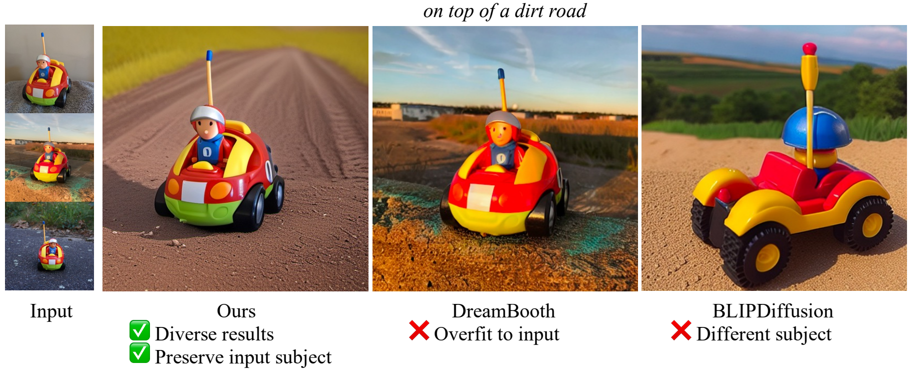
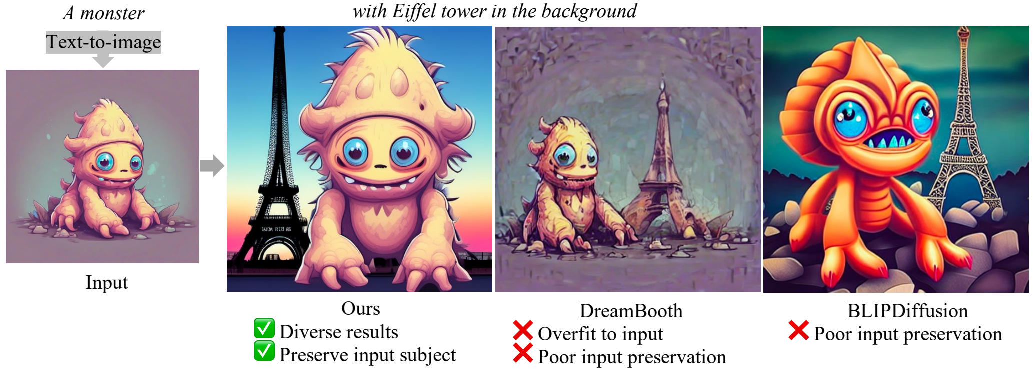
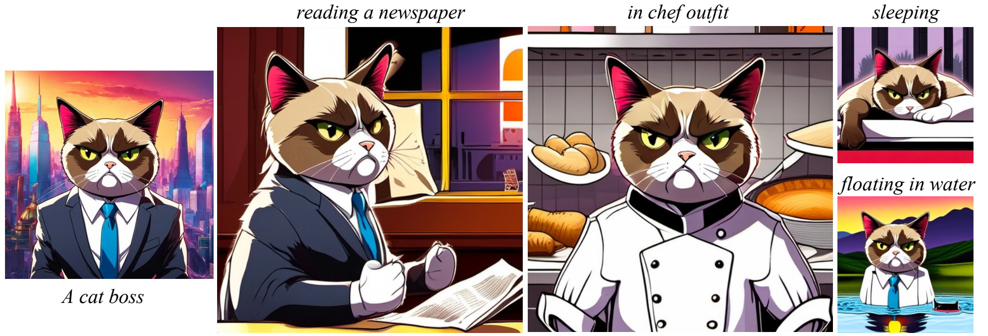
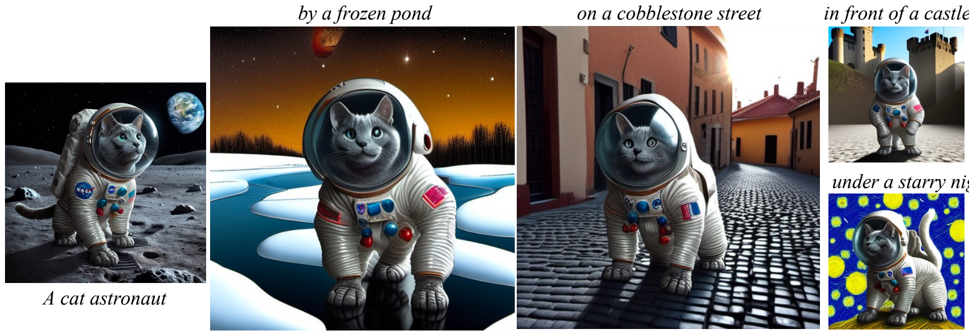
JeDi: Joint-image Diffusion Models for Finetuning-free Personalized Text-to-image Generation

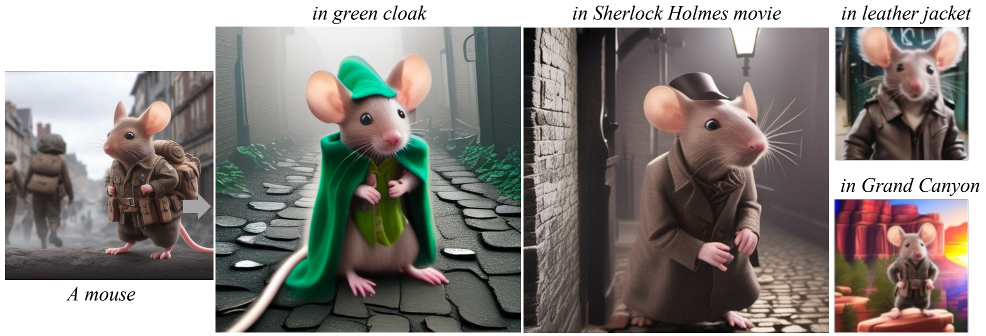
Yu Zeng, Vishal M. Patel, Haocheng Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, Yogesh Balaji

