FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention

Guangxuan Xiao* **Tianwei Yin*** **William T. Freeman** **Frédo Durand** **Song Han**
Massachusetts Institute of Technology
https://fastcomposer.mit.edu

Contents

1 Broader Impact.

FastComposer provides a fast and effective approach to personalized multi-subject text-to-image generation, thus democratizing AI-driven content creation by reducing the demand for computational resources and hardware. However, the utilization of this model may lead to unexpected consequences. For instance, the simplified process of creating personalized multi-subject images could enable malicious activities such as the production of deepfakes or other deceptive content. Concerns may also arise regarding data privacy and consent in the context of using reference subjects to generate new images. It is critical that the deployment of such methods is approached with a keen awareness of ethical considerations and their potential societal impacts. In response to these challenges, the advancement of deepfake detection methods [1] could provide a countermeasure, ensuring that image generation techniques are used responsibly and beneficially for society.

2 Experiment Details

We use baselines implementations from the `diffusers` library [6]. Each baseline employs the same StableDiffusion v1-5 model [4] that we use in our approach. All baselines, by default, run on the standard set of hyperparameters, with the exception being that we've adjusted the number of training steps to 1000 for both DreamBooth [5] and Custom-Diffusion [2]. This adjustment was made due to noticeable underfitting when using the original number of steps, 250 for Custom-Diffusion and 400 for DreamBooth. All baselines are trained with 5 images per subject. In our experiments, we use $\alpha = 0.5$ for single object generation, and $\alpha = 0.6$ for multi-subject generation. We use PNDM sampling [3] with 50 steps and a classifier-free guidance scale of 5 across all methods.

* Equal Contribution.
Correspondence to: Guangxuan Xiao <xgx@mit.edu>, Tianwei Yin <tianweiy@mit.edu>.

Preprint. Under review.

