

# Controllable Text-to-Image Generation with Customized Guidance on Appearance and Position

Ruiling Pu

Supervisor: Dr. Kai Han, Department of Statistics and Actuarial Science



## Problem Definition

**Objective:** This work aims to conveniently **reuse satisfying elements during image generations**. Our pipeline enables easy control of target items between generated images, in respect of appearance and user-defined position.



### Motivation:

- Given the same prompt, output images can vary completely between generations.
- Users may need to repeatedly regenerate images to achieve a desired object configuration.

## Methodology

- Main Idea:** extract target properties from cross-attention layers in Stable Diffusion, then construct customized energy function as guidance.

### How to add guidance:

The energy function  $E$  is acting as a loss function and its gradient is computed with respect to latent  $z_t$  using the equation:

$$z_t \leftarrow z_t - \sigma_t \eta \nabla_{z_t} \sum_{y \in \Gamma} E(A^{(y)})$$

at certain timestep  $t$  before feeding it to the denoiser  $\mathcal{D}$ .

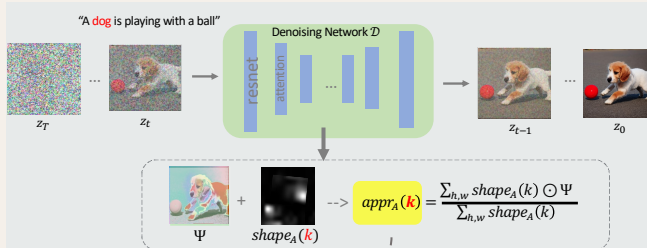
### Examples of energy function:

For appearance:  $E_{appr}(A^{(y)}, k) = \|appearance_A(k) - appearance_B(k)\|_1$

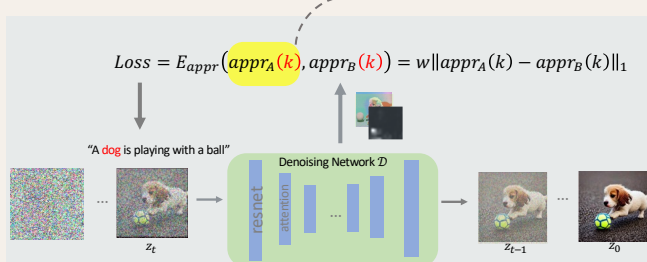
For position:  $E_{posi}(A^{(y)}, k, u) = \left(1 - \frac{\sum_{u \in B} A_{u,k}^{(y)}}{\sum_{u \in H \times W} A_{u,k}^{(y)}}\right)^2$ ,  $u$ =position box coordinates

## Pipeline

### Step 1: Reconstruct source image $image_A$



### Step 2: Generate new image $image_B$



## Experiments & Results

### Appearance guidance:



### Appearance + Positional guidance:

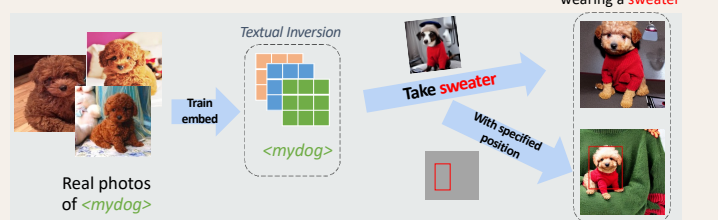


### Copy more than one targets:



### Personalized target control w/ Textual Inversion:

Main idea: use ONE WORD to represent a real personal concept



## Discussion & Future work

### Pose-appearance entanglement:

Cross-attention layers at different depth refer to different meanings (structural or details). We empirically found that dropping certain layers for appearance guidance leads to more flexibility of target pose. However, the output quality is unstable. Future work could explore quantifying task difficulty to better handle the varying complexity of different appearance-copy scenarios.



## Acknowledgement:

I especially appreciate Shaozhe Hao, a PhD candidate in the Department of Computer Science for his kindly help. I am also grateful to Xinci Ooi, a PhD candidate from the School of Biological Sciences for her support throughout the project.

## References:

- Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023.
- Natanuel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023.
- Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023.
- Dave Epstein, Allan Jabri, Ben Poole, Alexei A Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *arXiv preprint arXiv:2306.00986*, 2023.
- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023.