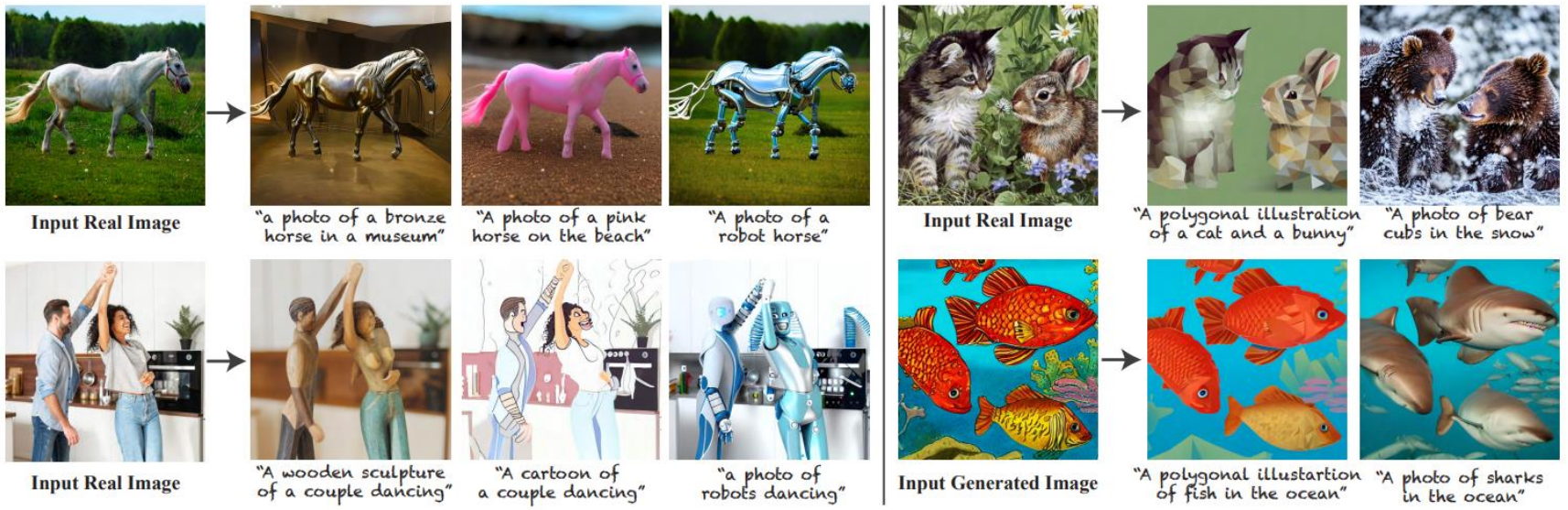
Medical Imaging Intelligent Reality Lab

Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation

Narek Tumanyan, Michal Geyer, Shai Bagon, Tali Dekel

Weizmann Institute of Science

Accepted in CVPR 2023



Content

1. Introduction

- With the rise of text-to-image foundation models, it seems that we can translate our imagination into high-quality images through text.
- Their power and expressivity come at the expense of user controllability, which is largely restricted to guiding the generation solely through an input text.
- We focus on attaining control over the generated structure and semantic layout of the scene.
- The goal is to take text-to-image generation to the realm of text-guided Image-to-Image translation.
- Not require any training or fine-tuning, but leverage a pre-trained and fixed text-to-image diffusion model (Stable Diffusion).

1. Introduction

- Fundamental question: How is structure information internally encoded in such a model?
- We dive into the intermediate spatial features that are formed during the generation process.
- We devise a new framework that enables fine-grained control over the generated structure by applying simple manipulations.
- Spatial features and self-attentions are extracted from the guidance image, and are directly injected into the text-guided generation process of the target image.

