Supplementary Material for Magic Insert: Style-Aware Drag-and-Drop

Nataniel Ruiz[1] Yuanzhen Li[1] Neal Wadhwa[2] Yael Pritch[2]
Michael Rubinstein[1] David E. Jacobs[2] Shlomi Fruchter[1]
[1]Google DeepMind [2]Google

Demo For a pre-computed demo of the method with more than 160 random results please check our attached demo.html file inside the **demo** folder. Once open, select a subject and select a background and then drag-and-drop the subject into the background on either side of the image. To try new combinations click on the green reset button.

Style Fidelity and Comparisons with Related Work For style fidelity, InstantStyle ControlNet can sometime quantitatively outperform our variants using the automatic metrics shown in the main paper. Even so, we strongly observe that subject details and contrast is lost in many of these samples as shown in Figure 1.

Finding strong quantitative metrics for subject fidelity and for style fidelity is an open problem in the field, and metrics have biases that make them suboptimal. Again, we show examples for our proposed style-aware personalization, along with top baseline contenders in Figure 1. We observe that the generation quality of our variants is stronger than the benchmarks, with both strong style fidelity and subject fidelity. Our Magic Insert + ControlNet variant is powerful given that it exactly follows the outline of the character, and thus has the strongest subject fidelity over all approaches, although it does not have the desirable properties of our method w/o ControlNet which include pose, form and attribute modification of the subject. We further the discussion on subject fidelity vs. editability tradeoff below.

Semantic Modifications of Subject Our method inherits all benefits of DreamBooth [1] and thus allows for modification of subject characteristics such as pose, adding accessories, changing appearance, shapeshifting and hybrids. We show some examples in Figure 2. The generated subjects can then be inserted into the background image.

Editability / Fidelity Tradeoff Our method (w/o ControlNet) also inherits DreamBooth's editability / fidelity tradeoff. Specifically, the longer the personalization training, the stronger the subject fidelity but the lesser the editability. This phenomenon is shown in Figure 3. In most cases a sweet spot can be found for different applications.

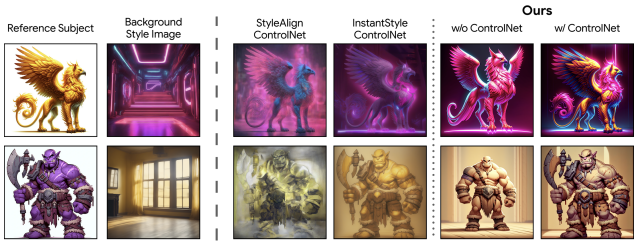

Figure 1. **Style-Aware Personalization Baseline Comparison:** We show some comparisons of our style-aware personalization method with respect to the top performing baselines StyleAlign + ControlNet and InstantStyle + ControlNet. We can see that the baselines can yield decent outputs, but lag behind our style-aware personalization method.

Figure 2. **Style-Aware Personalization with Attribute Modification:** Our method allows us to modify key attributes for the subject, such as the ones reflected in this figure, while consistently applying our target style over the generations. This allows us to reinvent the character, or add accessories, which gives large flexibility for creative uses. Note that when using ControlNet this capability disappears.

For our work we use 600 iterations with batch size 1, a learning rate of 1e-5 and weight decay of 0.3 for the UNet. We also train the text encoder with a learning rate of 1e-3 and weight decay of 0.1.

Figure 3. **Editability / Fidelity Tradeoff:** We show the phenomenon of editability / fidelity tradeoff by showing generations for different finetuning iterations of the space marine (shown above the images) with the "green ship" stylization and additional text prompting "sitting down on the floor". When the style-aware personalized model is finetuned for longer on the subject, we get stronger fidelity to the subject but have less flexibility on editing the pose or other semantic properties of the subject. This can also translate to style editability.

References

[1] 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. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510. IEEE, 2023. 1