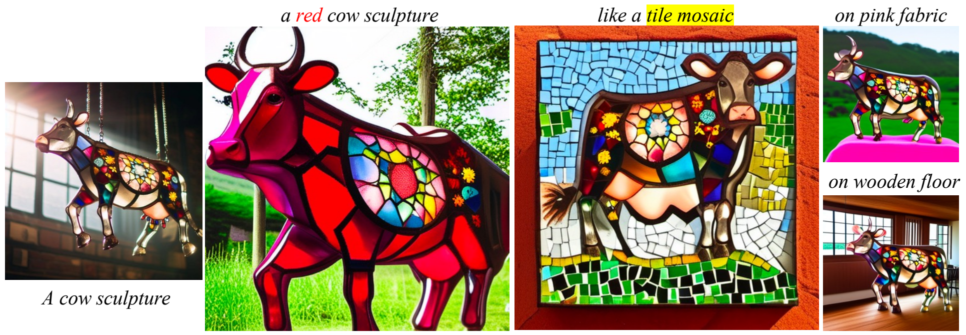






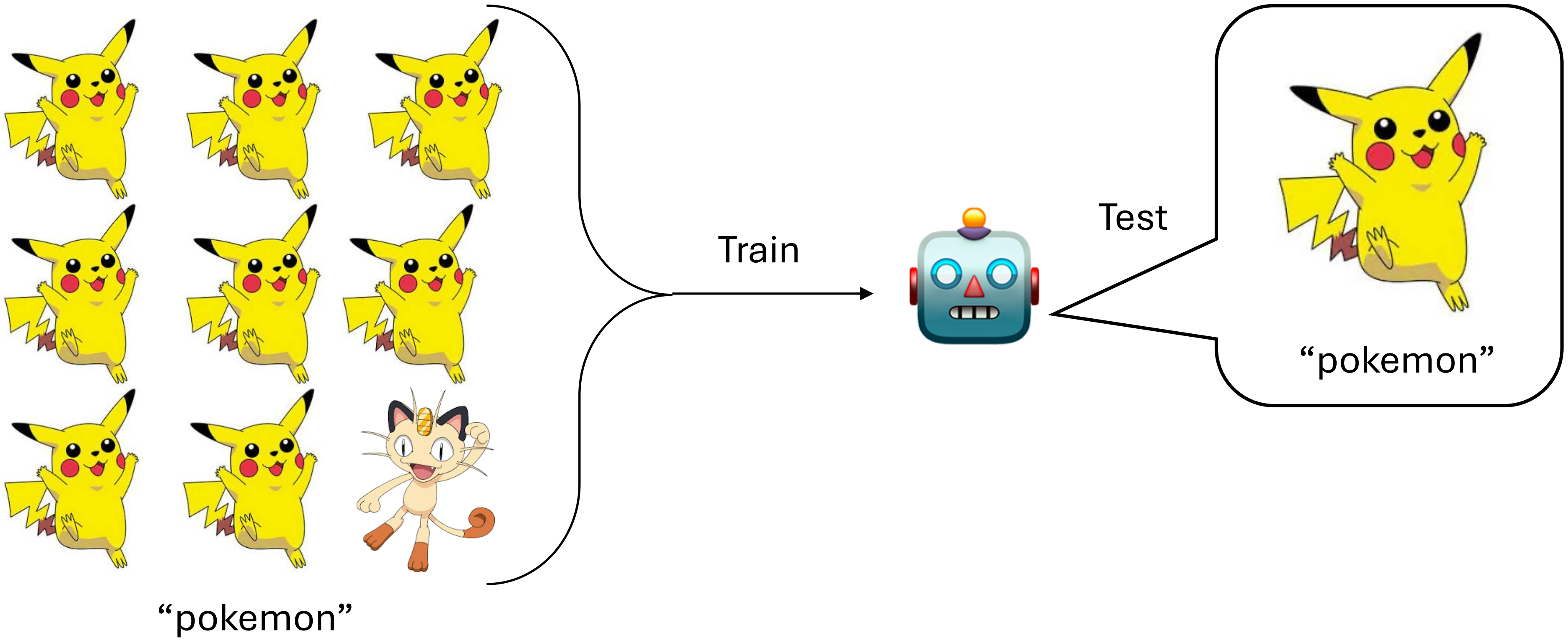






Limitation of learning-based generation

- Challenges of rare concepts and novel concepts



“pokemon”

Train

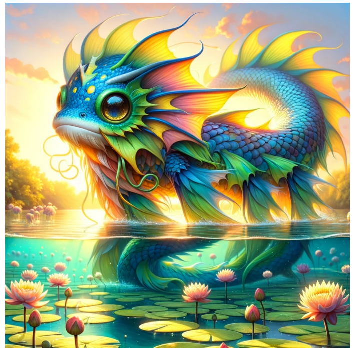
Test

“pokemon”

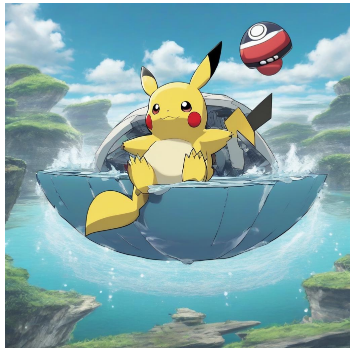
Limitation of learning-based generation

Pokemon meowth

Dalle3 (ChatGPT)



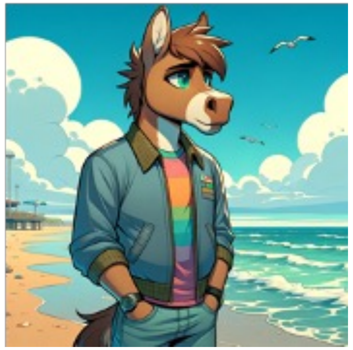
SDXL



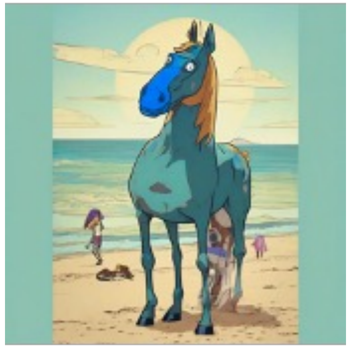
Pokemon meowth floating on top of the water

Bojack horseman

Dalle3 (ChatGPT)