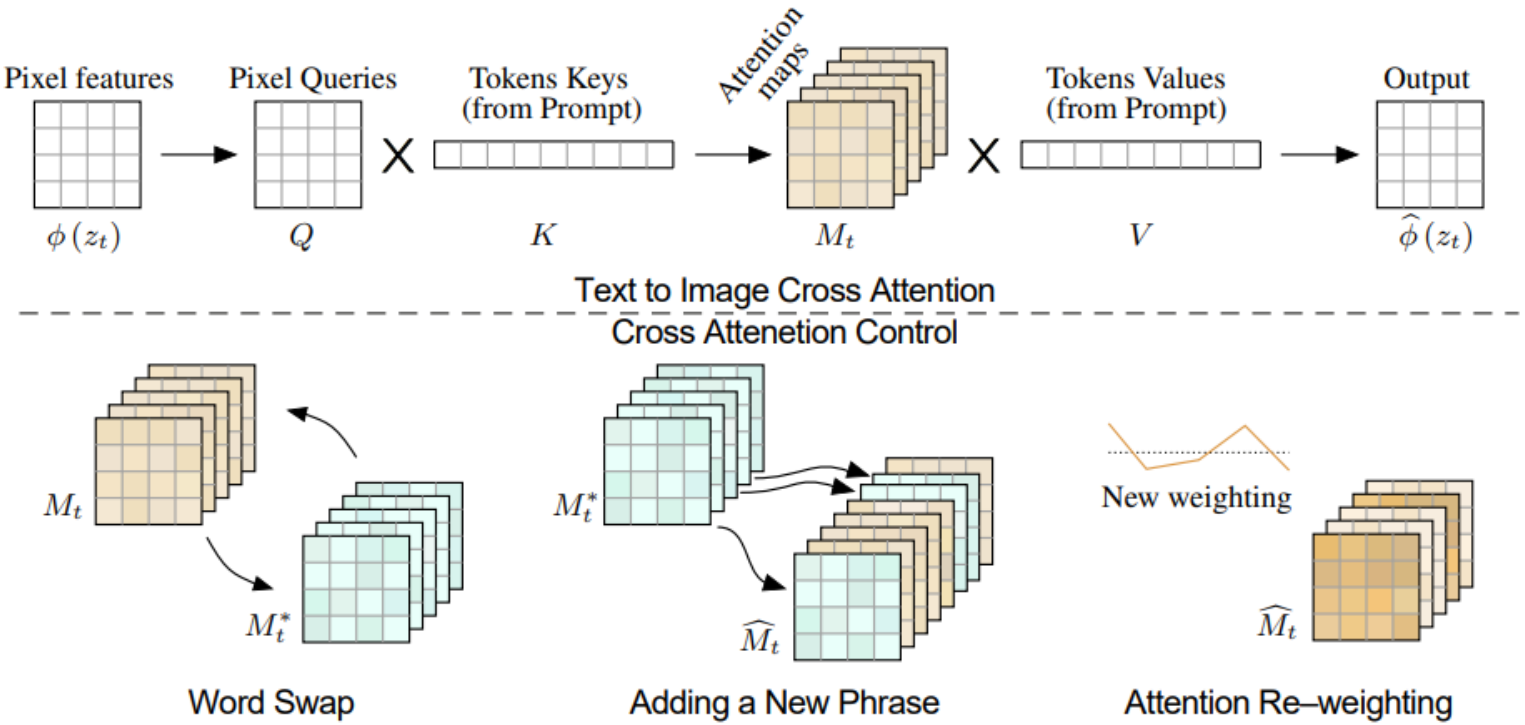
2. Related work

- **Image-to-image translation.** Estimating a mapping of an image from a source domain to a target domain, while preserving the domain-invariant characteristics of the input image, e.g., objects' structure or scene layout.

- **Text-guided image manipulation.** Various methods have proposed to combine CLIP with a pre-trained unconditional image generator.
  DiffusionCLIP uses CLIP to fine-tune a diffusion model.
  Text2LIVE trains a generator on a single image-text pair, without additional training data.
  There is still a gap between the generative prior that is learned solely from visual data, and the rich CLIP text-image guiding signal.
  Recently, text-to-image generative models have closed this gap by directly conditioning image generation on text during training.
  Nevertheless, such models offer little control over the generated content.

2. Related work

- Methodological approach is related to Prompt-to-Prompt (P2P).
- This allows to use arbitrary text-prompts to express the target translation; in contrast to P2P that requires word-to-word alignment between a source and target text prompts.



3. Preliminary

- Diffusion models are probabilistic generative models in which an image is generated by progressively removing noise from an initial Gaussian noise image.

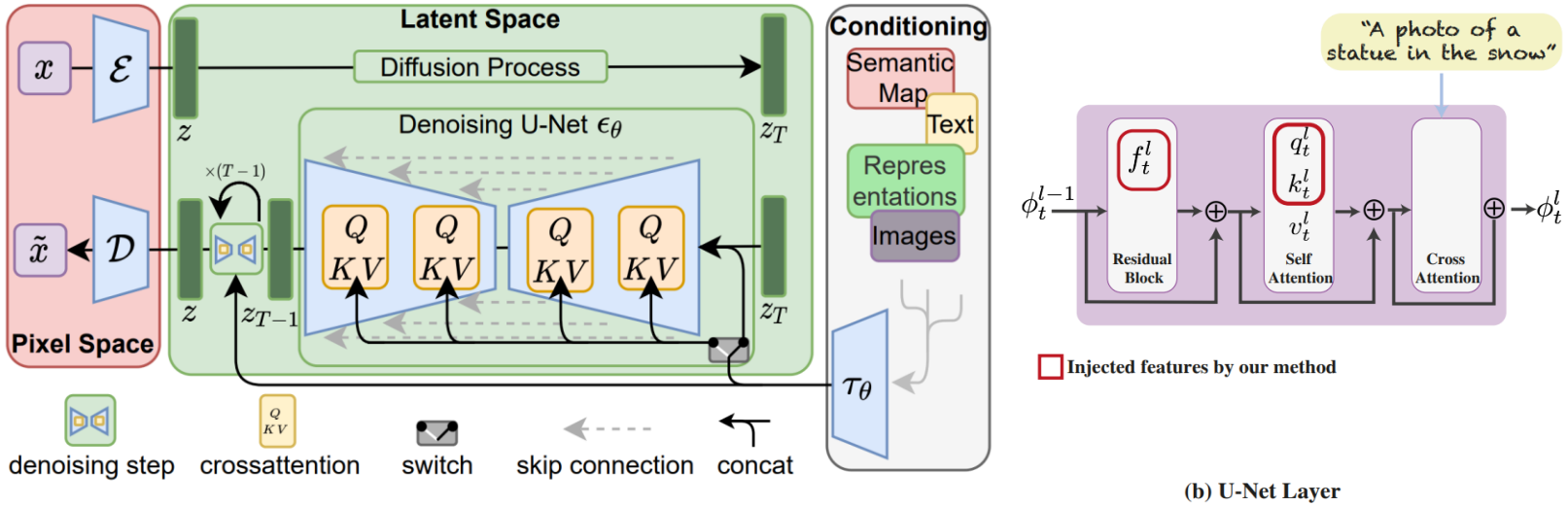
- The *forward* process: Gaussian noise is progressively added to a clean image, x_0:

$$x_t = \sqrt{\alpha_t} \cdot x_0 + \sqrt{1 - \alpha_t} \cdot z$$

- The *backward* process: gradually denoising x_T, where at each step a cleaner image is obtained. This process is achieved by a neural network $\epsilon_\theta(x_t, t)$ that predicts the added noise z. → conditioned on a guiding signal $\epsilon_\theta(x_t, y, t)$.