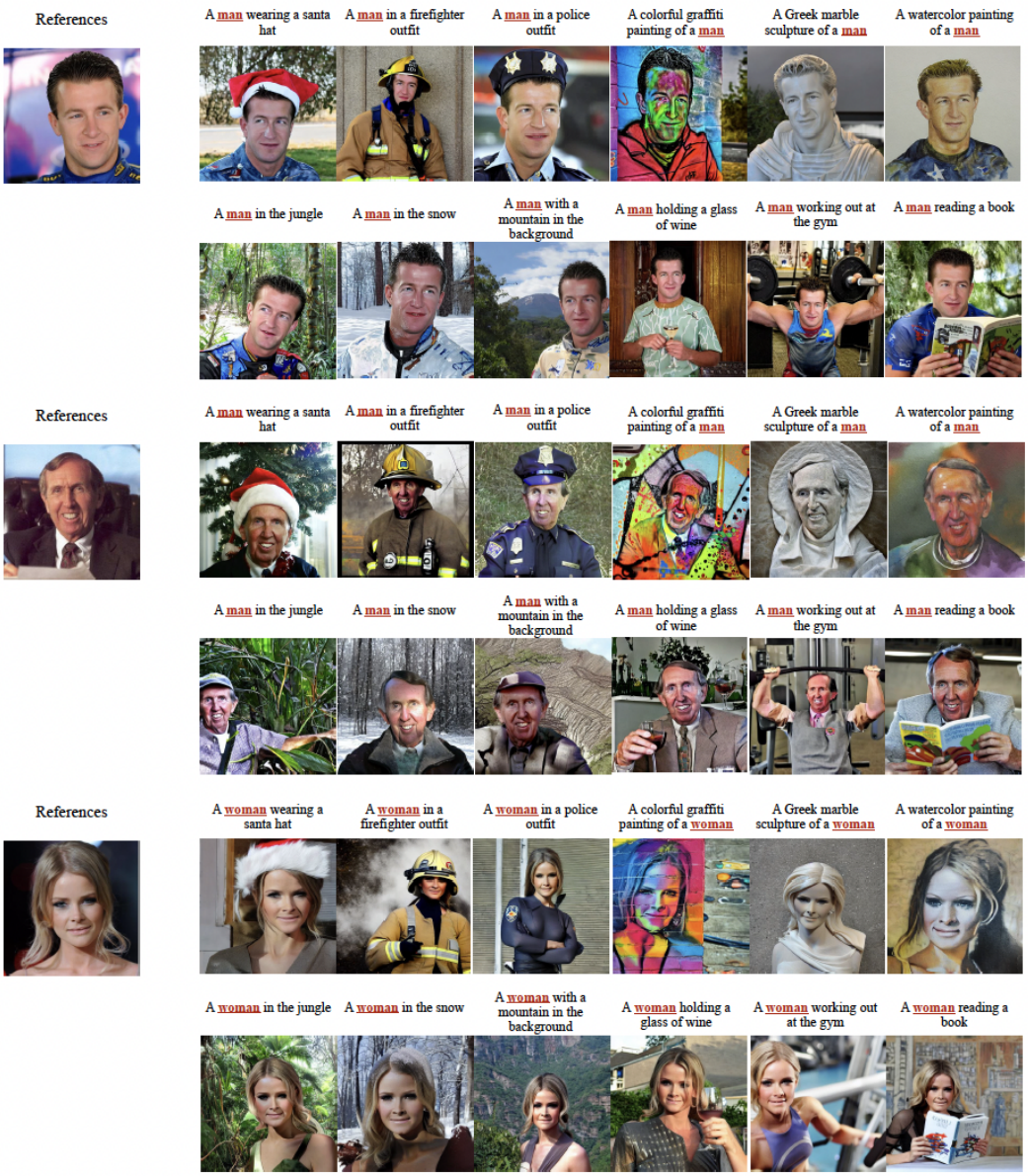

Figure 1: Uncurated single-subject generation results.

3 Additional Results

In Figure 1 and Figure 2, we present uncurated results for single-subject and multiple-subject generation, respectively. We will make the complete set of model predictions publicly available for easy comparisons.

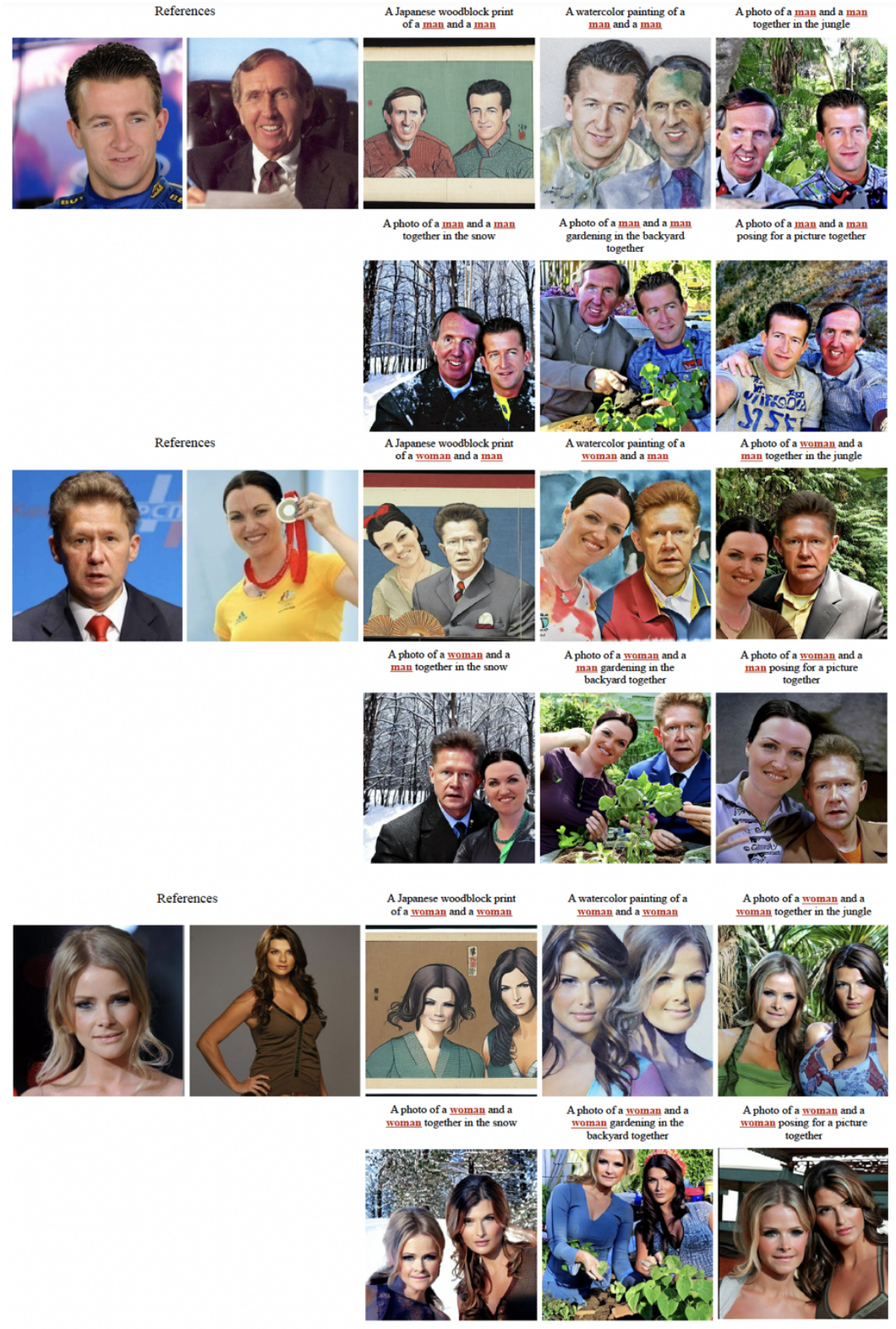

Figure 2: Uncurated multiple-subject generation results.

