The CLIP Model is Secretly an Image-to-Prompt Converter

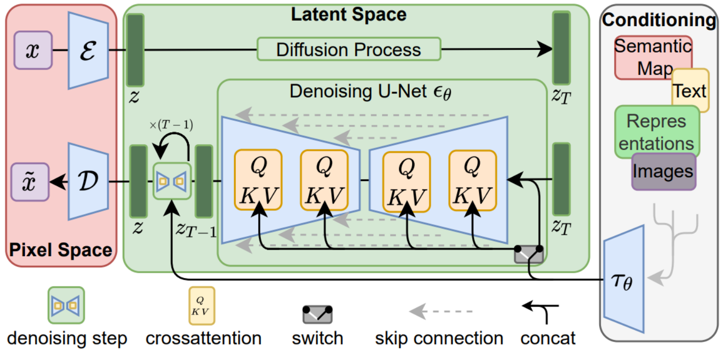
Yuxuan Ding[1], Chunna Tian[1*], Haoxuan Ding[2], and Lingqiao Liu[3*]

[1] School of Electronic Engineering, Xidian University, Xi'an, China
[2] Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an, China
[3] Australian Institute for Machine Learning, The University of Adelaide, Adelaide, Australia

yxding@stu.xidian.edu.cn, chnatian@mail.xidian.edu.cn, haoxuan.ding@mail.nwpu.edu.cn, lingqiao.liu@adelaide.edu.au

Introduction -- Background

- **Stable Diffusion**
  - As one of the most popular text-to-image generators, *Stable Diffusion* is built on the Latent Diffusion Model (LDM), which consists of a VAE compressor, a condition encoder, and a U-Net denoiser.
- **Text Encoder**
  - Stable Diffusion utilizes the *CLIP model* as its condition encoder, the text prompt is coded by the CLIP text transformer and then input into the cross-attention layers of U-Net.

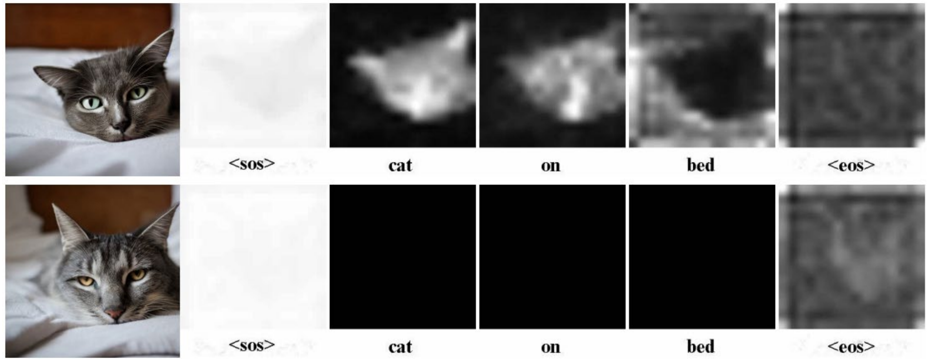


Latent Diffusion Model
(Rombach, Robin, *et al.*, 2022, *CVPR*)

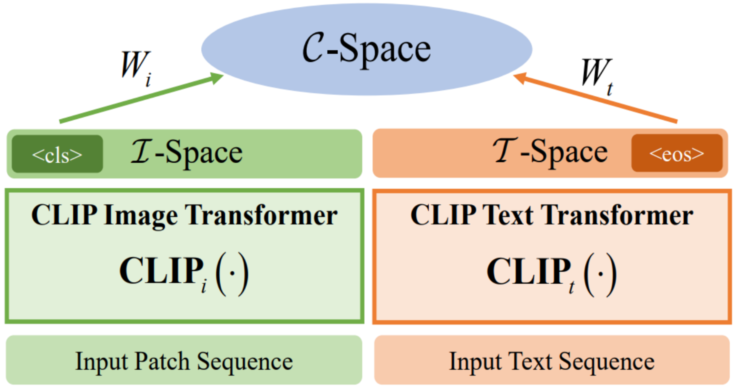
Stable Diffusion
(Rombach, Robin, *et al.*, 2022, *CVPR*)

Introduction -- Findings

- **Stable Diffusion Generation**
  - The generation results are highly related to the *end-token embedding*.
  - Masking the word-tokens in a sentence does not influence the generation results severely.
- **Embedding Conversion in CLIP**
  - Image embeddings and text embeddings are projected into a common space in the CLIP pipeline.
  - The image embedding can be converted into text embedding space with just a *pseudo-inverse matrix*.



