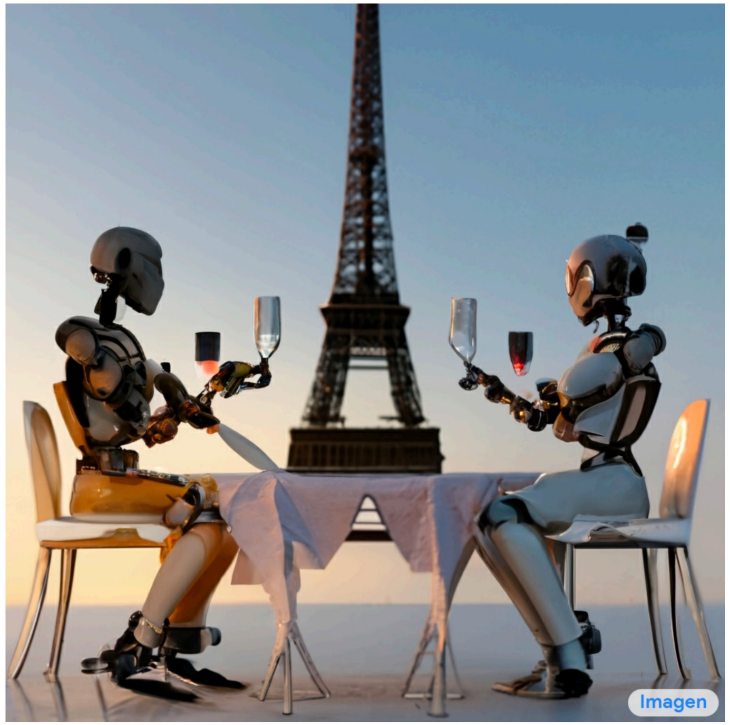
Unleashing the Power of Synthetic Data in GenAI

Speaker: Wenhu Chen

Background and Motivation

- Text-to-Image Generation
  - 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.

Controllable Image Generation

- How to control the model to generate specific subject?
  - Subject-Level Control
    - A specific dog or a specific person in different scenarios.
  - Background-Level Control
    - A specific scene with different subjects in the front.

- How to generate an image by editing another image?
  - Local-Level Control
    - Addition/Removal/Swap of Subjects
  - Global-Level Control
    - Scene/Style/Attribute Modification
  - Environment Modification

