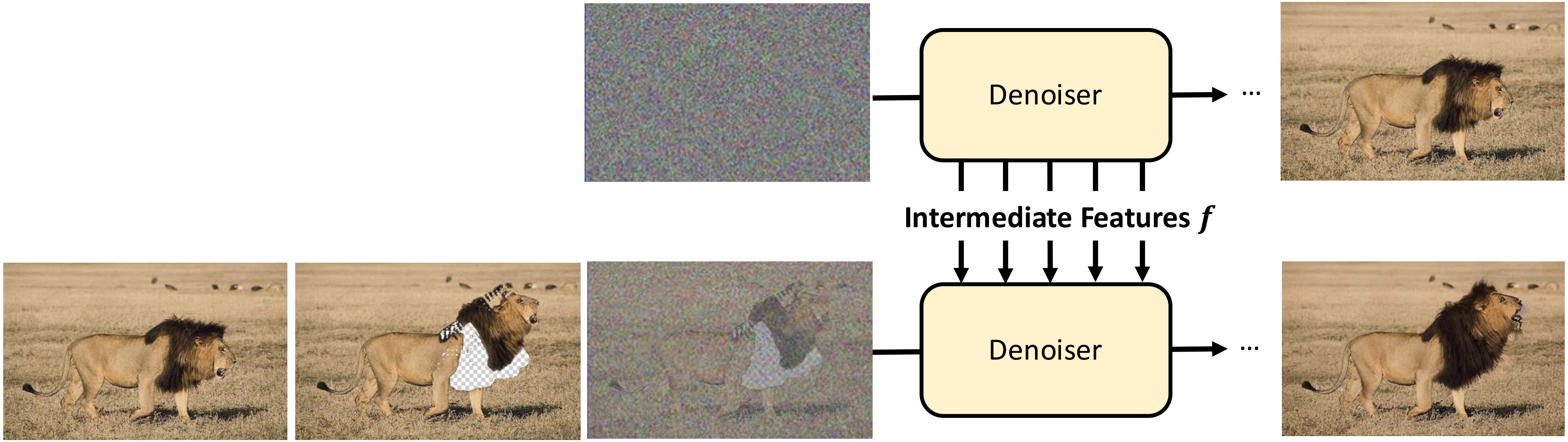
"A statue of Lincoln."

… Features f_i Features f_{i+1} …

Fine-tuned
Feature Adapter

Adding Conditional Control to Text-to-Image Diffusion Models, Zhang et al., ICCV 2023 (a.k.a. ControlNet)
LooseControl: Lifting ControlNet for Generalized Depth Conditioning, Bhat et al., ArXiv Dec. 2023

Edit Control Through the Noisy Input



Denoiser

Intermediate Features *f*

Denoiser

Input Image

Coarse Edit

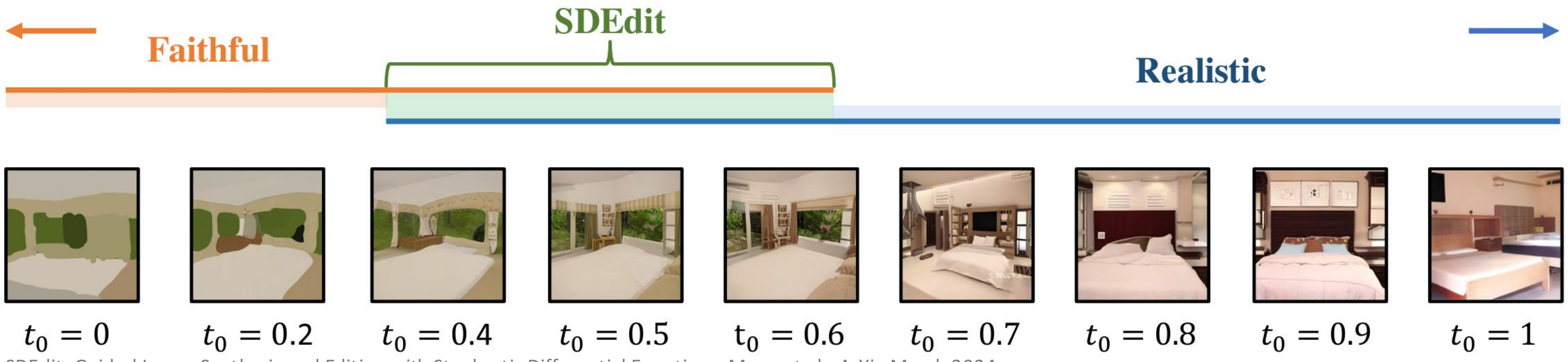
Coarse Edit + Noise

Only attention-based feature injection possible.
-> Training required with pairs of (coarse edit, ground truth).

SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations, Meng et al., ArXiv March 2024
Magic Fixup: Streamlining Photo Editing by Watching Dynamic Videos, AlZayer et al., ArXiv March 2024
Image Sculpting: Precise Object Editing with 3D Geometry Control, Yenphraphai et al., CVPR 2024

Edit Control Through the Noisy Input

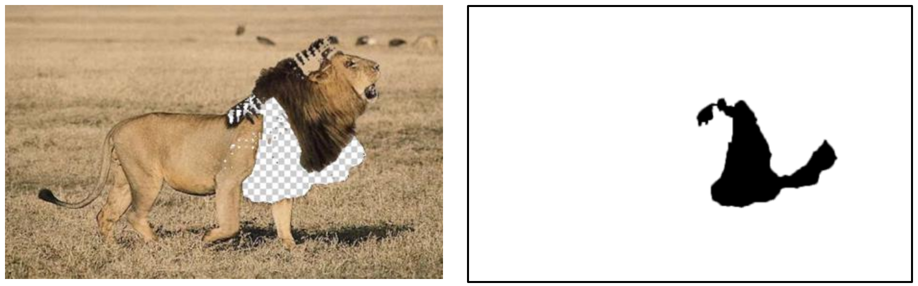
How much noise?



SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations, Meng et al., ArXiv March 2024

Disocclusions?

Fine-tune generator to use masks of disoccluded regions.



Magic Fixup: Streamlining Photo Editing by Watching Dynamic Videos, AlZayer et al., ArXiv March 2024

Use coarse estimate of disoccluded regions in coarse edit.

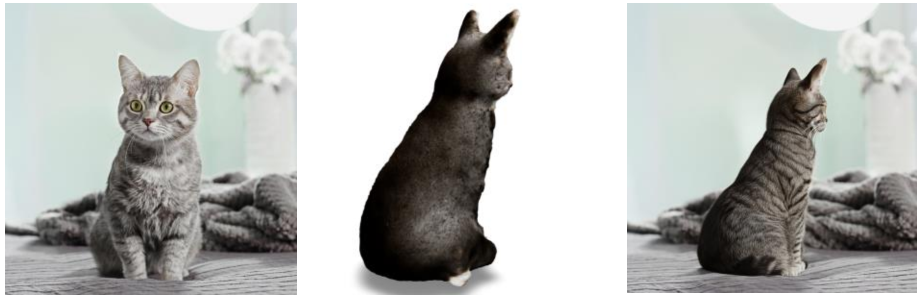


Image Sculpting: Precise Object Editing with 3D Geometry Control, Yenphraphai et al., CVPR 2024

Edit Control Through the Noisy Input

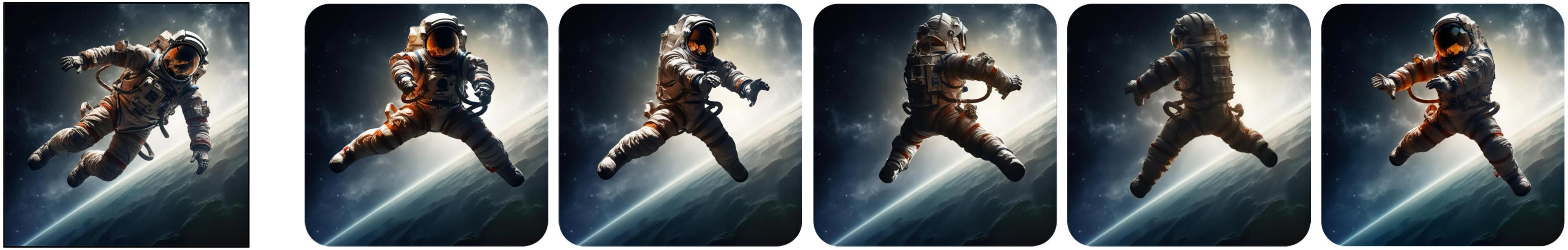
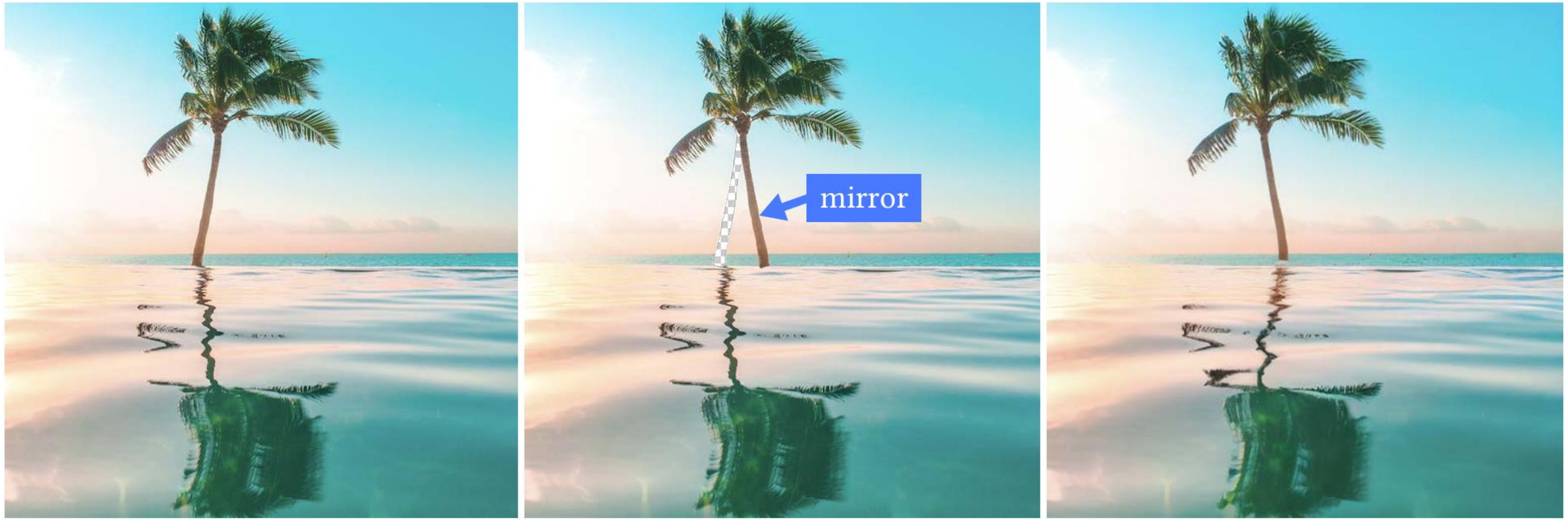


