Customizable Image Synthesis with Multiple Subjects
– Supplementary Material –

Anonymous Author(s)
Affiliation
Address
`email`

A Experiment details

We supplement the experimental details of each method (Textual Inversion [1], DreamBooth [2], Custom Diffusion [3], and Cones [4]) in this section. For better generation quality, we use Stable Diffusion v2-1-base[1] as the pretrain model. For a fair comparison, we use 50 steps of DDIM [5] sampler with a scale of 7.5 for all above methods. All experiments are conducted using one A-100 GPU.

A.1 Textual Inversion

We use the third-party implementation of huggingface [6] for Textual Inversion. We train each subject-specific token with the recommended[2] batch size of 4 and a learning rate of 0.002 for 3000 steps. In particular, we initialize the subject-specific token with the corresponding class token. For example, to customize a specific cat, we initialize the subject-specific token "<cat>" with the original "cat" token.

A.2 DreamBooth

We use the third-party implementation of huggingface [6] for DreamBooth. Training is with a batch size of 2, learning rate 5×10^{-5}, and training steps of $800 \times$ number of subjects.

A.3 Custom Diffusion

We use the official implementation[3] for Custom Diffusion. Training is with a batch size of 2, learning rate 1×10^{-5} and training steps of $250 \times$ number of subjects.

A.4 Cones

We use the official implementation for Cones. Training is with a batch size of 2, learning rate 4×10^{-5} and training steps of $1200 \times$ number of subjects.

A.5 Our Approach

For our approach, We train each subject-specific residual token embedding with a batch size of 1 and a learning rate of 1×10^{-6} for 3,000 steps. At inference time, the layouts are appointed by bounding boxes given by the users to indicate the location of each subject. We use a positive value of $+2.5$ to strengthen the signal of the target subject and we use a negative value of -1×10^{-5} to weaken

[1] https://huggingface.co/stabilityai/stable-diffusion-2-1
[2] https://github.com/rinongal/textual_inversion
[3] https://github.com/adobe-research/custom-diffusion

the signal of irrelevant subjects. Furthermore, we guide all 50 steps with the layout guidance in the whole generation process to get good customized generation results.

A.6 User Study

For two- to four-subject generation tasks, we design four different subject combinations for each task. This will yield 12 subject combinations in total. For each subject combination, we design four different text prompts to generate images with 5 random seeds. We conduct this procedure to all four methods. With such settings, each method generates 80 different images for each task. We give each generated image 4-8 questions for testing image alignment (2-4 questions) and text alignment (2-4 questions). The number of questions is proportional to the number of subjects used to customize the image (average 6 questions per generated image). Finally, we mess up the order of all the image-question pairs and assigned them to 25 different users for scoring, and finally summarized the results. In detail, every user needs to score $4 \times 4 \times 5 \times 6 = 480$ questions for each task and for each method.

B More comparisons

In this section, we conducted further comparisons between our approach and three other baselines. As shown in Fig. S1, regarding the generation of single subjects, the four methods exhibited similar performance. However, when dealing with semantically similar subjects, such as a dog and a cat, as well as scenarios involving three or more subjects, our approach clearly exhibited superior performance. Moreover, as shown in Fig. S2, we further showcase additional generated results, providing further evidence of the robustness of our approach.

