CS230

Final Report: Achieving Global and Object Consistency in Stable Diffusion Model

Weilun Chen, Lizhong Zhang
chen1108@stanford.edu, lizhongz@stanford.edu

1 Introduction

Stable Diffusion is a widely recognized technique for generating high-quality images. It reconstructs images by employing text guidance derived from Gaussian noise. Nonetheless, maintaining consistency in the generated images presents a significant challenge. This inconsistency originates from two distinct sources:

- The stateless nature of each generation leads to generated images that do not adhere to the visual characteristics of preceding images. Consequently, substantial variations in artistic style between batches of generations are observed.
- Object inconsistency: Images generated to represent the same object, for instance, "a little girl," may display considerable discrepancies in their visual appearances.

In this project, we propose a new method that can tackle consistency in global and object level. To demonstrate our work, we will build a model to generate children story books.

2 Related work

Method	Training Time (min)	Storage	Fidelity	required examples
Dreambooth	30	Few GB	High	10-20
LoRa	<10	Few hundreds MB	Medium	10-20
Textual Inversion	40	Few KB	Low	10-20
Custom Diffusion	40	Few hundreds MB	High	5-10
Encode based domain fine tuning	1	Few GB	High	1

Dreambooth [6] addresses model consistency by fine-tuning the entire stable diffusion with multiple concept images, demonstrating high fidelity but with significant time, space complexity, and reliance on existing domain knowledge.

LoRA [2] offers a more efficient alternative, though it also necessitates existing images and considerable space.

Textual Inversion [1] identifies optimal text embeddings to infer images but is generally slow due to required full back-propagation.

ControlNet [7] introduces conditional control to the Stable Diffusion model, transferring knowledge from the original network to a new domain, which we intend to employ for object-level consistency.

Custom Diffusion [3] targets the Key and Value matrices in the attention layer, thereby reducing output, and introduces a technique for merging trained weights through a combined matrix for feature projection.

A method [4] using extensive domain information for fine-tuning with a single image, similar to Custom Diffusion, focuses on cross-attention layer fine-tuning.

For our work, we experimented ControlNet and Custom Diffusion for local character-level consistency.

3 Dataset

Intriguingly, we explore training using character generation. Leveraging the established domain knowledge of the stable diffusion model, we generate and subsequently partition new character concepts, exemplified in Figure 1. Custom Diffusion is chosen for this unique dataset due to its faster training speed and smaller weight output.

Figure 1: generated character sheet with prompt *consistent character concept on (white background), multiple poses and expressions, small robot wearing a orange overalls, child story book illustration*

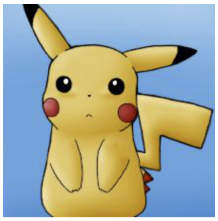
We also explore the style capture. The style is captured using a few pages of a child story book with the hypothesis that training style as a concept will make the model capture the underlying style. Figure 2 shows some of the pages used.

Figure 2: Style images for reference

4 Approaches

4.1 Failed Approach: ControlNet

For ControlNet, we train with x iterations with the aforementioned triplet sampled from the Pokemon dataset. The intuition is that by training ControlNet of concept to concept mapping, it will be able to understand that the input image is providing a concept to the network, and learn to adapt the provided concept into a new setting. However, naively training the network in such a approach generate very unconvincing results which is presented in Figure 3.



(a) Input

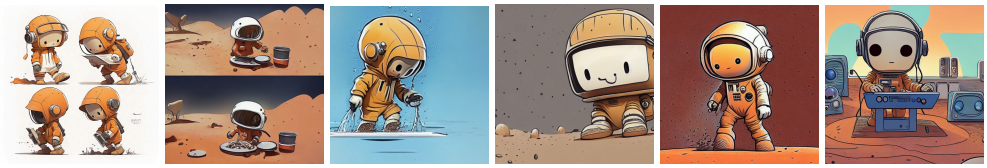
(b) A S^* cat

(c) A S^* cat

Figure 3: ControlNet output

4.2 Custom Diffusion

ControlNet's results underscore the difficulty of inference time concept learning for the model, which instead appears to adapt an image on-the-fly rather than encode the provided concept within its weightings. We addressed this complexity by leveraging model training for concept embedding. Custom Diffusion was chosen due to its compact output size and speedy fine-tuning capabilities. The observation is based on the fact that during the fine tuning process, most of the the weight changes happen in the cross attention layer, and it can achieve comparable result by using Four concept images, generated by the Stable Diffusion model, were used to train the Custom Diffusion model using a learning rate of $1e-5$ for 2500 steps. The process, executed on an RTX 4090 graphics card, took approximately 30 minutes. The results are displayed in Figure 4. By incorporating a degree of training, the model was able to retain the style of the input character and successfully adapt it to new environmental settings.



(a) Input (b) Output 1 (c) Output 2 (d) Output 3 (e) Output 4 (f) Output 5

Figure 4: The input image and the output images.

5 Experiment

Each model we examine is trained utilizing Stable Diffusion 2, as delineated in the study by Rombach et al., 2022 [5]. We establish a comparison base using three primary models: the original Stable Diffusion, Stable Diffusion supplemented by image guidance, and the Dreambooth model. Our input training data is the character stylesheet, as displayed in Figure 1. The evaluation process includes both quantitative and qualitative methods. For the training loss, we use the same loss function proposed by [6] with prior preservation to prevent overfitting[6]:

$$\mathbb{E}_{x,c,\epsilon,\epsilon,t}[w_t\|\hat{x}_\theta(\alpha_t+\sigma_t\epsilon,c)-x\|_2^2+\lambda w_{t'}\|\hat{x}_\theta(\alpha_{t'}x_{pr}+\sigma_{t'}\epsilon',c_{pr})-x_{pr}\|_2^2] \quad (1)$$

All trainings are performed on a RTX 4090 graphic card, for both our method and dreambooth, we use a learning rate of 1e-6 and train for 2500 steps. The concept and style are jointly trained with different identifiers. The final result are produced by using the prompt *A S^* cartoon character on mars in the style of V^**, where S^* is the token for concept capture and V^* represents the style.