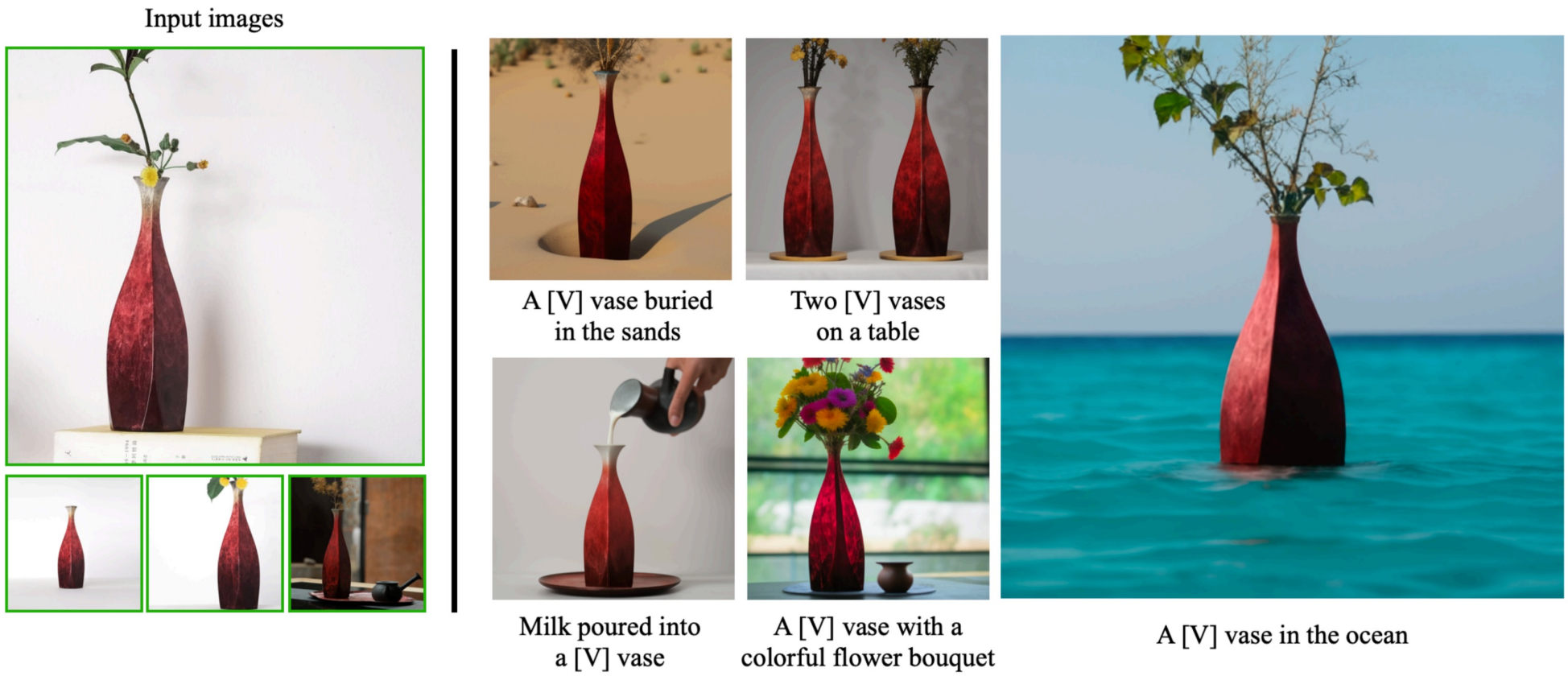
Google DeepMind

Subject-driven Text-to-Image Generation via Apprenticeship Learning

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, William W. Cohen

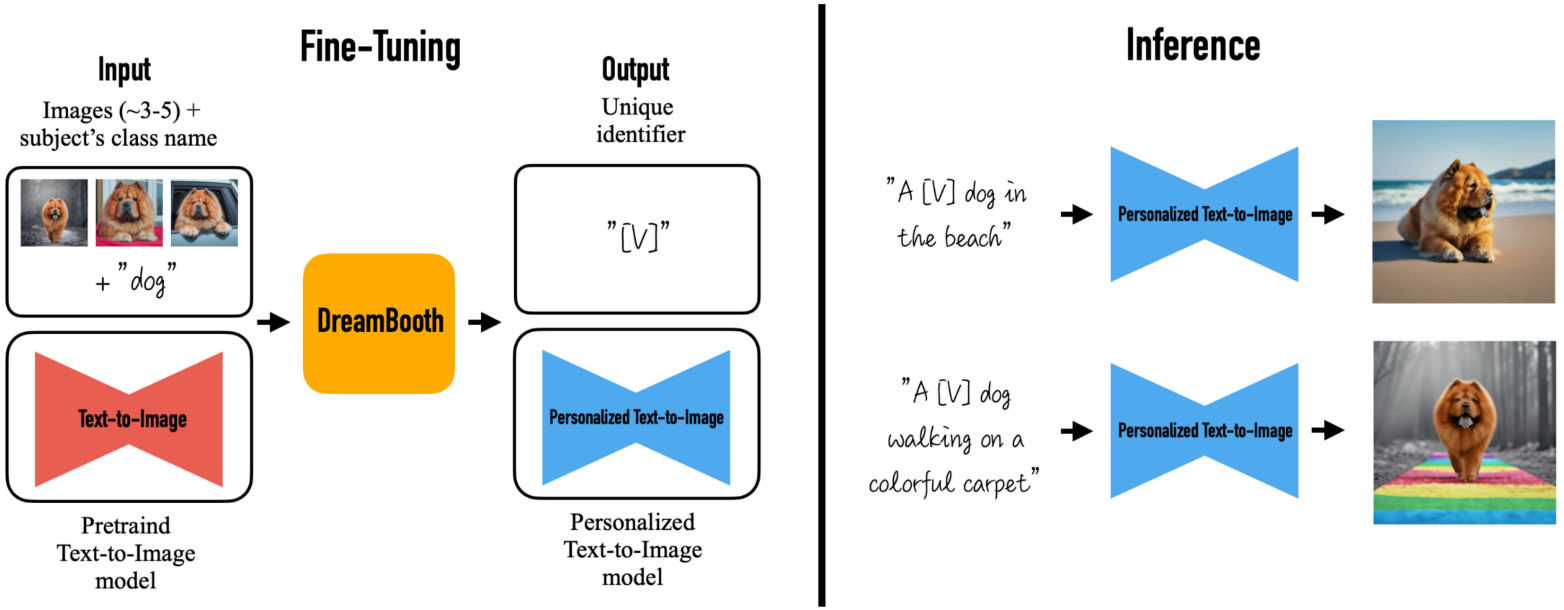
Published at NeurIPS 2023

Subject-Level Control



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 it to generate images with [V]



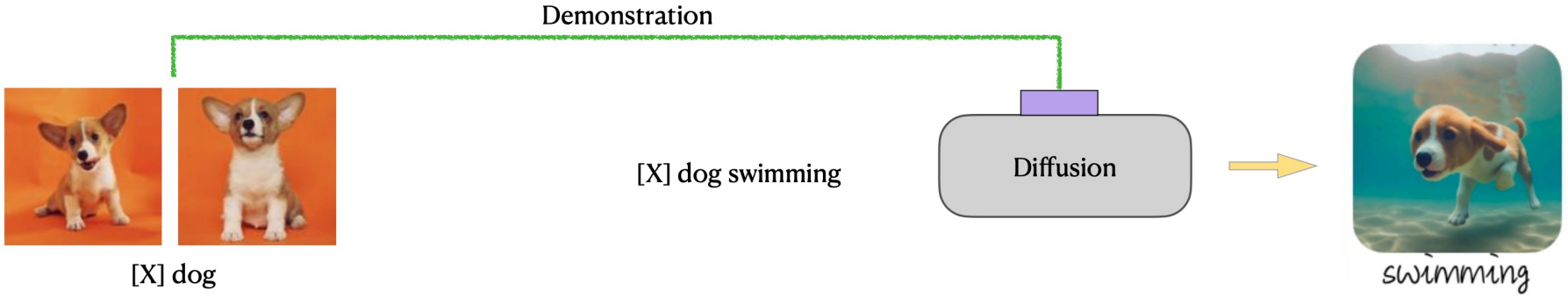
DreamBooth: Fine Tuning Text-to-Image Diffusion Models

- 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.

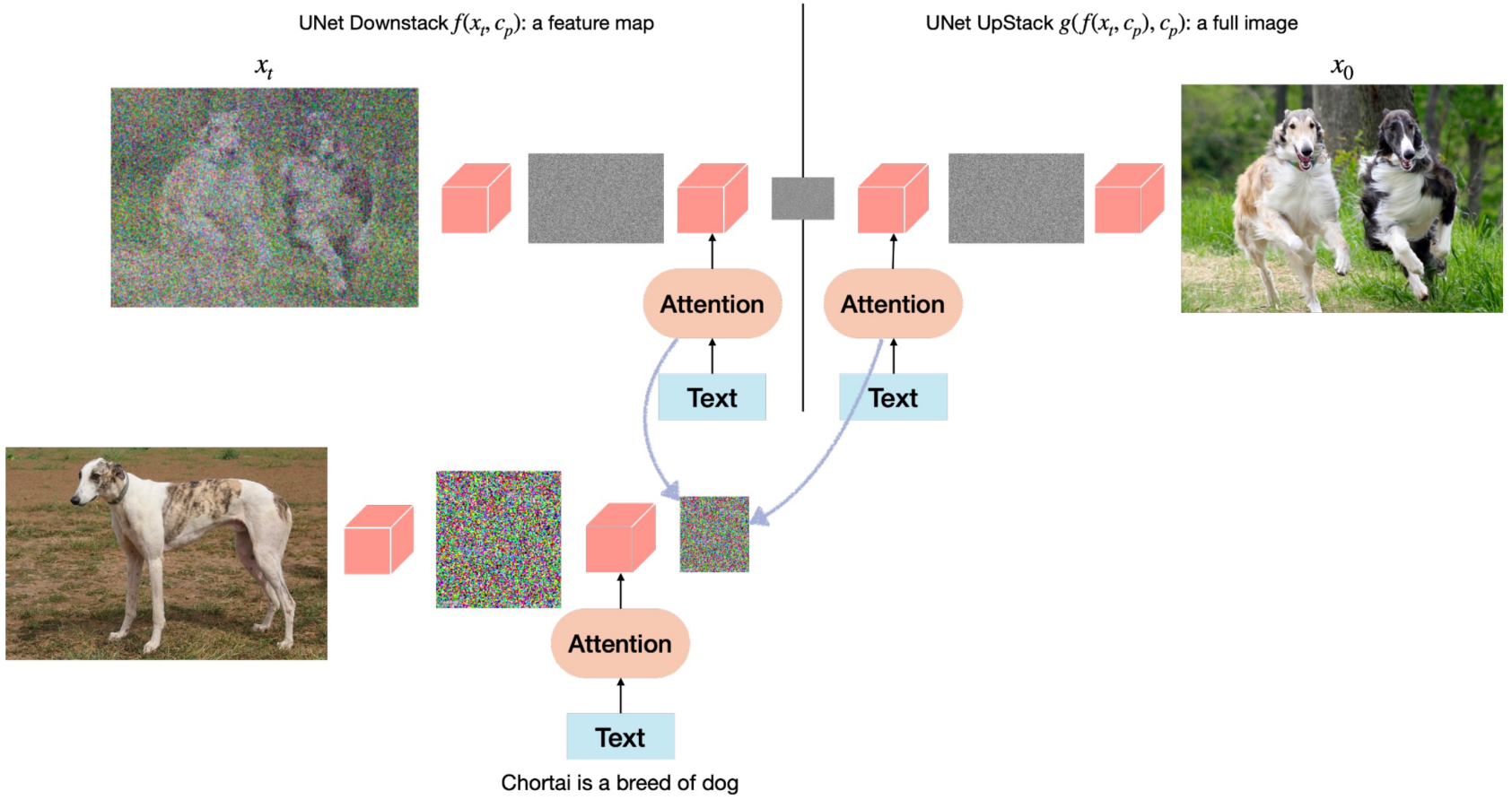


What do we need to achieve In-Context Learning?