3. Preliminary

- We consider StableDiffusion, pre-trained and fixed text-to-image LDM model, denoted by $\epsilon_\theta(x_t, P, t)$, P is the text prompt.



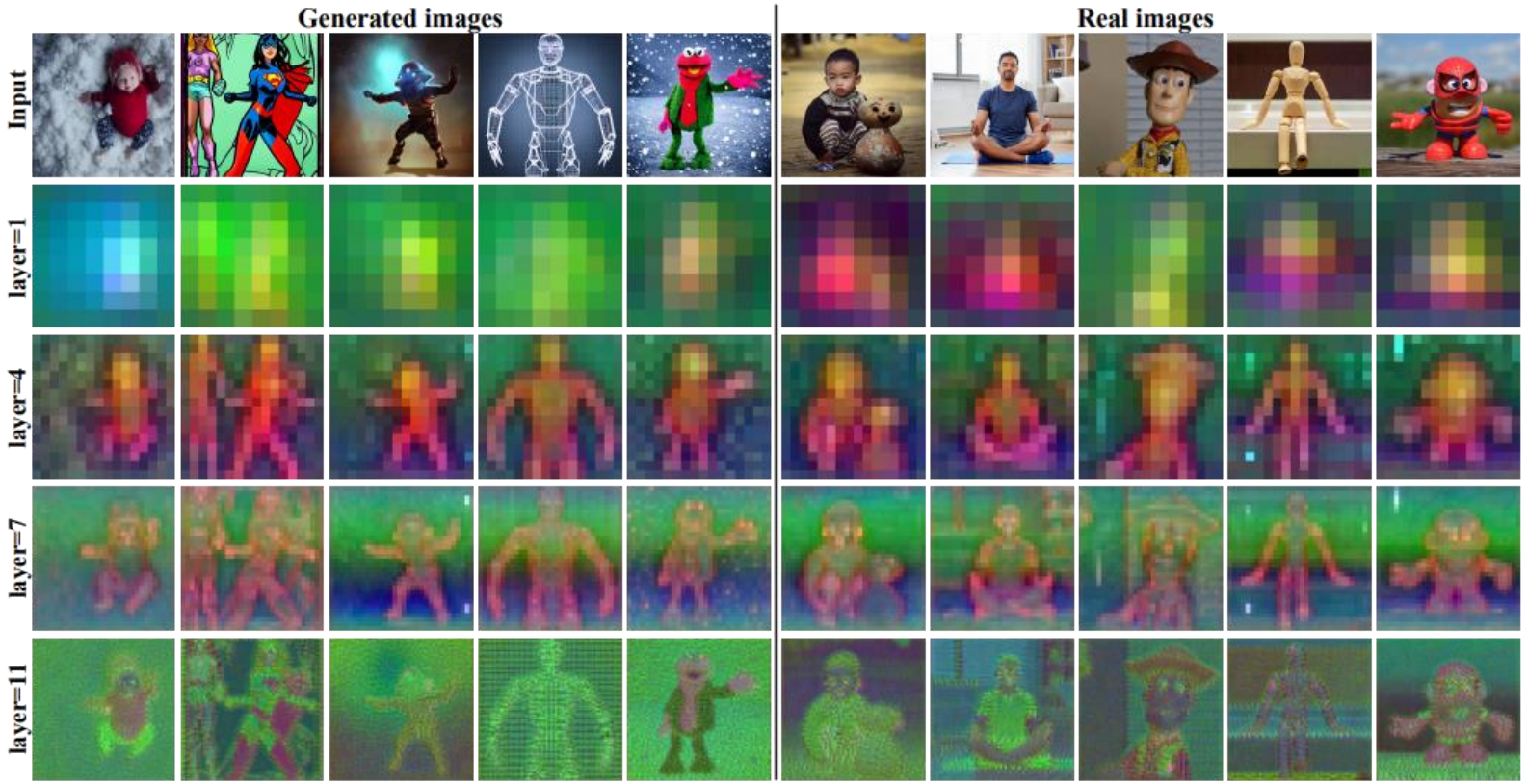
- The key is that fine-grained control over the generated structure can be achieved by manipulating spatial features inside the model during the generation process.

- We can observe and empirically demonstrate that:
  1. spatial features extracted form intermediate decoder layers encode localized semantic information and are less affected by appearance information.
  2. the self-attention, representing the affinities between the spatial features, allows to retain fine layout and shape details.

4. Method

Spatial features.