SDXL



Bojack horseman on the beach

Represent new concepts

- Use example images

Pokemon floating on top of the water

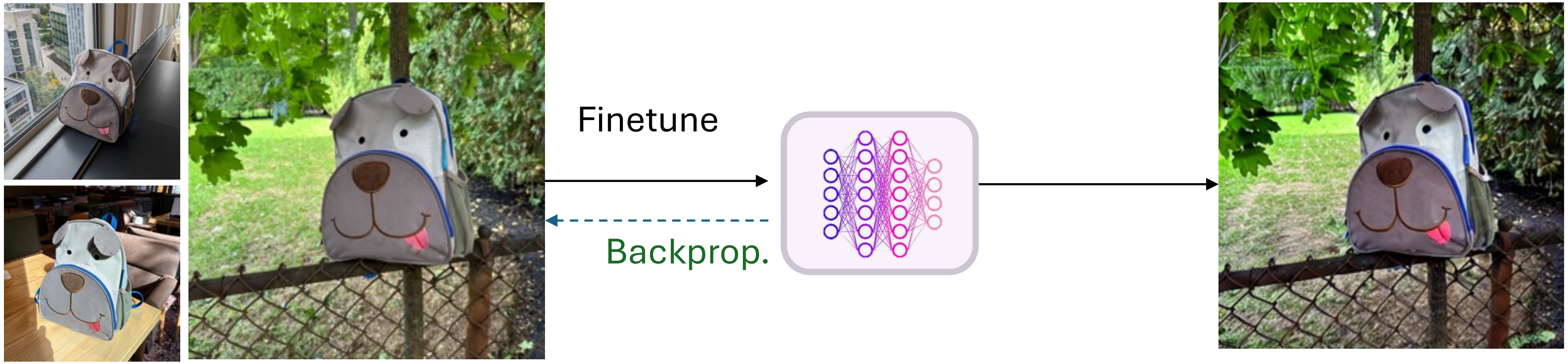
Input

Bojack horseman on the beach

Input

Adapt to new concepts

- Finetuning ?
  - Time- and resource- consuming
  - Overfitting to finetuning samples



Finetune

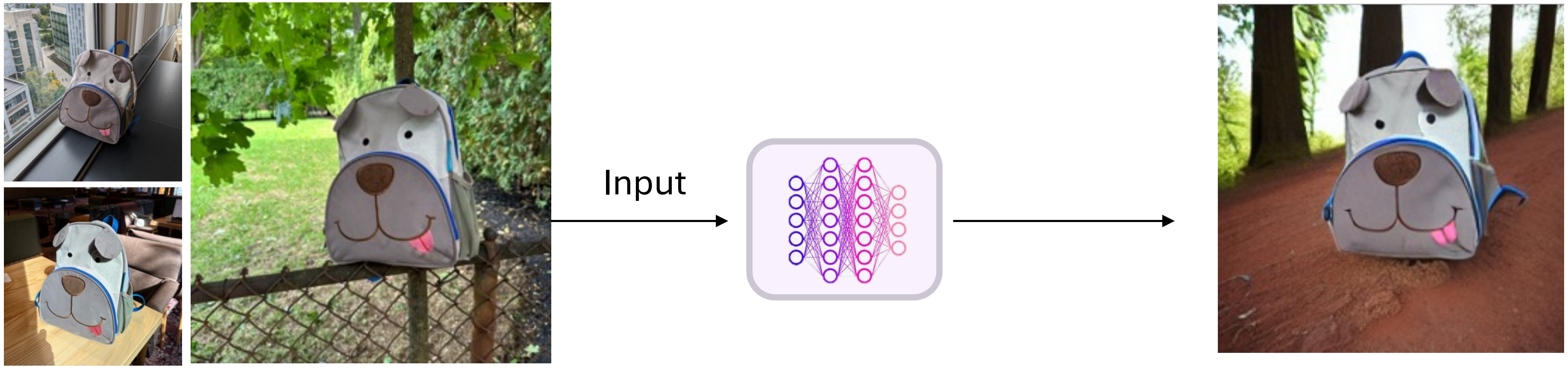
Backprop.

“photo of a backpack”

“a backpack on top of a dirt road”

Training-free fast adaptation

- Finetuning
    - Time- and resource- consuming
    - Overfitting to finetuning samples
- Reference-based fast adaptation



“photo of a backpack”

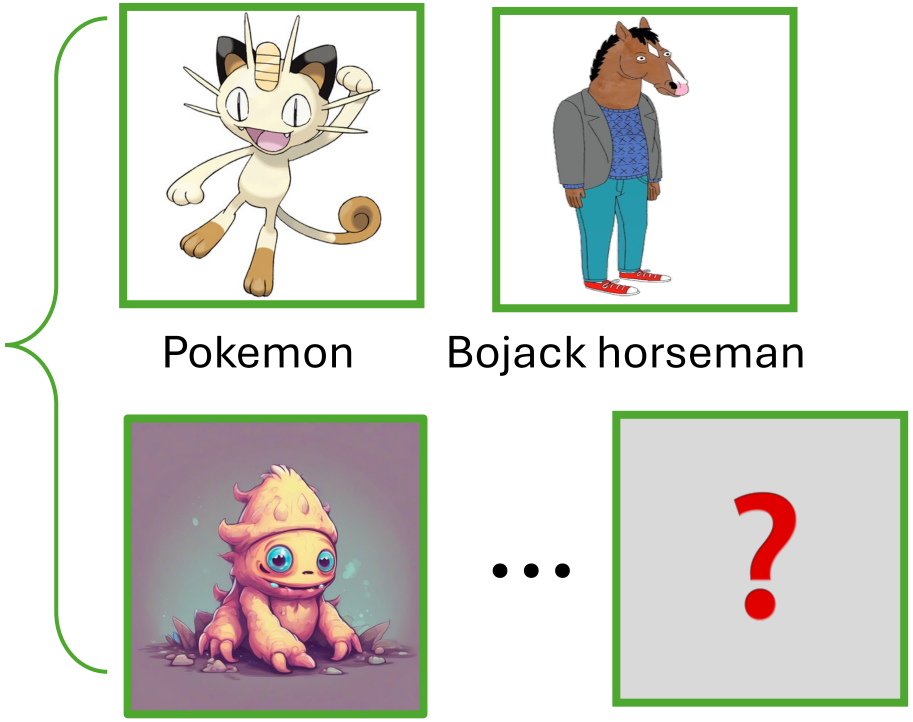
“a backpack on top of a dirt road”

Y. Zeng, V. Patel, et al, “JeDi: Joint-image diffusion models for finetuning-free personalized text-to-image generation”, CVPR, 2024.