- 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 from the input

- 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: cross attention layer



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 through the PaLI model inside Google.

How is the clusted data quality?

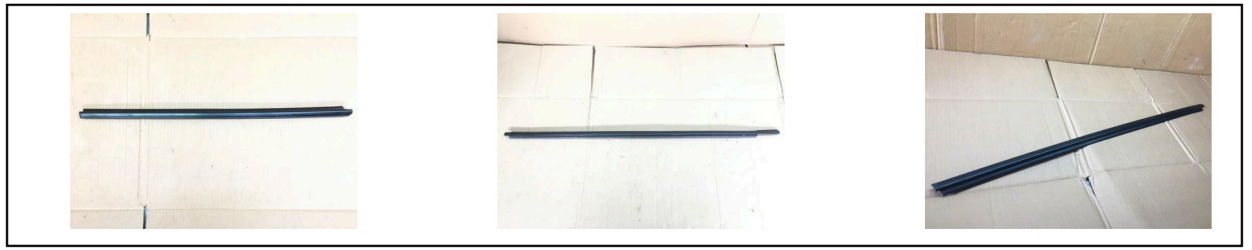
A limousine parked in a parking lot

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

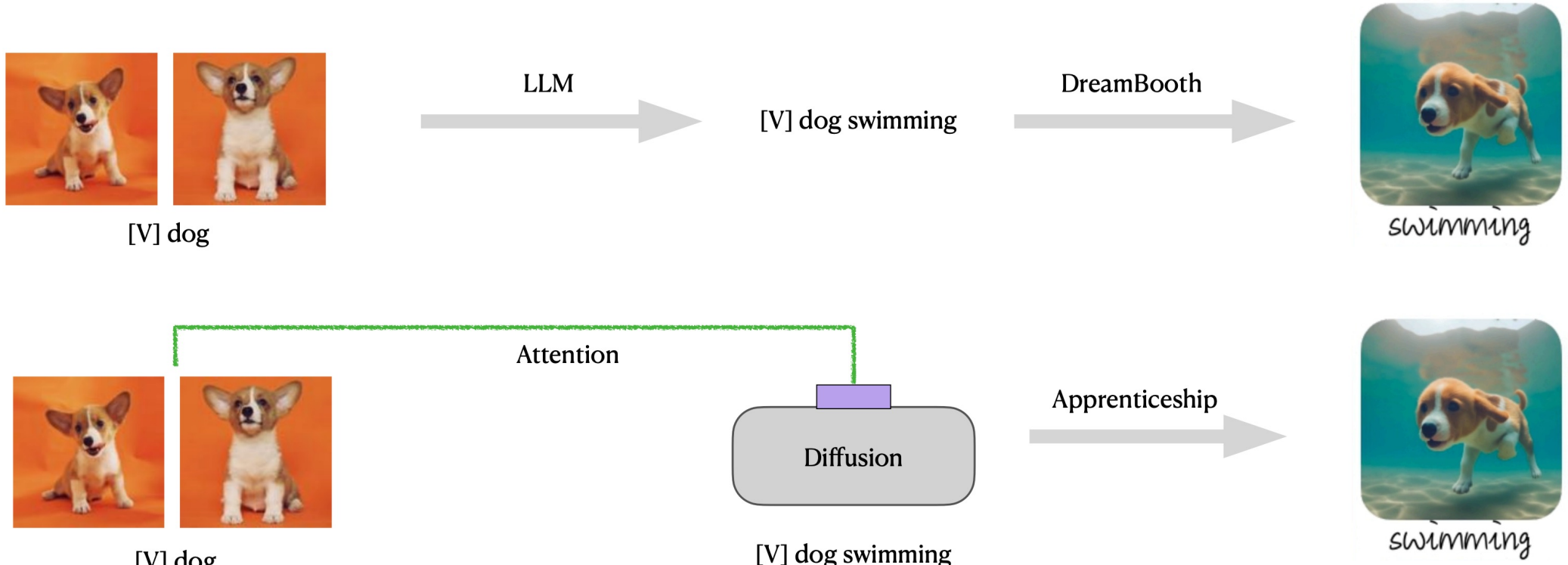
- Reasons:
  - The target and demonstrations images are too similar
  - The model falls into a copy-paste local optima

How can we make it better?

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

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

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 to complete the training
  - The apprentice model learns surprisingly fast

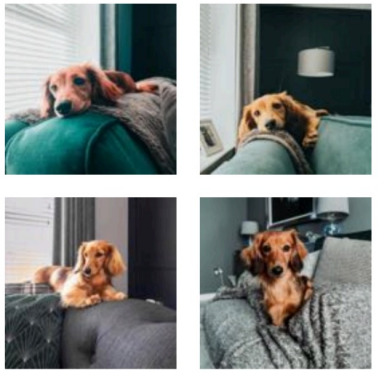
- Skillset of the apprentice model
  - Stylization: changing the style of the subject
  - Recontextualization: changing the scene of the subject
  - Multi-View synthesis: changing the view perspective of the subject
  - Attribute Modification: changing the color, textual,emotion,etc
  - Compositional: Stylization + Recontextualization