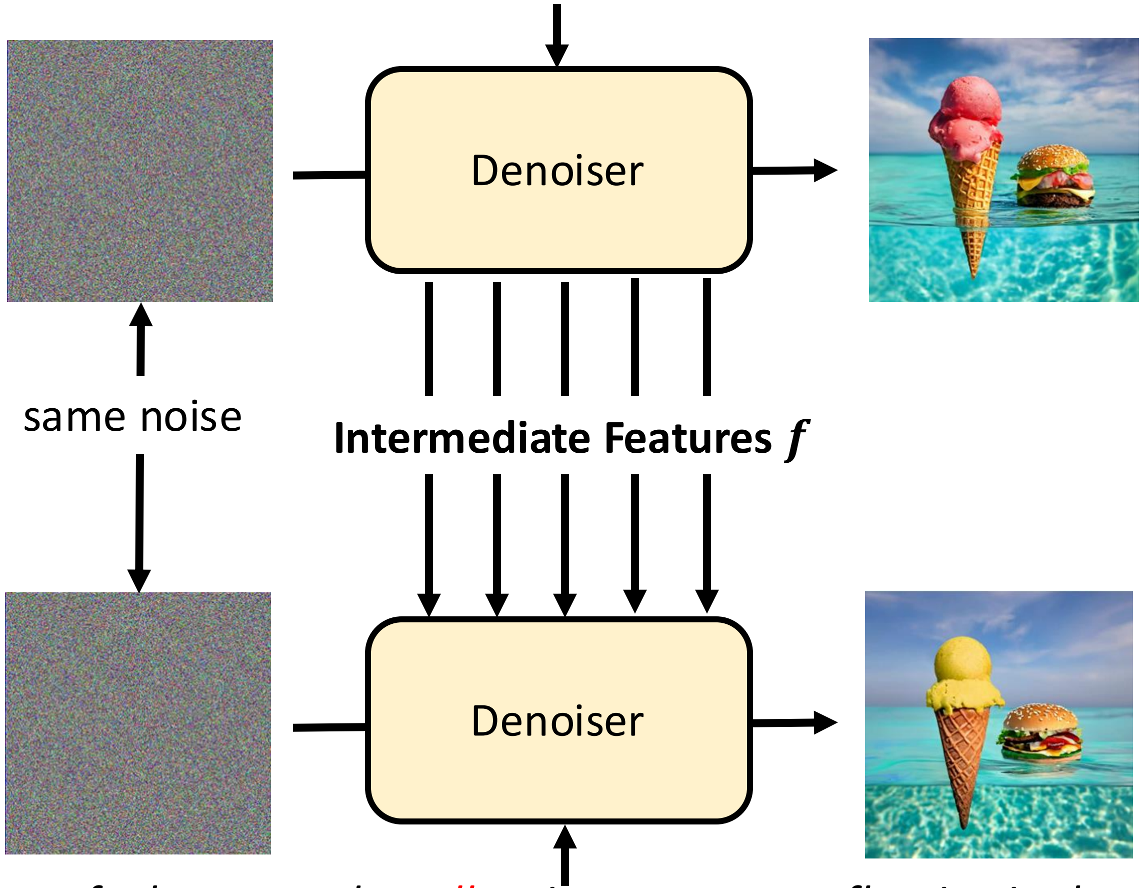
Image Sculpting: Precise Object Editing with 3D Geometry Control, Yenphraphai et al., CVPR 2024



Magic Fixup: Streamlining Photo Editing by Watching Dynamic Videos, AlZayer et al., ArXiv March 2024

Edit Control by Moving Intermediate Features

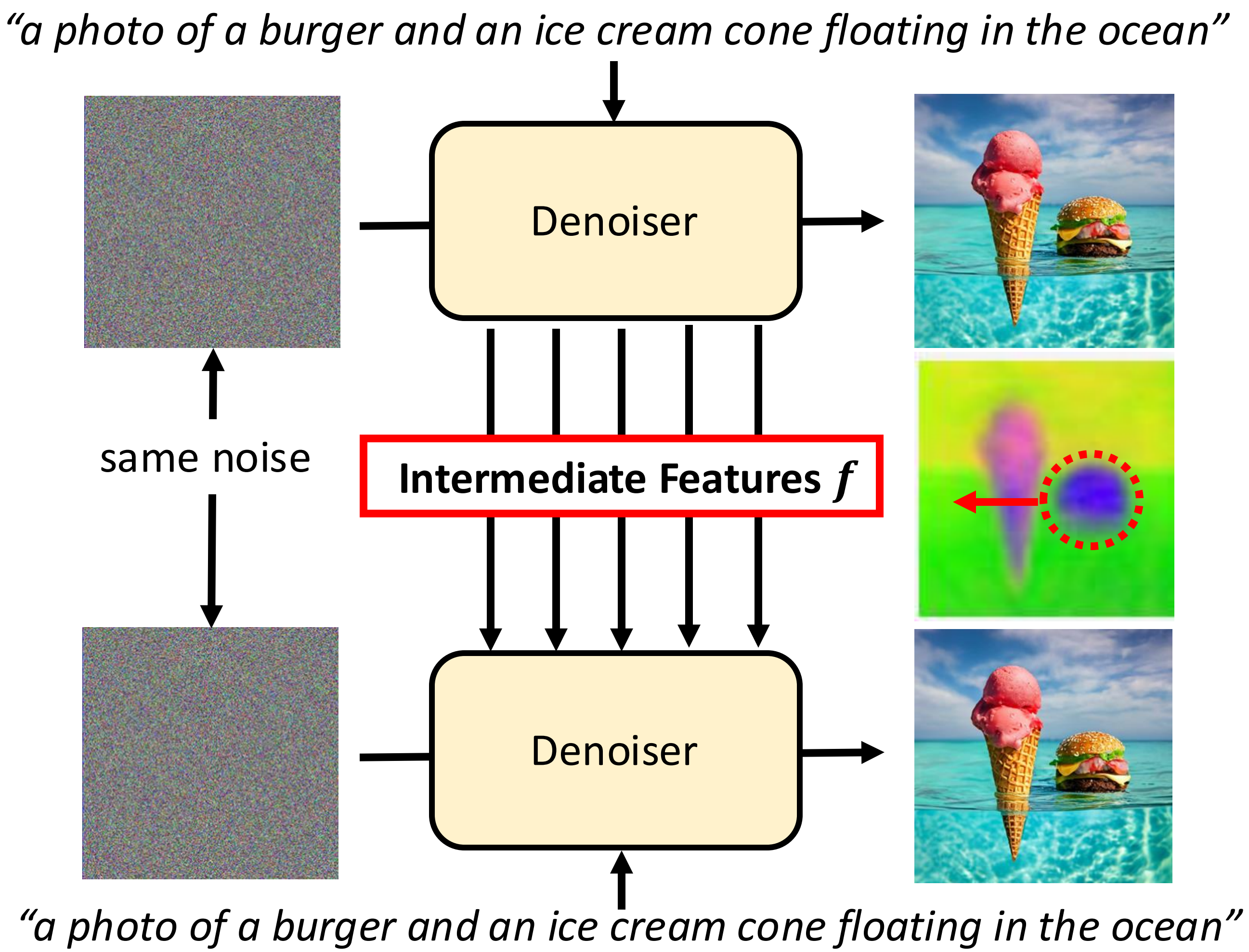
“a photo of a burger and an ice cream cone floating in the ocean”



