Towards more Controllable Text-to-Image Generation

Wenhu Chen
Assistant Professor at University of Waterloo
Faculty at CIFAR AI Chair
Researcher at Google Deepmind

VECTOR INSTITUTE

Jul/23/2023

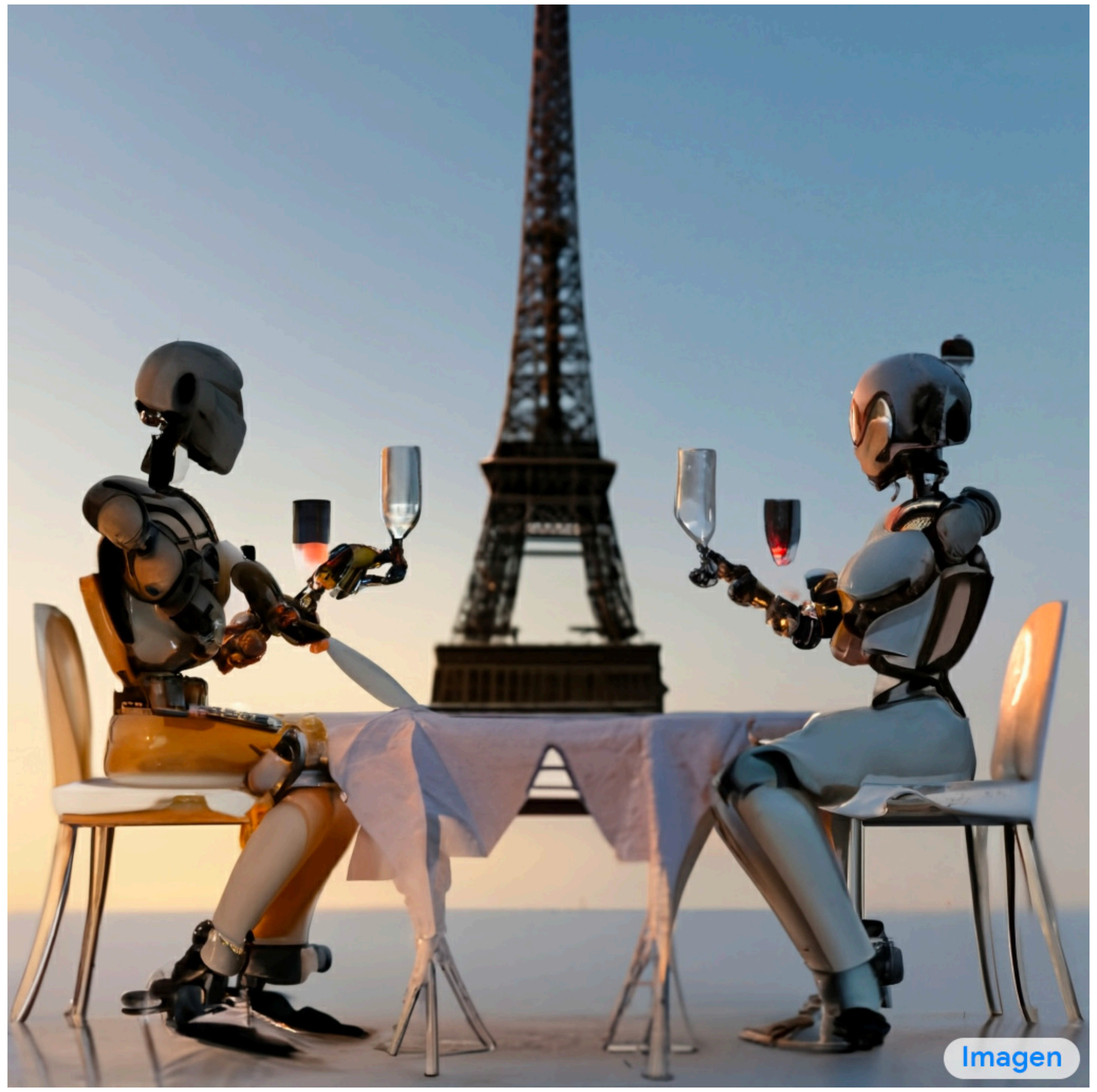
Outline

- Background and Motivation
  - Building text-to-image model with more controllability
- Subject-level Control for Text-to-Image Generation
  - Subject-driven Text-to-Image Generation via Apprenticeship Learning
  - With Hexiang Hu, William Cohen, etc at Google DeepMind
- Subject and Background-level Control for Text-to-Image Generation
  - DreamEdit: Subject-driven Image Editing
  - With Tianle Li, Max ku, Cong Wei at University of Waterloo
- Conclusion and Future Work

Background and Motivation

Text-to-Image Generation

- Text-to-Image Generation has achieved great success
  - Text-image alignment is high
  - Images are creative
  - Resolution is also high
- However, it's only controlled by text
  - Text is known to be ambiguous
  - Subject, Pose, Background, View, etc



A Robo couple fine dining with Eiffel tower in the background.

Background and Motivation

Controllability in Text-to-Image Generation

- How can we control the model to generate a specific subject
  - Subject-Level Control
    - A specific dog or a specific person in different scenarios.

- How can we control the model to generate a specific subject in a specfic scene
  - Background Control
    - A specific scene like a garden, a yard, etc.