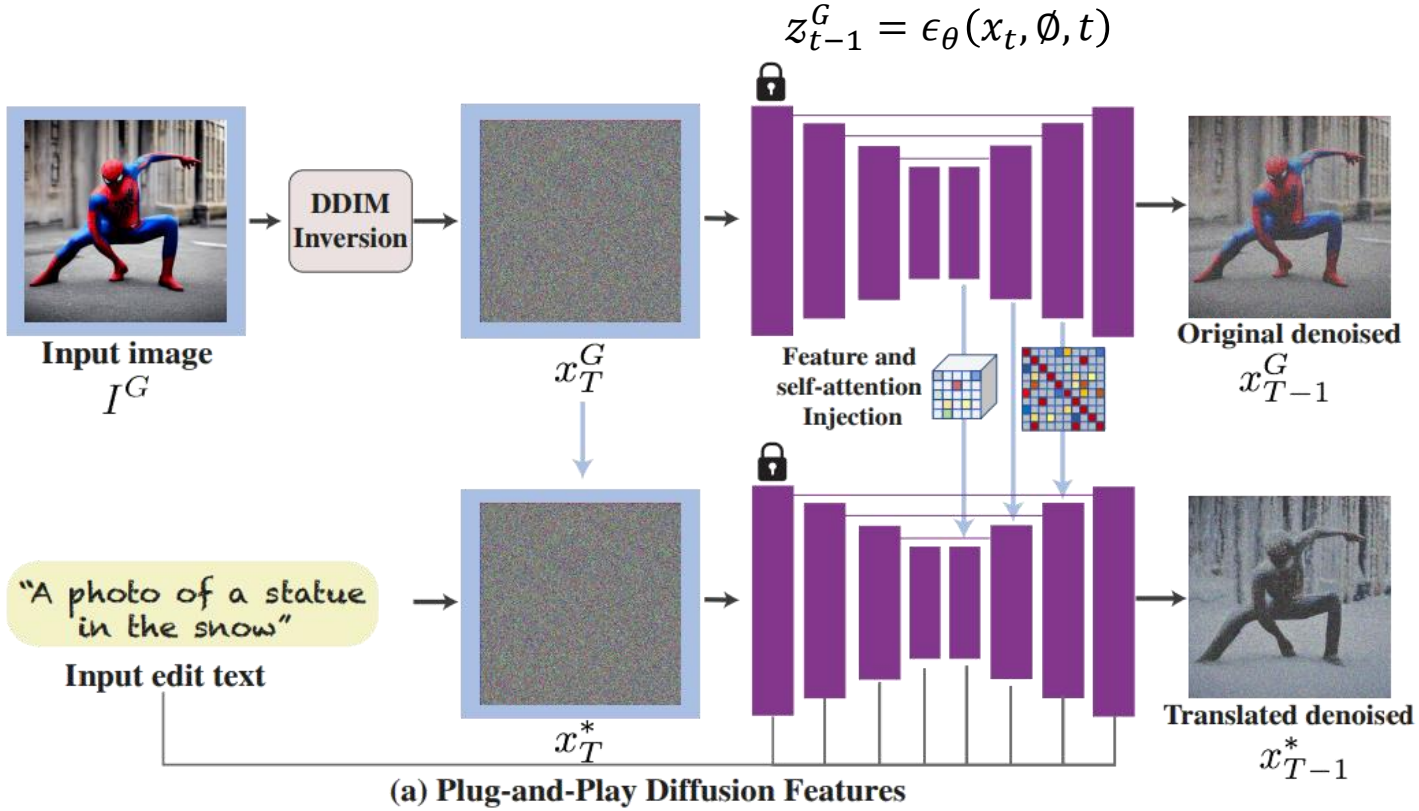
- In text-to-image generation, descriptive text prompts specify various scene and object properties. However, they often significantly vary across generated images from the same prompt under different initial noise x_T. → This suggests that the diffusion process itself and the resulting spatial features have a role in forming such fine-grained spatial information.



< Spatial features PCA>

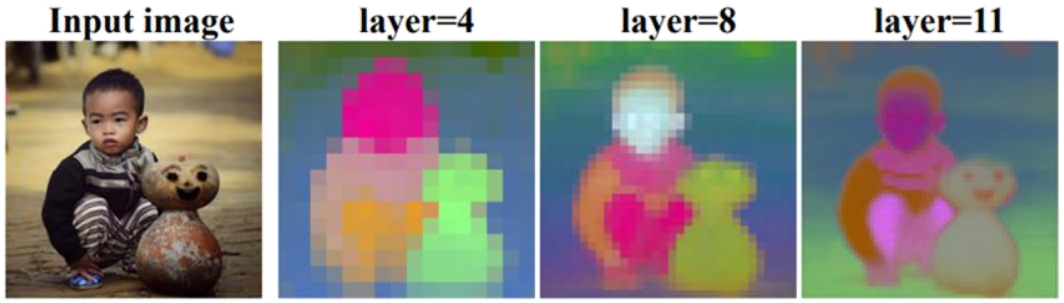
4. Method

Feature injection.