“a photo of a burger and a yellow ice cream cone floating in the ocean”

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

Edit Control by Moving Intermediate Features



"a photo of a burger and an ice cream cone floating in the ocean"

Denoiser

same noise

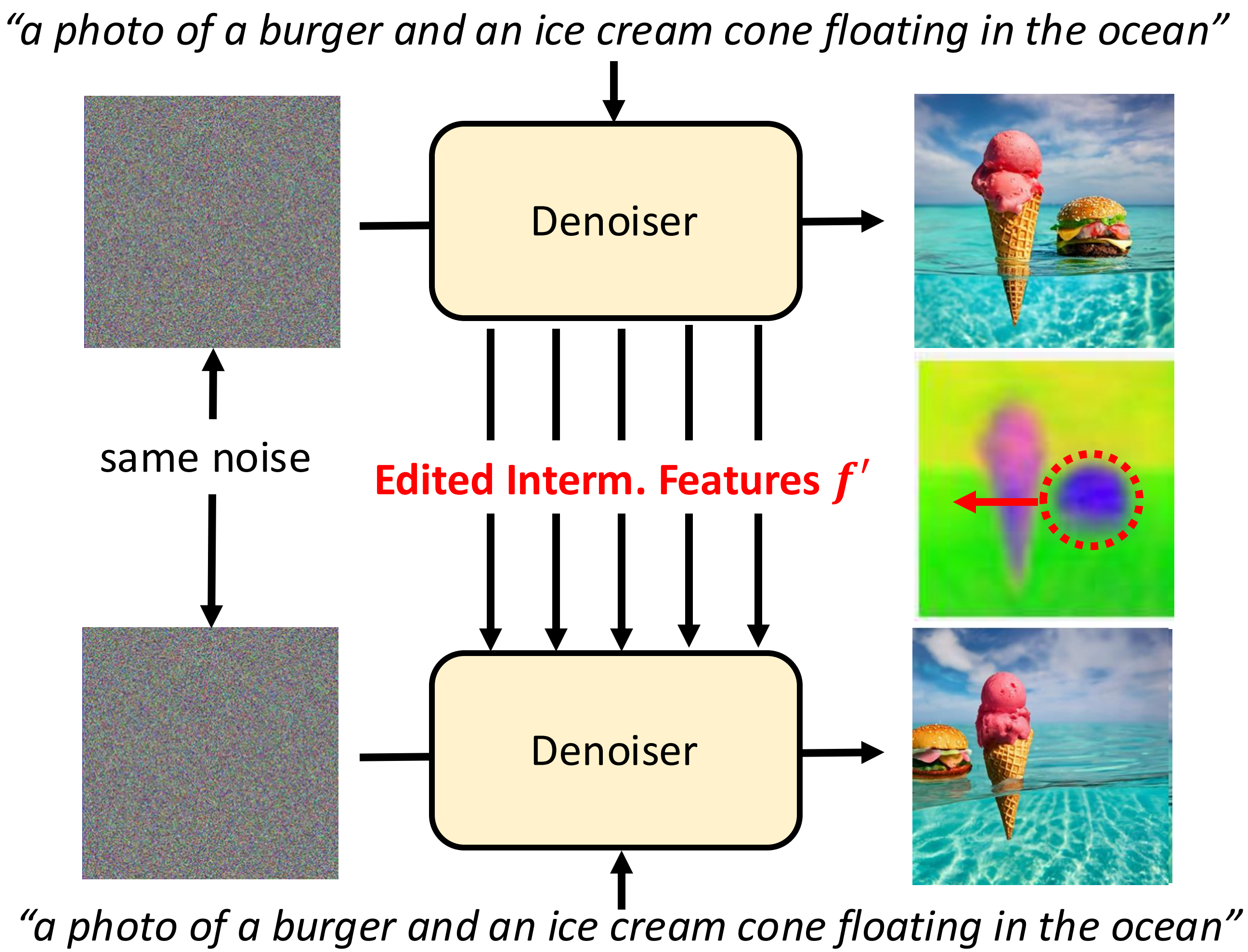
Intermediate Features f

Denoiser

"a photo of a burger and an ice cream cone floating in the ocean"