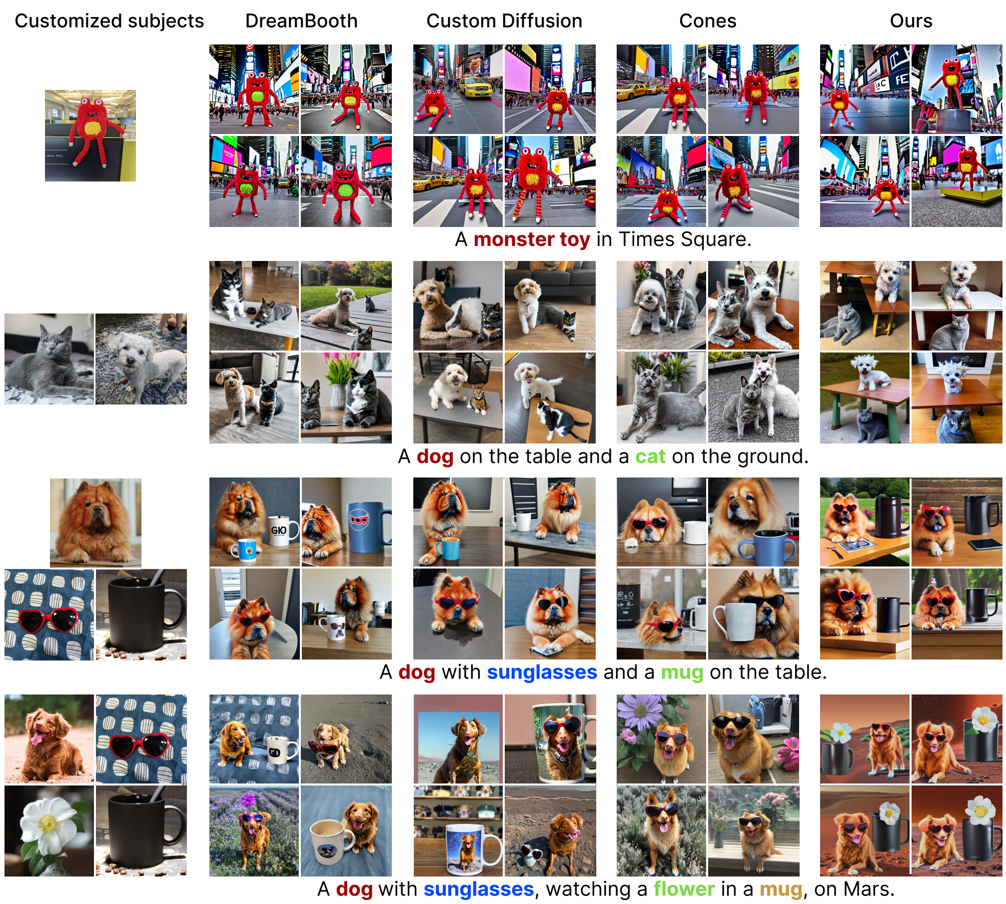

Figure S1: More comparison of our approach and other baselines.

Table S1: Quantitative comparisons between our method (learning a residual token embedding) and learning a token embedding directly.

	Single Subject		Two Subjects		Three Subjects		Four Subjects	
	Text Alignment	Image Alignment	Text Alignment	Image Alignment	Text Alignment	Image Alignment	Text Alignment	Image Alignment
Our	0.330	0.725	0.309	0.708	0.304	0.689	0.299	0.673
Token embedding	0.324	0.720	0.291	0.686	0.292	0.669	0.281	0.651

C More challenging cases

As shown in Fig. S3, we present a larger number of images generated by our approach, featuring a greater diversity of customized subjects. In comparison with other methods, we observe that when the number of customized subjects reaches four, the performance of other methods significantly deteriorates. In contrast, our approach can generate a larger number of customized subjects, exemplifying the superiority of our approach.

D Importance of residual token embedding

To demonstrate the superior generalization of the residuals, we conduct comparative experiments. As shown in Tab. S1, compared to directly updating the class embedding parameters in a single text embedding, our method, which involves updating the text encoder and calculating the average shift from the class to the specific subject based on a certain number of text templates, outperforms in both textual and visual similarity.

E Generated results of textual inversion

As shown in Fig. S4We observe that Textual Inversion struggles with the generation of complex single subject and multiple subjects.

F Social impact and limitations

F.1 Social impact.

While training individual large-scale diffusion models remains prohibitively expensive, advancements in fine-tuning techniques have enabled individual users to customize their own models. Our technology empowers users to linearly combine their personalized single-subject models, generating high-quality images with multiple customized subjects, while maintaining significant advantages in terms of computation and storage efficiency. Furthermore, there is a growing need for more reliable detection techniques to identify and mitigate the presence of fake data.