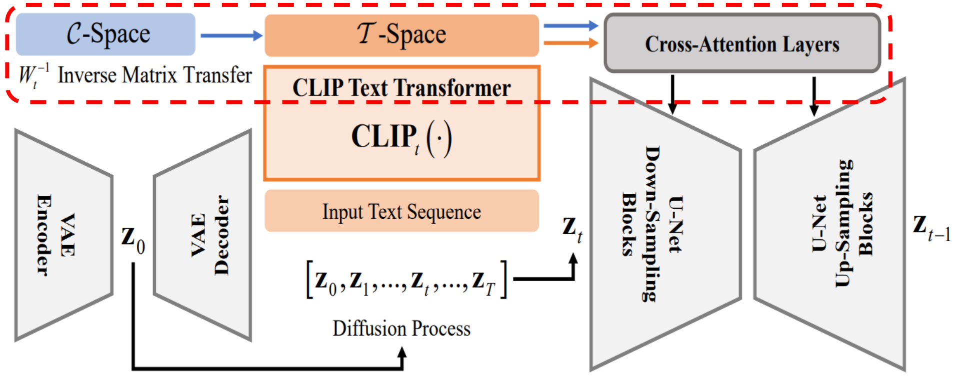
Attention Visualization of Stable Diffusion.



The architecture of CLIP model.

Introduction -- Motivation

- **Image Input for Stable Diffusion**
  - Findings: 1) image embedding can be converted to text end-token. 2) generation can just rely on end-token embedding.
  - A naïve intuition is that the *image can directly input into the Stable Diffusion*.
- **Stable Diffusion Reimagine (SD-R)**
  - Generating multiple variations from an uploaded image.
  - The algorithm is built on the Stable-unCLIP model, which *fine-tunes the Stable Diffusion to adapt to the CLIP visual embeddings*.



Inputting image to Stable Diffusion.

SD-R is an algorithm for image variation.
https://stability.ai/news/stable-diffusion-reimagine

Related Works – Image Variation & Customized Generation

- **Image Variation**
  - Generating images similar to the reference image.
  - SD-R (Rombach, Robin, *et al.*, 2022, *CVPR*) needs *expensive fine-tuning*, which requires 200,000 GPU hours.
- **Customized Generation**
  - Synthesizing *specific objects or persons*.
  - DreamBooth (Ruiz, Nataniel, *et al.*, 2023, *CVPR*), Textual Inversion (Gal, Rinon, *et al.*, 2022, *ICLR*), and Custom Diffusion (Kumari, Nupur, *et al.*, 2023, *CVPR*) are recent methods.
- **Image Editing**
  - *Attention-based methods*: Prompt-to-Prompt (Hertz, Amir, *et al.*, 2022, *ICLR*), Plug-and-Play (Tumanyan, Narek, *et al.*, 2023, *CVPR*), *etc*.
  - *Inversion-based methods*: Null-Text Inversion (Mokady, Ron, *et al.*, 2023, *CVPR*), Pix2Pix-Zero (Parmar, Gaurav, *et al.*, 2023, *SIGGRAPH*), and *etc*.
  - *InstructPix2Pix (Brooks, Tim, et al., 2023, CVPR)* creates a dataset of image editing and fine-tunes Stable Diffusion for editing.

Methodology – SD-IPC

- **Image-to-Prompt Conversion (SD-IPC)**
  - Moore-Penrose pseudo-inverse.

$$\frac{\mathbf{f}^{c}_{img}}{\|\mathbf{f}^{c}_{img}\|} \approx \frac{\mathbf{f}^{c}_{txt}}{\|\mathbf{f}^{c}_{txt}\|}, \text{ with } \mathbf{f}^{c}_{txt} = W_t \mathbf{f}^{t,\langle eos\rangle}_{txt},$$

$$\mathbf{f}^{t,\langle eos\rangle}_{txt} \approx \frac{\|\mathbf{f}^{c}_{txt}\|}{\|\mathbf{f}^{c}_{img}\|} W_t^{+} \mathbf{f}^{c}_{img} := \mathbf{f}^{cnvrt}_{txt}, \text{where } W_t^{+} = \left(W_t^{\top} W_t\right)^{-1} W_t^{\top}.$$

  - Constructing *converted image prompt*.