Subject-Level Control

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

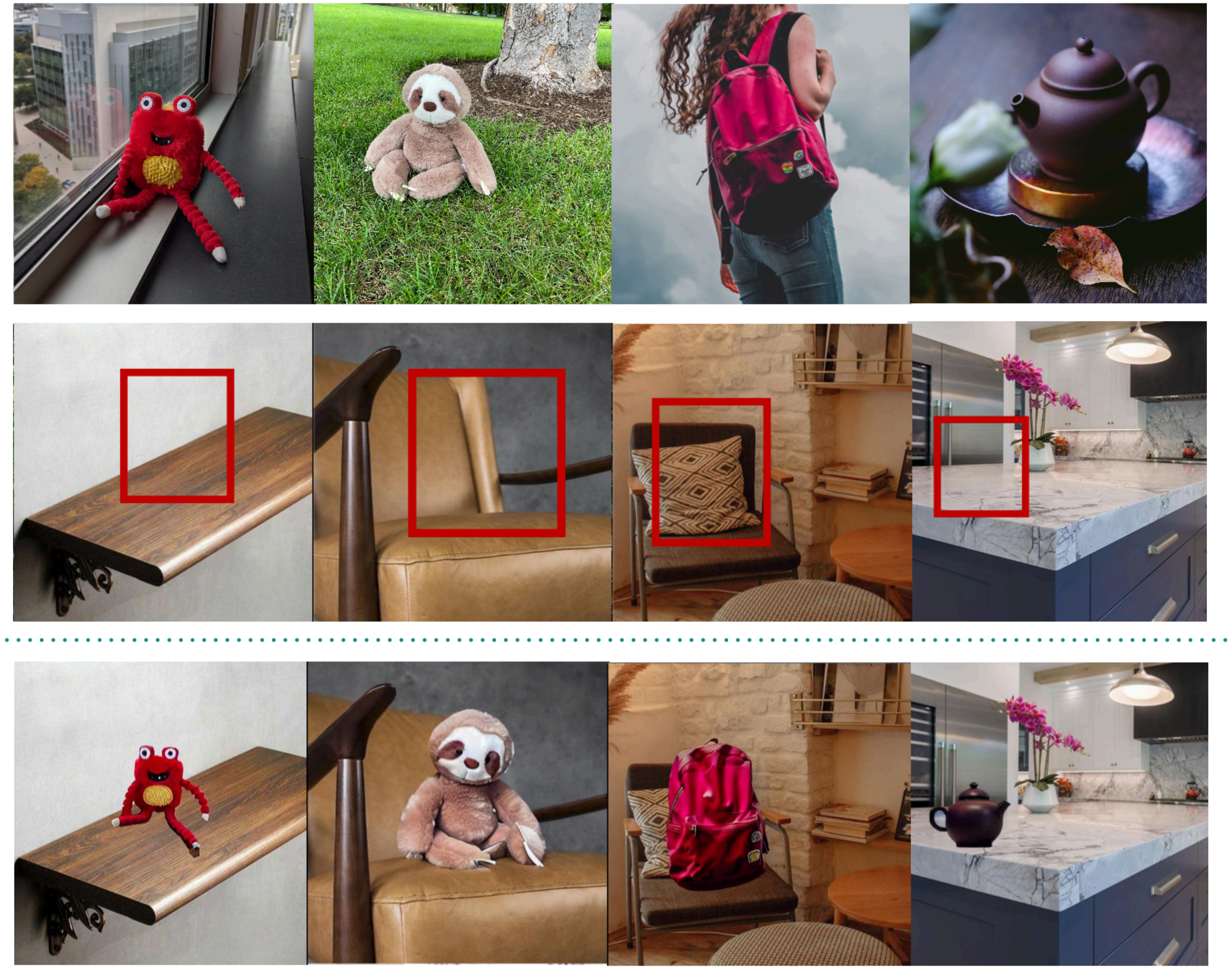
A [V] vase in the ocean

Subject and Background-Level Control

Subject

Background

Output

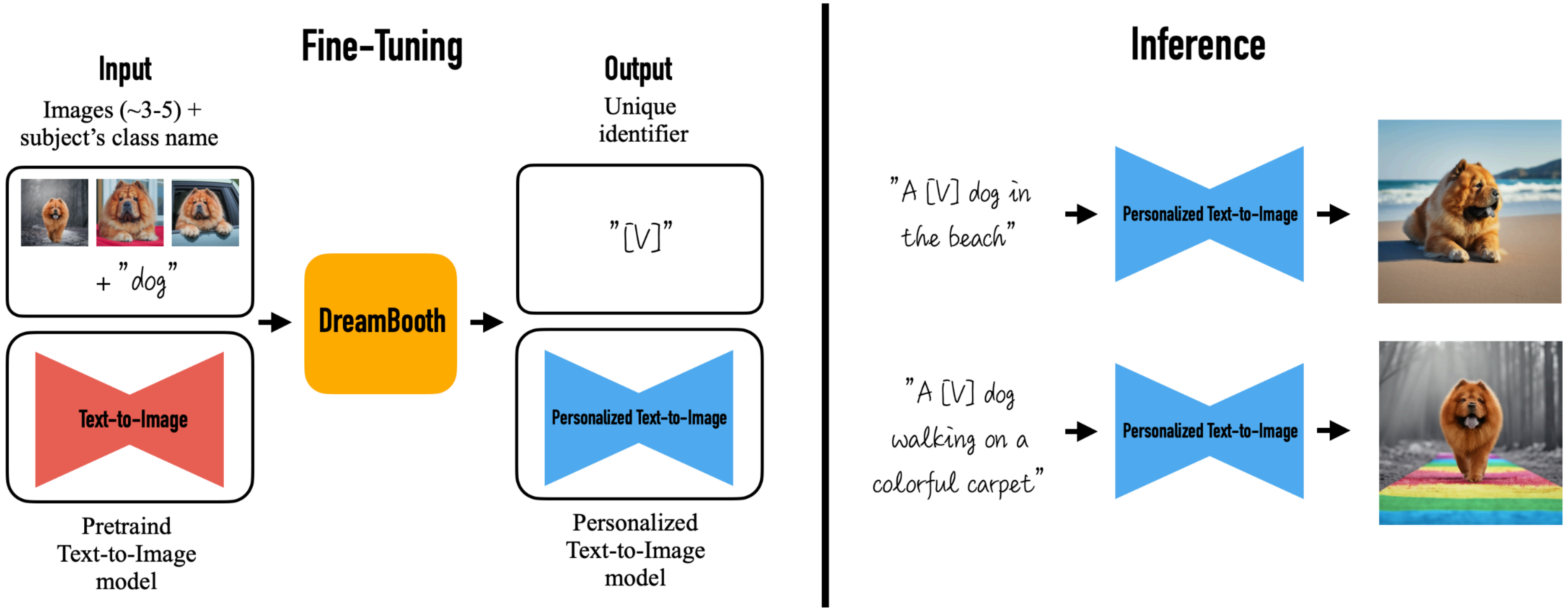


Outline

- Background and Motivation
  - Building text-to-image model with more controllability
- Subject-level Control for Text-to-Image Generation
  - Subject-driven Text-to-Image Generation via Apprenticeship Learning
  - With Hexiang Hu, William Cohen, etc at Google DeepMind
- Subject and Background-level Control for Text-to-Image Generation
  - DreamEdit: Subject-driven Image Editing
  - With Tianle Li, Max ku, Cong Wei at University of Waterloo
- Conclusion and Future Work

DreamBooth: Fine Tuning Text-to-Image Diffusion Models

- Finetune on 3-5 images regarding the subjects for 1000 steps.
- Maximize the diffusion model's likelihood p( | [V] dog).
- Save the checkpoint, then use the checkpoint to generate images with [V].

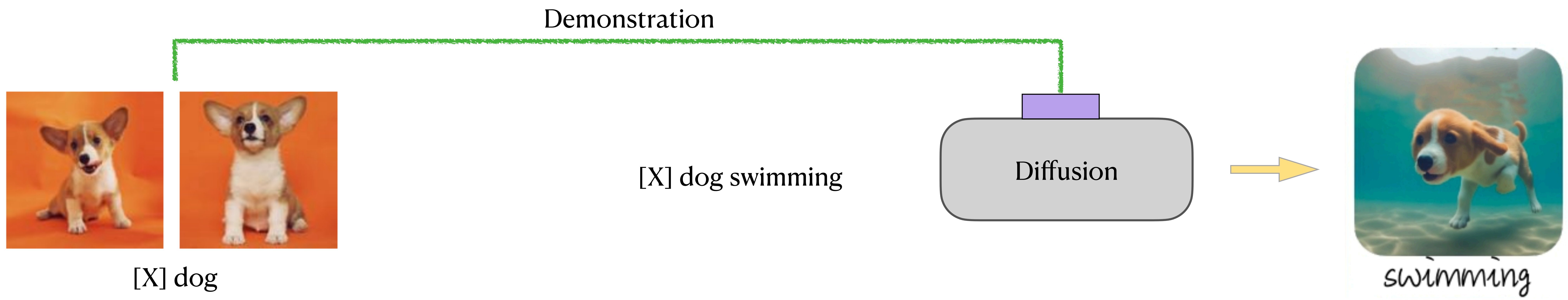


DreamBooth: Fine Tuning Text-to-Image Diffusion Models

- It requires fine-tuning the model
  - It consumes a lot of time. Normally 5-10 minutes to generate 1 image, which is 50x slower than normal text-to-image generation.
  - Saving one checkpoint per subject requires lots of disk space.
  - Therefore, this approach cannot scale up

In-Context Learning for Subject-Driven image generation

- Can we avoid fine-tuning?
- A single model to ace it all:
  - In-context demonstration without gradient descent.
  - Adapt to any subject quickly within 30 seconds.

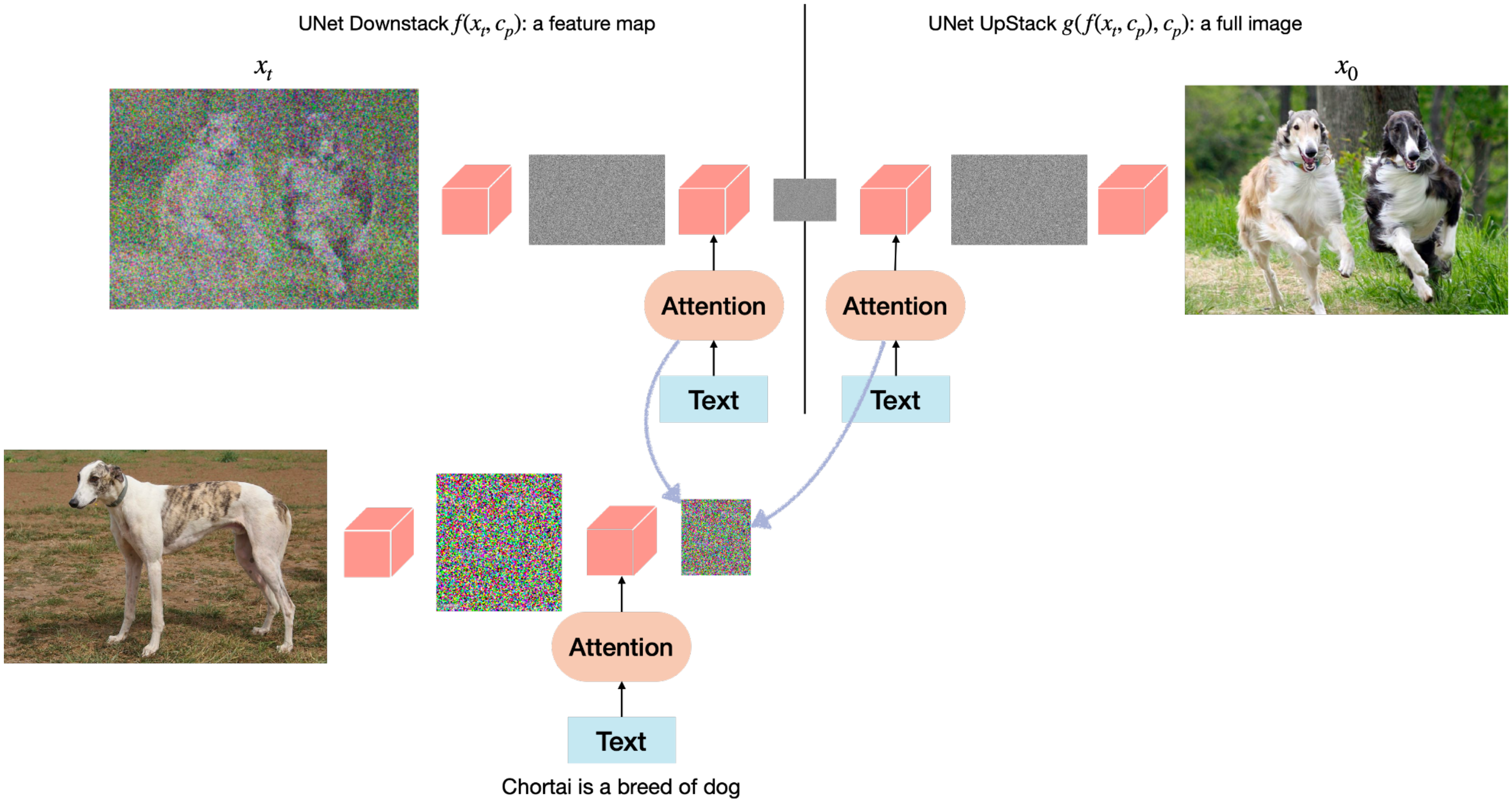


What do we need to achieve In-Context Learning?

- We need to change the diffusion model architecture
  - The current architecture only supports image input
  - The model needs to attend to demonstration of multiple (image, text) pairs

- We also need to construct new dataset to train the model
  - ((subject image1, subject image2, …, subject text) => (new text, new image))
  - The diffusion model attends to these subject and generalize it to new scenario

Architecture: Adding additional attention layer



Dataset: how can we obtain such data?

- Desired format
  - (text_1, Image_1), (text_2, image_2), … (text_t, image_t), where these group of image-text pairs share the same subject.

- Challenge
  - However, such data does not exist on the web!
  - The existing dataset consists of standalone (image, text) pairs.

Web Image-Text Data Clustering

- Clustering
  - We group (image, text) pairs based on their URLs
    - We assume (image, text) pairs mined from the same URL are more likely to contain the same subject, like Amazon shopping site, etc.
  - We filter the groups based on the inter-image similarity to remove the low-quality clusters containing highly different images.

- Re-Annotating Text Caption
  - The crawled alt text is noisy, we group these images to generate caption jointly

How is the clusted data quality?