Model Outputs

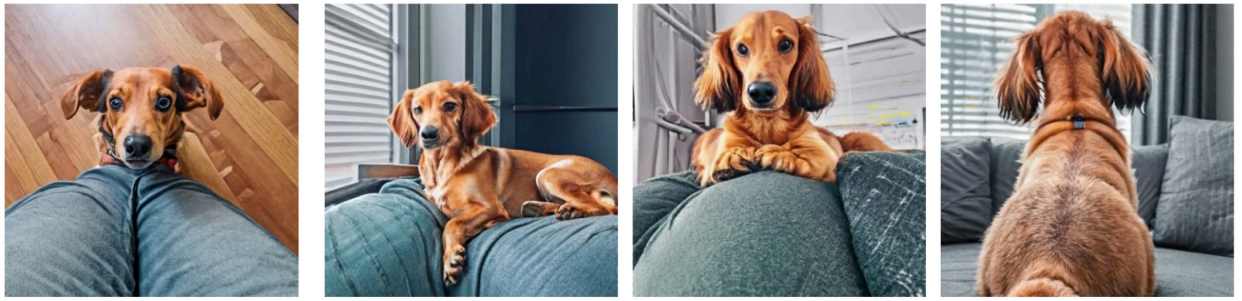
A duck toy

Pablo Picasso | Rembrandt | Rene Magritte | Vincent van Gogh

A dog



Top-down view | Side view | Bottom view | Back view



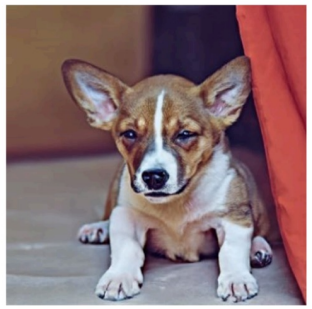
Model Outputs

A dog

Depressed

Joyous

Sleepy



Screaming

A monster toy

Blue

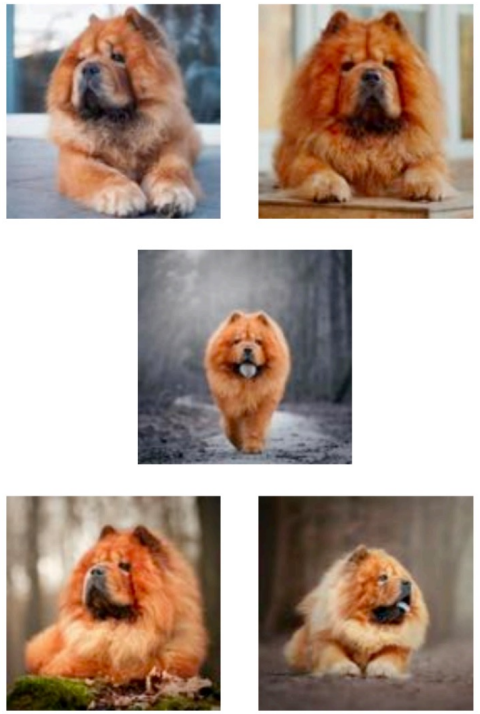
Green

Purple

Pink

Model Outputs

A dog



Chef outfit

Police outfit

Nurse outfit

Fire-Fighter outfit

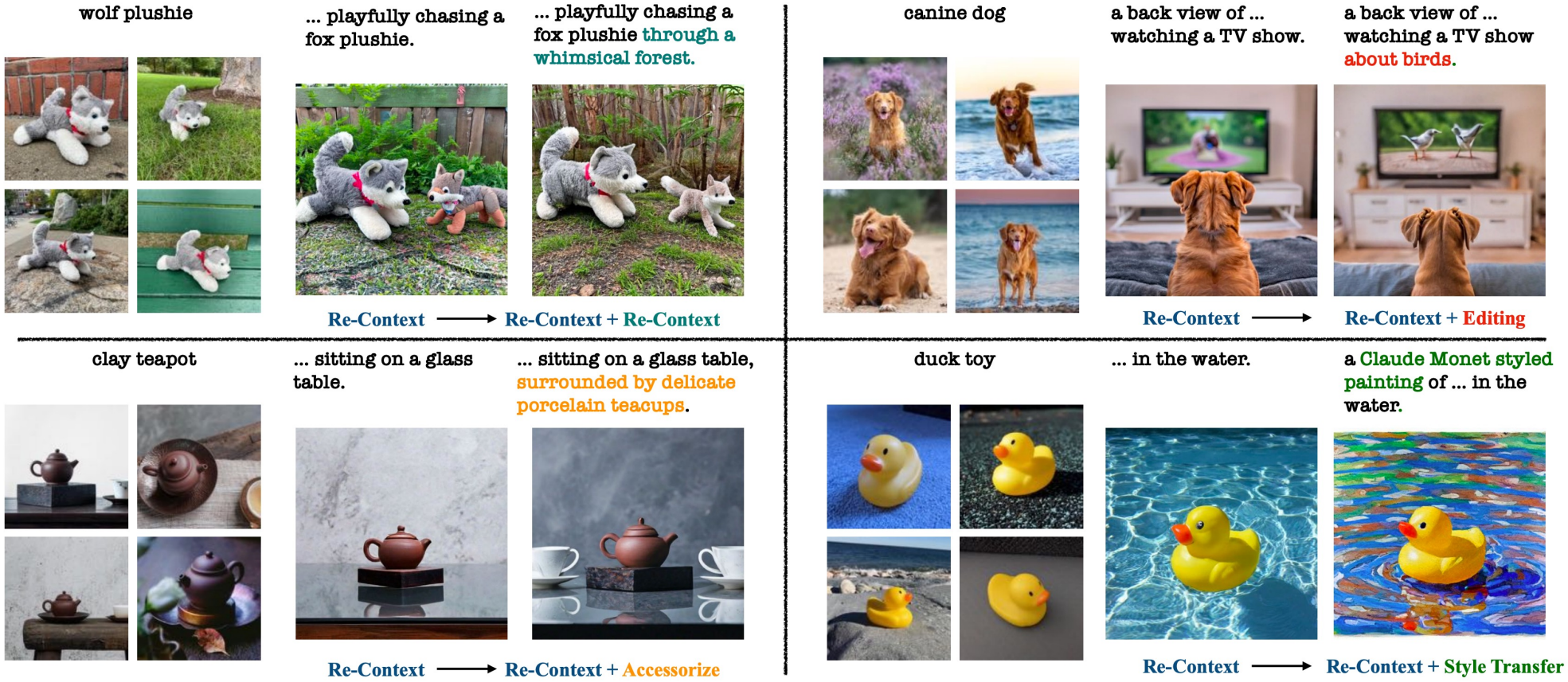
Ironman outfit

Witch outfit

Superman outfit

Angel outfit

Compositional Model Outputs



Human Evaluation

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

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**

VIEScore: Towards Explainable Metrics for Conditional Image Synthesis Evaluation

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, Wenhu Chen

Published at ACL 2024 Main Conference

Image Generation Evaluation

- Image generation evaluation
  - LPIPS
  - DINO
  - CLIP
  - DreamSim
  - FID/KID