4 Evaluation Text Prompts

4.1 Single-Subject Prompts

"a painting of a [class noun] <A*> in the style of Banksy"

"a painting of a [class noun] <A*> in the style of Vincent Van Gogh"

"a colorful graffiti painting of a [class noun] <A*>"

"a watercolor painting of a [class noun] <A*>"

"a Greek marble sculpture of a [class noun] <A*>"

"a street art mural of a [class noun] <A*>"

"a black and white photograph of a [class noun] <A*>"

"a pointillism painting of a [class noun] <A*>"

"a Japanese woodblock print of a [class noun] <A*>"

"a street art stencil of a [class noun] <A*>"

"a [class noun] <A*> wearing a red hat"

"a [class noun] <A*> wearing a santa hat"

"a [class noun] <A*> wearing a rainbow scarf"

"a [class noun] <A*> wearing a black top hat and a monocle"

"a [class noun] <A*> in a chef outfit"

"a [class noun] <A*> in a firefighter outfit"

"a [class noun] <A*> in a police outfit"

"a [class noun] <A*> wearing pink glasses"

"a [class noun] <A*> wearing a yellow shirt"

"a [class noun] <A*> in a purple wizard outfit"

"a [class noun] <A*> riding a horse"

"a [class noun] <A*> holding a glass of wine"

"a [class noun] <A*> holding a piece of cake"

"a [class noun] <A*> giving a lecture"

"a [class noun] <A*> reading a book"

"a [class noun] <A*> gardening in the backyard"

"a [class noun] <A*> cooking a meal"

"a [class noun] <A*> working out at the gym"

"a [class noun] <A*> walking the dog"

"a [class noun] <A*> baking cookies"

"a [class noun] <A*> in the jungle"

"a [class noun] <A*> in the snow"

"a [class noun] <A*> on the beach"

"a [class noun] <A*> on a cobblestone street"

"a [class noun] <A*> on top of pink fabric"

"a [class noun] <A*> on top of a wooden floor"

"a [class noun] <A*> with a city in the background"

"a [class noun] <A*> with a mountain in the background"

"a [class noun] <A*> with a blue house in the background"

"a [class noun] <A*> on top of a purple rug in a forest"

4.2 Multiple-Subject Prompts

"a painting of a [class noun] <A*> and a [class noun] <A*> together in the style of Banksy"

"a painting of a [class noun] <A*> and a [class noun] <A*> together in the style of Vincent Van Gogh"

"a watercolor painting of a [class noun] <A*> and a [class noun] <A*> together"

"a street art mural of a [class noun] <A*> and a [class noun] <A*> together"

"a black and white photograph of a [class noun] <A*> and a [class noun] <A*> together"

"a pointillism painting of a [class noun] <A*> and a [class noun] <A*> together"

"a Japanese woodblock print of a [class noun] <A*> and a [class noun] <A*> together"

"a photo of a [class noun] <A*> and a [class noun] <A*> gardening in the backyard together"

"a photo of a [class noun] <A*> and a [class noun] <A*> cooking a meal together"

"a photo of a [class noun] <A*> and a [class noun] <A*> sitting in a park together"

"a photo of a [class noun] <A*> and a [class noun] <A*> working out at the gym together"

"a photo of a [class noun] <A*> and a [class noun] <A*> baking cookies together"

"a photo of a [class noun] <A*> and a [class noun] <A*> posing for a selfie together"

"a photo of a [class noun] <A*> and a [class noun] <A*> making funny faces for a photo booth together"

"a photo of a [class noun] <A*> and a [class noun] <A*> playing a musical duet together"

"a photo of a [class noun] <A*> and a [class noun] <A*> together in the jungle"

"a photo of a [class noun] <A*> and a [class noun] <A*> together in the snow"

"a photo of a [class noun] <A*> and a [class noun] <A*> together on the beach"

"a photo of a [class noun] <A*> and a [class noun] <A*> together with a city in the background"

"a photo of a [class noun] <A*> and a [class noun] <A*> together with a mountain in the background"

"a photo of a [class noun] <A*> and a [class noun] <A*> together with a blue house in the background"

References

[1] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 1

[2] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *CVPR*, 2023. 1

[3] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *ICLR*, 2022. 1

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[5] 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. *CVPR*, 2023. 1

[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. 1