Y. Zeng, Y. Balaji, T. Wang, X. Huang, M. Liu, “Neural networks to generate objects within different images,” US Patent App. 18/518,430, Dec. 2023.

Training-free fast adaptation

What's changing



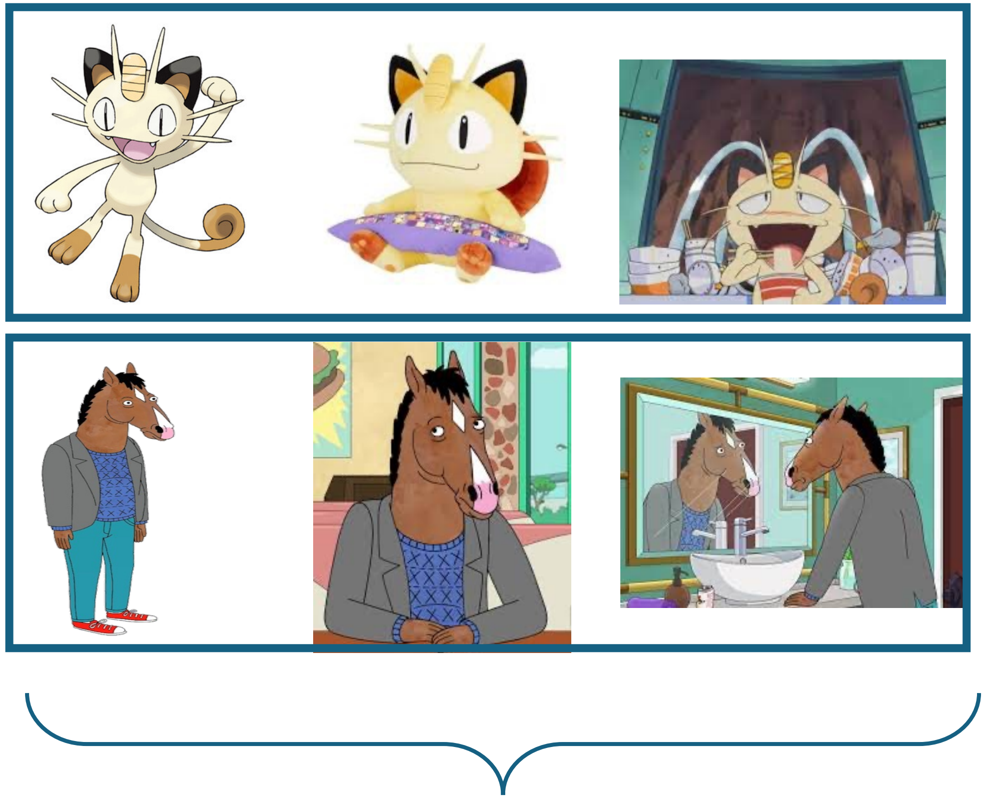
Individual characters

Pokemon

Bojack horseman

New made-up monster

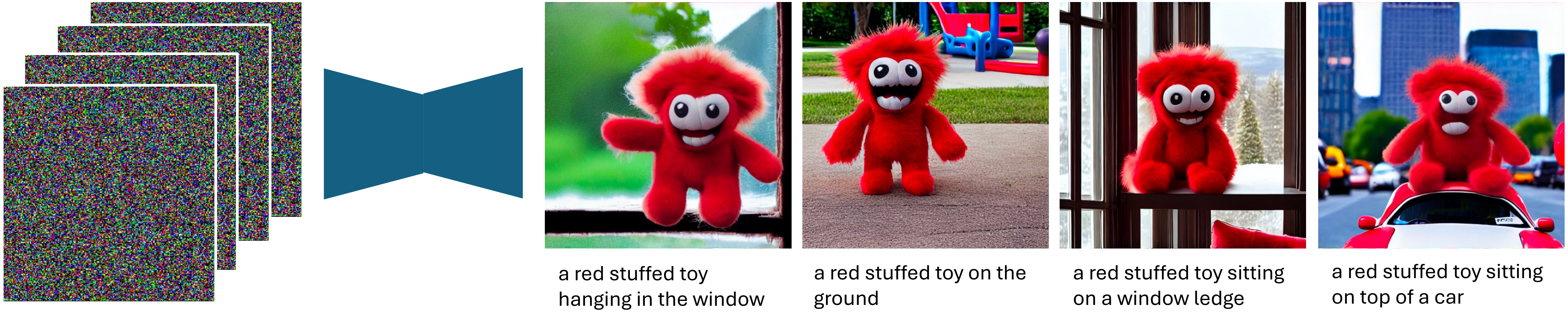
What's invariant



Same-subject relationship

Joint-image diffusion

- Modeling the joint distribution of multiple images sharing the same concept
- Joint-image denoising diffusion