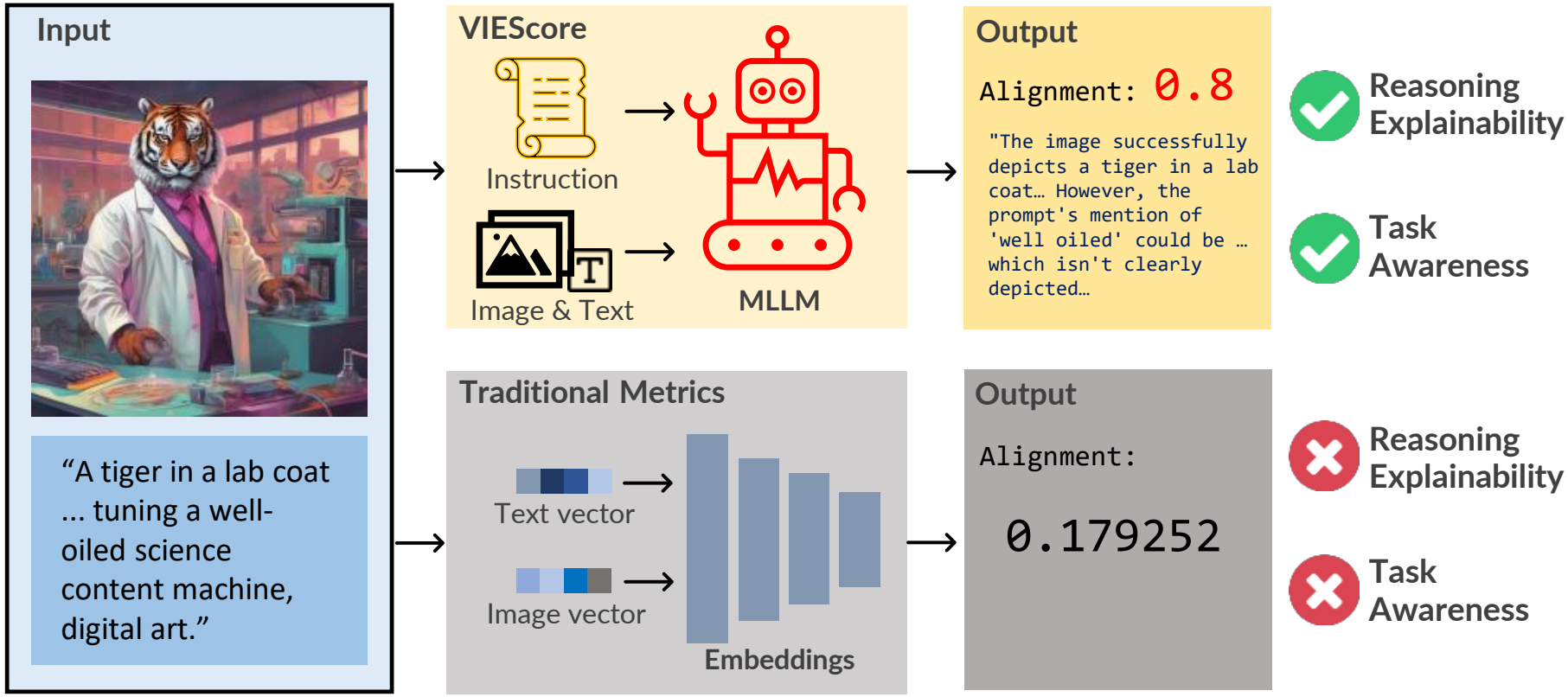
- These scores are only capable of evaluating text-to-image.

How good are they with Humans

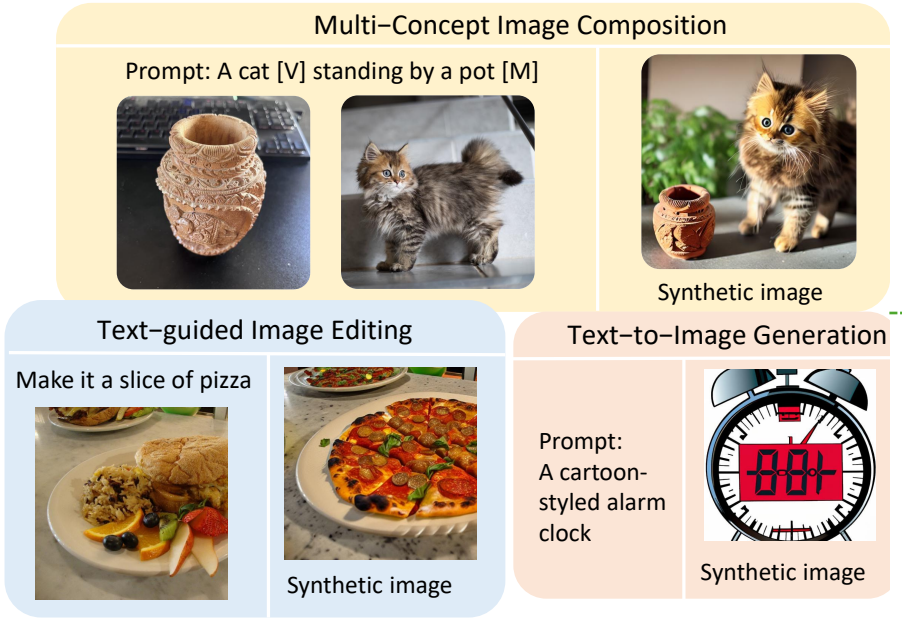
These automatic scores have shown very low correlation with humans on wider image generation tasks.

	PQ ↑				SC ↑
Task	corr(LPIPS)	corr(DINO)	corr(CLIP-I)	corr(DreamSim)	corr(CLIP)
Text-guided Image Generation	N/A	N/A	N/A	N/A	-0.0819
Mask-guided Image Editing	-0.0502	-0.0421	-0.0308	-0.0294	-0.0257
Text-guided Image Editing	0.0439	0.0449	0.0765	0.0432	0.0437
Subject-driven Image Generation	0.0905	0.1737	0.1204	0.0943	-0.0117
Subject-driven Image Editing	0.2417	0.1690	0.3770	0.3846	-0.0443
Multi-concept Image Composition	-0.1018	-0.1657	-0.0965	0.0140	-0.0107
Control-guided Image Generation	0.1864	0.1971	-0.0985	-0.0045	-0.2056

