CONCEPT CONCERTO:

Orchestrating Personalized Styles and Characters in Text-to-Image Diffusion

DLCV 2024 Group TBD

林孟璇 資管五 (B09303009). 郭沛孚 資工一 (R12922217). 楊鈞安 電機二 (R12921A25).

陳柏霖 生機五 (B09611007). 涂銘洋 電信二 (R12942053).

INTRODUCTION

Generating multiple personalized text-to-image concepts with specific styles remains a challenging task. In this project, we introduce a novel two-stage diffusion pipeline with a style injection module and attention separation loss. Compared to previous works, our approach generates images more closely aligned with target styles while preserving object features and overall structure, achieving superior results in both quantitative and qualitative analyses.

METHODOLOGY

TWO-STAGE DIFFUSION PIPELINE

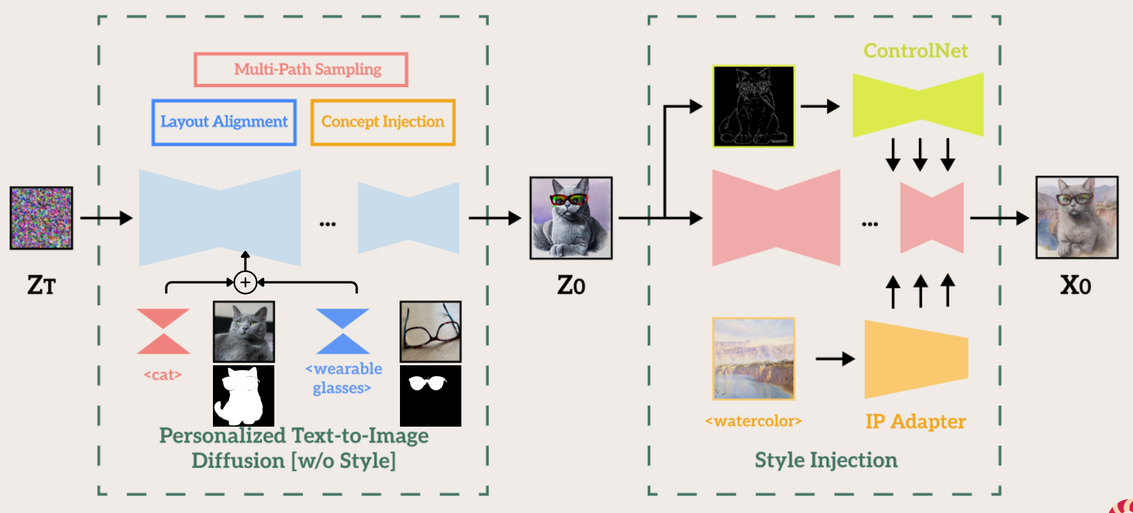


Figure 1. Our two-stage diffusion framework for multi-concept text-to-image personalization

Our method consists of two separable stages. In the first stage, the Concept Conductor framework fuses multiple concepts into a coherent image. The ED-LoRAs were trained separately, and were combined with Stable Diffusion v1.5 while sampling. During inference, we employed Dall-E 3 model to generate reference images, and used segment anything model to obtain masks of reference images and data images, which are used to manipulate the attention layer in the model.

The second stage subsequently infuses $\mathbf{Z_0}$ with a new style using IP-Adapter and preserves its structural layout through ControlNet. The IP-Adapter extracts stylistic features from reference images, while ControlNet leverages control maps to ensure layout fidelity. Figure 2 demonstrates how these conditioning modules enable flexible style injection at varying levels while maintaining spatial coherence.

Figure 2. Style injection using a single style reference

LAYOUT ALIGNMENT LOSS

Building on the layout loss from Concept Conductor, we propose an attention separation loss based on IOU to ensure spatial separation between concepts. F_t^{m} are self-attention features from model m.