F.2 Limitations.

Our method is limited by the inherent capabilities of the base model. Specifically, when it comes to combining more than six subjects, our method may not be able to consistently generate satisfactory results. In order to achieve the desired generation results, the provided layout by the user needs to be roughly consistent with the textual description.

Figure S2: More results of our approach.

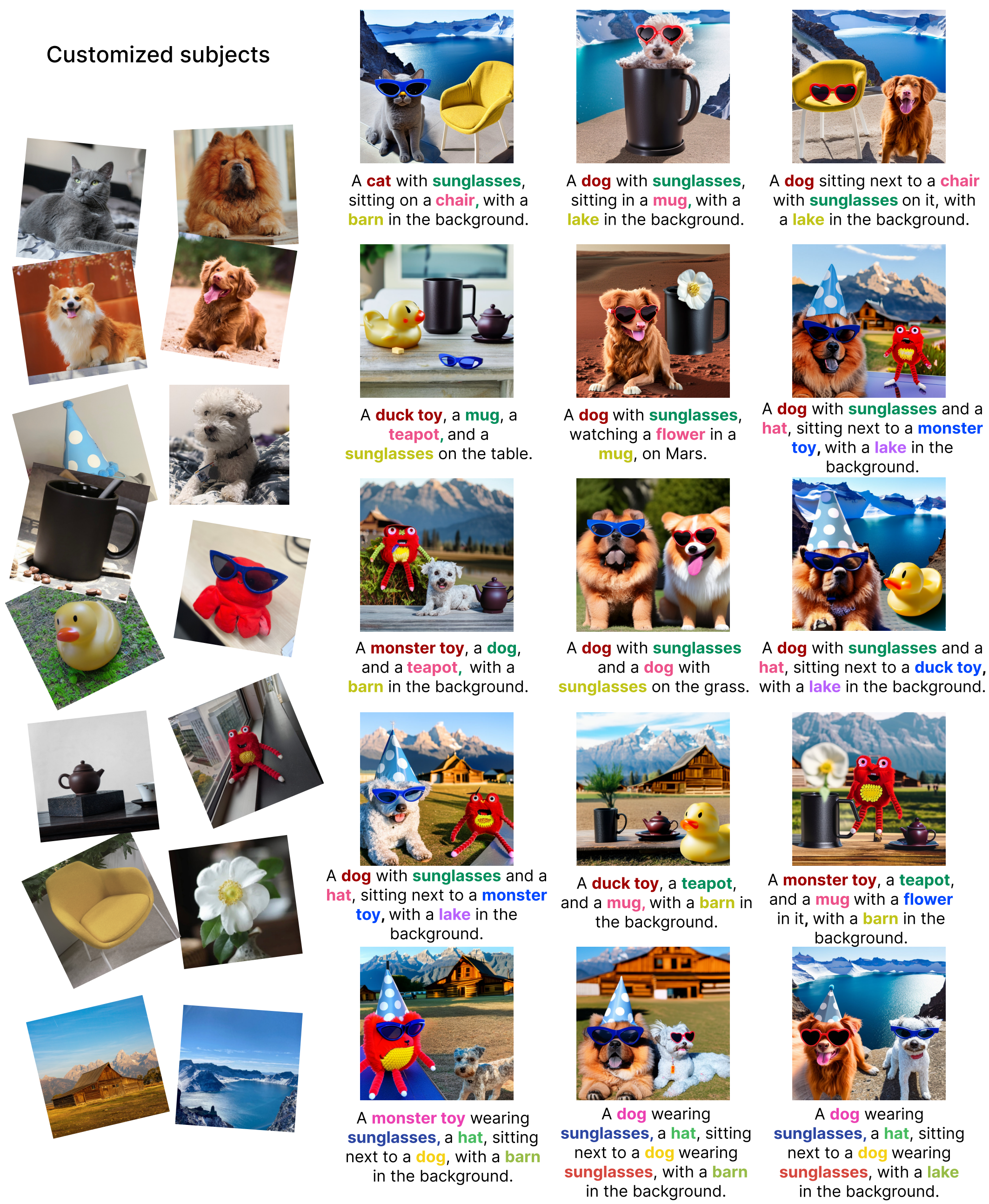

Figure S3: More results of multi-subject generation.