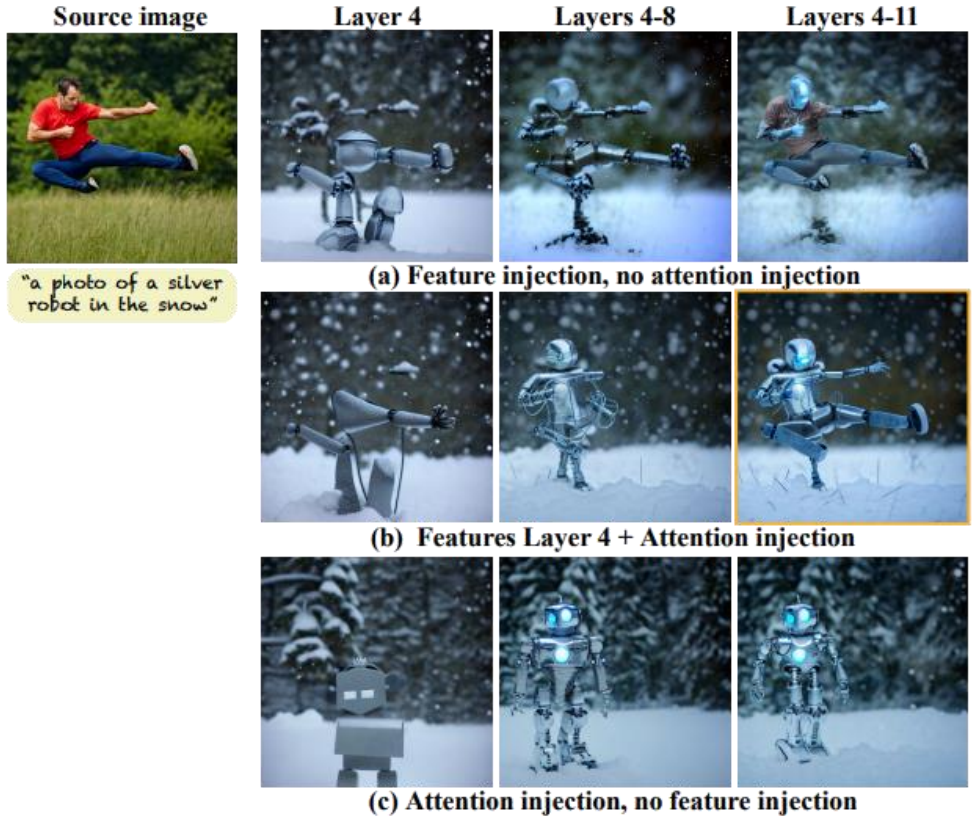


4. Method

Self-attention.



< Self-attention features >

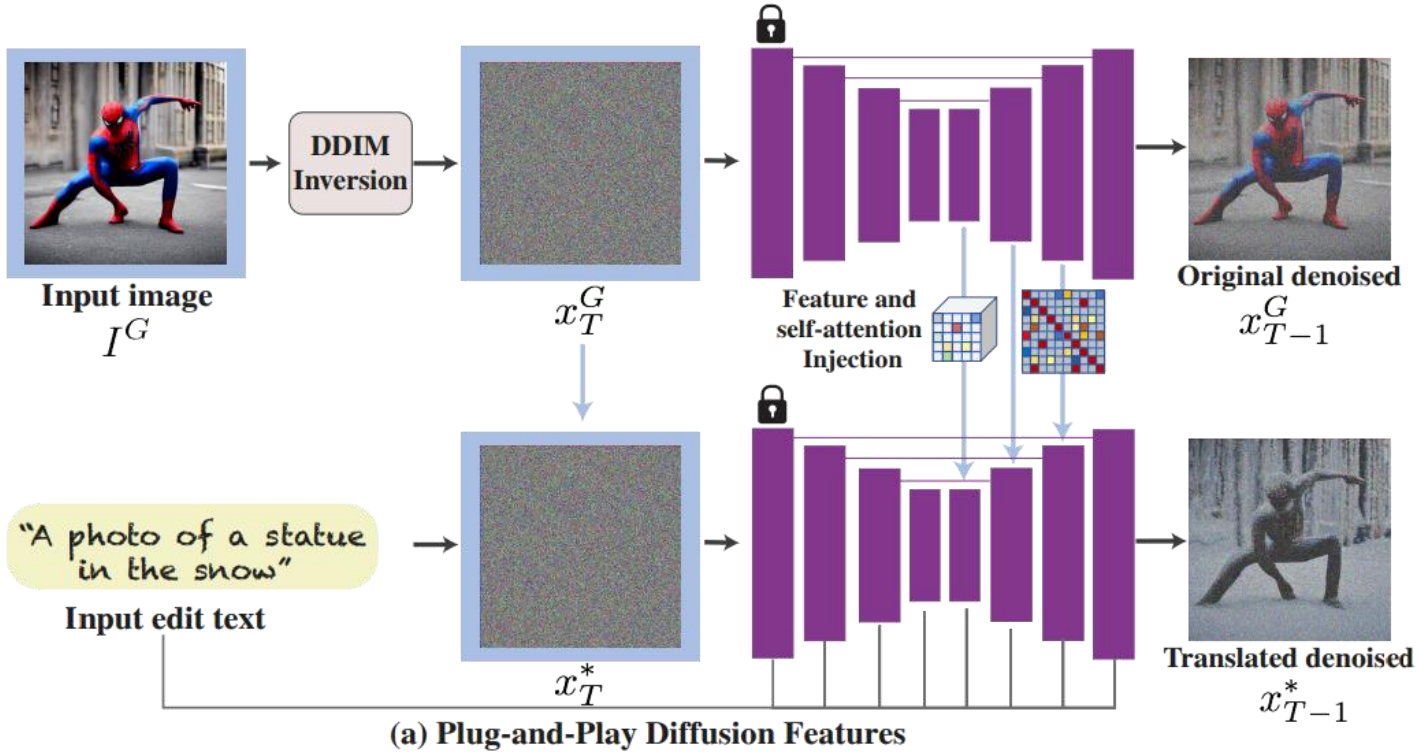


< Ablation studies >

4. Method

Self-attention.

$$z_{t-1}^{G} = \epsilon_\theta(x_t, \emptyset, t)$$



(a) Plug-and-Play Diffusion Features

$$z_{t-1}^{*} = \hat{\epsilon}_\theta\left(x_t^*, P, t;\ \{f_t^l\}\right) \rightarrow z_{t-1}^{*} = \hat{\epsilon}_\theta\left(x_t^*, P, t;\ f_t^4, \{A_t^l\}\right)$$

4. Method

Negative-prompting.