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

Edit Control by Moving Intermediate Features

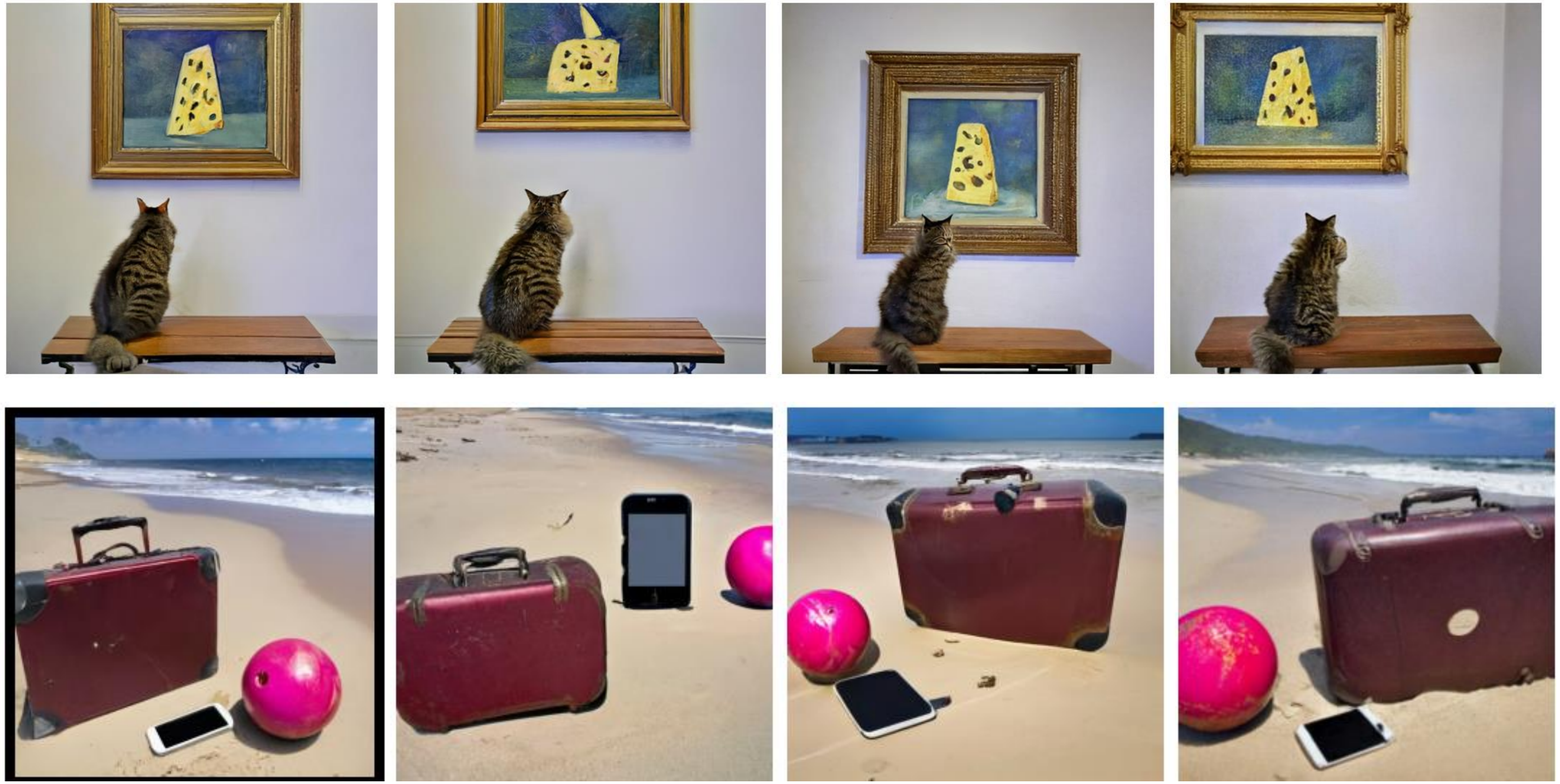


"a photo of a burger and an ice cream cone floating in the ocean"
Denoiser
same noise
Edited Interm. Features f'
Denoiser
"a photo of a burger and an ice cream cone floating in the ocean"

Overwrite/guidance-based feature injection possible.
-> Training-free approach.

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

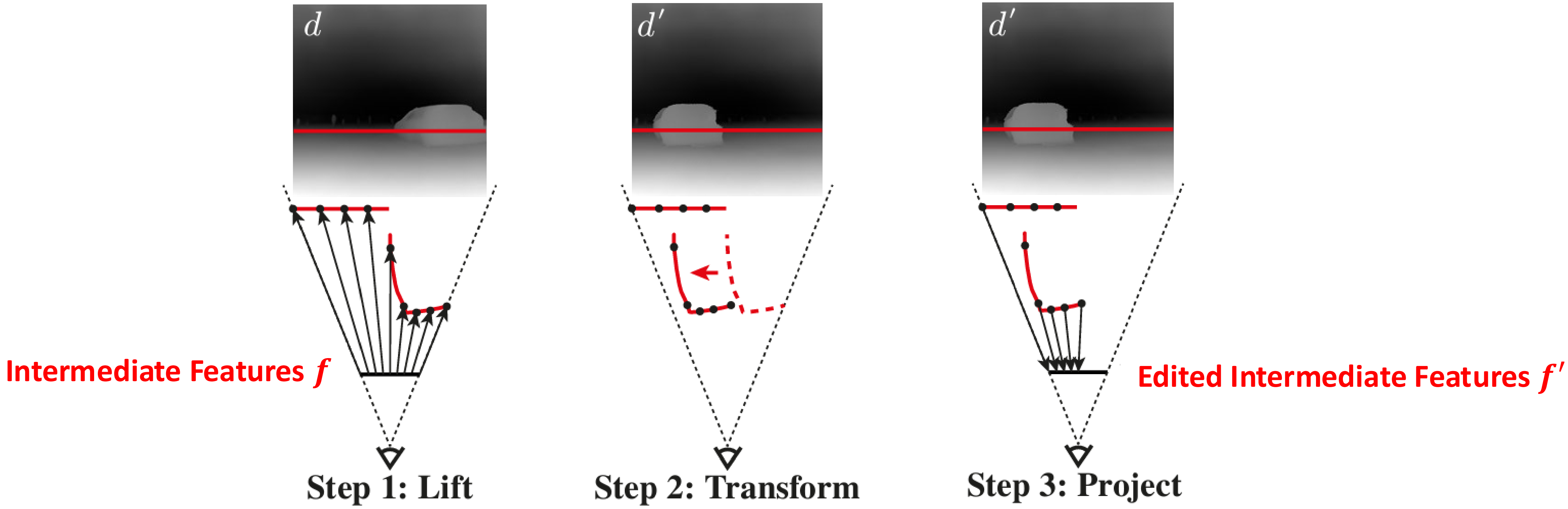
Edit Control by Moving Intermediate Features