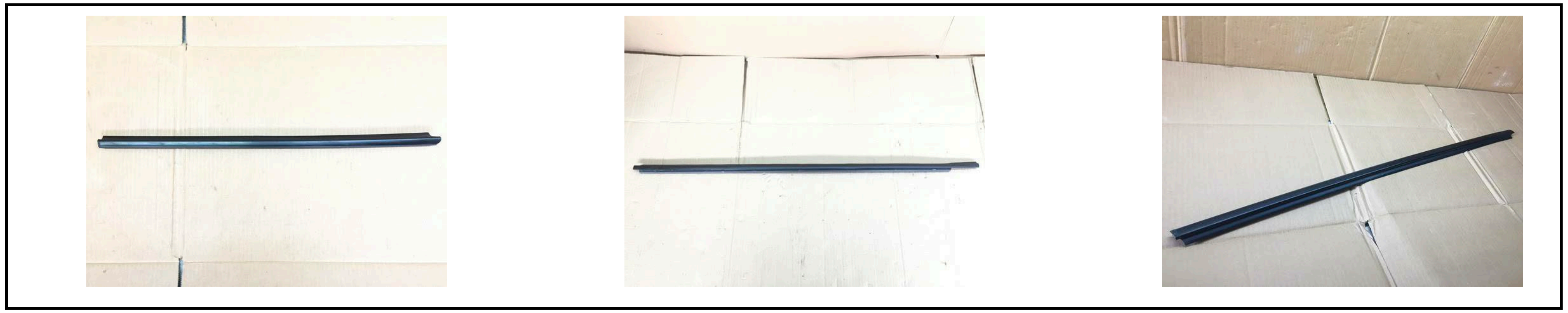
A limousine parked in a parking lot

A couple of birds standing in the water

A gold cross with diamonds

A pair of shorts

A pair of sneakers



A dirty picture of a window seal

How is the clustered data quality?

- The data quality is reasonably good
  - The grouped images are mostly about a single subject
  - If not, it's mostly about the same type of subject.

- Can we use the clustered dataset to train the model?

A limousine parked in a parking lot

A limousine in a parking lot

How well does the trained model work?

- We train the first version to train our model
  - The model does not view the text prompt
  - Only copy-paste demonstration

- Reason:
  - The target and demonstrations images are too similar
  - The model falls into a copy-paste local optima

How can we make it better?

- Make the target (image, text) highly different from the demonstration!

A pair of shorts

A man wearing a pair of shorts

- How can we obtain such diverse target (image, text) pair?
    - Use LLM to imagine a new prompt
    - Then use DreamBooth to fine-tune on the demonstration and then generate.

Apprenticeship Learning

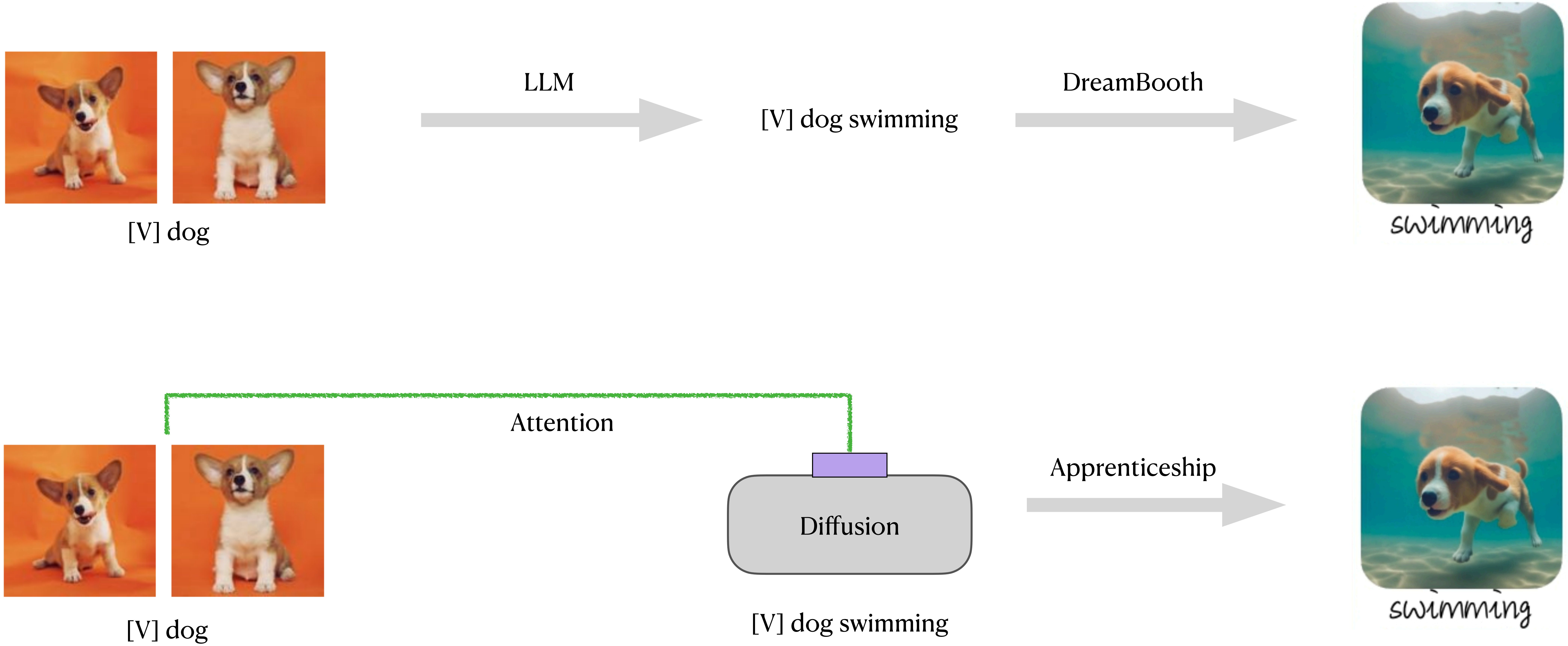
[V] dog

LLM

[V] dog swimming

DreamBooth

Apprenticeship Learning



Apprenticeship Learning

- DreamBooth as the experts to demonstrate the output
  - We have 2M subjects, i.e. 2M DreamBooth experts
  - Parallelized Training, each takes 5 minutes
  - We use 800 v4 TPUs and run for 1-2 week to store all the DreamBooth outputs
  - Once and for all

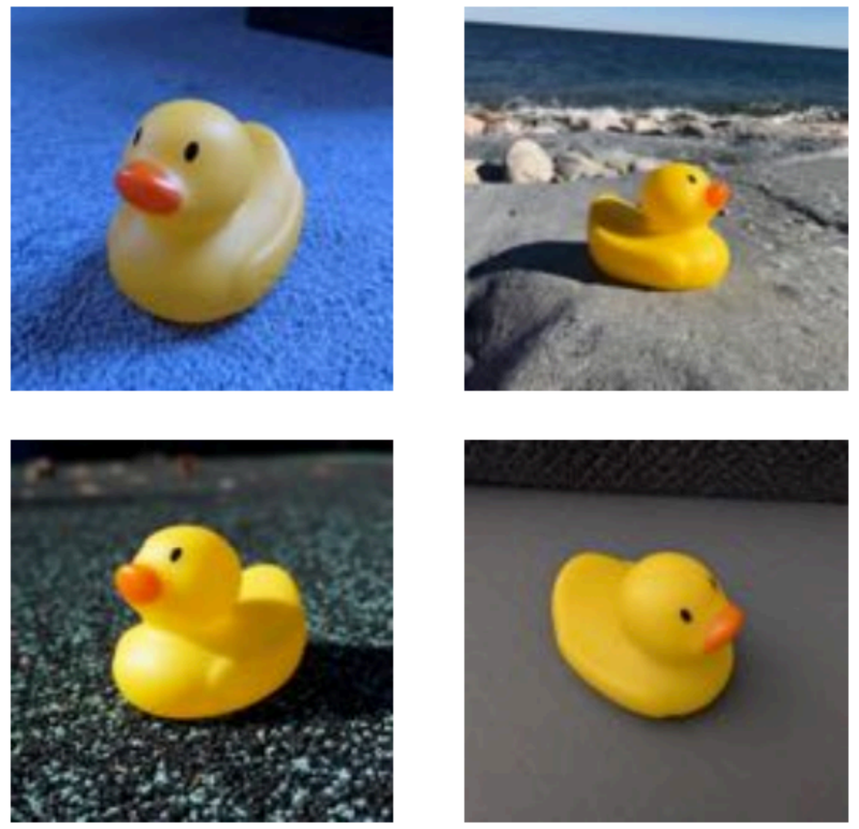
- The apprentice model (SuTI) follows the DreamBooth experts
  - Distill from millions of experts!

