InstantPainting: Expanding GANs for Efficient Text-Conditioned Image Generation Platform

Bing-Kun Bao[1]*, Yefei Sheng[1], Jie Wang[1], Yaning Li[1], Sisi You[1]

[1]Nanjing University of Posts and Telecommunications
bingkunbao@njupt.edu.cn, ysheng990618@gmail.com, jwang_c@njupt.edu.cn, liyaning@njupt.edu.cn, ssyou@njupt.edu.cn

Abstract

Text-conditioned image generation enables cross-modal comprehension. Recent emergence of many platforms have found applications in diverse domains like assisted designing and video gaming. However, there still exist challenges in existing platforms due to their expensive training and time-consuming generation processes. In this paper, we introduce an efficient text-conditioned image generation platform, termed InstantPainting. Unlike existing platforms based on large-scale pre-trained diffusion models, InstantPainting expands generative adversarial networks (GANs) to achieve efficient generation by using only about 3% pre-training data of other platforms. Compared to existing platforms, InstantPainting achieves the following functions at a very low deployment cost and approximately 4 to 5 times faster generation speeds: (1) Multi-category and multi-size image generation (2) Image stylization and controlled generation (3) Creative generation, including the generation of poetry pictures and counterfactual images. The proposed platform provides web application implementations for PC and mobile, users can create high-quality images directly through the user interface.

Introduction

Text-conditioned image generation serves as a crucial task in multi-modal information processing, it utility extends across various of applications, including virtual reality and computer-aided design. However, prevailing platforms grapple with efficiency issues stemming from the burdensome nature of model deployments and the time-intensive generation process. Presently, these platforms predominantly rely on large-scale diffusion models. These models necessitate extensive pre-training, coupled with the consumption of costly computational resources and hiring of model training practitioners. Consequently, the generation process is often characterized by sluggish speeds, with deployment costs at least 256 Tesla V100, which is about $ 1.35 million . The excessively high deployment costs not only impede the flexibility of model updates but also create significant barriers to entry for smaller-scale projects and research endeavors. Consequently, there is an urgent need for more streamlined and cost-effective solutions.

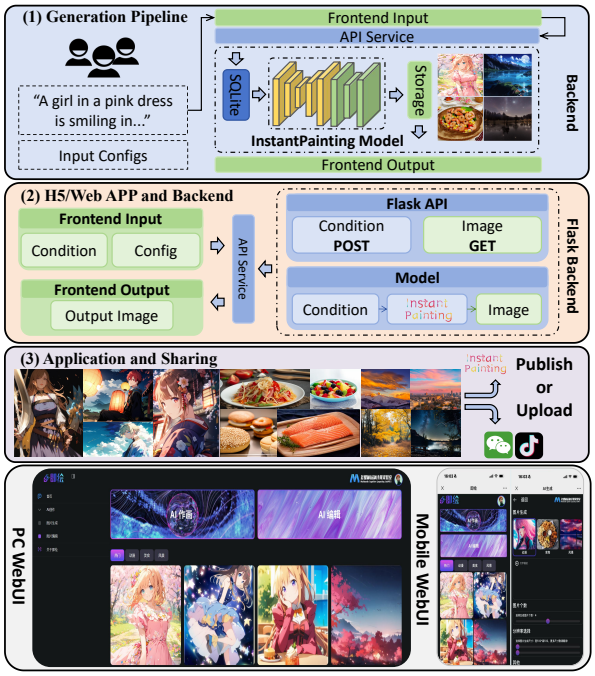

Figure 1: System Framework of InstantPainting

To address the imperative of enhancing generation efficiency and mitigating deployment costs in image generation platforms, we introduce InstantPainting, built upon our cross-model generative model GALIP(Tao et al. 2023). In contrast to existing platforms, InstantPainting excels in delivering faster, more diverse, and meticulously controlled high-fidelity image generation processes. Specifically, the platform supports multiple categories of image generation, including anime-style images, recipe and food generation, landscape and scenery generation. It can also support custom sizes ranging from 512*512 to 1024*1024. Additionally, the platform supports the stylization of generated images or the direct use of style qualifiers for multi-style generation. For complex text input such as "mountains covered with snow and clouds at sunset, sky in the background, flaming clouds, steep cliffs, bright golden sun dappled", InstantPainting excels in controlling details well. Furthermore, users can uti-

lize InstantPainting to unleash their creativity by generating images from poetry or counterfactual scenarios. It even has the capability to generate some dark dishes.