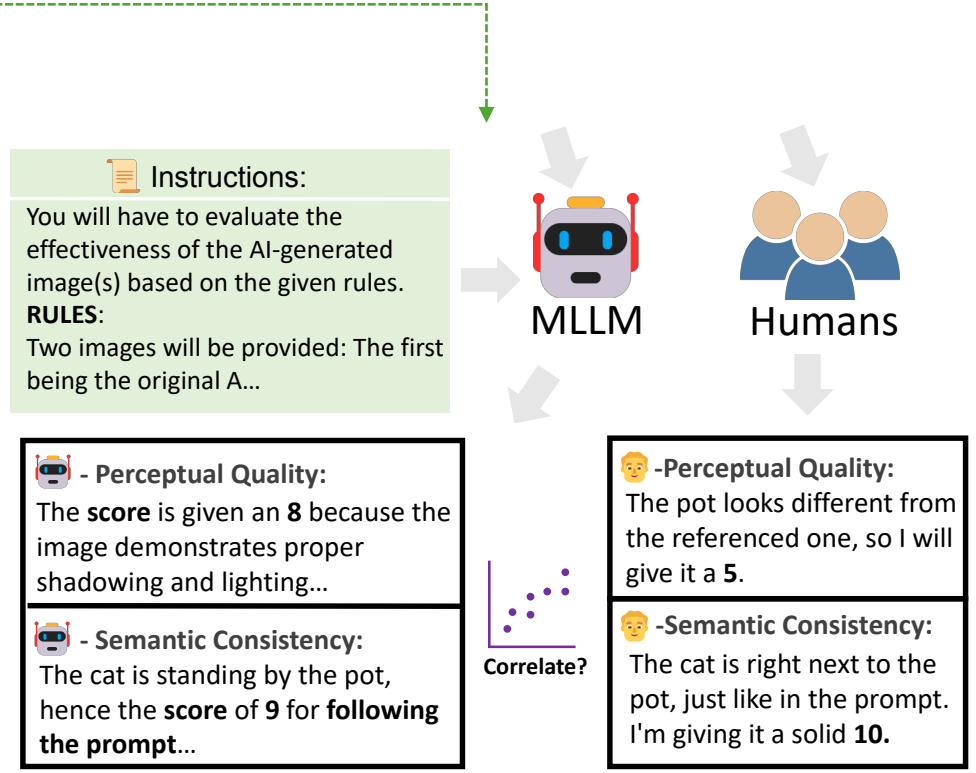
Multimodal Model Evaluation



Multimodal Model Evaluation



Multimodal Conditions



Correlation with Humans

- Some large multimodal models are highly effective in judging generation quality.

- GPT-4o achieves quite promising alignment with humans on semantic consistency.

- Open-source multimodal models have almost zero performance on all the different tasks.

- GPT-4o and upcoming models could be a good alternative for automatic evaluation.

Backbone	M-H^{SC}_{corr}	M-H^{PQ}_{corr}	M-H^{O}_{corr}
Across All 7 Tasks			
Human Raters	0.4700	0.4124	0.4558
VIESCORE			
GPT-4o$_{0shot}$	0.4459	0.3399	0.4041
GPT-4o$_{1shot}$	0.4309	0.1167	0.3770
Gemini-Pro$_{0shot}$	0.3322	0.2675	0.3048
Gemini-Pro$_{1shot}$	0.3094	0.3070	0.3005
GPT-4v$_{0shot}$	0.3655	0.3092	0.3266
GPT-4v$_{1shot}$	0.2689	0.2338	0.2604
LLaVA$_{0shot}$	0.1046	0.0319	0.0925
LLaVA$_{1shot}$	0.1012	0.0138	0.0695
Qwen-VL$_{0shot}$	0.0679	0.0165	0.0920
BLIP2$_{0shot}$	0.0504	-0.0108	0.0622
InstructBLIP$_{0shot}$	0.0246	0.0095	0.0005
Fuyu$_{0shot}$	-0.0110	-0.0172	0.0154
CogVLM$_{0shot}$	-0.0228	0.0514	-0.0050
OpenFlamingo$_{0shot}$	-0.0037	-0.0102	-0.0122

Table 1: Correlations across all tasks with different backbone models. We highlight the highest correlation numbers in green. See Appendix C for details.

OmniEdit: Building Image Editing Generalist Models Through Specialist Supervision

Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, Wenhu Chen

Work in Progress

Image Editing

- Image Editing
- Multi-Skill
- Multi-Resolution
- Instruction

Background

- Instruction-guided Image Editing Models
  - Inversion-based methods
    - SDEdit
    - Prompt2Prompt
    - Plug-and-Play
    - Null Text Inversion
  - End-to-end methods
    - InstructPix2Pix
    - MagicBrush
    - HQEdit
    - HIVE
    - UltraEdit

Pro and Cons

- Inversion-based methods
  - Zero-shot without tuning
  - Slow
  - Sensitive to hyper-parameters

- End-to-end methods
  - Requires Tuning
  - Fast
  - Stable w.r.t hyper-parameters

Issues with end-to-end methods

- Trained entirely with synthetic data, which suffers from the bias or limitation in the synthesis process.
  - InstructPix2Pix was synthesized with P2P, which can barely handle local editing like addition/removal/swap
  - MagicBrush was synthesized with DALLE-inpainting, which can barely handle global editing like style/environment
  - The same happens to UltraEdit, CosXL-Edit, etc

- Quality Control is very crude
  - Mostly using CLIP-score, DINO-score ,etc
  - These scores can only measure the semantic correspondence

End-to-end Methods

Preliminary Results on 40 prompts across different editing tasks

Table 1: Comparison of OMNI-EDIT with all the existing end-to-end image editing models.

Property	InstructP2P	MagicBrush	UltraEdit	MGIE	HQEdit	CosXL	OMNI-EDIT
Training Dataset Properties							
Real Image?	✗	✓	✓	✓	✗	✗	✓
Any Res?	✗	✗	✗	✗	✗	✗	✓
High Res?	✗	✗	✗	✗	✓	✗	✓
Fine-grained Image Editing Skills							
Obj-Swap	★★☆	★★☆	★★☆	★½☆	★★☆	★☆☆	★★½
Obj-Add	★☆☆	★★☆	★☆☆	★½☆	★☆☆	★☆☆	★★½
Obj-Remove	★☆☆	★★☆	★☆☆	★½☆	★☆☆	★☆☆	★★½
Attribute	★★☆	★☆☆	★★☆	★½☆	★☆☆	★☆☆	★★½
Back-Swap	★★☆	★★☆	★★☆	★½☆	★★☆	★★☆	★★½
Environment	★☆☆	★☆☆	★☆☆	★½☆	★☆☆	★★☆	★★½
Style	★★☆	★☆☆	★★☆	★½☆	★☆☆	★★½	★★½

Motivation

- We need to build more omnipotent image editing models to handle all the editing tasks
  - Balanced Skillset
  - Better Quality Control
  - Support for different aspect ratio
  - Support for high resolution images.

- Instead of restricting to a fixed data synthesis process, we could learn from data synthesized from multiple specialist.

Formulation

Assuming that we have an optimal distribution $p(x'|x, c)$

$$L(\theta) := \sum_{\mathbf{x},c} D_{KL}(p(\mathbf{x}'|\mathbf{x}, c)||p_\theta(\mathbf{x}'|\mathbf{x}, c)) = -\sum_{\mathbf{x},c}\sum_{\mathbf{x}'} p(\mathbf{x}'|\mathbf{x}, c) \log p_\theta(\mathbf{x}'|\mathbf{x}, c) + C \quad (4)$$

where, x' is the target image, x is the source image and c is the instruction

Formulation

Assuming that we have an optimal distribution $p(x'|x,c)$

$$L(\theta) := \sum_{\mathbf{x},c} D_{KL}(p(\mathbf{x}'|\mathbf{x},c)\|p_\theta(\mathbf{x}'|\mathbf{x},c)) = -\sum_{\mathbf{x},c}\sum_{\mathbf{x}'} p(\mathbf{x}'|\mathbf{x},c)\log p_\theta(\mathbf{x}'|\mathbf{x},c) + C \quad (4)$$