Training Details

- We use the synthesized data to train the apprentice model for 1 day
  - The apprentice model learns surprisingly fast
- Skillset of the apprentice model:
  - Stylization: changing the style of the subject
  - Recontextualization: changing the scene of the subject
  - Multi-View synthesis: changing the view perspective of the subject
  - Attribute Modification: changing the color,textual,emotion,etc of the subject
  - Compositional: Stylization + Recontextualization

Model Outputs

A duck toy



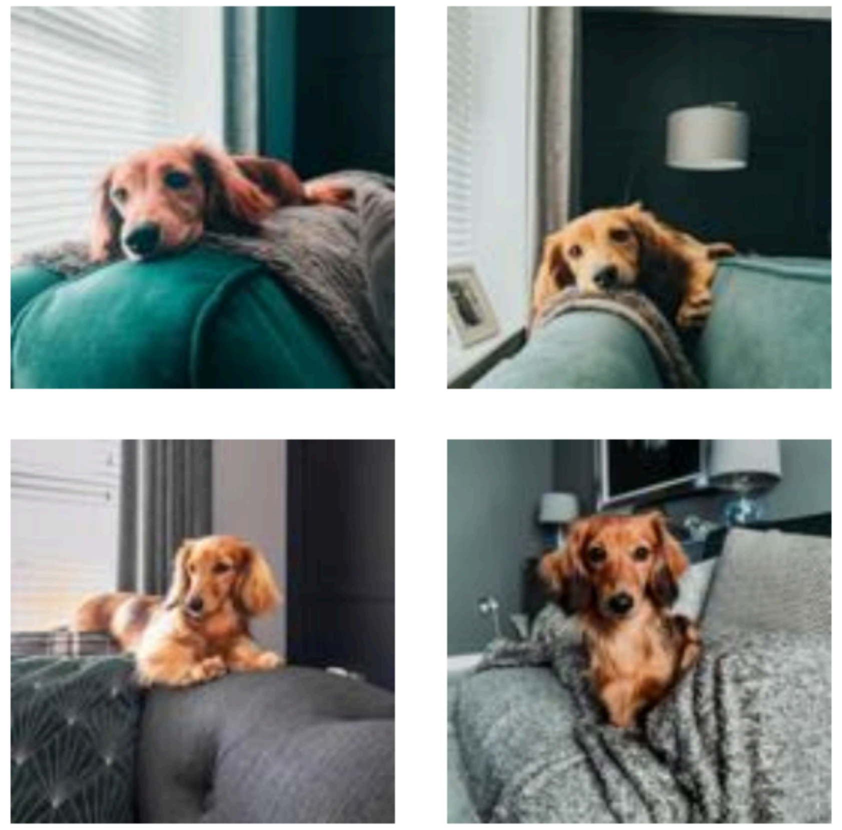
Pablo Picasso

Rembrandt

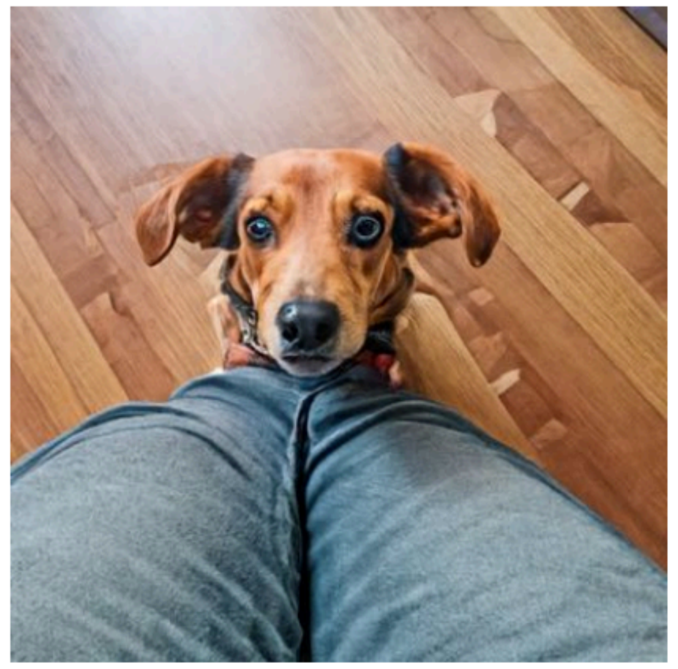
Rene Magritte

Vincent van Gogh

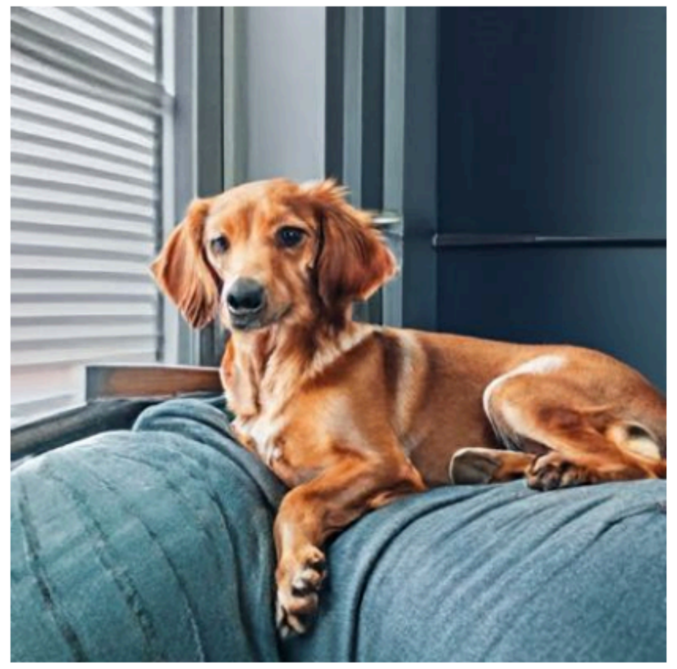
A dog



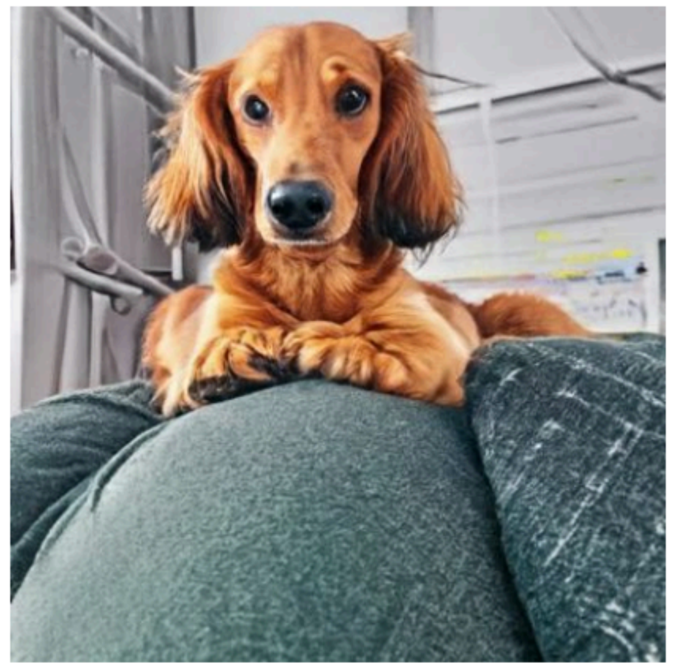
Top-down view



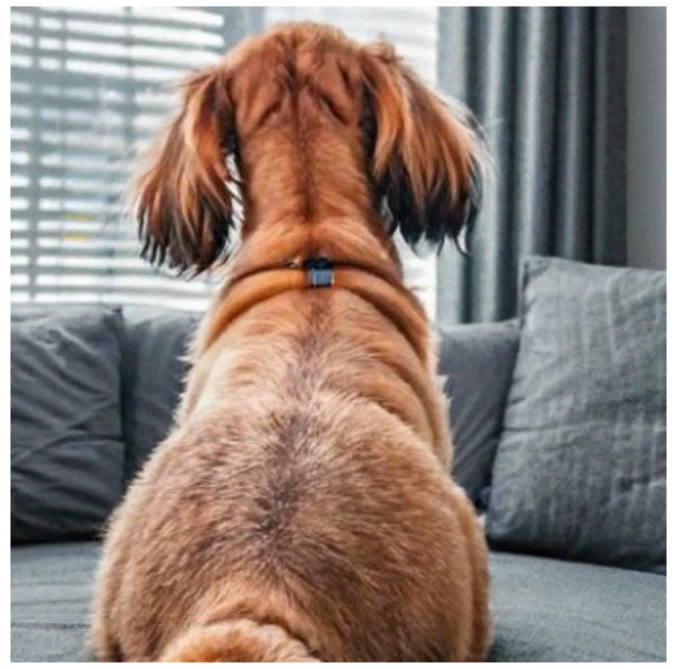
Side view



Bottom view



Back view



Model Outputs

A dog

Depressed

Joyous

Sleepy

Screaming

A monster toy

Blue

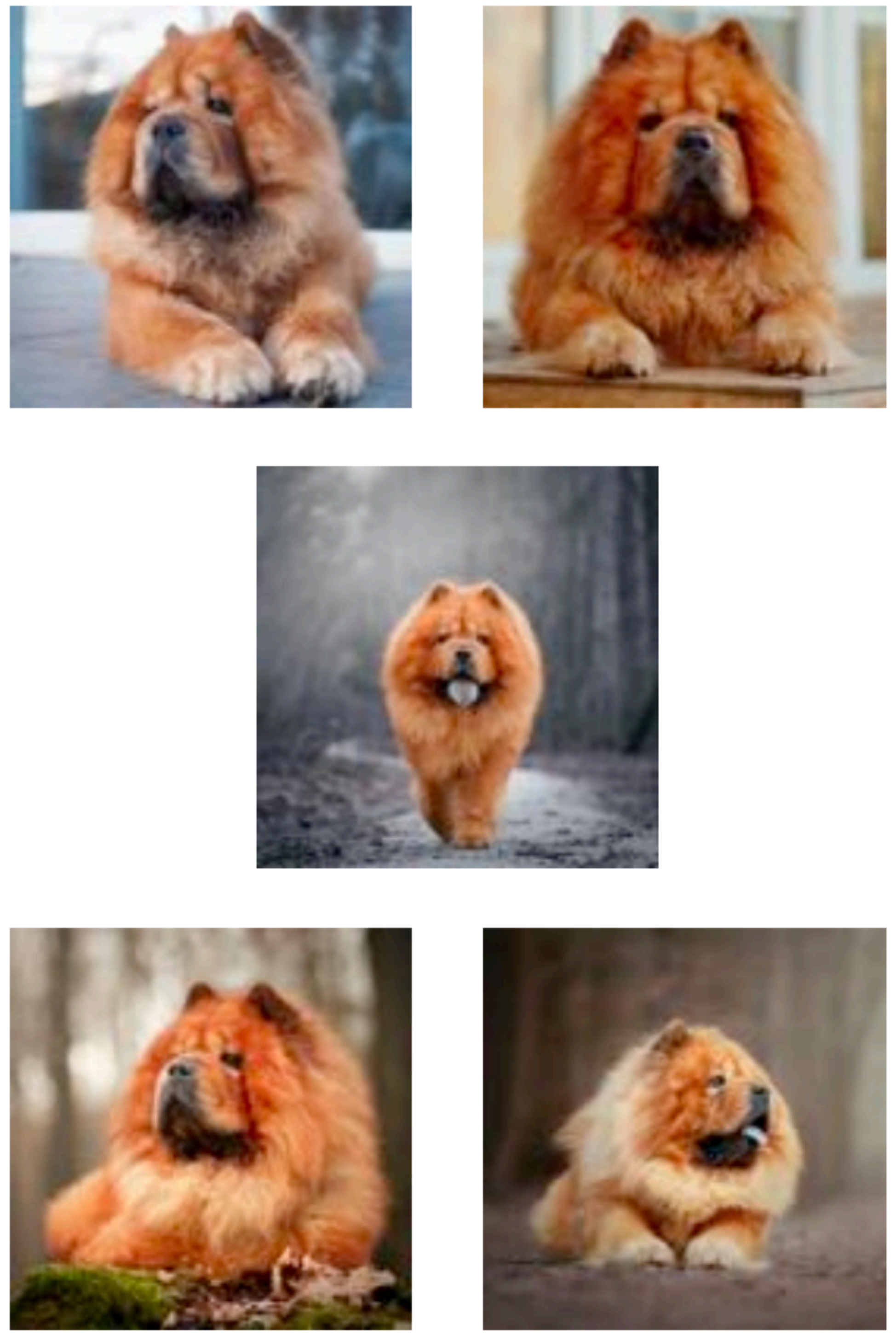
Green

Purple

Pink

Model Outputs

A dog



Compositional Model Outputs

wolf plushie

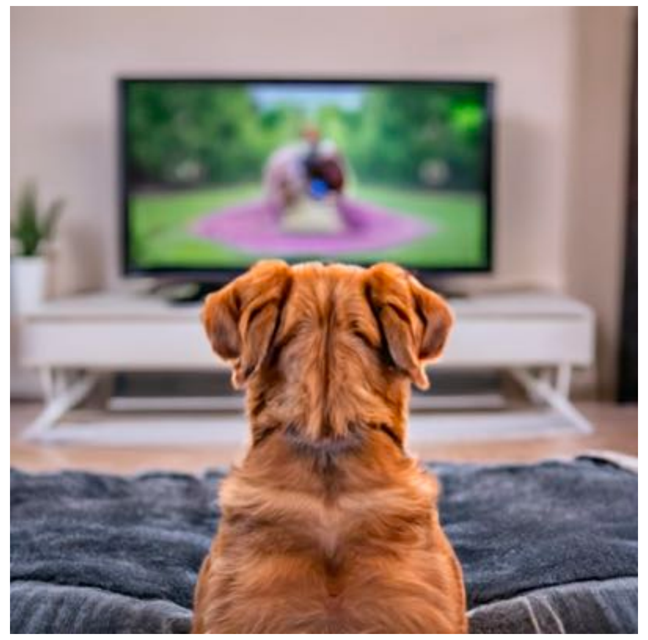
... playfully chasing a fox plushie.

... playfully chasing a fox plushie through a whimsical forest.

Re-Context → Re-Context + Re-Context

canine dog

a back view of ... watching a TV show.



a back view of ... watching a TV show about birds.

Re-Context → Re-Context + Editing

clay teapot

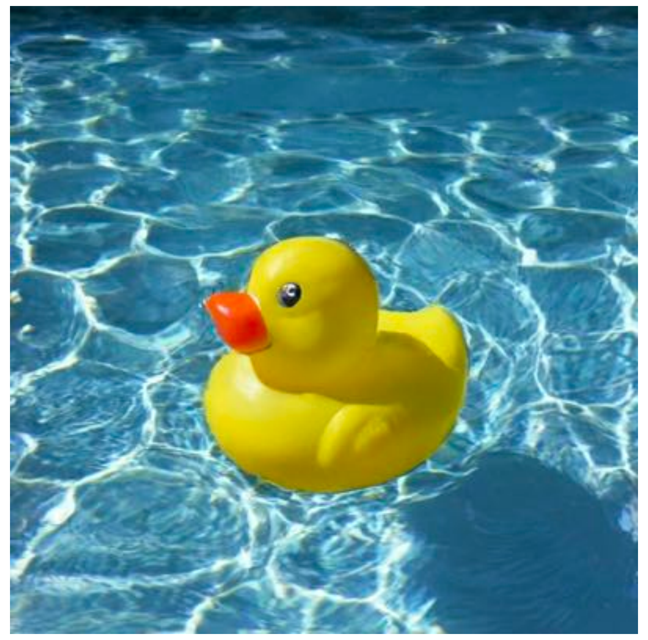
... sitting on a glass table.