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

Edit Control by Moving Intermediate Features

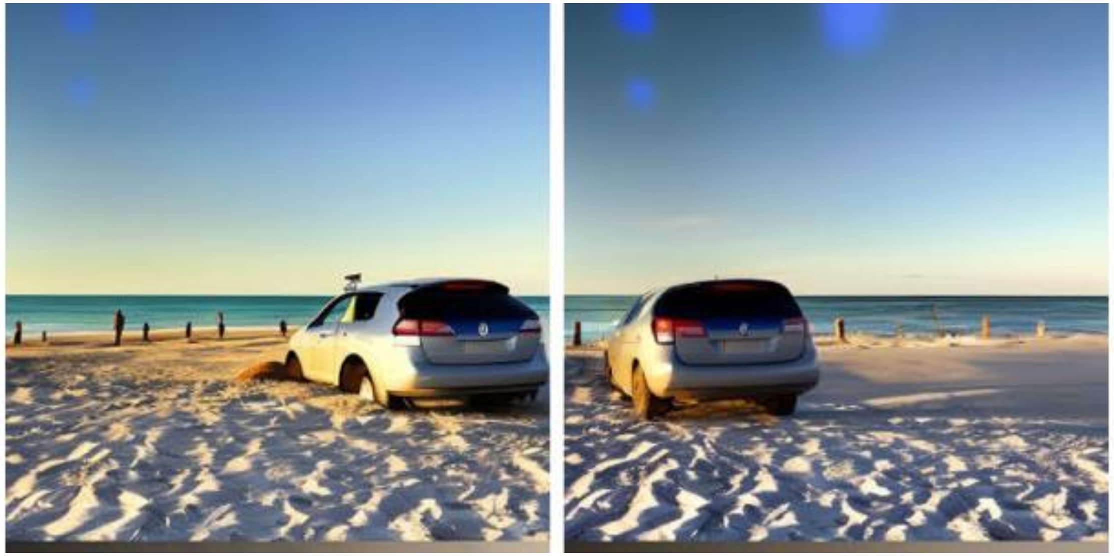
Intermediate features can be 3D-transformed using monocular depth estimates.



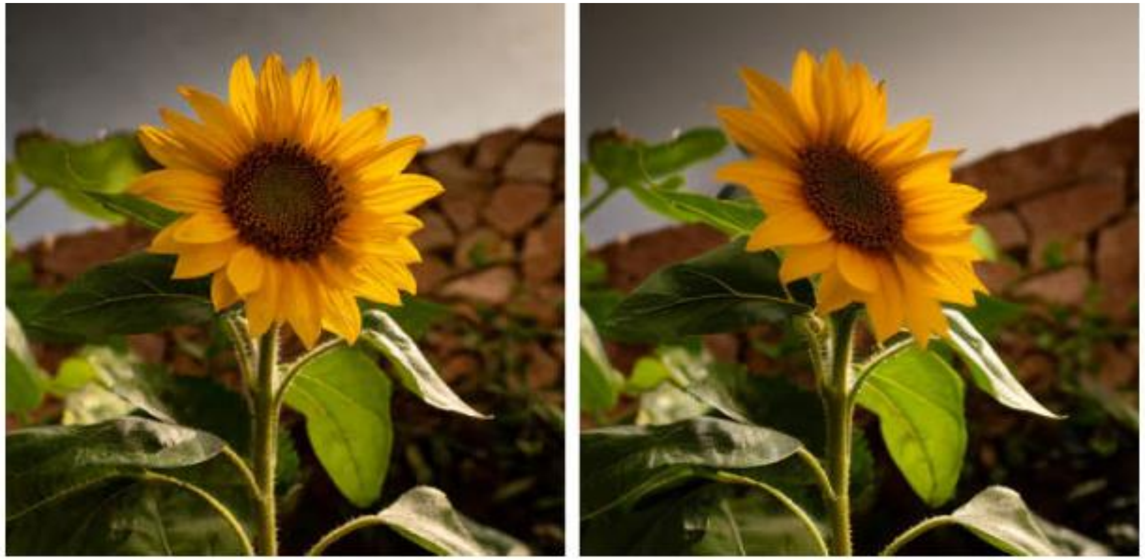
Diffusion Handles Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D, CVPR 2024