a red stuffed toy hanging in the window

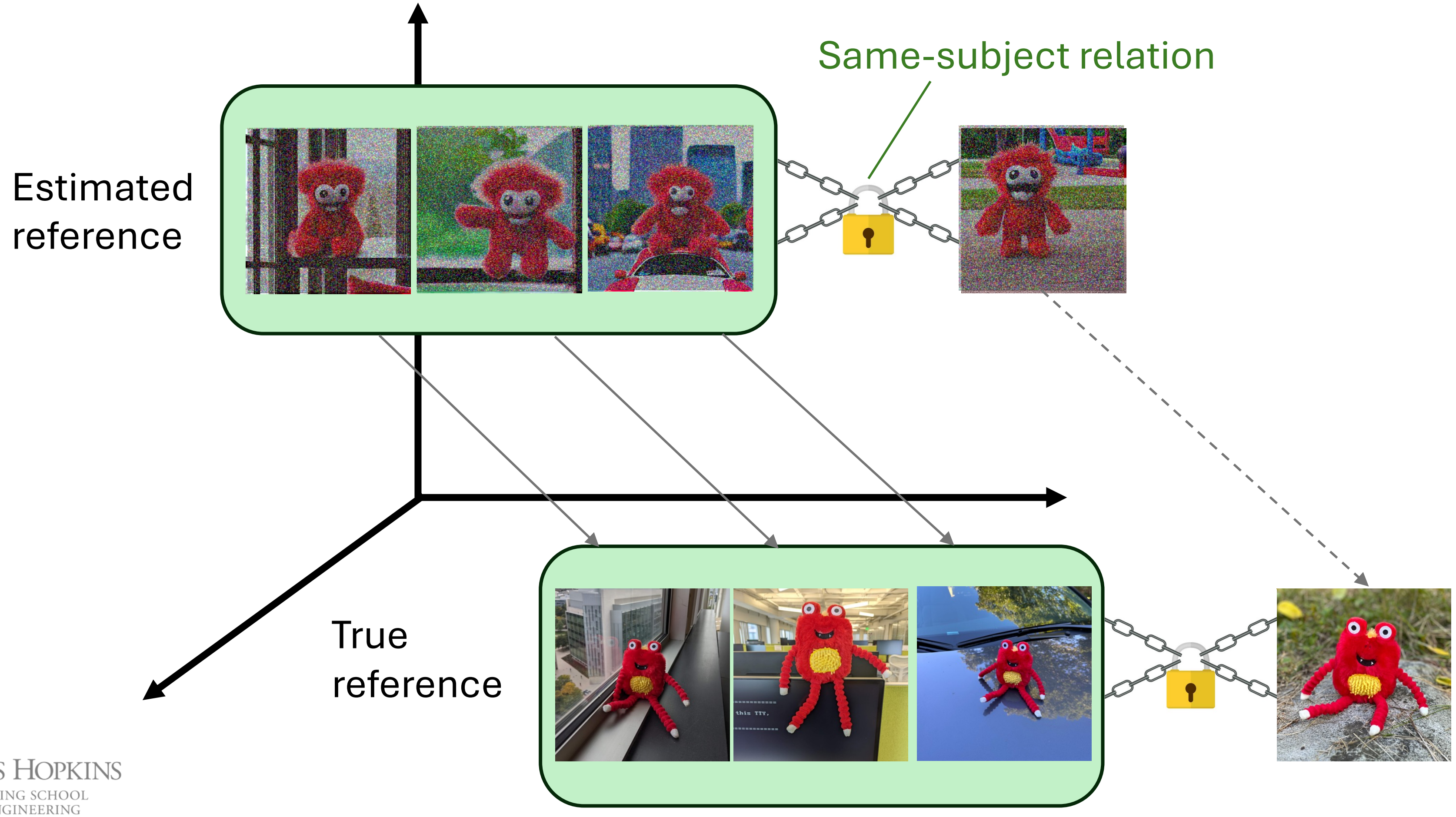
a red stuffed toy on the ground

a red stuffed toy sitting on a window ledge

a red stuffed toy sitting on top of a car

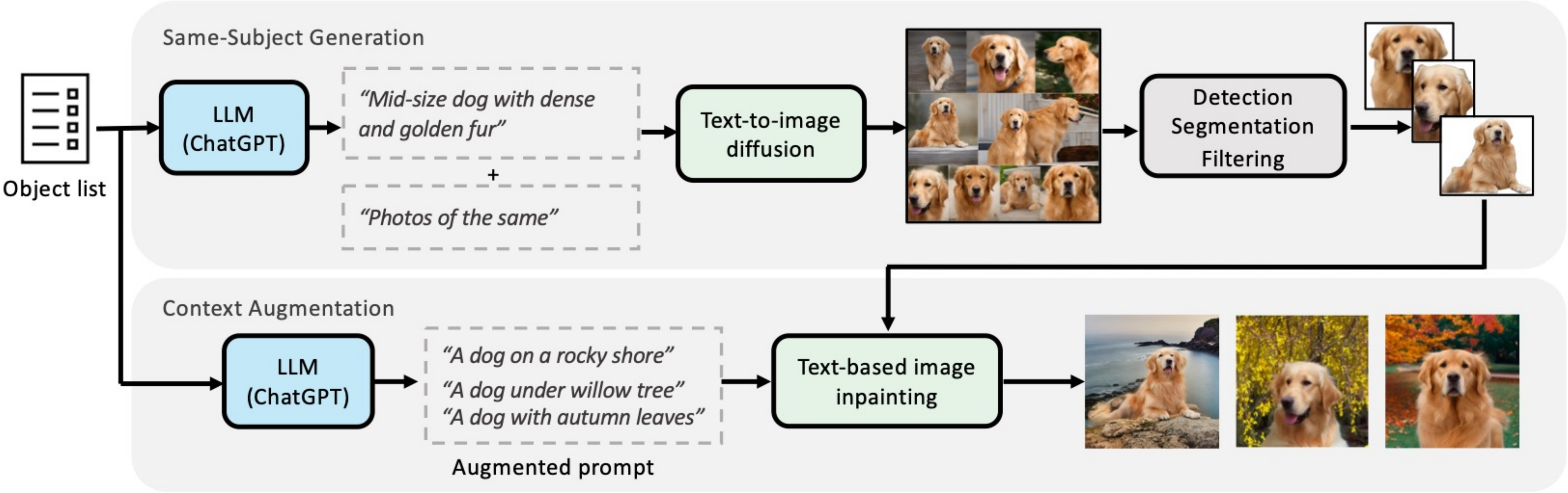
Generation guided by reference

- Generate both the reference images and target image
- Using the true reference images as guidance