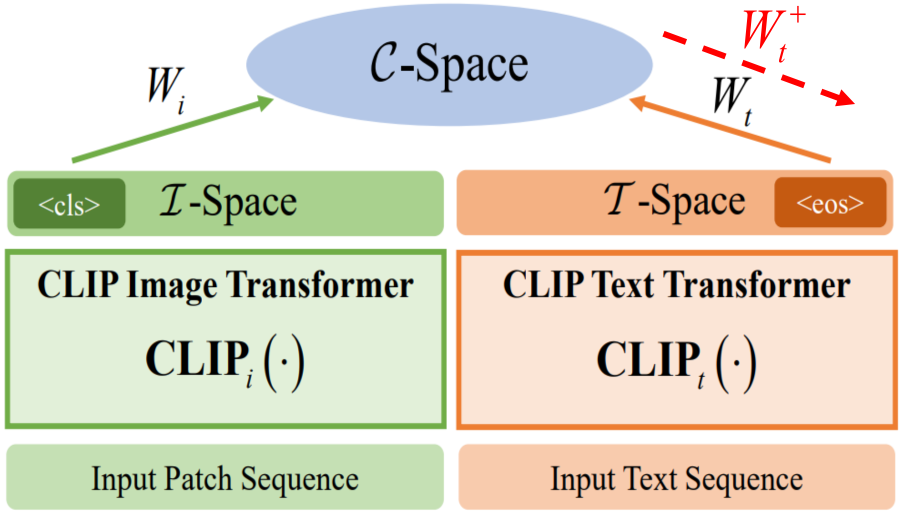
$$\mathbf{f}_{txt} := \left[\mathbf{f}^{0,\langle sos\rangle}_{txt}, \mathbf{f}^{1,w_0}_{txt}, \ldots, \mathbf{f}^{t,\langle eos\rangle}_{txt}, \ldots, \mathbf{f}^{76,\langle eos\rangle}_{txt}\right],$$

$$\mathbf{f}'_{txt} := \left[\mathbf{f}^{0,\langle sos\rangle}_{txt}, \varnothing, \ldots, \mathbf{f}^{t,\langle eos\rangle}_{txt}, \ldots, \mathbf{f}^{76,\langle eos\rangle}_{txt}\right],$$

$$\mathbf{f}''_{txt} := \left[\mathbf{f}^{0,\langle sos\rangle}_{txt}, \mathbf{f}^{1,\langle eos\rangle}_{txt}, \ldots, \mathbf{f}^{76,\langle eos\rangle}_{txt}\right],$$

$$\tilde{\mathbf{f}}_{txt} := \left[\mathbf{f}^{0,\langle sos\rangle}_{txt}, \mathbf{f}^{1,cnvrt}_{txt}, \ldots, \mathbf{f}^{76,cnvrt}_{txt}\right],$$

$$\tilde{\mathbf{f}}^{edit}_{txt} = \left[\mathbf{f}^{0,\langle sos\rangle}_{txt}, \mathbf{f}^{1,w_0}_{txt}, \ldots, \mathbf{f}^{t,comb}_{txt}, \ldots, \mathbf{f}^{76,comb}_{txt}\right].$$



Converting image embedding to text space by a pseudo-inverse matrix.

