InstructPix2Pix: Learning to Follow Image Editing Instructions

Jun Hyung Lee

Korea University Department of Computer Science and Engineering

IE LAB Interlligence Engineering

Before we get into the paper... Let’s check out some recent trends

- Diffusion based Audio Generation is popping.
- Text to Image Generation -> Text to Audio Generation

Audio AI Timeline

Here we will keep track of the latest AI models for audio generation, starting in 2023!

2023

Date	Release	Paper	Code	Trained Model
30.01	SingSong: Generating musical accompaniments from singing	arXiv	-	-
30.01	AudioLDM: Text-to-Audio Generation with Latent Diffusion Models	arXiv	*	-
30.01	Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion	arXiv	GitHub	-
29.01	Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models	PDF	-	-
28.01	Noise2Music	-	-	-
27.01	RAVE2	arXiv	GitHub	-
26.01	MusicLM: Generating Music From Text	arXiv	-	-
18.01	Msanii: High Fidelity Music Synthesis on a Shoestring Budget	arXiv	GitHub	Hugging Face Colab
16.01	ArchiSound: Audio Generation with Diffusion	PDF	GitHub	-
05.01	VALL-E: Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers	arXiv	-	-

References

- **Denoising Diffusion Probabilistic Models**, Jonathan Ho (2020)
- **SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations**, Chenlin Meng (2021)
- **Diffusion Models Beat GANs on Image Synthesis**, Prafulla Dhariwall (2021)
- **High-Resolution Image Synthesis with Latent Diffusion Models**, Robin Rombach (2022)
- **Prompt-to-Prompt Image Editing with Cross Attention Control**, Amir Hertz (2022)
- https://www.youtube.com/watch?v=7y1z-eGuV2Q&t=1157s&ab_channel=DoyupLee

Contribution

① Given an input image and a written instruction that tells the model what to do, the model follows these instructions to edit the image.
② - Editing performs in the forward pass, so it does not require per-example fine-tuning, inversion, user-drawn mask, and additional images.
- Therefore, it can edit images quickly (seconds)
③ - This model does not need a full description of any image. It only requires a single image and an instruction on how to edit the image. -
- Enabling editing from just the instructions can give users the benefit of telling the model what to do in natural written text.
④ Capable of replacing objects, changing the style of an image, changing the setting, the artistic medium, among others.