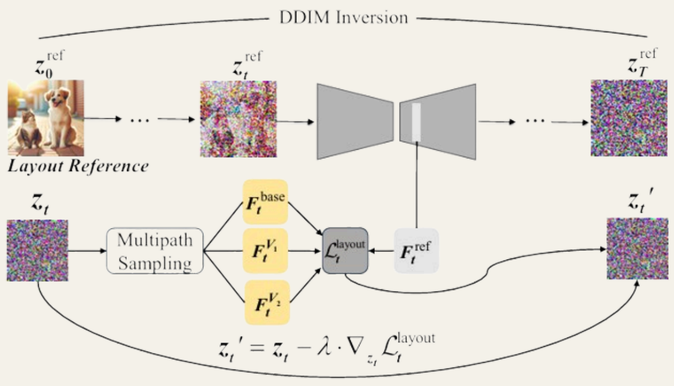


Figure 3. Illustration of layout alignment

$$\mathcal{L}_t^{\text{layout}} = |F_t^{\text{base}} - F_t^{\text{ref}}|_2^2 + \alpha \frac{1}{N} \sum_{i=1}^{N} |F_t^{V_i} - F_t^{\text{ref}}|_2 + \beta \frac{1}{M} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{F_t^{V_i} \cdot F_t^{V_j}}{|F_t^{V_i}|_1 + |F_t^{V_j}|_1 - F_t^{V_i} \quad F_t^{V_j}}$$

Where $M = N(N-1)/2$ represents the number of pairs.

EXPERIMENTS & ABLATION STUDY

QUANTITATIVE COMPARISON

- ***Style-Free Personalization***

	CLIP-I	CLIP-T
Custom Diffusion	76.48	29.33
Concept-Conductor	77.37	**31.30**
Ours (w/ loss aug)	**77.49**	30.99

- ***Style-Conditioned Personalization***

	CLIP-I	CLIP-T
Custom Diffusion	64.87	37.99
LoRA	**66.02**	37.05
Ours	65.79	**38.81**

QUALITATIVE COMPARISON

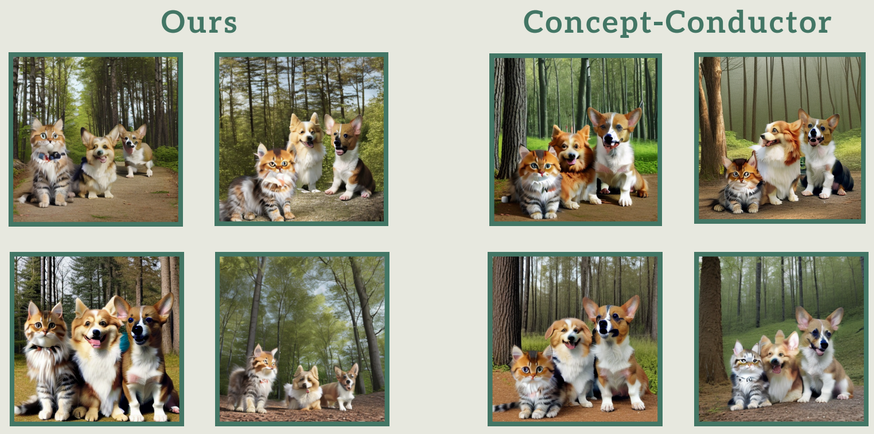
- ***Style-Free Personalization***

"A <cat2> on the right and a <dog6> on the left." | "A <flower_1> in a <vase>." | "A <dog>, a <pet_cat1> and a <dog6> near a forest."

- ***Style-Conditioned Personalization***

"A <cat2> wearing <wearable_glasses> in a <watercolor> style."

- ***Effects of Layout Alignment Loss***



CONCLUSION

In this work, we present a novel two-stage diffusion framework designed for style-adaptive multi-concept personalization, achieving high visual fidelity and consistent style representation. Moreover, we propose an attention separation loss to ensure spatial separation between concepts. Our framework outperforms Custom Diffusion and LoRA in multi-concept text-to-image personalization.

REFERENCE

- Yao et al. "Concept Conductor: Orchestrating Multiple Personalized Concepts in Text-to-Image Synthesis". arXiv preprint arXiv:2408.03632
- Kumari, Nupur, et al. "Multi-concept customization of text-to-image diffusion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
- Ye, Hu, et al. "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models." arXiv preprint arXiv:2308.06721 (2023).