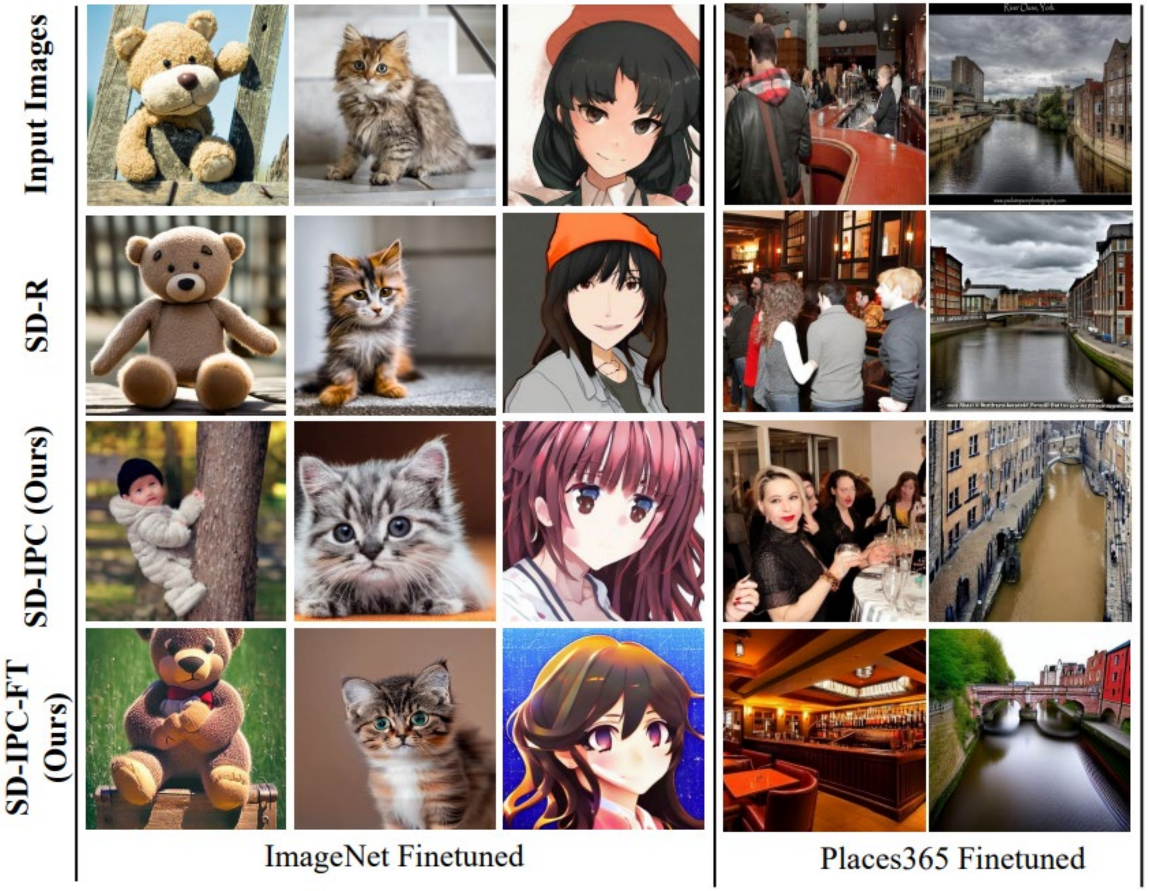
Emb. Space	Acc@1	Acc@5	TR@1	TR@5	IR@1	IR@5
$\mathcal{C}$-space	71.41	91.78	74.58	92.98	55.54	82.39
$\mathcal{T}$-space	69.48	90.62	71.62	92.06	54.82	82.20

No performance loss after conversion to text embedding space.

Methodology – SD-IPC-FT

- **Fine-tuning with Image-to-Prompt Conversion**
  - Approximation error in SD-IPC.
  - It is crucial to have a method that allows *control of the content* we wish to preserve, *e.g.* objects, scenes, styles, or identities.
  - CLIP *prompt tuning* & U-Net *cross-attention layers finetuning*.

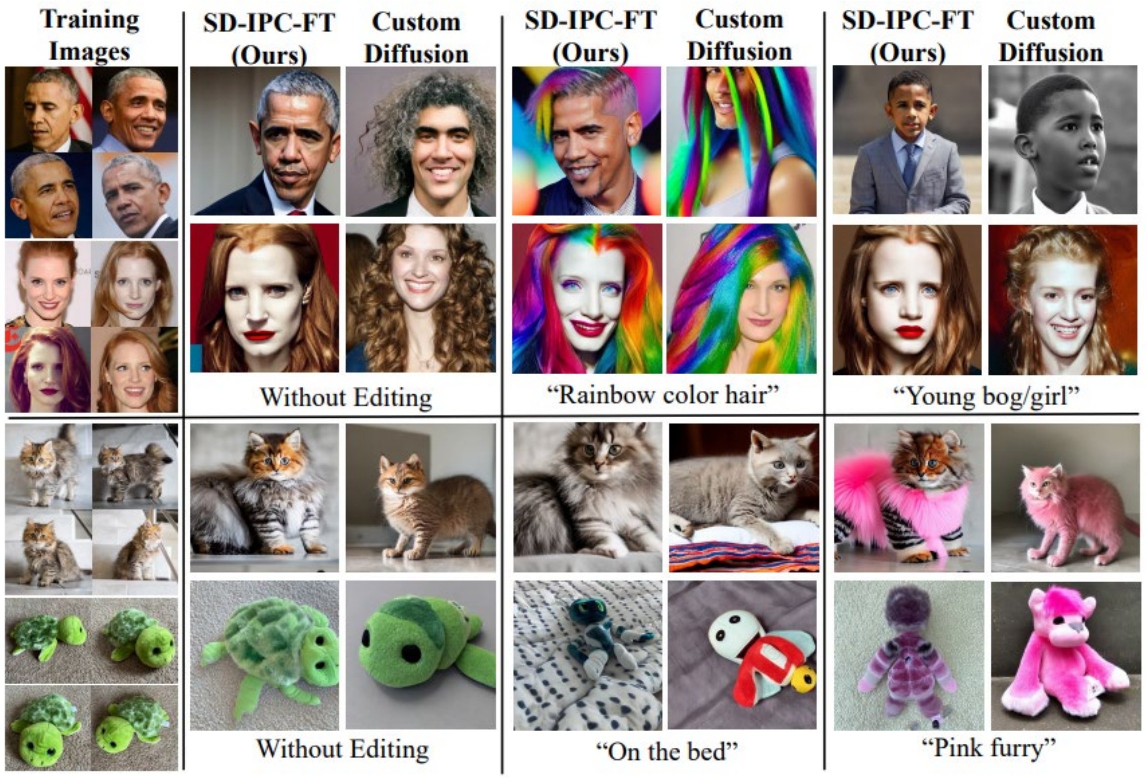
$$\underbrace{\mathbb{E}_{\epsilon,\mathbf{z},x_{\text{ref}},t}\left[\left\|\epsilon-\epsilon_\theta\left(\mathbf{z}_t,c_{img}(x_{\text{ref}}),t\right)\right\|^2\right]}_{\text{Finetuning with SD-IPC}}+\underbrace{\mathbb{E}_{\epsilon,\mathbf{z},p_{\text{txt}},t}\left[\left\|\epsilon-\epsilon_\theta\left(\mathbf{z}_t,c_{\text{txt}}(p_{\text{txt}}),t\right)\right\|^2\right]}_{\text{Regularization term with text}}.$$