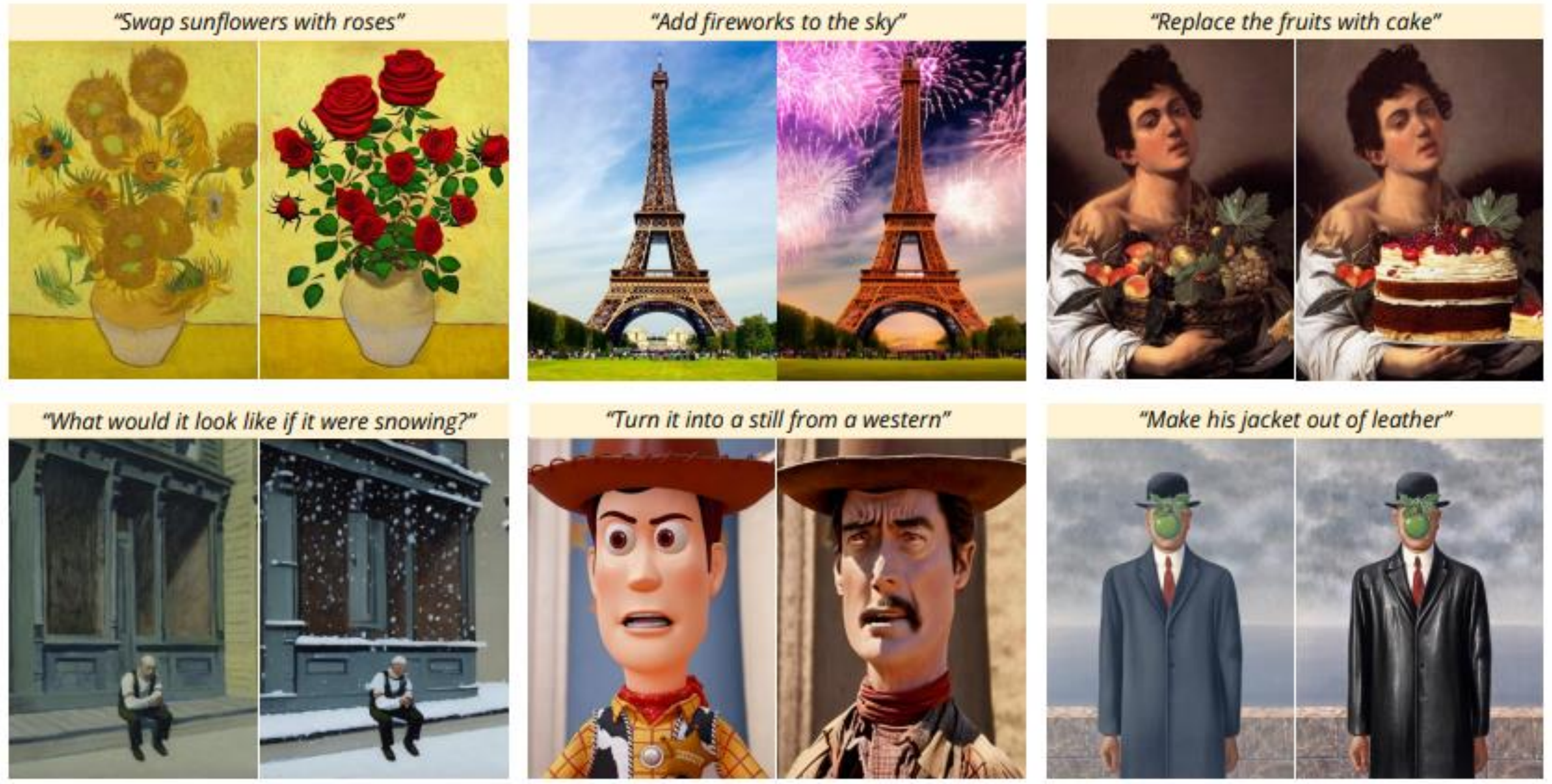


Figure 1. Given **an image** and **an instruction** for how to edit that image, our model performs the appropriate edit. Our model does not require full descriptions for the input or output image, and edits images in the forward pass without per-example inversion or fine-tuning.

Overview

- InstructPix2Pix follows the recent trend of combining large pre-trained models to solve multimodal tasks that no one model can perform alone. (In this work, GPT-3, Stable Diffusion + Prompt2Prompt) -> <u>However, the difference is that it uses to generate paired multi-model training data.</u>
- InstructPix2Pix is trained in the supervised manner
- Method:
  (1) Generate a paired training dataset of text editing instructions and images before/after the edit.
  (2) Train an image editing diffusion model on this generated dataset.
  -> Despite being trained entirely on synthetic examples editing instructions, the model is able to generalize to editing real images using arbitrary human-written instructions.