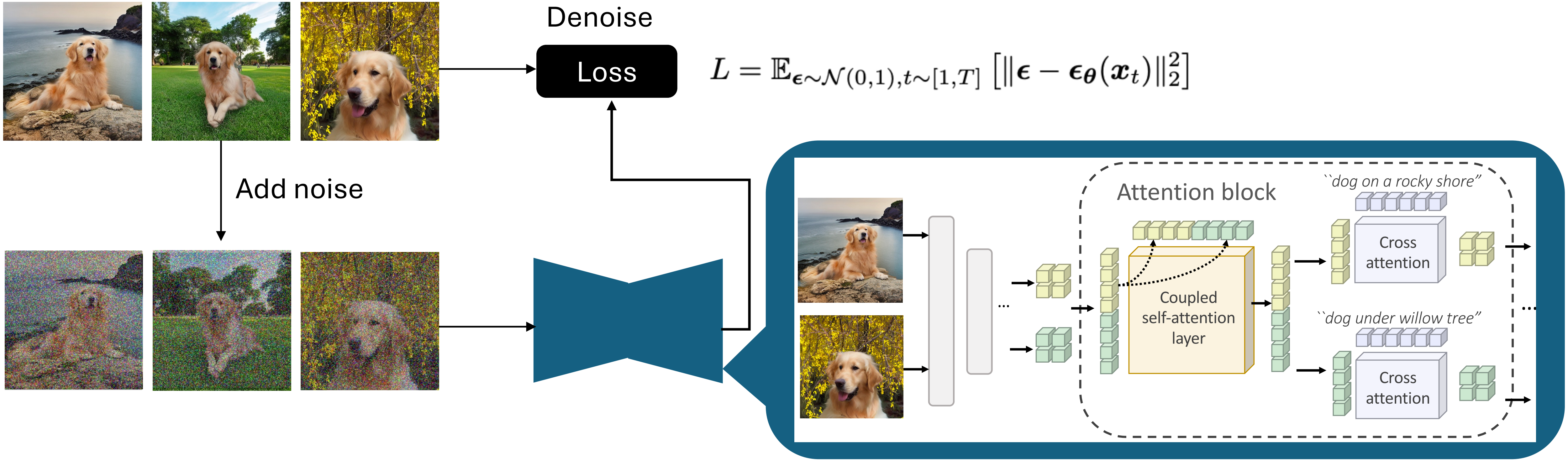
Joint-image diffusion: training

- Training data: image sets



Joint-image diffusion: training

- Modeling the joint distribution of an image set sharing the same concept
- Joint-image denoising diffusion

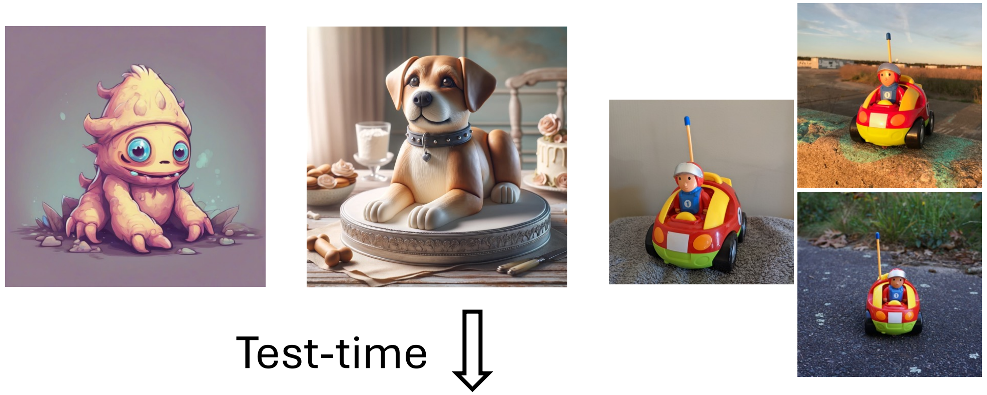


$$L = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),t\sim[1,T]}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t)\|_2^2\right]$$