SD-IPC-FT can alleviate the error and preserve specific content.

Methodology – SD-IPC-CT

- **Fast Update for Customized Generation**
  - Achieving customized generation by *online update with SD-IPC*.
  - Benefiting from the good initialization of SD-IPC, our method can generate customized images with *much fewer updates* (30 iterations *vs.* 250 iterations).
  - Quantitative analysis with the benchmark in DreamBooth.



SD-IPC-CT can get better performance with few updates.

Experimental Results – Image Variation

- Our SD-IPC holds the *same performance with text-to-image* Stable Diffusion.

Methods	FID	CLIP-Score
SD w/ Text	23.65	70.15
SD-IPC (Ours)	24.78	73.57

Our SD-IPC is close to the original Stable Diffusion on FID and CLIP-Score.

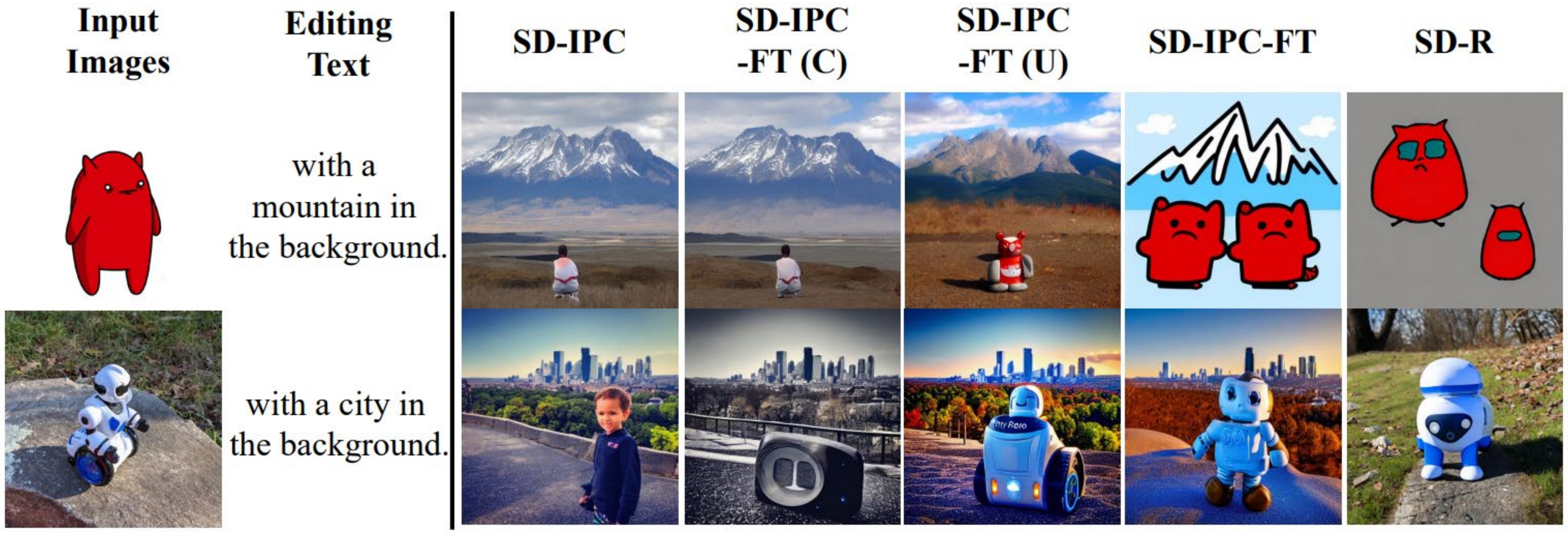
Image variation examples.

