StyleBoost: A Study of Personalizing Text-to-Image Generation in Any Style using DreamBooth

(1) Goals and Focus:
The "StyleBoost" paper focuses on fine-tuning a text-to-image latent diffusion model that enables style personalization. It proposes a methodology to address the lack of knowledge on how to learn the abstract concept of "style" in pre-trained models by projecting additional segment information of the style. It also suggests a method for selecting images. In contrast, "StyleForge" generalizes the approach by experimenting with various styles and comparison groups, and it addresses the issues in the image selection methodology from the previous StyleBoost research by focusing on a more detailed representation of styles.

(2) Methodological differences: "StyleBoost" constructs 20 images consisting of people and backgrounds, along with text ("A [V] style"), to learn abstract styles. However, due to the limitations of expressing detailed styles with abstract segment knowledge, it proposes a dual-binding approach. To enhance the boosting effect of style learning and focus on fine details such as human poses and facial expressions, StyleBoost introduces autogenous class-specific prior preservation, originally adopted in DreamBooth to prevent overfitting and language drift, by directly selecting images of people in the target style.

On the other hand, "StyleForge" encounters poor text alignment when fine-tuning a pair of 20 images and text featuring people and backgrounds for style learning. To address this, it presents the Multi-StyleForge method, which divides the people and background images into two pairs and fine-tunes them simultaneously. While this is similar to the parameter-efficient learning method proposed by "Custom-Diffusion," the key difference is that "StyleForge" recommends full fine-tuning for abstract subject learning due to the limitations of parameter-efficient models in learning capability.

In the previous study, "StyleBoost" evaluated three styles based on FID and CLIP scores by fine-tuning an abstract style personalization model built on DreamBooth, a text-to-image latent diffusion model for object personalization, and comparing it with the original DreamBooth model. In contrast, "StyleForge" evaluates image output performance with FID and KID scores, and text alignment with CLIP scores across six styles, using various personalization-capable text-to-image models like "Custom-Diffusion," "LoRA," "Textual Inversion," and "DreamBooth." Additionally, it analyzes style segment information through attention maps, demonstrating the generalization of style learning and the ability to express more detailed styles.

(3) Experiments and results: Since "StyleBoost" is a modified version of the DreamBooth model, it compares image evaluation and text alignment related to style learning with DreamBooth. On the other hand, "StyleForge" demonstrates that among the evaluations of six different styles, the text alignment of multi-StyleForge is the most superior.

(4) Conclusive difference: "StyleBoost" presents a learning method based on DreamBooth for learning the ambiguous concept of style, while "StyleForge" proposes Multi-StyleForge to improve text alignment and generalizes its approach by comparing various styles and personalization models.

I²AM: Interpreting Image-to-Image Latent Diffusion Models via Bi-Attribution Maps

(1) Goals and Focus: The "I²AM" paper focuses on visualizing attention maps using existing image-to-image latent diffusion models, aiming to analyze and improve how these models function. In contrast, "StyleForge" emphasizes the detailed reproduction of various artistic styles in text-to-image synthesis and proposes a methodology for personalizing existing text-to-image latent diffusion models.

(2) Methodological differences: "I²AM" visualizes attention maps using three inpainting models—Paint-by-Example, DCI-VTON, and StableVITON—and a super-resolution model called PASD. By utilizing the VITON-HD dataset, it conducts traditional performance evaluations using 13,679 image pairs with FID, KID, SSIM, and LPIPS metrics. Additionally, it assesses model performance as explainable AI (XAI) through the IMACS metric proposed in the paper. In contrast, "StyleForge" introduces two methodologies: Single-StyleForge and Multi-StyleForge, focusing on reproducing a target style with only 15 to 20 images. Single-StyleForge not only learns the target style but also uses auxiliary images to help the model understand complex style relationships. Multi-StyleForge improves alignment between the text prompt and the generated image by separating the attributes of the target style into background and person components. "StyleForge" compares its performance with various models like Textual Inversion, DreamBooth, LoRA, and CustomDiffusion, using FID, KID, and CLIP scores as evaluation metrics.

(3) Experiments and results: The experiments in "I²AM" focus on demonstrating that visualizing attention maps can help understand the generation process of models and improve their performance. It primarily uses the VITON-HD dataset to address high-resolution virtual clothing try-on scenarios. In contrast, "StyleForge" conducts experiments on six different artistic styles (e.g., realism, SureB, anime, romanticism, cubism, pixel-art) and compares the performance of personalized text-to-image models, emphasizing the superiority of style learning.

(4) Conclusive difference: "I²AM" focuses on visualizing relationships between images and emphasizes interpretability, while "StyleForge" aims to enhance alignment between text and images through the refined reproduction of artistic styles, thereby improving personalized user experiences. These differences highlight how the goals and approaches of the two studies differ, providing a clear understanding of the problems each model seeks to address.