- In classifier-free guidance, the predicted noise ϵ at each sampling step is given by:

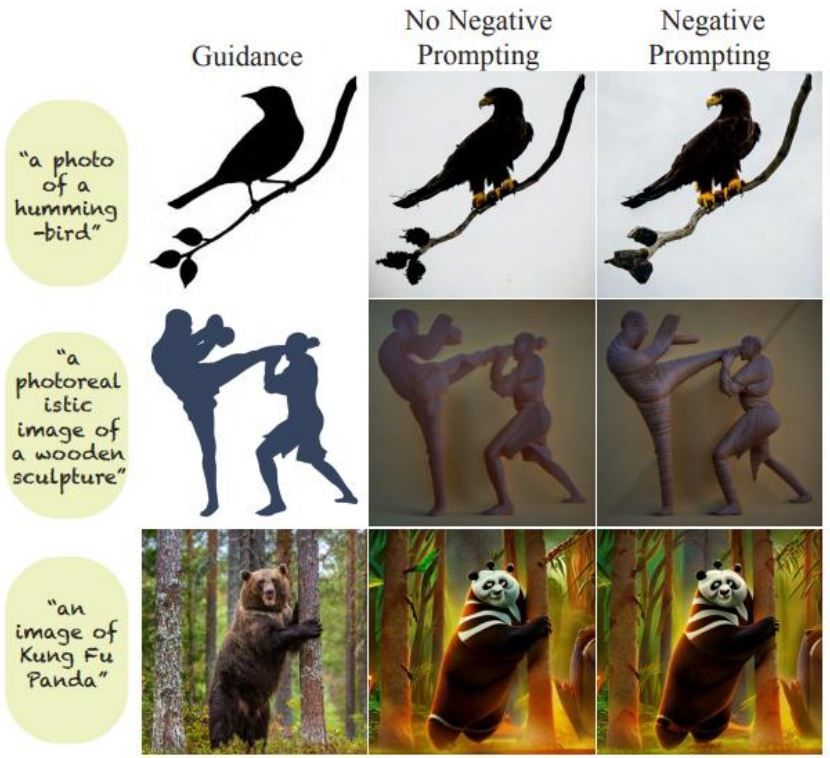
$$\epsilon = \omega\epsilon_\theta(x_t, P, t) + (1 - \omega)\epsilon_\theta(x_t, \emptyset, t)$$

- Similarly, by replacing the empty prompt with a "negative" prompt P_n:

$$\tilde{\epsilon} = \alpha\epsilon_\theta(x_t, \emptyset, t) + (1 - \alpha)\epsilon_\theta(x_t, P_n, t),$$

  pluging $\tilde{\epsilon}$ instead of $\epsilon_\theta(x_t, \emptyset, t)$. That is, $\epsilon = \omega\epsilon_\theta(x_t, P, t) + (1 - \omega)\tilde{\epsilon}$.

- In practice, negative-prompting is beneficial for handling textureless "primitives" guidance images.



5. Results

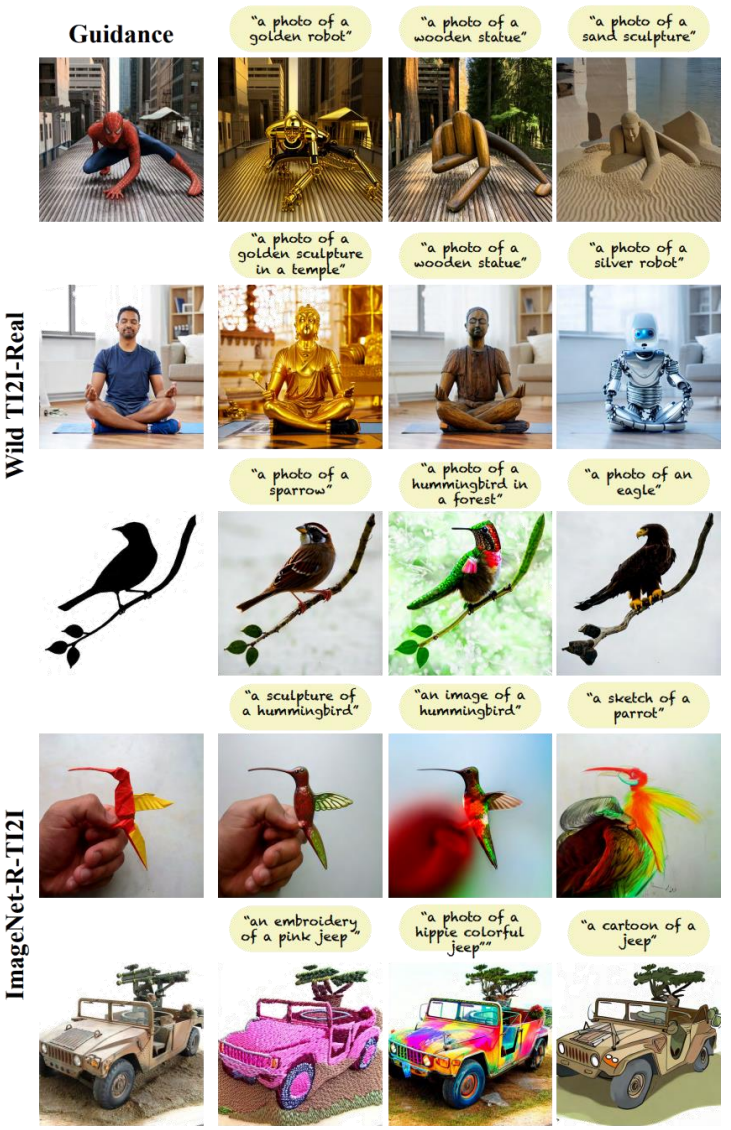
Datasets.

- Since there is no existing benchmark, we created two new datasets:
  1. *Wild-TI2I*: comprises of 148 diverse text-image pairs, 53% of which consists of real guidance images.
  2. *ImageNet-R-TI2I*: comprises of various renditions of ImageNet object classes.

- To adopt this dataset for our purpose, we manually selected 3 high-quality images from 10 different classes.