Experimental Results – Text-edited Image Variation

- Our SD-IPC-FT *gets superior editing performance* compared to SD-R.
- SD-R *fails in image editing*, the results only show the variation but without editing. Even SD-IPC slightly outperforms SD-R.

Method	CLIP-T
SD-IPC	26.84
SD-IPC-FT	**28.69**
SD-R	26.01

Superior editing performance of SD-IPC-FT.



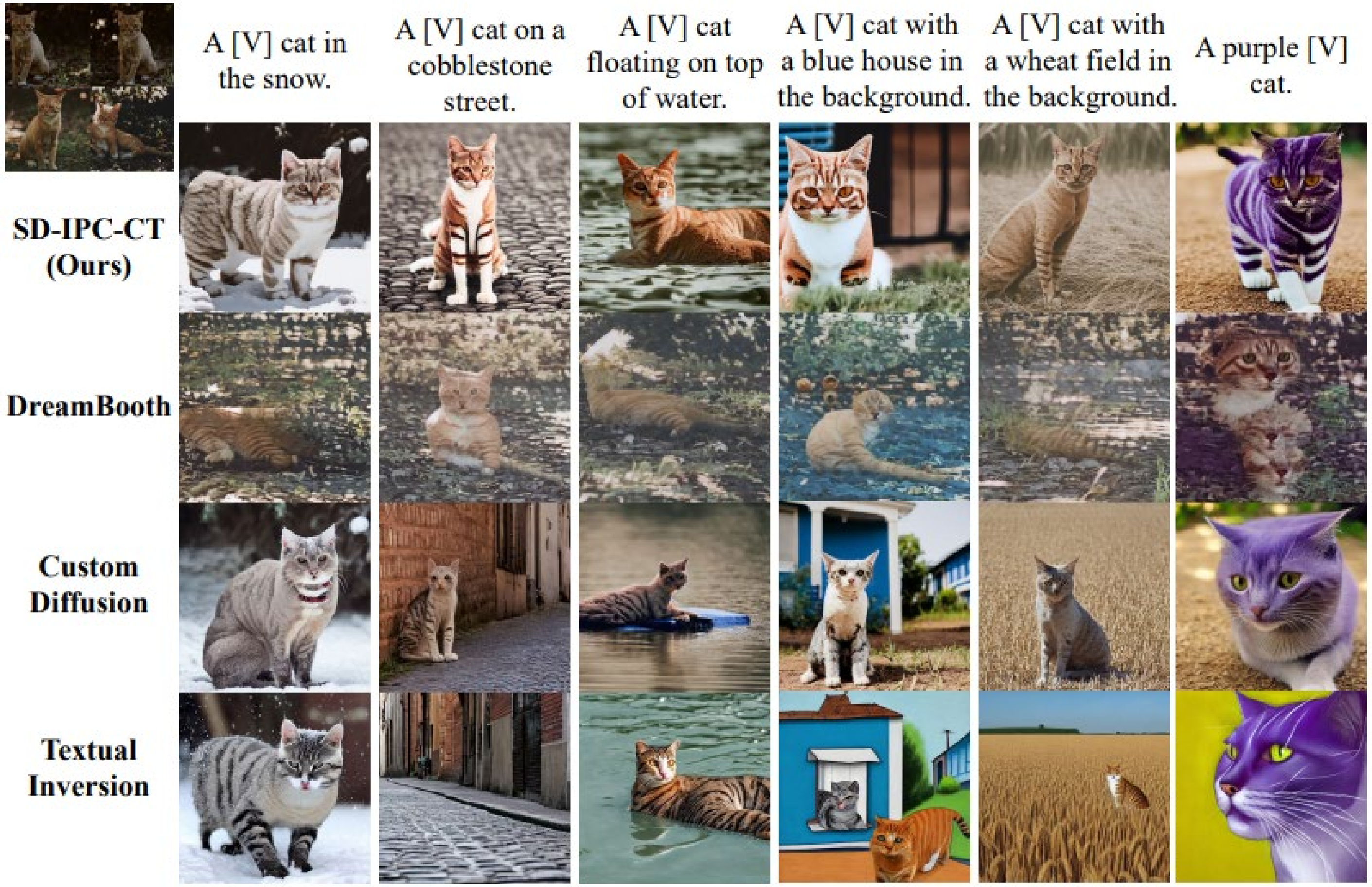
Text editing performance. SD-R is prone to ignore the text condition.