(step 1) Original caption ⟶ human instruction & edited caption

(step 2) Original caption & Edited caption ⟶ Original image & edited image

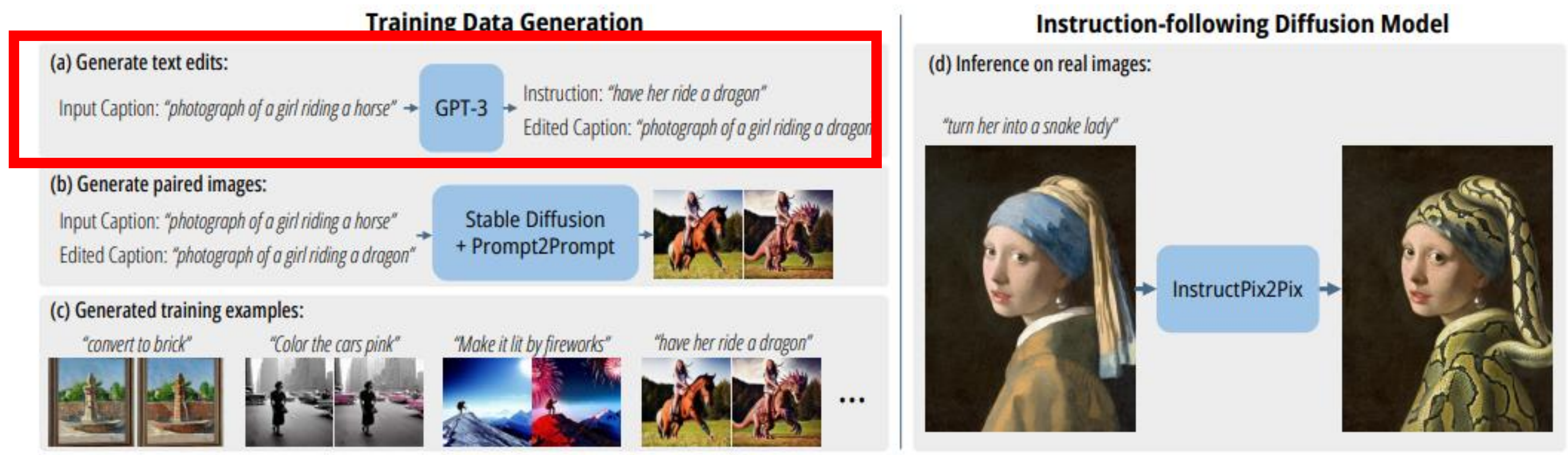
InstructPix2Pix

Generating Instructions and Paired Captions

- Fine-tuning GPT-3 to generate instructions and edited captions
- We call it editing "Triplets"

  (1) Input Caption: "*photograph of a girl riding a horse*"
  (2) Edit instruction: "*have her ride a dragon*"
  (3) Output caption: "*photograph of a girl riding a dragon*"

- Around 700 captions of (LAION-Aesthetics V2 6.5+ dataset) human-written edit instruction and output caption is required.
- Training : GPT-3 Davinci model is fine-tuned for a single epoch using the default training parameters
- Inference: For a randomly sampled caption, the finetuned GPT-3 generates instructions and captions.



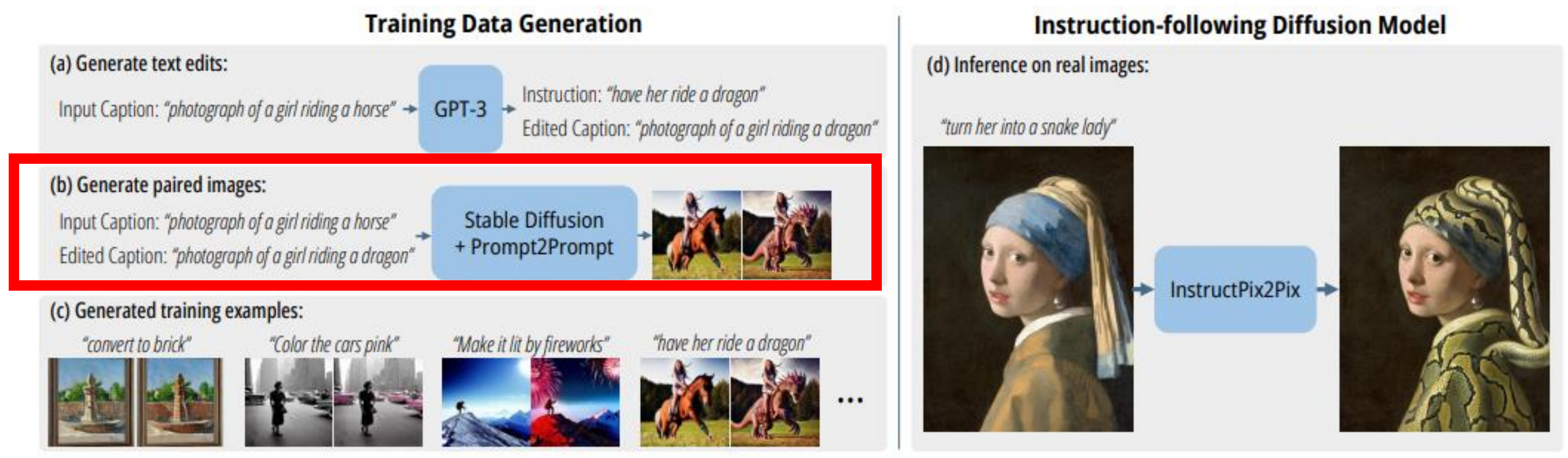
InstructPix2Pix

	Input LAION caption	Edit instruction	Edited caption
Human-written (700 edits)	*Yefim Volkov, Misty Morning*	*make it afternoon*	*Yefim Volkov, Misty Afternoon*
	girl with horse at sunset	*change the background to a city*	*girl with horse at sunset in front of city*
	painting-of-forest-and-pond	*Without the water.*	*painting-of-forest*
	...	...	...
GPT-3 generated (>450,000 edits)	*Alex Hill, Original oil painting on canvas, Moonlight Bay*	*in the style of a coloring book*	*Alex Hill, Original coloring book illustration, Moonlight Bay*
	The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it	*Add a giant red dragon*	*The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it with a giant red dragon flying overhead*
	Kate Hudson arriving at the Golden Globes 2015	*make her look like a zombie*	*Zombie Kate Hudson arriving at the Golden Globes 2015*
	...	...	...