Edit Control by Moving Intermediate Features

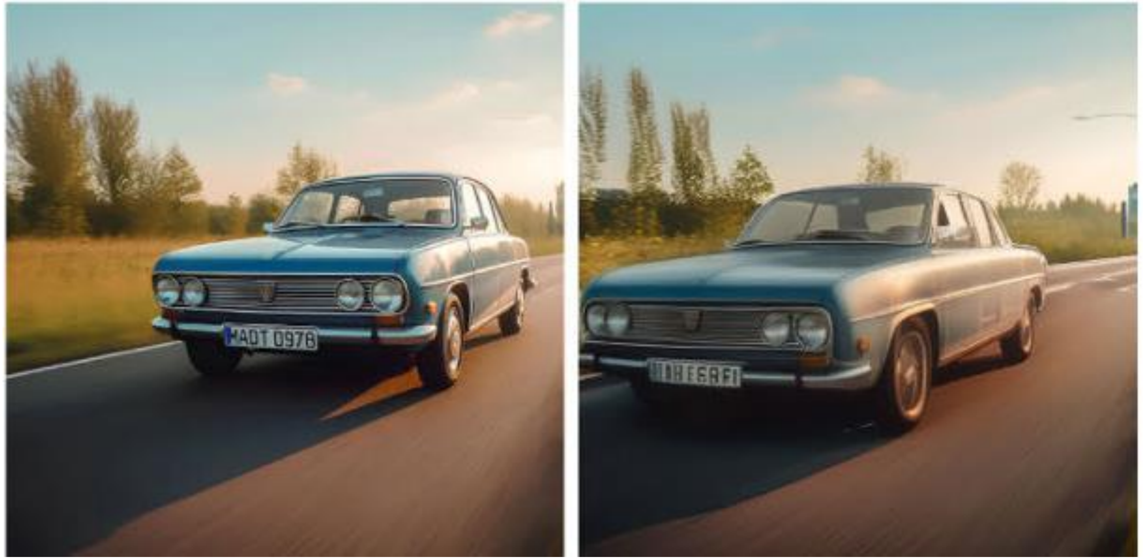
Attention maps / intermediate features can be 3D-transformed using monocular depth estimates.



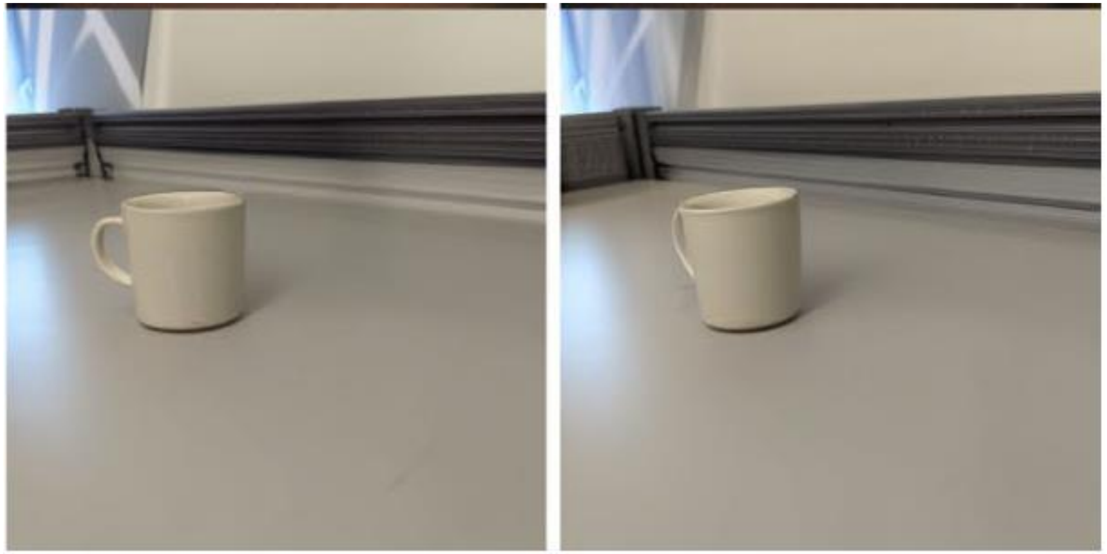
Diffusion Handles Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D, CVPR 2024



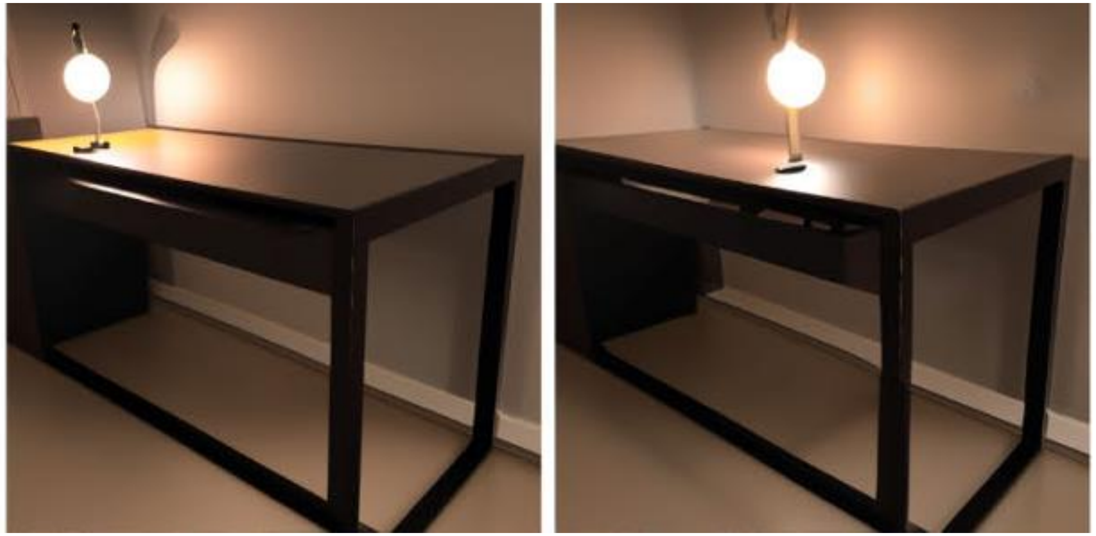
Edit Control by Moving Intermediate Features