where, x' is the target image, x is the source image and c is the instruction

Importance Sampling Approach with a proposal distribution $q(x'|x,c)$

$$\begin{aligned} L(\theta) &= -\sum_{\mathbf{x},c}\sum_{\mathbf{x}'} q(\mathbf{x}'|\mathbf{x},c)\frac{p(\mathbf{x}'|\mathbf{x},c)}{q(\mathbf{x}'|\mathbf{x},c)}\log p_\theta(\mathbf{x}'|\mathbf{x},c) \\ &\approx -\mathbb{E}_{(\mathbf{x},c)\sim D}\left[\mathbb{E}_{\mathbf{x}'\sim q(\mathbf{x}'|\mathbf{x},c)}\left[\lambda(\mathbf{x}',\mathbf{x},c)\log p_\theta(\mathbf{x}'|\mathbf{x},c)\right]\right] \\ &\approx -\mathbb{E}_{(\mathbf{x},c)\sim D}\left[\mathbb{E}_{\mathbf{x}'\sim q_s(\mathbf{x}'|\mathbf{x},c)}\left[\lambda(\mathbf{x}',\mathbf{x},c)\log p_\theta(\mathbf{x}'|\mathbf{x},c)\right]\right] \end{aligned}$$

$q(x'|x,c)$ can be represented as an ensemble of lots of specialists $q_s(x'|x,c)$

Formulation

- We adopt importance sampling function $\lambda(x', x, c)$

$$
\begin{aligned}
L(\theta) &= -\sum_{\mathbf{x},c}\sum_{\mathbf{x}'} q(\mathbf{x}'|\mathbf{x},c)\frac{p(\mathbf{x}'|\mathbf{x},c)}{q(\mathbf{x}'|\mathbf{x},c)}\log p_\theta(\mathbf{x}'|\mathbf{x},c) \\
&\approx -\mathbb{E}_{(\mathbf{x},c)\sim D}\left[\mathbb{E}_{\mathbf{x}'\sim q(\mathbf{x}'|\mathbf{x},c)}\left[\lambda(\mathbf{x}',\mathbf{x},c)\log p_\theta(\mathbf{x}'|\mathbf{x},c)\right]\right] \\
&\approx -\mathbb{E}_{(\mathbf{x},c)\sim D}\left[\mathbb{E}_{\mathbf{x}'\sim q_s(\mathbf{x}'|\mathbf{x},c)}\left[\lambda(\mathbf{x}',\mathbf{x},c)\log p_\theta(\mathbf{x}'|\mathbf{x},c)\right]\right]
\end{aligned}
$$

- We use large multimodal models (GPT-4o) to approximate λ

$$
\lambda(\mathbf{x}',\mathbf{x},c) = \begin{cases} 1, & \text{if LMM}(\text{prompt},\mathbf{x}',\mathbf{x},c) \geq 9 \\ 0, & \text{otherwise} \end{cases}
$$

Building Swap Specialist

- Mask Generation: Use GroundingDINO and SAM to generate a mask for swapping candidate
- Mask Dilation: Dilate the boundary
- Image Editing: Apply BrushNet (our own version) to generate the edited image x_{edit} by replacing the source image with:

$$\mathbf{x}_{\text{edit}} = q_{\text{obj\_replace}}\left(\mathbf{x}_{\text{src}} \odot (1 - M_{\text{src\_obj}}),\ M_{\text{src\_obj}},\ C_{\text{trg\_obj}}\right)$$

- Rewriting instruction: We apply large multimodal models to rewrite the “replace A with B” with more detailed instruction.

Building Removal Specialist

- Mask Generation: Use GroundingDINO and SAM to generate a mask for swapping candidate
- Prompt GPT-4o to predict what background should fill in that swapped bounding box
- Image Editing: Apply BrushNet (our own version) to generate the edited image x_{edit} by replacing the source image with:

$$\mathbf{x}_{\text{edit}} = q_{\text{obj\_removal}}\left(\mathbf{x}_{\text{src}} \odot (1 - M_{\text{src\_obj}}),\ M_{\text{src\_obj}},\ C_{\text{trg\_background}}\right)$$

- Rewriting instruction: We apply large multimodal models to rewrite the “Remove A from image” with more detailed instruction.

Building Attribute Modify Specialist

- We first let the image generation model to generate a source image x_{src} using c_{src}
- Mask Generation: Use GroundingDINO and SAM to generate a mask for modifying candidate
- Utilizing Prompt2Prompt with mask guidance to generate edited images for attribute modification

$$M_{\text{obj}} \odot x_{\text{edited},t} + (1 - M_{obj}) \odot x_{\text{input},t}$$

- Rewriting instruction: We apply large multimodal models to rewrite the “make … ” with more detailed instruction.

Data Synthesis

- The input images are sampled from LAION-5B and OpenImageV6. We sample images with aspect ratios of 1:1, 2:3, 3:2, 3:4, 4:3, 9:16 and 16:9.

- We utilize these specialists with designated pipelines to generate 200K – 1M pairs as the candidates.

- However, we do observe plenty of artifacts and low-quality pairs.

Strict Quality Control

- Unlike previous research to use CLIP-Score, we prompt large multimodal models to assign quality score.

- We adopt the VIEScore prompt to assign “semantic consistency”, “visual quality” scores to each pairs with GPT-4o.

- To save cost, we distill GPT-4o rationale to InternVL2-8B with 200K examples and then apply that to filter.

Data Statistics

Task	Pre-Filtering Number	After-Filtering Number
Object Swap	1,500,000	150,000
Object Removal	1,000,000	100,000
Object Addition	1,000,000	100,000
Background Swap	500,000	50,000
Environment Change	500,000	50,000
Style Transfer	250,000	25,000
Object Property Modification	300,000	30,000
Total	**5,050,000**	**505,000**