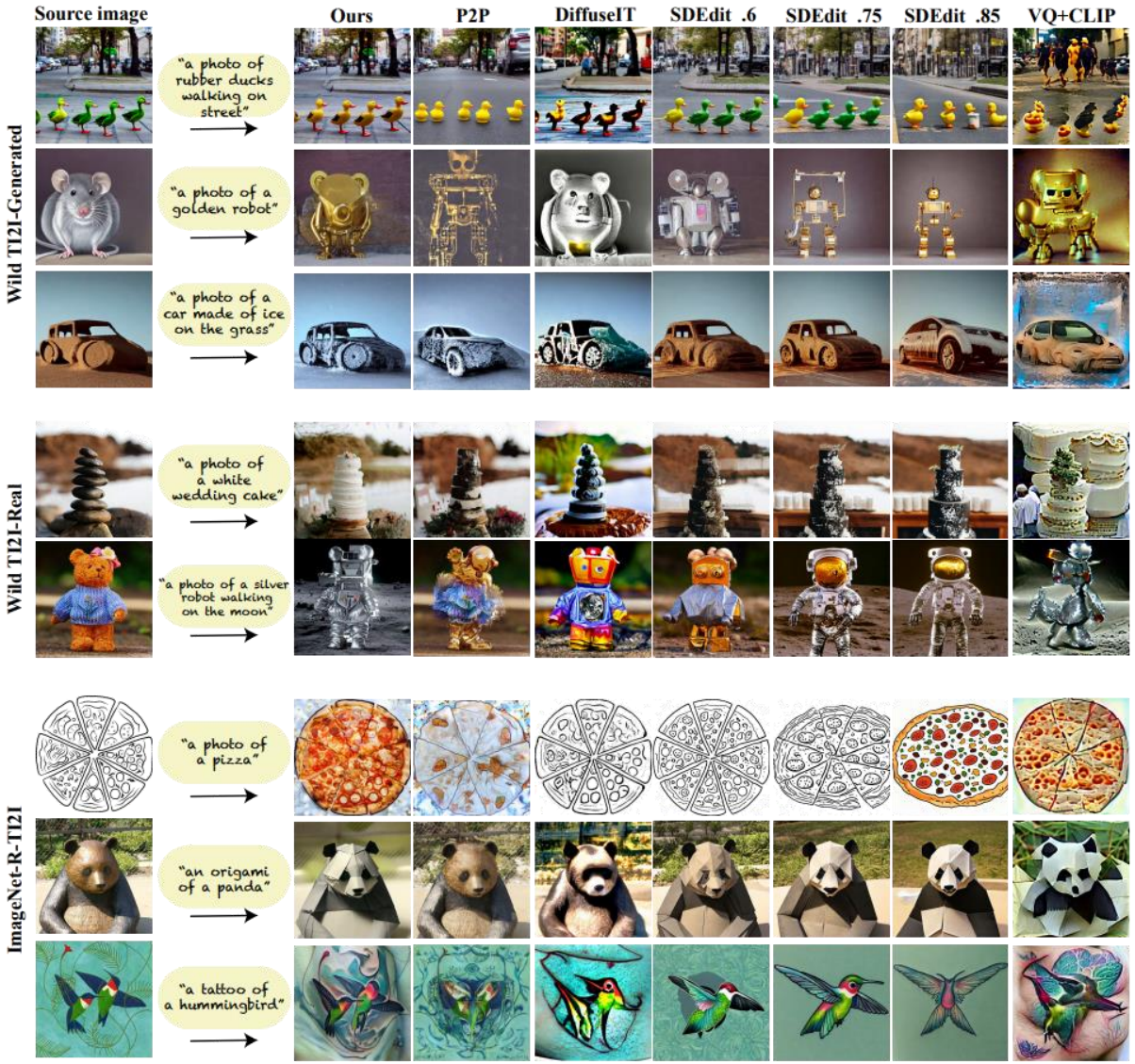
- To generate our image-text examples, we created a list of text templates by defining for each source class target categories and styles, and automatically sampled their combinations. (e.g. ‘a paint’ of ‘a bird’, ‘a photo’ of ‘car’)

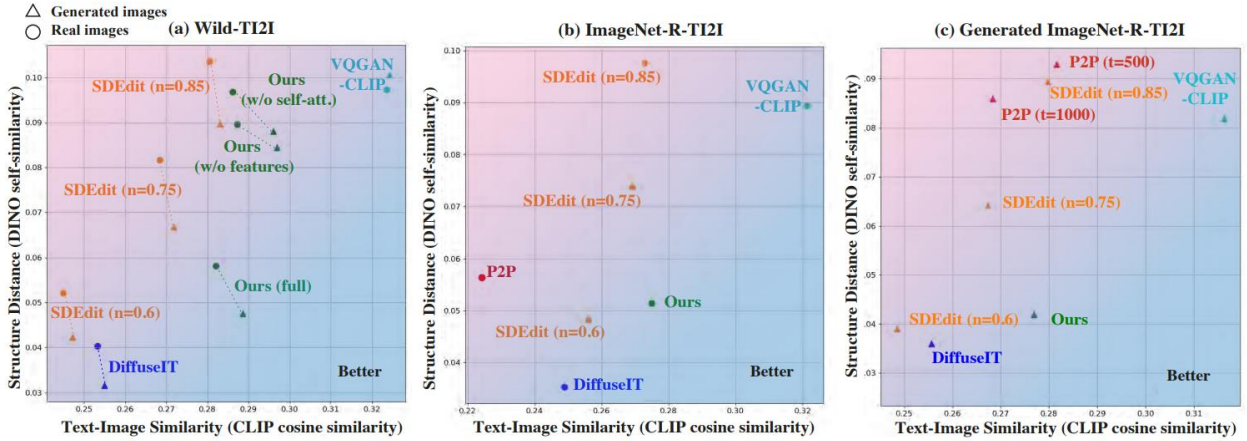
- This results in total of 150 image-text pairs.