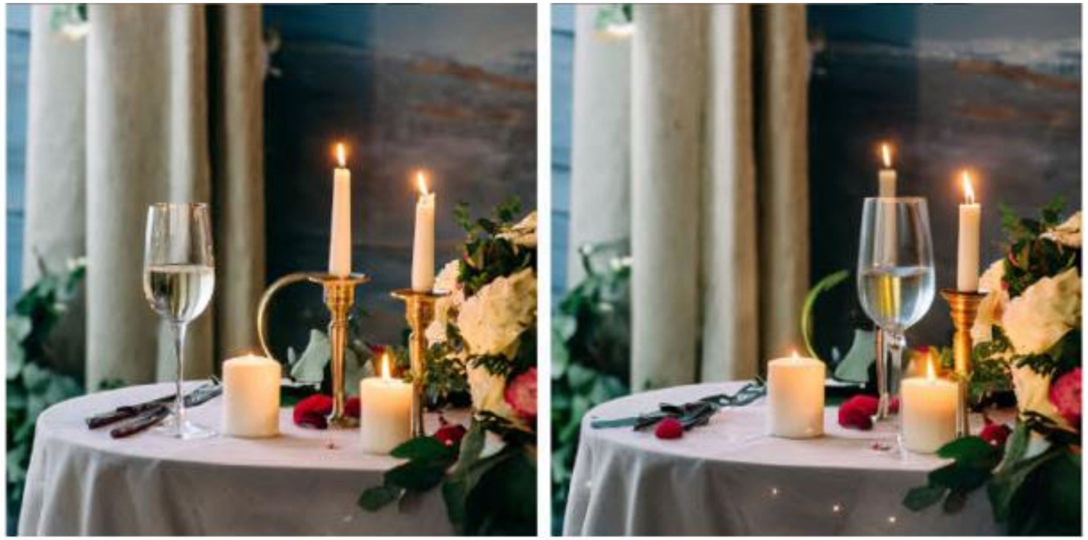
Diffusion Handles Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D, CVPR 2024



Edit Control by Moving Intermediate Features



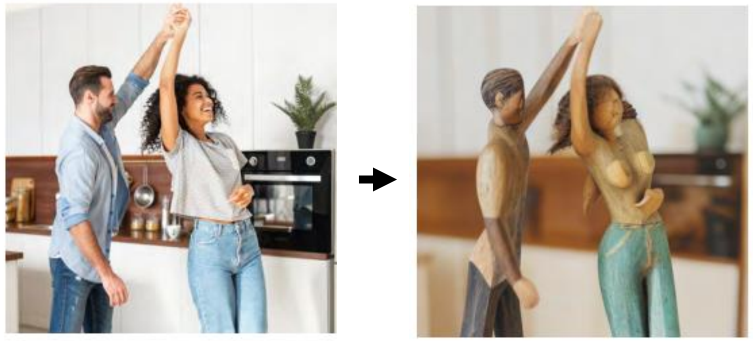
Diffusion Handles Enabling 3D Edits for Diffusion Models by Lifting Activations to 3D, CVPR 2024



Edit Control - Summary

Text

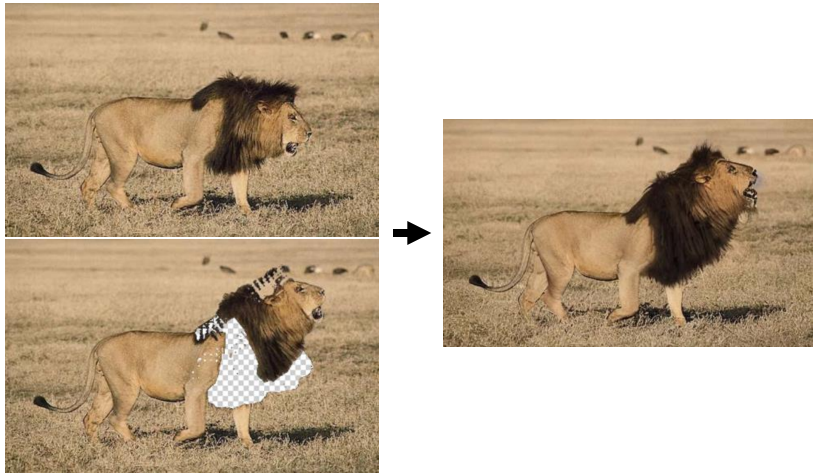
- Most widely-used form of control.
- Very general in what it can control.
- Only coarse control.
  (No detailed control over locations/layouts/amounts/degrees.
)



“A wooden sculpture of a couple dancing”

Noisy Input

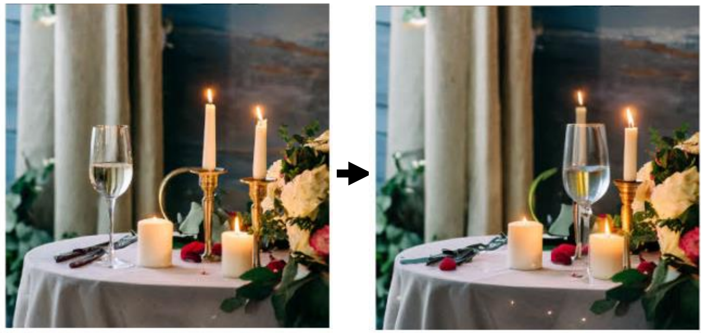
- More detailed control.
- Some strategy required
  to create coarse input.
- Typically requires training/fine-tuning.



Coarse Input

Moving Intermediate Features

- More detailed control.
- Edits can only move objects.
- Can be training-free.



Presentation Schedule

Introduction to Diffusion Models

Guidance and Conditioning Sampling

Attention

Break

Personalization and Editing

Beyond Single Images

Diffusion Models for 3D Generation

End of Part 4 – Personalization & Editing