... sitting on a glass table, surrounded by delicate porcelain teacups.

Re-Context → Re-Context + Accessorize

duck toy

... in the water.



a Claude Monet styled painting of ... in the water.

Re-Context → Re-Context + Style Transfer

Human Evaluation

- We collect 220 prompts regarding 30 different subjects.
- We feed the (subject image, text) -> (prompt, ?) to different models for genertation

Methods	Backbone	Space	Time	Subject ↑	Text ↑	Photorealism ↑	Overall ↑
Models requiring test-time tuning							
Textual Inversion [10]	SD [25]	$	30 mins	0.22	0.64	0.90	0.14
Null-Text Inversion [19]	Imagen [28]	$$	5 mins	0.20	0.46	0.70	0.10
Imagic [15]	Imagen [28]	$$$$	70 mins	0.78	0.34	0.68	0.28
DreamBooth [27]	SD [25]	$$$	6 mins	0.74	0.53	0.85	0.47
DreamBooth [27]	Imagen [28]	$$$	10 mins	0.88	0.82	**0.98**	0.77
InstructPix2Pix [4]	SD [25]	-	10 secs	0.14	0.46	0.42	0.10
Re-Imagen [6]	Imagen [28]	-	20 secs	0.70	0.65	0.64	0.42
Ours: `SuTI`	Imagen [28]	-	30 secs	**0.90**	**0.90**	0.92	**0.82**

Outline

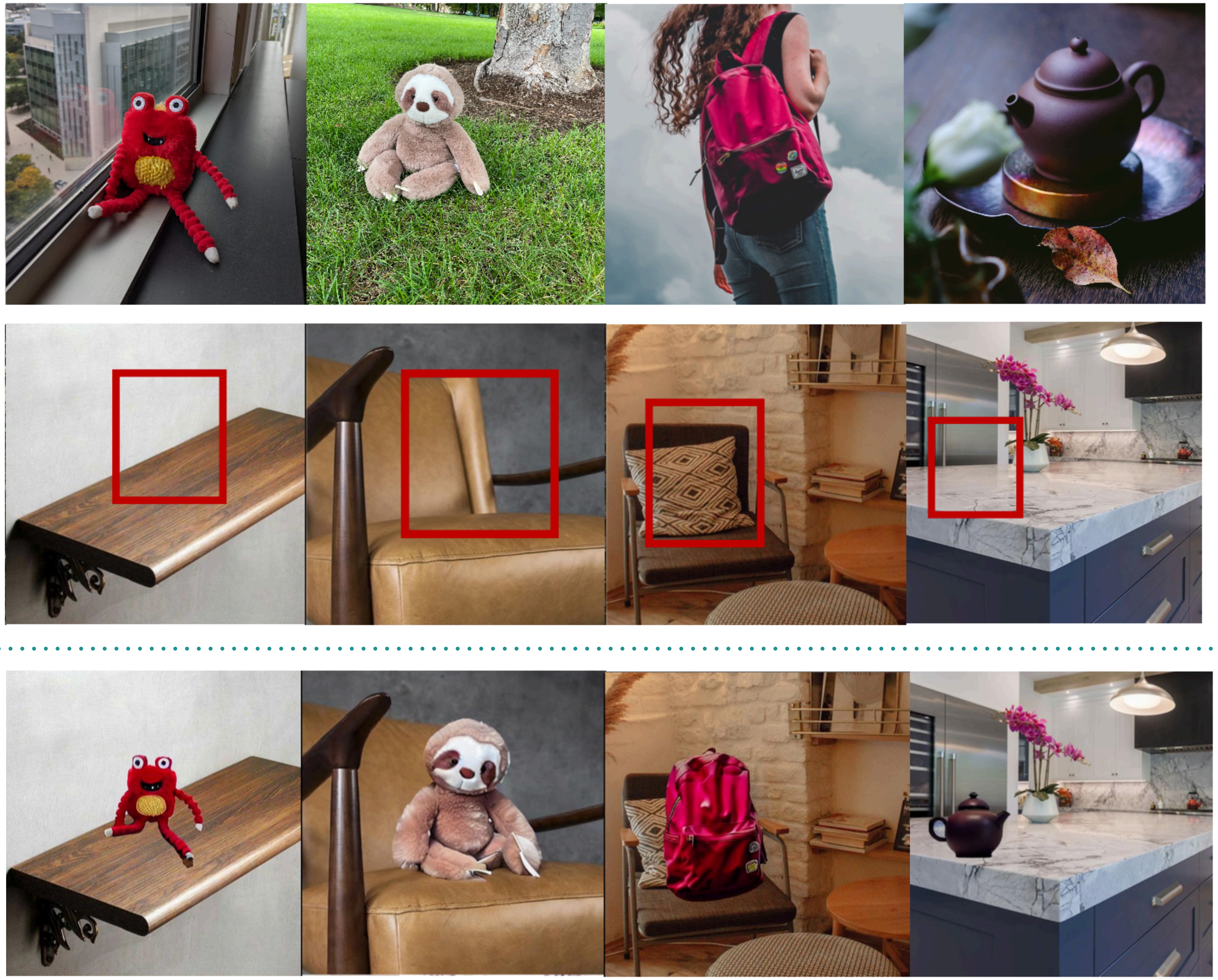
- Background and Motivation
    - Building text-to-image model with more controllability
- Subject-level Control for Text-to-Image Generation
    - Subject-driven Text-to-Image Generation via Apprenticeship Learning
    - With Hexiang Hu, William Cohen, etc at Google DeepMind
- Subject and Background-level Control for Text-to-Image Generation
    - DreamEdit: Subject-driven Image Editing
    - With Tianle, Max ku, Cong Wei at University of Waterloo
- Conclusion and Future Work