5. Results

5.1. Comparison to Prior/Concurrent Work.



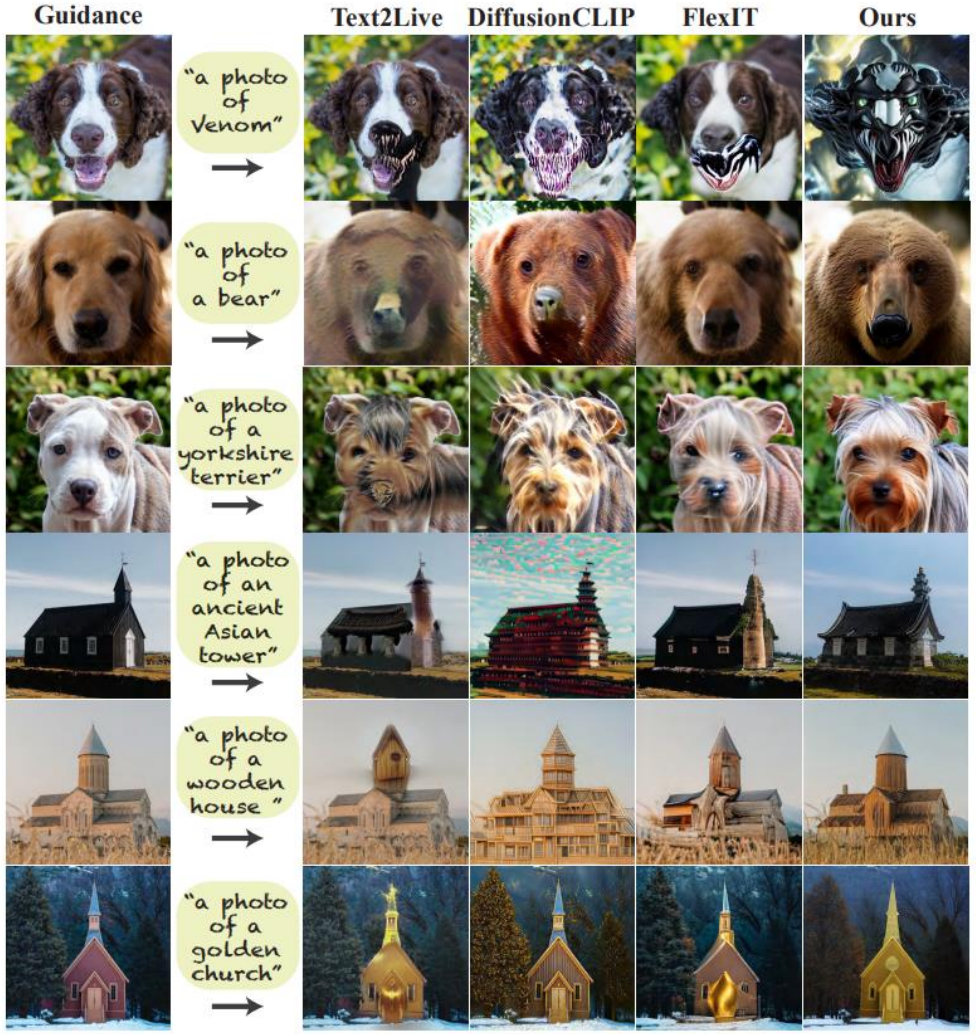


5. Results

5.1. Comparison to Prior/Concurrent Work.

- **Extended comparison to P2P.**

- **Additional baselines.**

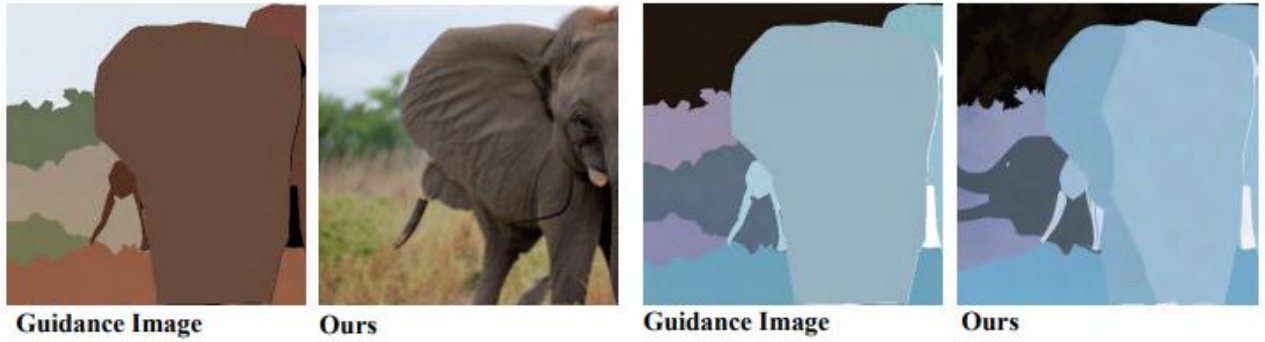


6. Discussion and Conclusion

- We presented a new framework for diverse text-guided image-to-image translation, founded on new insights about the internal representation of a pre-trained text-to-image diffusion model.

- Our method outperforms existing baselines, achieving a significantly better balance between preserving the guidance layout and deviating from its appearance.

- As for limitations, our method relies on the semantic association between the original and translated content in the diffusion feature space. Thus, it does not work well on detailed label segmentation masks where regions are colored arbitrarily. In addition, we are relying on DDIM inversion, which we found to work well in most of our examples.



- Our work demonstrates the yet unrealized potential of the rich and powerful feature space spanned by pre-trained text-to-image diffusion models.