5.1 Quantitative Evaluation

We adopt the same strategy employeed by [4], where a method is evaluated using *image alignment* for concept capturing and *text alignment* for generalization capability. The image alignment is generated by computing the average pair-wise identity similarity between each concept's training set and the generated image, while text alignment is obtained by calculating the average CLIP-space similarity between the generated image and its original concept-less prompt. Figure 5 shows the result, where

our method achieves better score in both metrics. It's worth noting that, compared to Dreambooth, which outputs the entire fine-tuned model with 3.5GB size even after model pruning, our method only outputs 98MB output weights. This is because custom diffusion only fine-tunes the cross attention layer in the diffusion UNet, thus resulting a smaller parameter set.

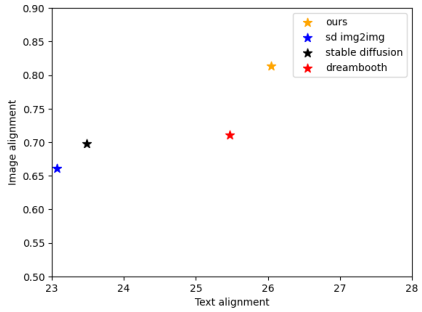


Figure 5: our method achieves best performance in both image alignment and

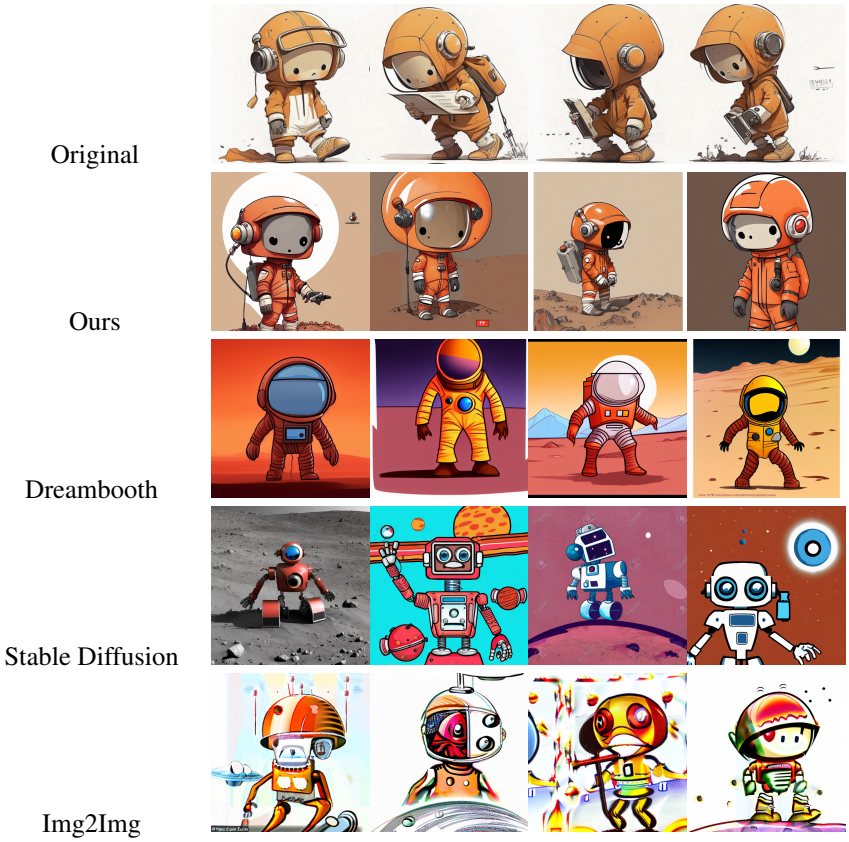


Figure 6: Qualitative comparison of different methods. For Dreambooth and our method, the prompt we're using is *A S^* cartoon character on mars in the style of V^**

5.2 Qualitative Evaluation

Figure 6 shows the qualitative result with the original training set. The results here shows that methods that involves training will produce concepts that better captures the style and concept, compared to pure descriptive methods such as pure stable diffusion model. When provided with an image, Img2Img does capture certain shape and color style, but fail to render the concept in a wholistic way.

6 Code

GitHub Repo: https://github.com/lizhongz/consistent-diffusion

7 Conclusion

In this work, we presented our evaluation for different methods of maintaining concept consistency in image generation, focusing on the global and object level. Our model demonstrated that diffusion model has the ability to produce cohesive, high-quality images that adhere to the visual characteristics of training images, addressing the key problem of inconsistency typically found in generated images.

We employed Custom Diffusion model to ensure both character-level and style-level consistency, and proved that it achieves both small weight output and higher consistency. Our training method enabled the model to retain the concept of the input character and adapt it to new settings effectively.

8 Contributions

Lizhong contributed on implementation of the model we used by adapting open source implementation on our own training dataset, while Weilun contributed by exploring and finding dataset and evaluate the model performance. Both authors equally contributed on exploring existing techniques and reading research paper to propose the training techniques.

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, 2022.

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021.

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

[4] Yuval Atzmon Amit H. Bermano Gal Chechik Daniel Cohen-Or Rinon Gal, Moab Arar. Encoder-based domain tuning for fast personalization of text-to-image models. 2023.

[5] 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 (CVPR)*, pages 10684–10695, June 2022.

[6] 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. 2022.

[7] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.