Background Control

Subject

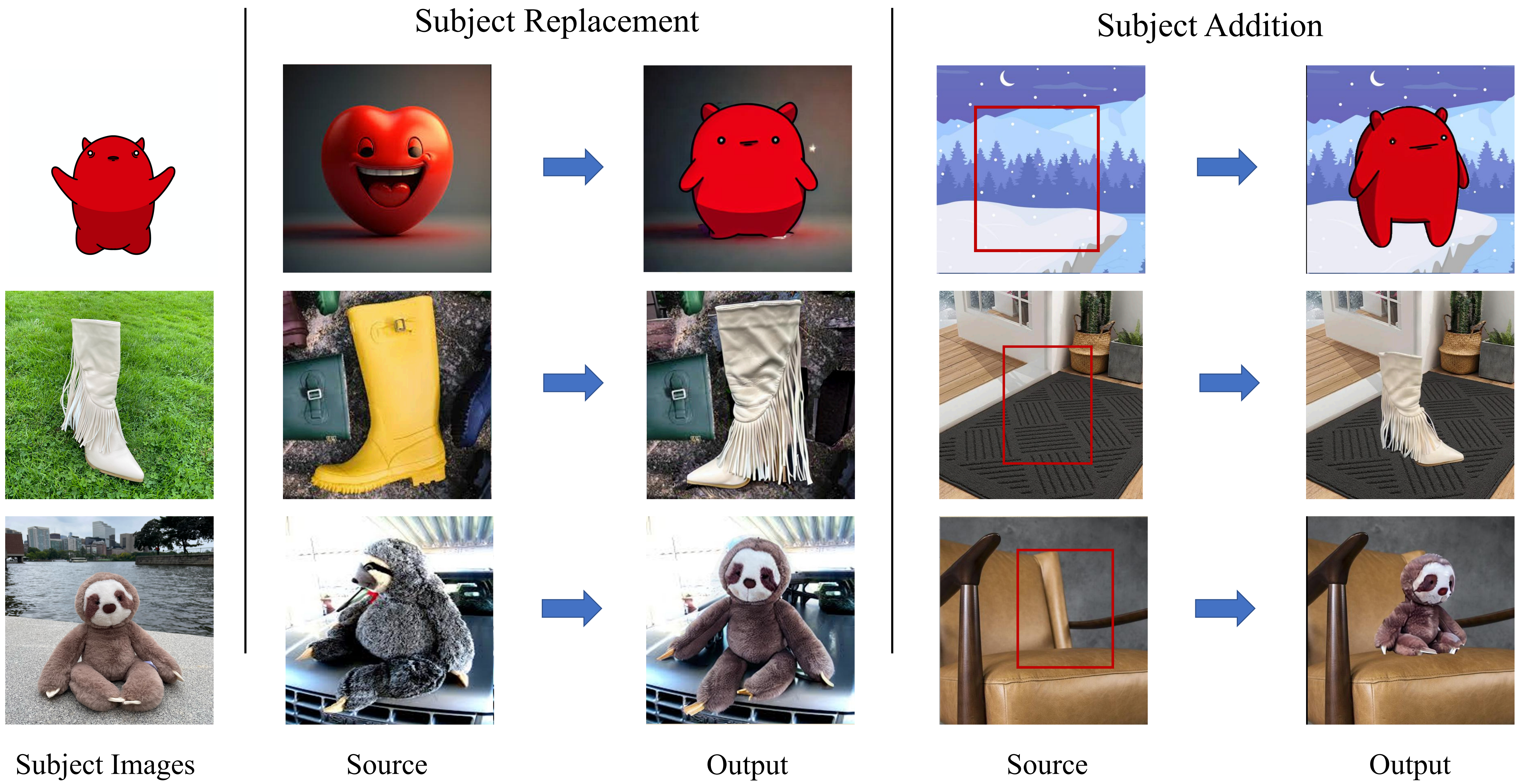
Background

Output



Task Definition

- Subject Replacement
  - Replace the subject in the source image with the customised subject
- Subject Addition
  - Add the customised subject to the designated position in a given background



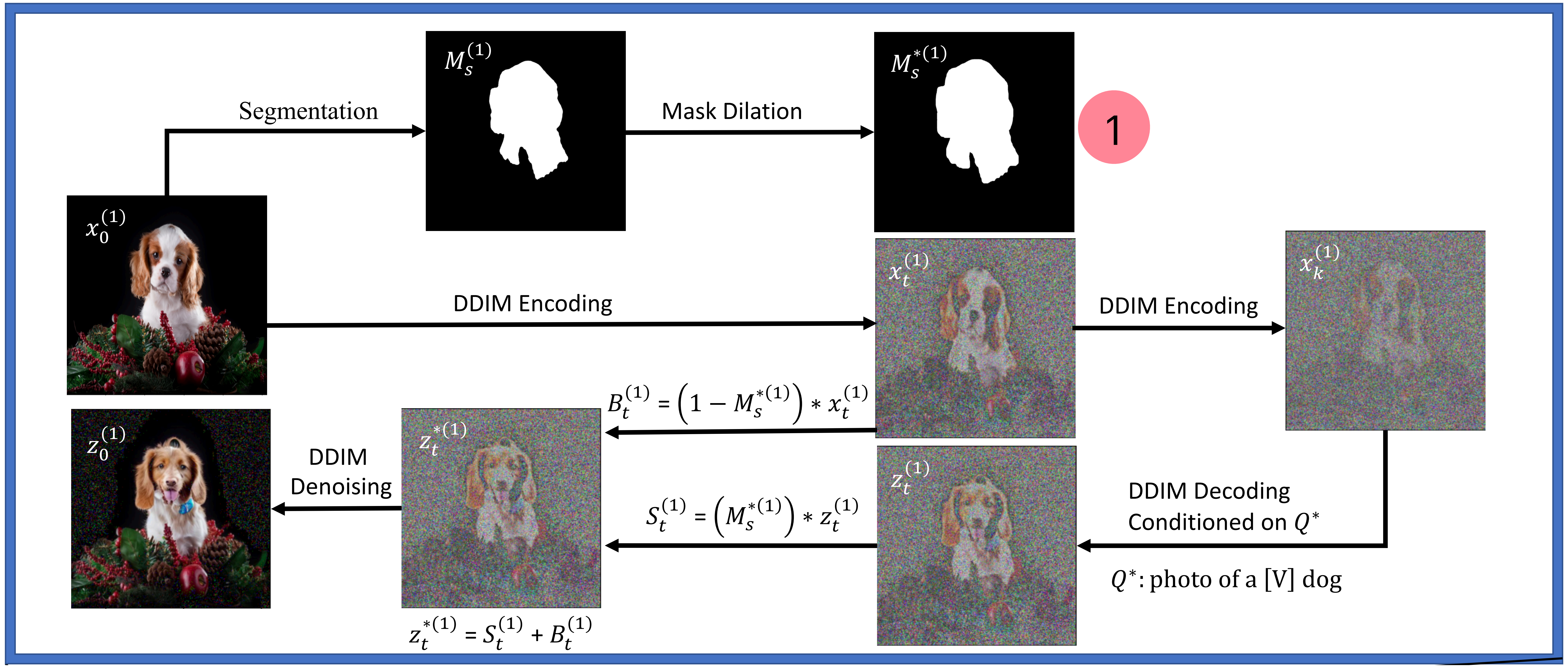
Iteartive Mask-based In-painting

- Challenges
  - How to replace the subject differs dramatically from the target subject?
  - How to blend the added subject naturally in the designated environment?
- Solution:
  - Iterative generation: Gradual adaptation to the customized subject

Iteration 1 Iteration 2 Iteration N

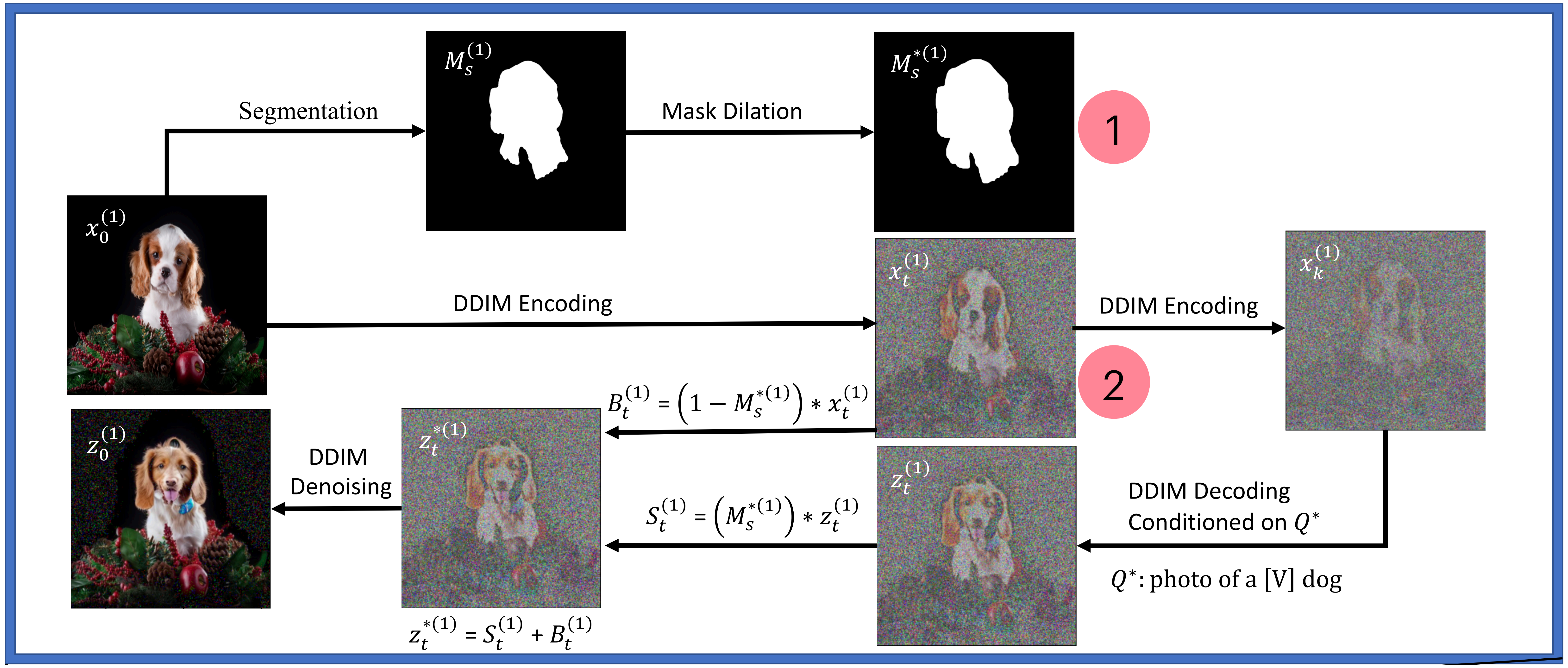
Mask-based Inpainting

- Customized In-painting
  - Fine-tuning with model with [V] token
  - Subject segmentation mask dilation
  - In-painting guided by dilated mask and special token [V]