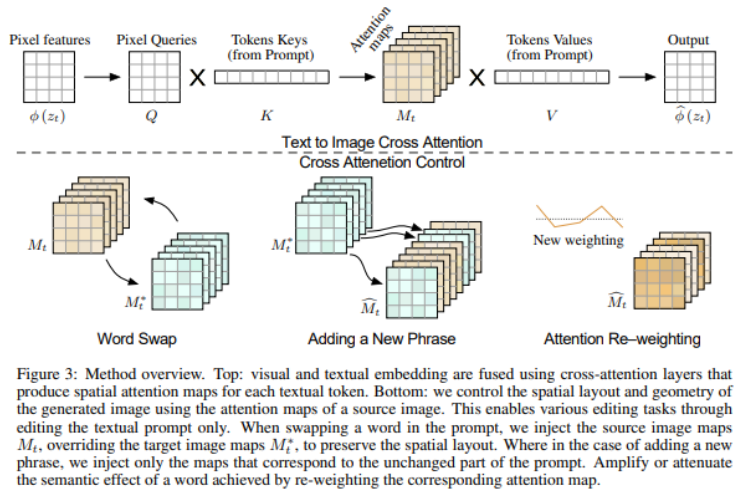
Table 1. We label a small text dataset, finetune GPT-3, and use that finetuned model to generate a large dataset of text triplets. As the input caption for both the labeled and generated examples, we use real image captions from LAION. Highlighted text is generated by GPT-3.

Generating Paired Images from Paired Captions

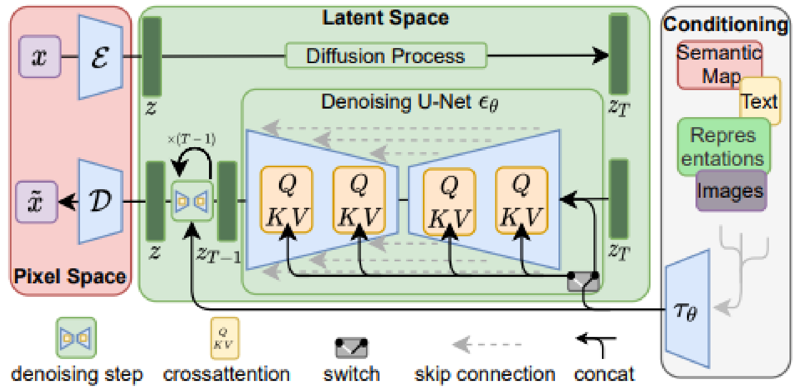
- Prompt2Prompt is a technique for image editing without masking & fine-tuning.
- Attention injection steps p is uniformly and randomly sampled from [0.1,0.9].
- 100 sample pairs are first generated per a caption pair, and filtered with
  - Filtering Criteria: Image-Image (>0.75), Image-Caption(>0.2), Directional CLIP similarity (>0.2)



InstructPix2Pix



Prompt2Prompt



Stable Diffusion

- Comparison between (a) not using Prompt to Prompt / (b) using Prompt to Prompt

(1) Input Caption: "*photograph of a girl riding a <u>horse</u>*"
(2) Edit instruction: "*have her ride a <u>dragon</u>*"
(3) Output caption: "*photograph of a girl riding a <u>dragon"</u>*