EditNet

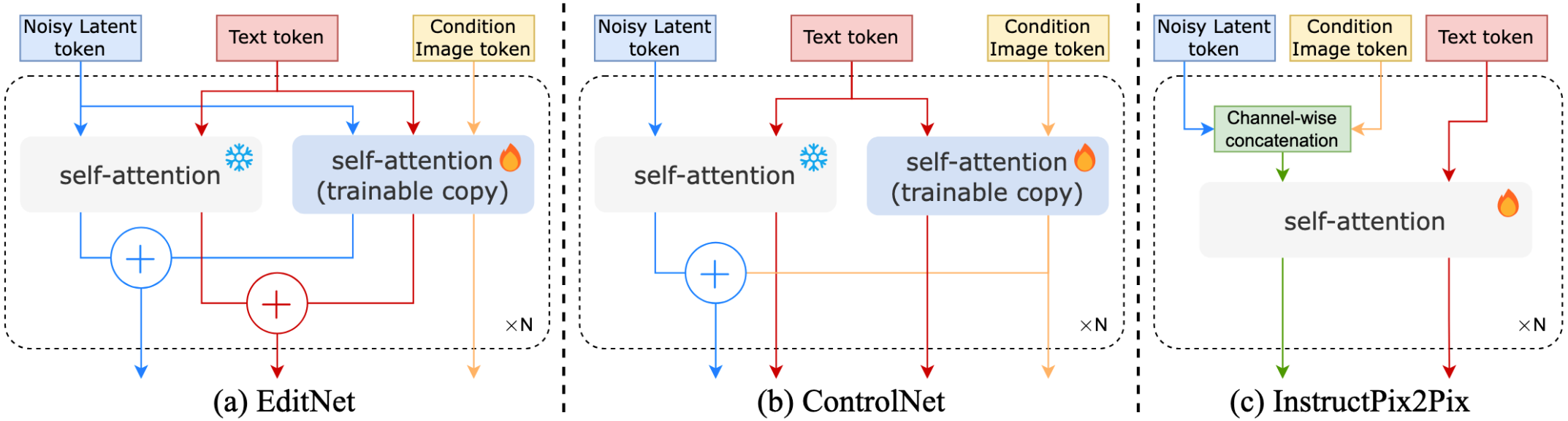


Figure 3: Architecture Comparison between **EditNet(ours)**, ControlNet and InstructPix2Pix for DiT models. Unlike ControlNet's parallel execution, EditNet allows adaptive adjustment of control signals by intermediate representations interaction between the control branch and the original branch. EditNet also updates the text representation, enabling better task understanding.

Evaluation Set

- To create a high-resolution, multi-aspect ratio, multi-task benchmark for instruction-based image editing, we manually collected 62 images from pexels and LAION-5B.

- These images cover a variety of aspect ratios, including 1:1, 2:3, 3:2, 3:4, 4:3, 9:16, and 16:9.

- We write prompts to cover all the mentioned 7 tasks.

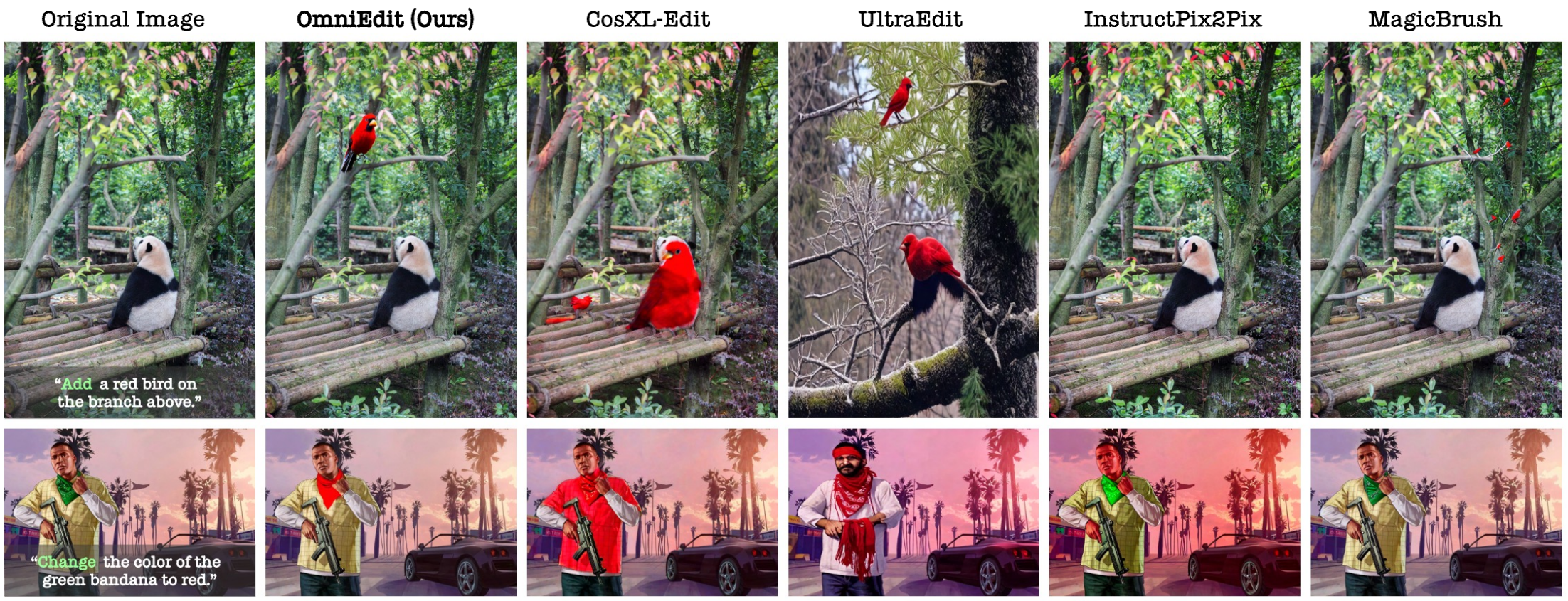
Experimental Results

Table 3: Main evaluation results on Omni-Edit-Bench. In each column, the highest score is bolded, and the second-highest is underlined.

Models	VIEScore (GPT4o)			VIEScore (Gemini)			Human Evaluation			
	$PQ_{avg}\uparrow$	$SC_{avg}\uparrow$	$O_{avg}\uparrow$	$PQ_{avg}\uparrow$	$SC_{avg}\uparrow$	$O_{avg}\uparrow$	$PQ_{avg}\uparrow$	$SC_{avg}\uparrow$	$O_{avg}\uparrow$	$Acc_{avg}\uparrow$
Inversion-based Methods										
DiffEdit	5.88	2.73	2.79	6.09	2.01	2.39	-	-	-	-
SDEdit	6.71	2.18	2.78	6.31	2.06	2.48	-	-	-	-
End-to-End Methods										
InstructPix2Pix	7.05	3.04	3.45	6.46	1.88	2.31	-	-	-	-
MagicBrush	6.11	3.53	3.60	6.36	2.27	2.61	-	-	-	-
UltraEdit(SD-3)	6.44	4.66	4.86	6.49	4.33	4.45	0.72	0.52	0.57	0.20
HQ-Edit	5.42	2.15	2.25	6.18	1.71	1.96	0.80	0.27	0.29	0.10
CosXL-Edit	<u>8.34</u>	<u>5.81</u>	<u>6.00</u>	<u>7.01</u>	<u>4.90</u>	<u>4.81</u>	<u>0.82</u>	<u>0.56</u>	<u>0.59</u>	<u>0.35</u>
HIVE	5.35	3.65	3.57	5.84	2.84	3.05	-	-	-	-
OMNI-EDIT	**8.38**	**6.66**	**6.98**	**7.06**	**5.82**	**5.78**	**0.83**	**0.71**	**0.69**	**0.55**
Δ - Best baseline	+0.04	+0.85	+0.98	+0.05	+0.92	+0.97	+0.01	+0.15	+0.10	+0.20

Visualization

Visualization



Future Work

- Improving the existing image editing models.
- Explore the possibility of using them as world models.
- Tackle difficult tasks like state change/transition tasks.
- Editing artificial images like diagram, plots or icon, etc.