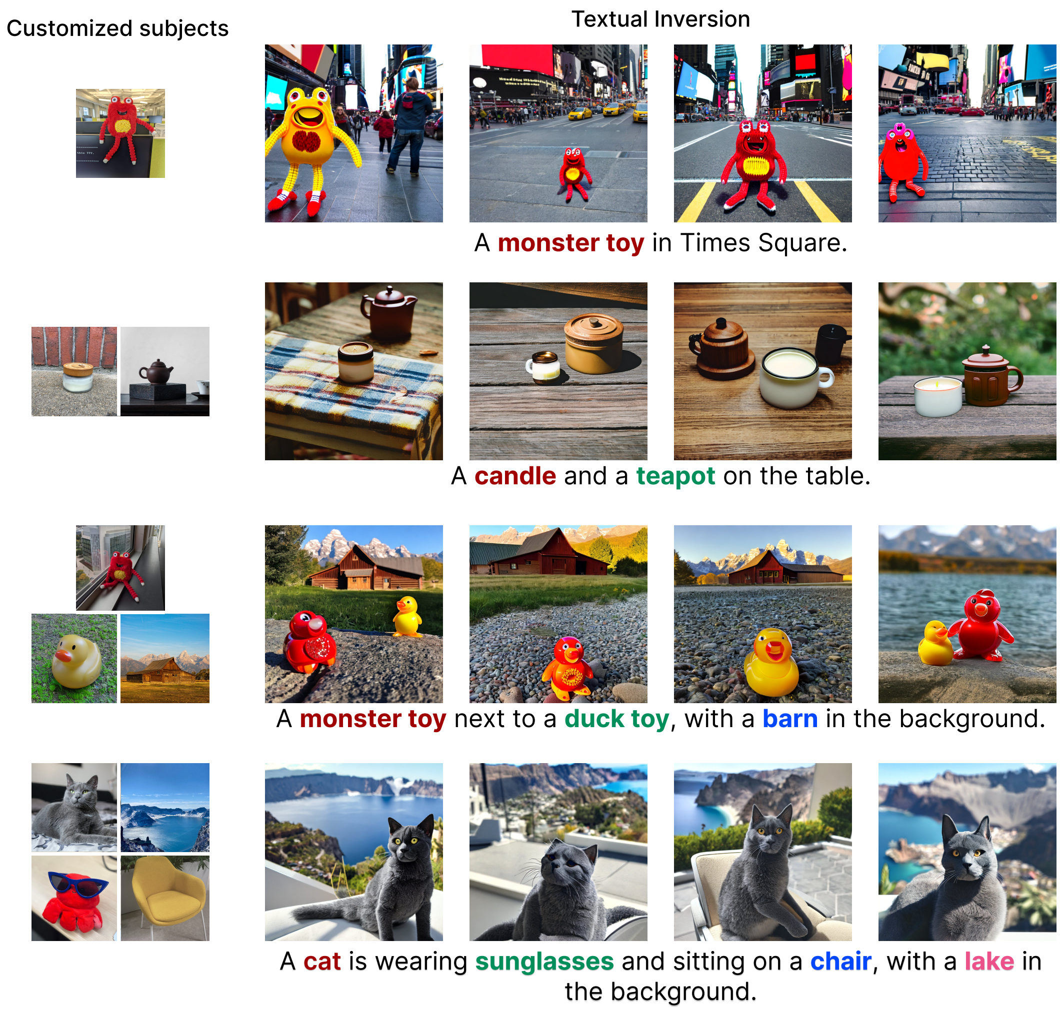

Figure S4: Generated results of Textual Inversion.

References

[1] 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. *arXiv preprint arXiv:2208.01618*, 2022.

[2] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

[3] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022.

[4] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023.

[5] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Int. Conf. Learn. Represent.*, 2021.

[6] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.