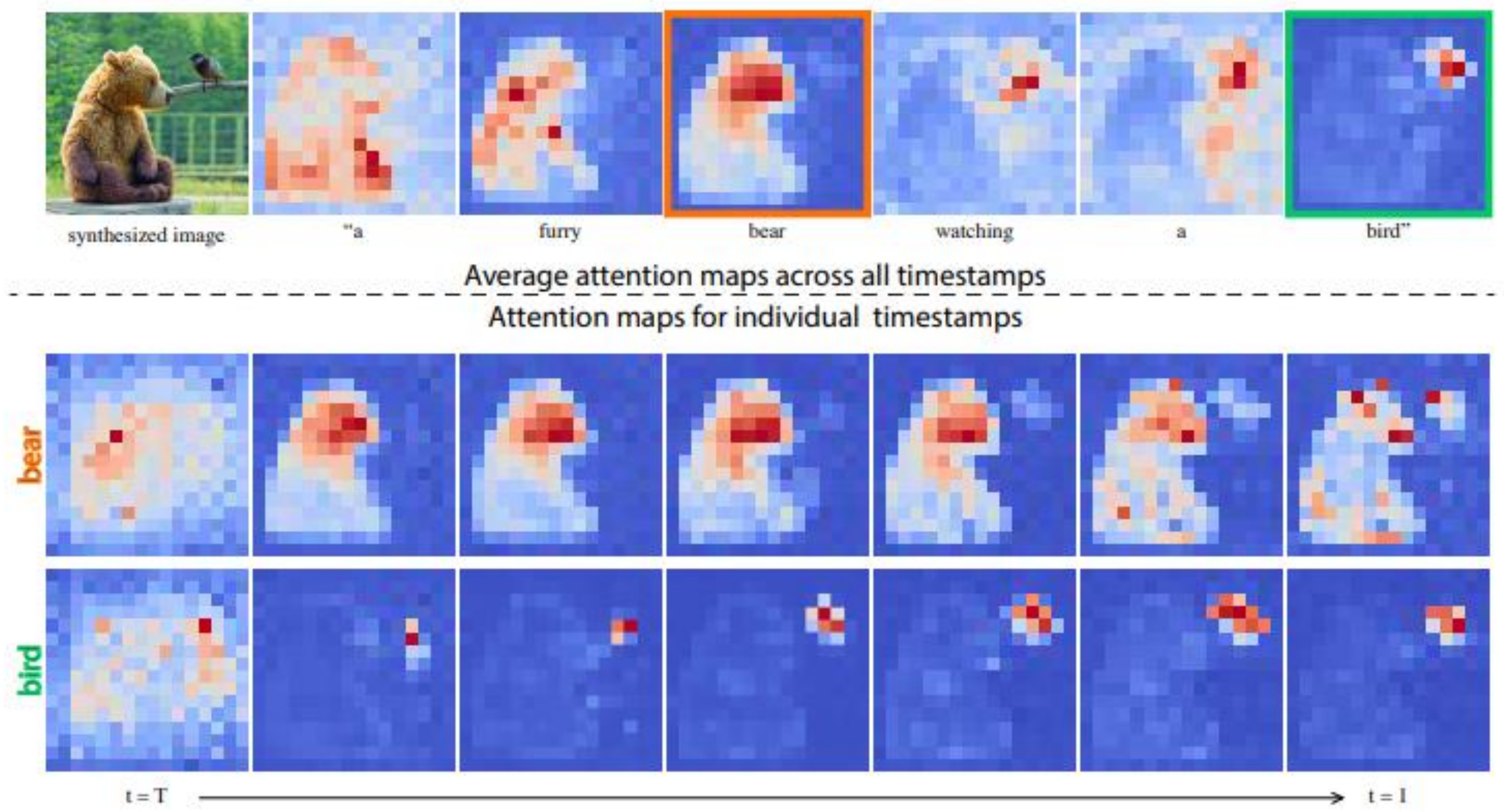
(a) Without Prompt-to-Prompt. (b) With Prompt-to-Prompt.

Figure 3. Pair of images generated using StableDiffusion [52] with and without Prompt-to-Prompt [17]. For both, the corresponding captions are *"photograph of a girl riding a horse"* and *"photograph of a girl riding a dragon"*.

Prompt2Prompt

- Prompt2Prompt is a technique for image editing without masking & fine-tuning.
- A cross attention (Q: pixels, K; texts) determines the structure of generated images during backward diffusion steps.
- Cross attention maps can be manually replaced or revised(Word Swap, Adding a New Phrase, Attention Re-weighting) during inference for image generation.