Experimental Results – Customized Generation

- DreamBooth is limited on editing, Textual Inversion and Custom Diffusion are challenging on subject details preservation.
- Our SD-IPC-CT strikes *a balance between subject fidelity and editing performance*.

Methods	DNIO	CLIP-I	CLIP-T
DreamBooth	**60.11**	**77.78**	25.81
Textual Inversion	25.11	62.44	29.53
Custom Diffusion	39.67	68.37	**30.90**
SD-IPC-CT (Ours)	50.25	74.59	28.14

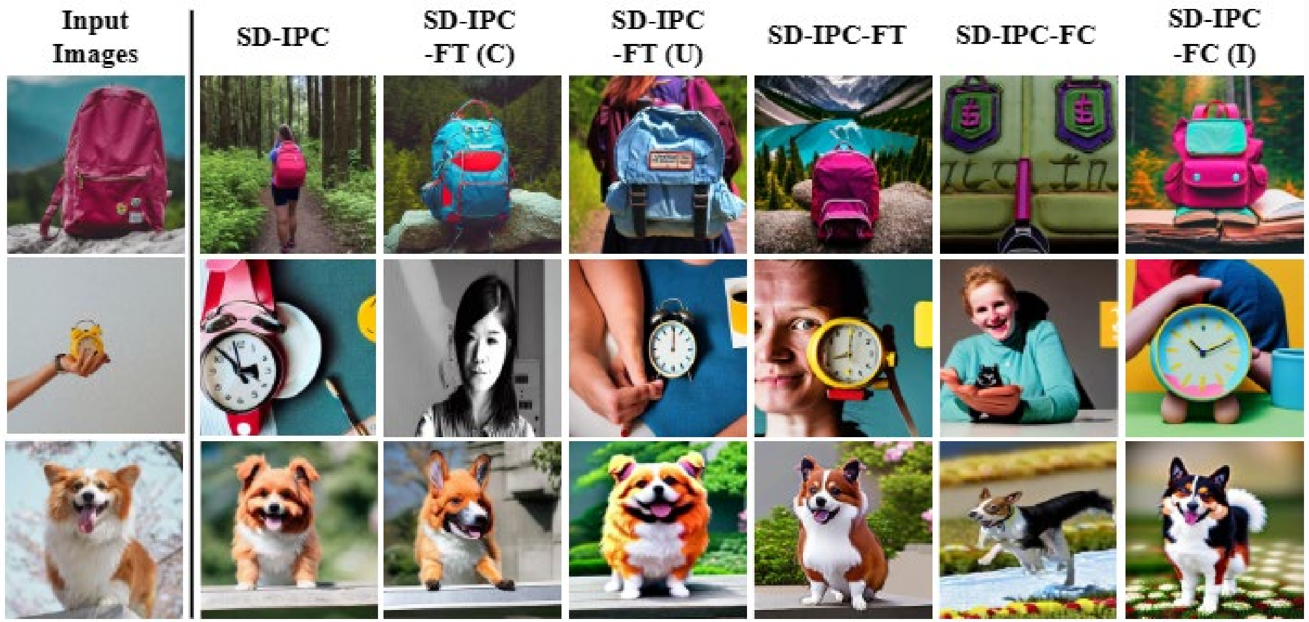
SD-IPC-CT shows both good identity preservation and good editing performance.



Example of DreamBooth benchmark. DreamBooth overfits the input images, while Textual Inversion and Custom Diffusion can not preserve the subject.

Experimental Results – Ablation Study

- CLIP prompt tuning & U-Net cross-attention layers finetuning *both contribute to extract correct information*.
- Replacing our pseudo-inverse matrix with a *FC layer leads to overfitting*.



Visualization of image variation with different fine-tuning settings.

Method	DNIO	CLIP-I	CLIP-T
SD-IPC	44.60	77.44	25.47
SD-IPC-FT (C)	49.11	76.51	25.82
SD-IPC-FT (U)	48.53	79.06	**26.17**
SD-IPC-FT	**52.03**	**79.59**	25.90

Quantitative results of image variation.