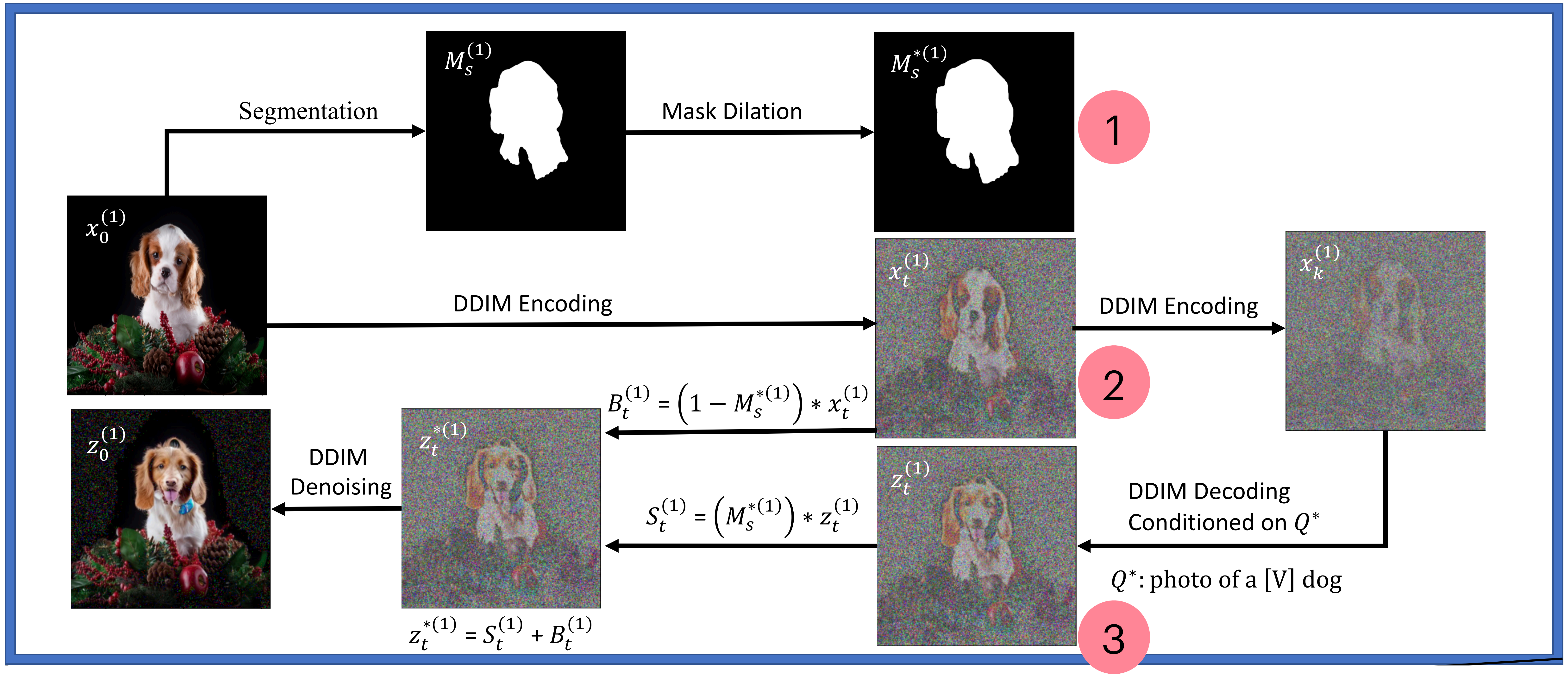
Mask-based Inpainting

- Customized In-painting
  - Fine-tuning with model with [V] token
  - Subject segmentation mask dilation
  - In-painting guided by dilated mask and special token [V]



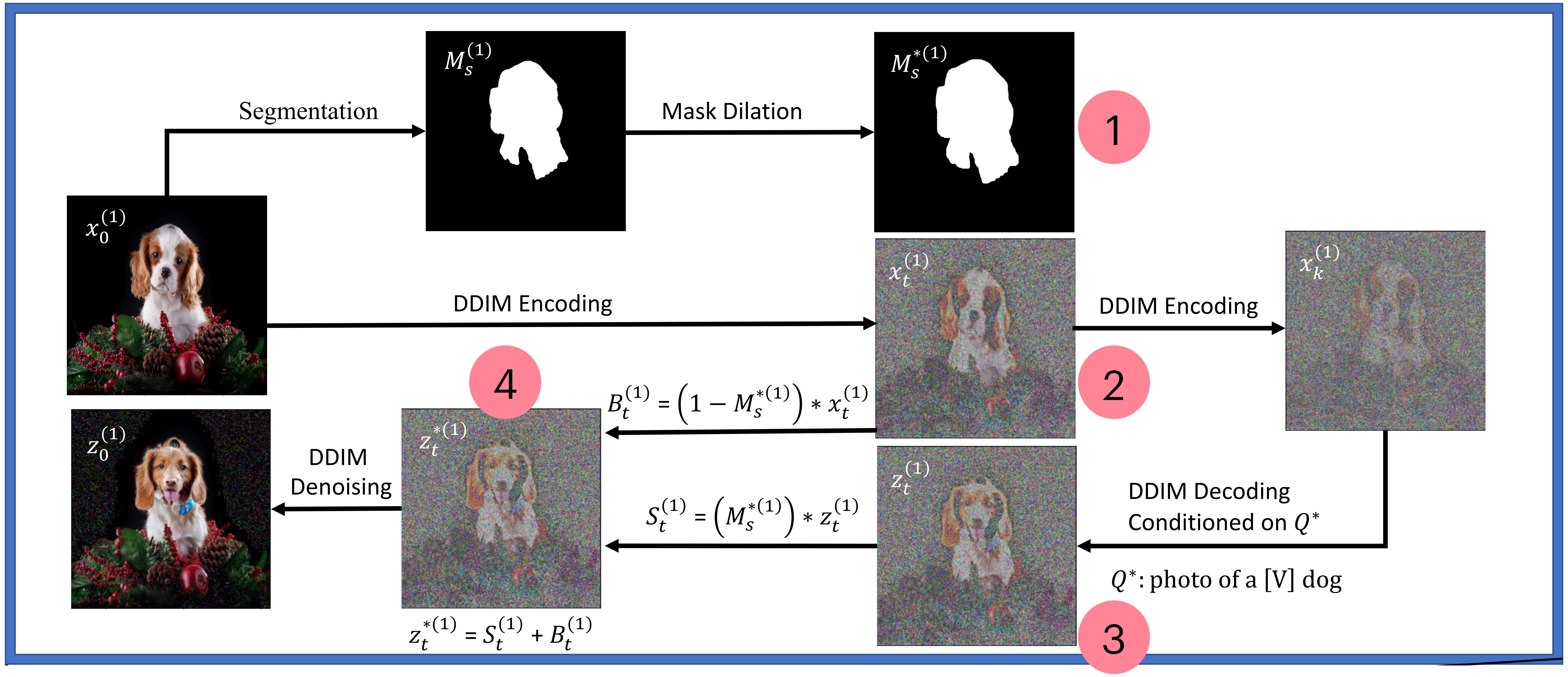
Mask-based Inpainting

- Customized In-painting
  - Fine-tuning with model with [V] token
  - Subject segmentation mask dilation
  - In-painting guided by dilated mask and special token [V]