Prompt2Prompt



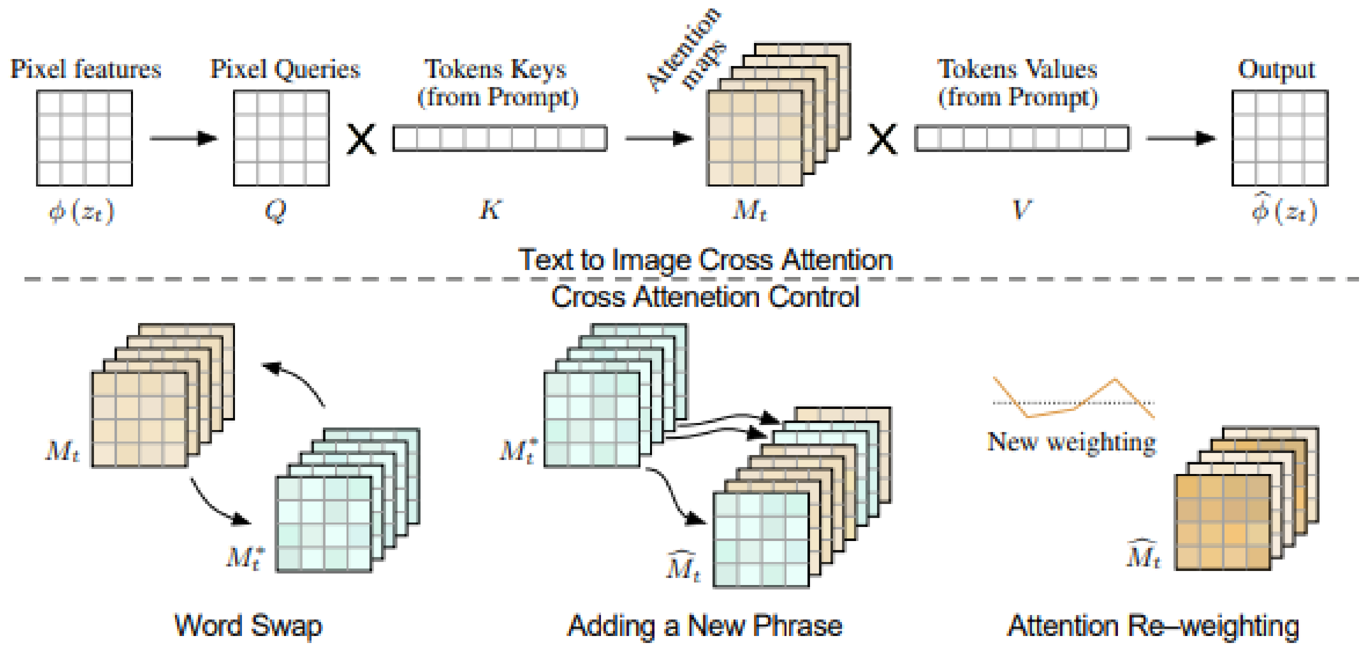
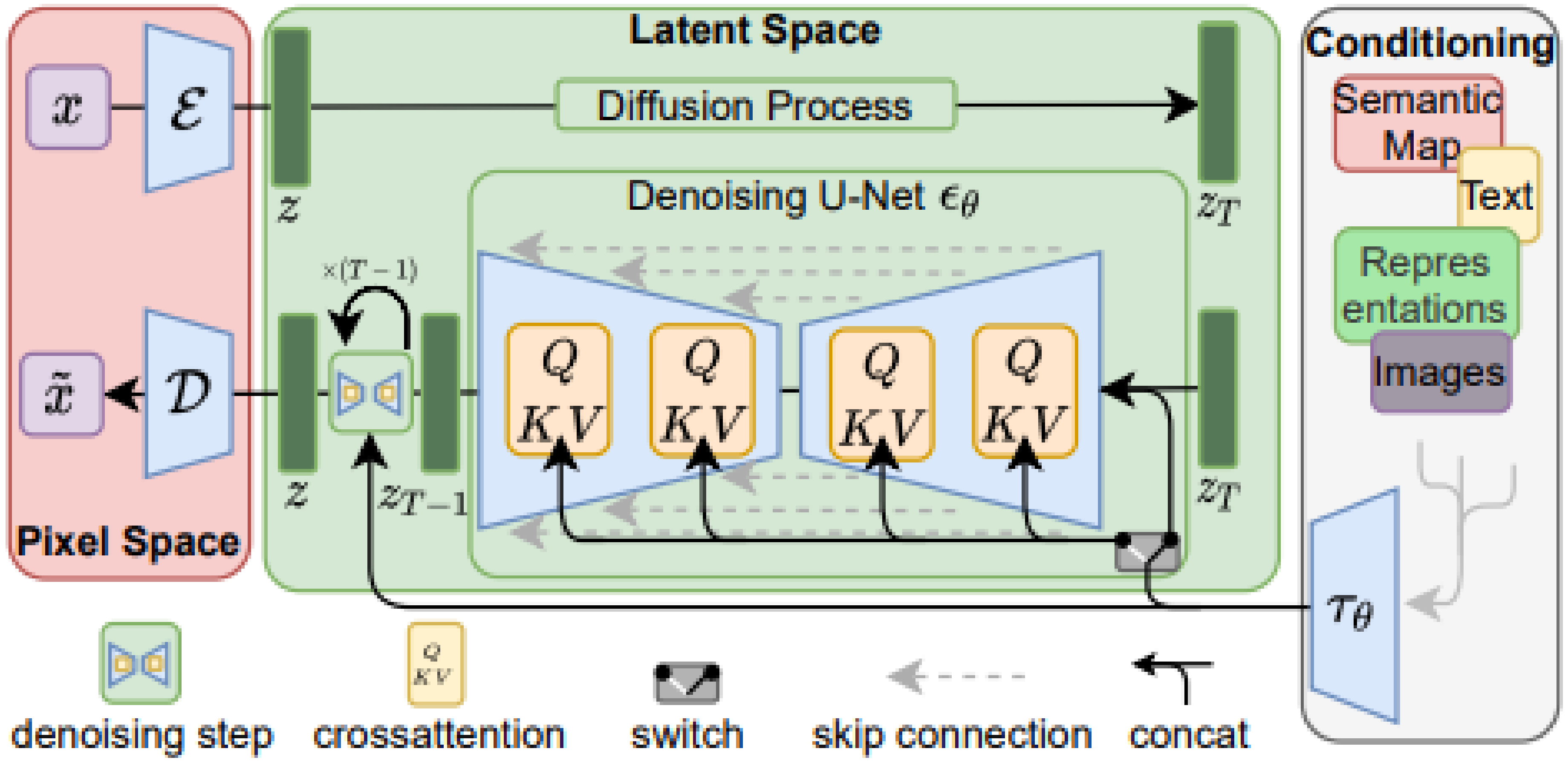