Method	DNIO	CLIP-I	CLIP-T
SD-IPC	31.09	68.66	26.84
SD-IPC-FT (C)	29.10	67.03	27.99
SD-IPC-FT (U)	35.21	69.99	28.56
SD-IPC-FT	**40.28**	**71.97**	**28.69**

Quantitative results of text-edited image variation.

XIDIAN UNIVERSITY

THE UNIVERSITY of ADELAIDE

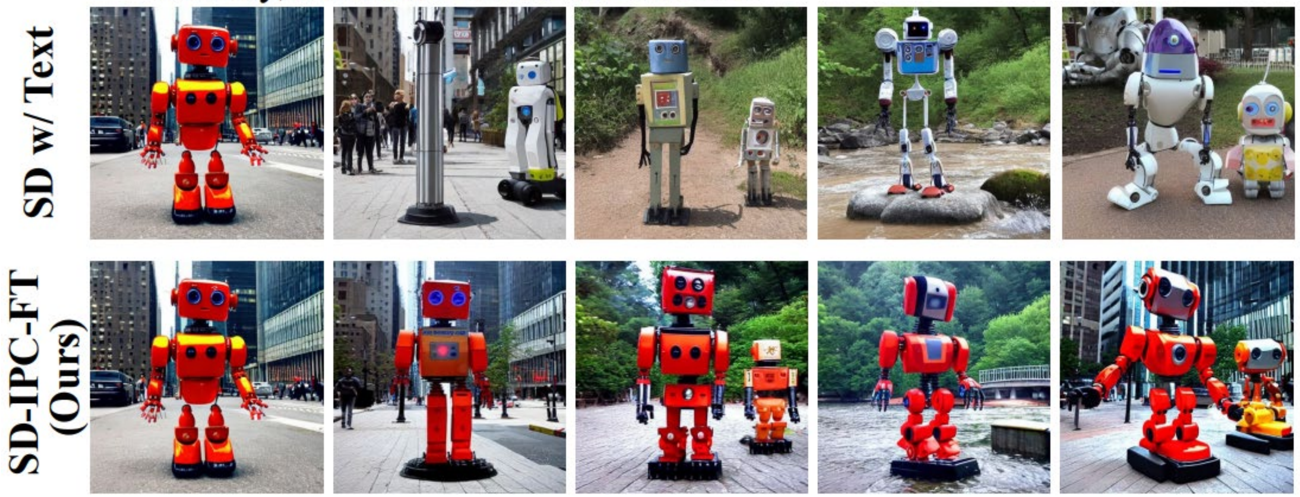
Australian Institute for Machine Learning

NEURAL INFORMATION PROCESSING SYSTEMS

Future Directions

- Better editing performance.
- Multi-concept generation.
- Story generation with consistency.
- Feature explainability of Stable Diffusion & CLIP.
- Image-to-prompt pathway in CLIP-based or LDM-based models.

- A little robot named Rusty went on an adventure to a big city.
- The robot found no other robot in the city but only people.
- The robot went to the village to find other robots.
- Then the robot went to the river.
- Finally, the robot found his friends.



Story generation example.

References

[1] Rombach, Robin, *et al.*, “High-resolution image synthesis with latent diffusion models.” in *CVPR*, 2022.

[2] Radford, Alec, *et al.*, “Learning transferable visual models from natural language supervision.” in *ICML*, 2021.

[3] Ruiz, Nataniel, *et al.*, “Dreambooth: Fine-tuning text-to-image diffusion models for subject-driven generation.” in *CVPR*, 2023.

[4] Gal, Rinon, *et al.*, “An image is worth one word: Personalizing text-to-image generation using textual inversion.” in *ICLR*, 2022.

[5] Kumari, Nupur, *et al.*, “Multi-concept customization of text-to-image diffusion." in *CVPR*, 2023.

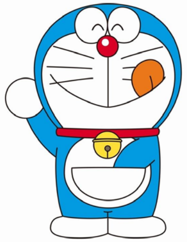
[6] Hertz, Amir, *et al.*, “Prompt-to-prompt image editing with cross-attention control." in *ICLR*, 2022.

[7] Tumanyan, Narek, *et al.*, "Plug-and-play diffusion features for text-driven image-to-image translation." in *CVPR*, 2023.

[8] Mokady, Ron, *et al.*, "Null-text inversion for editing real images using guided diffusion models." in *CVPR*, 2023.

[9] Parmar, Gaurav, *et al.*, "Zero-shot image-to-image translation." in *SIGGRAPH*, 2023.

[10] Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. "Instructpix2pix: Learning to follow image editing instructions." in *CVPR*, 2023.



Thank you!