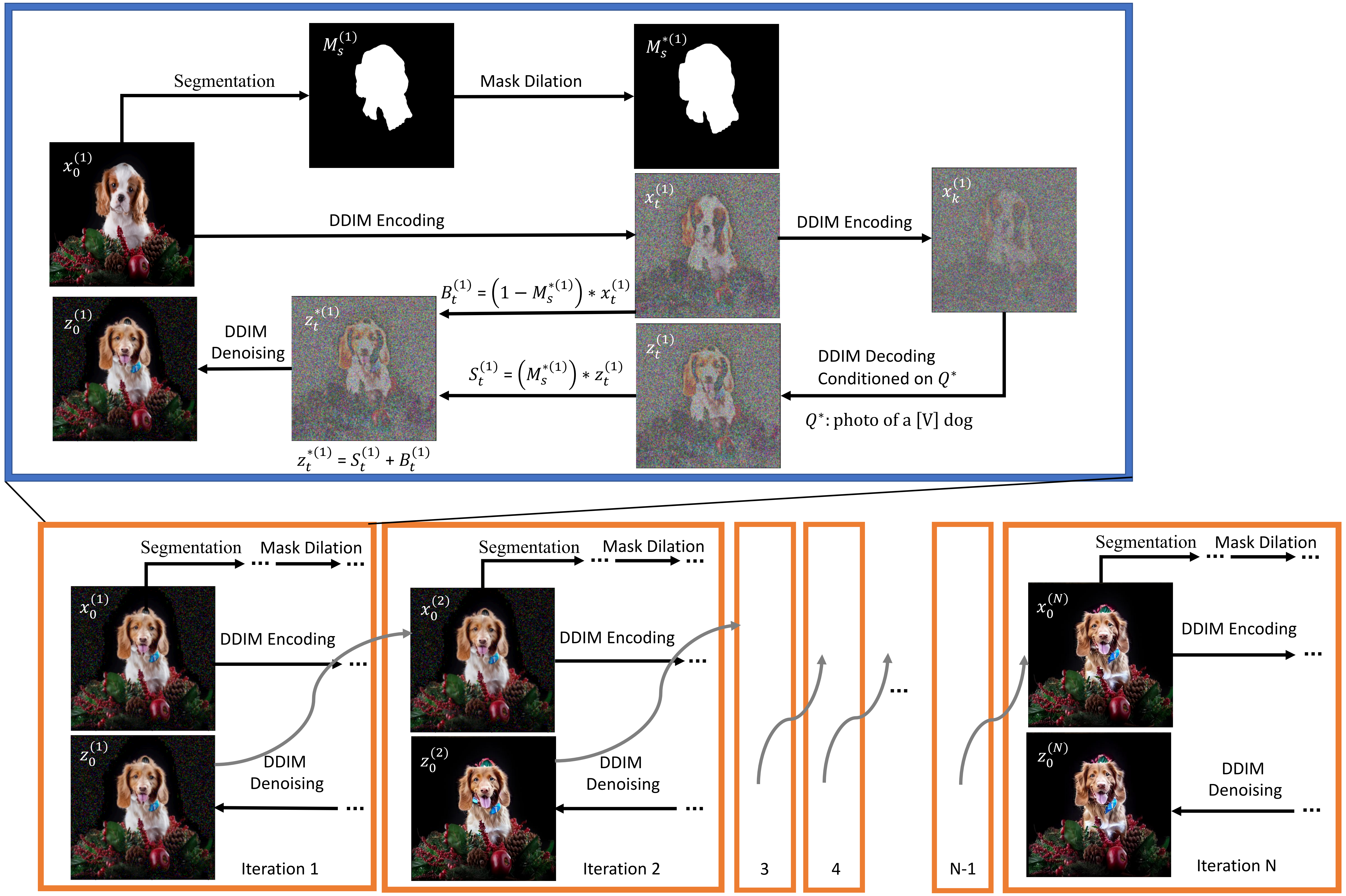
Mask-based Inpainting

- Customized In-painting
  - Fine-tuning with model with [V] token
  - Subject segmentation mask dilation
  - In-painting guided by dilated mask and special token [V]



Iterative Mask-based Inpainting

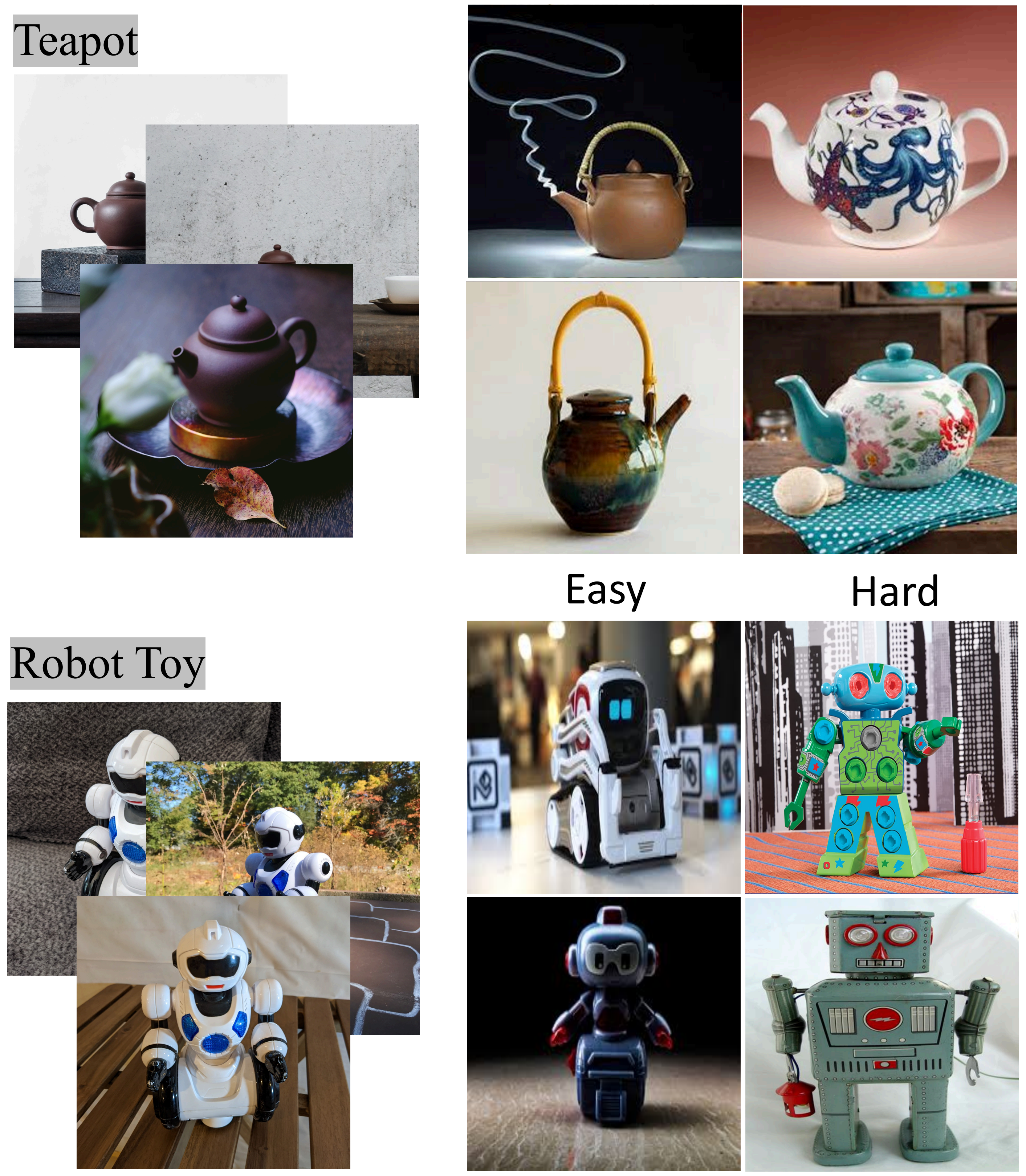
- Iterative Generation
  - The output of the current iteration is fed to the next iteration as the input
  - Easy examples: one iteration is enough
  - Hard examples: longer iterations



Dataset Curation

- DreamEditBench:
  - Manually collect 220 images of 22 subjects for each task
  - Easy and hard division based on difference

Subject Replacement



Subject Addition

Experimental Results

- Human evaluation result on curated dataset

Method	Initialization	Subject↑	Background↑	Realistic↑	Overall↑
		Subject Replacement			
DreamBooth	-	0.543	0.0	**0.707**	0.072
Customized-DiffEdit	-	0.21	**0.828**	0.668	0.488
CopyPaste	COPY	**1.00**	0.148	0.123	0.263
`DreamEditor` (1)	COPY	0.778	0.407	0.52	0.548
`DreamEditor` (5)	COPY	0.817	0.505	0.54	0.606
`DreamEditor` (1)	-	0.532	0.760	0.557	0.608
`DreamEditor` (5)	-	0.630	0.800	0.582	**0.664**
		Subject Addition			
DreamBooth	-	0.477	0.0	**0.635**	0.067
Customized-DiffEdit	GLIGEN	0.288	0.302	0.252	0.280
CopyPaste	COPY	**0.983**	**1.0**	0.033	0.319
`DreamEditor` (1)	COPY	0.635	0.978	0.265	0.548
`DreamEditor` (5)	COPY	0.633	0.973	0.393	0.623
`DreamEditor` (1)	GLIGEN	0.287	0.99	0.427	0.495
`DreamEditor` (5)	GLIGEN	0.478	0.972	0.528	**0.626**

Outline

- Background and Motivation
  - Building text-to-image model with more controllability
- Subject-level Control for Text-to-Image Generation
  - Subject-driven Text-to-Image Generation via Apprenticeship Learning
  - With Hexiang Hu, William Cohen, etc at Google DeepMind
- Subject and Background-level Control for Text-to-Image Generation
  - DreamEdit: Subject-driven Image Editing
  - With Tianle Li, Max ku, Cong Wei at University of Waterloo
- Conclusion and Future Work

Diverse Image Editing Tasks

- Subject-driven Image Generation [DreamBooth, SuTI]
  - Given reference image of a subject -> target image containing the subject
- Text-guided Image Editing [Imagic, Prompt2Prompt, InstructP2P]
  - Given an image and an instruction -> target image following the instruction
- Subject-driven Image Editing [DreamEdit]
  - Given a subject and image -> target image containing the subject and background
- Style-guided Image Generation [StyleDrop]
  - Given a style reference and a source image -> target image with the given style
- Control-guided Image Generation [ControlNet]
  - Given a keypoint, bbox, pose, layout -> target image following these signal
- Compositional multi-subject-driven Image Generation [Custom Diffusion]
  - Given reference of multiple subjects -> target image containing all of the input subjects

Standardized Image Editing Model Evaluation

- There are huge amount of image editing models
  - All the evaluation is done differently
  - The code and data are dispersed everywhere
  - It's hard to keep track of all the model performance, etc

- We plan to host a platform for Holistic Image Editing Evaluation
  - Comile a set of evaluation tasks, hire human raters
  - Standaridize the input formats
  - Continuously update the Benchmark (Like lmsys and HELM)

Instruction-tuned Foundation model

- Currently, specific model is designed for specific task.
  - It's hard to maintain so many individual models

- Can we compile all these skills into a single model?
  - We plan to develop FLAN-type instruction-tuned Image manipulation model
  - By training on a large set of image manipulation task, we hope it can generalize to new tasks

- One difficulty now is that we need to have better foundation vision-language models
  - Encoding interleaved images and text
  - Better architecture than UNet to digest these diverse instruction inputs

Thank You!