Figure 3: Method overview. Top: visual and textual embedding are fused using cross-attention layers that produce spatial attention maps for each textual token. Bottom: we control the spatial layout and geometry of the generated image using the attention maps of a source image. This enables various editing tasks through editing the textual prompt only. When swapping a word in the prompt, we inject the source image maps M_t, overriding the target image maps M_t^*, to preserve the spatial layout. Where in the case of adding a new phrase, we inject only the maps that correspond to the unchanged part of the prompt. Amplify or attenuate the semantic effect of a word achieved by re-weighting the corresponding attention map.

Training

- InsturctPix2Pix fully exploit the weights of Stable Diffusion (pretrained checkpoints), while adding the input channel for image conditions and set zero values on new added weights.
- InstructPix2Pix is trained for 10,000 steps, takes about 24hrs with A100x8 GPUs.
  - (Inference takes 9 seconds on an A100 GPU for single image editing).



Stable Diffusion

$$\tilde{e_\theta}(z_t, c) = e_\theta(z_t, \varnothing) + s \cdot (e_\theta(z_t, c) - e_\theta(z_t, \varnothing))$$

$$\begin{aligned}\tilde{e_\theta}(z_t, c_I, c_T) = \; & e_\theta(z_t, \varnothing, \varnothing) \\ & + s_I \cdot (e_\theta(z_t, c_I, \varnothing) - e_\theta(z_t, \varnothing, \varnothing)) \\ & + s_T \cdot (e_\theta(z_t, c_I, c_T) - e_\theta(z_t, c_I, \varnothing))\end{aligned} \quad (3)$$

Classifier-free Guidance

Limitations

- New edits : human-written instructions is required to fine-tune GPT-3
- Struggles with counting numbers of objects and with spatial reasoning

"Zoom into the image" *"Move it to Mars"* *"Color the tie blue"* *"Have the people swap places"*

Figure 13. Failure cases. Left to right: our model is not capable of performing viewpoint changes, can make undesired excessive changes to the image, can sometimes fail to isolate the specified object, and has difficulty reorganizing or swapping objects with each other.