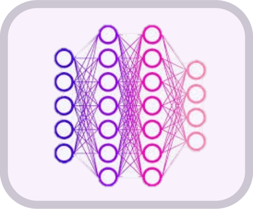
Training-free fast adaptation

- Image synthesis and manipulation
  - Training-free fast adaptation

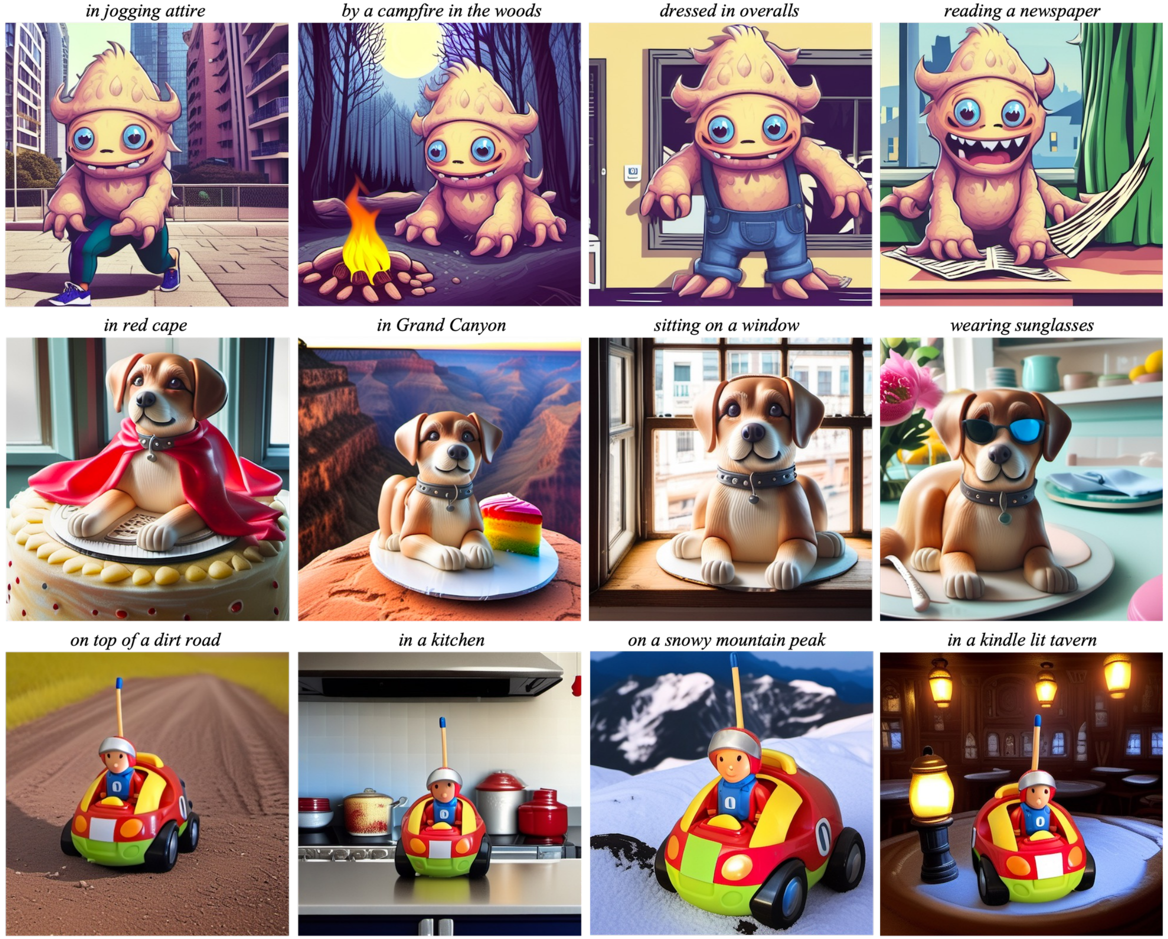
Unseen example images



Test-time



Pre-trained model



in jogging attire · *by a campfire in the woods* · *dressed in overalls* · *reading a newspaper*

in red cape · *in Grand Canyon* · *sitting on a window* · *wearing sunglasses*

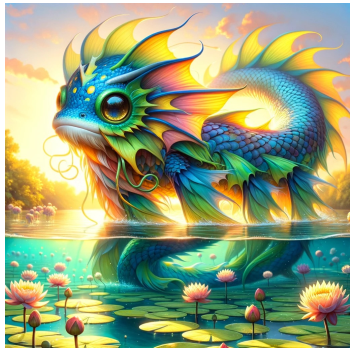
on top of a dirt road · *in a kindle* · *on a snowy mountain peak* · *in a kindle lit tavern*

Results

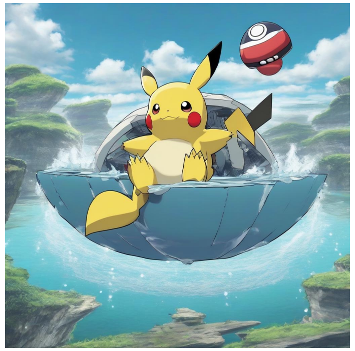
- Rare concept generation

Pokemon meowth

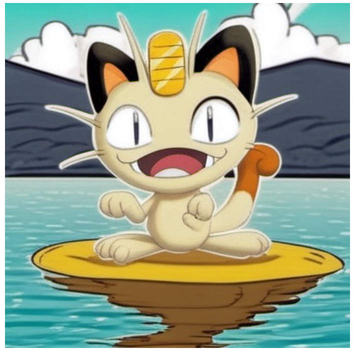
Pokemon meowth floating on top of the water



Dalle3 (ChatGPT)



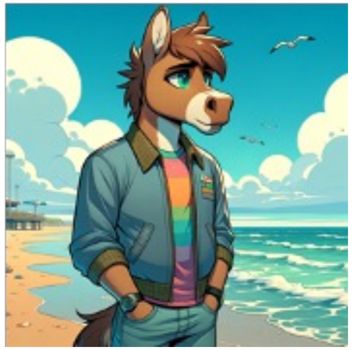
SDXL



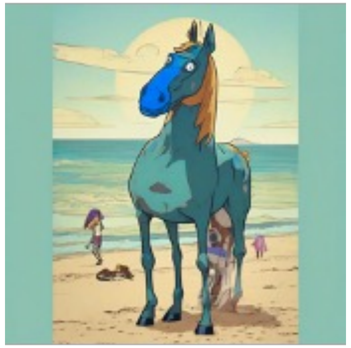
Ours

Bojack horseman

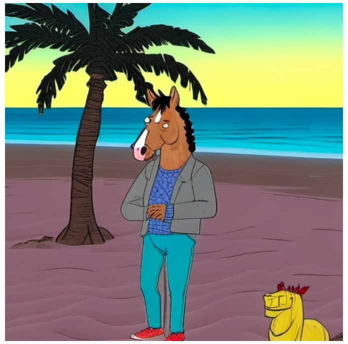
Bojack horseman on the beach



Dalle3 (ChatGPT)



SDXL



Ours

Results

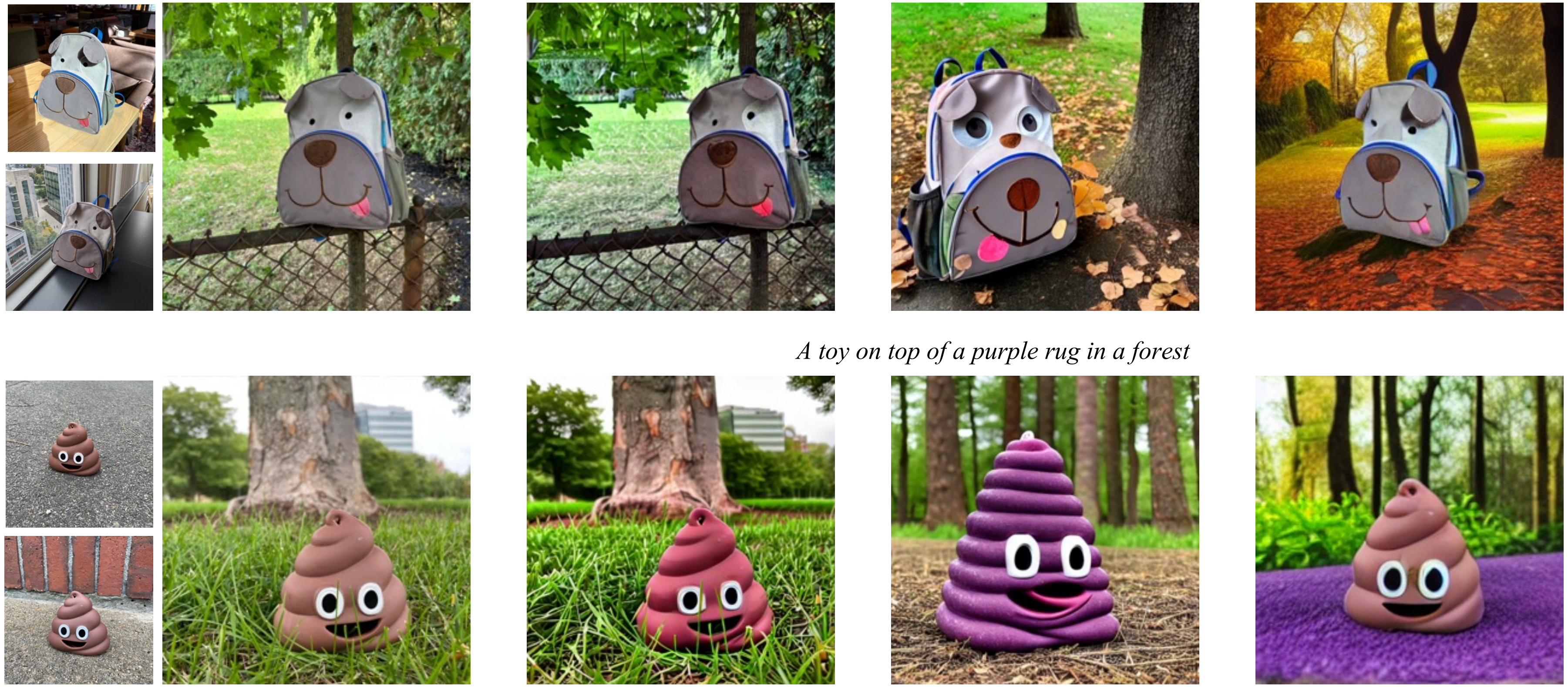
- Novel concepts generation

overfitting

Underfitting

A backpack with a tree and autumn leaves in the background

A toy on top of a purple rug in a forest



Reference images

Finetuning (CD, DB)

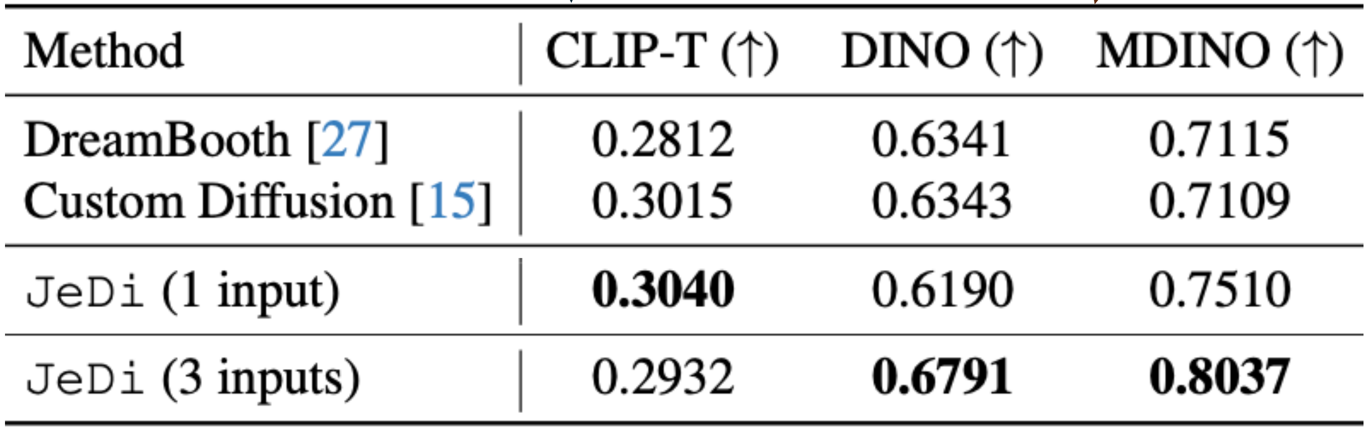
Ours

Results

- Novel concepts generation

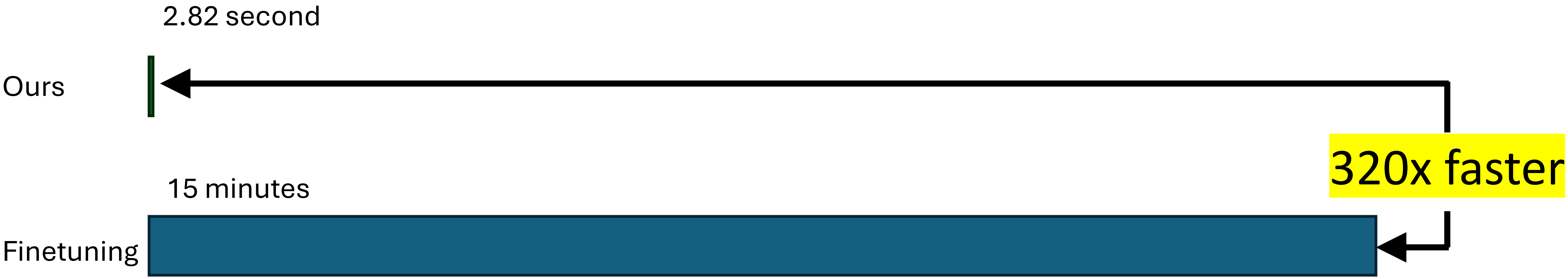
Text alignment

Concept preservation



	Method	CLIP-T (↑)	DINO (↑)	MDINO (↑)
Finetuning	DreamBooth [27]	0.2812	0.6341	0.7115
	Custom Diffusion [15]	0.3015	0.6343	0.7109
Ours	JeDi (1 input)	**0.3040**	0.6190	0.7510
	JeDi (3 inputs)	0.2932	**0.6791**	**0.8037**

13% improvement



2.82 second
Ours
320x faster
15 minutes
Finetuning