Notably, InstantPainting builds on only about 3% of other platforms' pre-training data and only 4% of the parameters. As shown in Table 1, the deployment cost has also been significantly reduced. In addition, InstantPainting provides an H5 Application for mobile terminals in addition to the PC services provided by the common platforms to provide users with a more convenient image generation services. (Rombach et al. 2022; Xue et al. 2023; Betker et al. 2023)

Method	Parameters	datasets	Deployment cost
SD-XL	6.6B	5B	>256 Tesla V100
RAPHAEL	3B	5B	1000 Tesla A100
DALL E-3	3.5B	5B	1024 Tesla V100
InstantPainting	**0.23B**	**400M**	**16 RTX A6000**

Table 1: The comparison between mainstream platforms

System Framework

Our system consists of two essential components. Firstly, the backend serves as the informational backbone for our applications. Secondly, our frontend application empowers users to interact directly with the interface.

Backend

Model Training: InstantPainting's backend model is based primarily on an extended GALIP (Tao et al. 2023) model. We extend the generator and discriminator, augmenting them with a convolutional network-based discriminator module and a noise selection module within the generator section. Additionally, we upgrade the CLIP model from the original version to VIT-16 (Patashnik et al. 2021). Concurrently, we selected the laion-400M (Schuhmann et al. 2021) as our pre-training dataset and curated a fine-tuning dataset consisting of approximately one million pairs for training each module of InstantPainting (Xu et al. 2024; Sheng et al. 2024; Tao et al. 2025). Furthermore, we devised a multi-stage fine-tuning scheme for InstantPainting, incorporating various screening strategies and training data at each stage to prevent the model from learning noisy pairs.

Model Deployment: Leveraging Flask, we deploy the comprehensive model on the backend, facilitating efficient interaction between the front and backend systems via SQLite Query statement, which is shown in Figure 1. This meticulous backend architecture ensures seamless operation and optimal performance across all aspects of text-conditioned image generation within the platform.

PC/Mobile Application

Our platform provides simultaneous access to H5 web pages on mobile terminals and web pages on PCs. It supports three primary functions: 1) anime-style generation, 2) recipe and food generation, and 3) landscape and scenery generation. On the front end, users input their prompts and configurations, which are then uploaded to our cloud server to update the SQLite database. Subsequently, the database issues instructions from the backend to generate images based on the input. These images are visually returned to the frontend, providing feedback to the user.

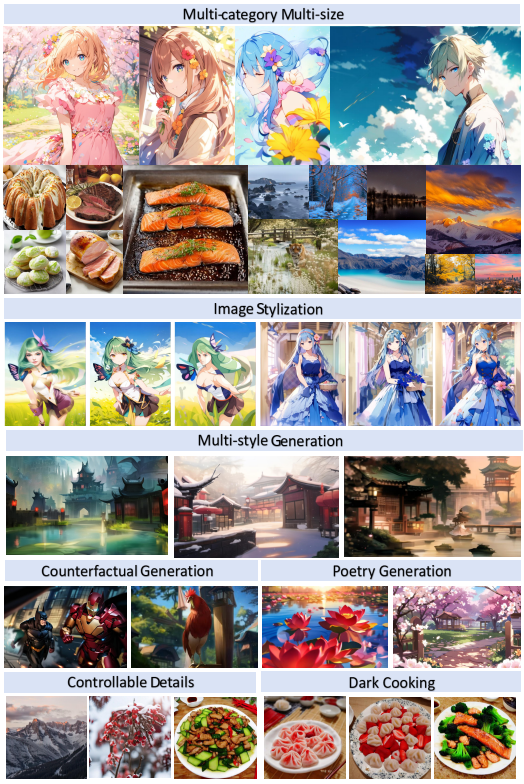

Figure 2: Samples on InstantPainting

User Interface: The application's user interaction page enables users to interact with backend-generated models in a low-latency manner, while preloaded models also enhance the user experience. This functionality can also be utilized to update our build database for real-time builds.

Publish and Share: This feature allows users to upload generated images to our website. Here, other users can view both the text descriptions and associated settings. Additionally, users have the option to share images with friends via social media platforms.

Conclusion

In this technical demo, we introduce an efficient image generation platform called InstantPainting. Our system integrates automated dataset construction with streamlined expanded adversarial generation networks. In contrast to existing platforms, InstantPainting excels in delivering faster, more diverse, and meticulously controlled high-fidelity image generation processes at a very low deployment cost and about 4-5 times faster generation speed.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants (No.62325206, 61936005, 62206132), and the Key Research and Development Program of Jiangsu Province under Grant BE2023016-4, Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX23_1023)

References

Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3): 8.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. 2085–2094.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Sheng, Y.; Tao, M.; Wang, J.; and Bao, B.-K. 2024. ISF-GAN: Imagine, Select, and Fuse with GPT-Based Text Enrichment for Text-to-Image Synthesis. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Tao, M.; Bao, B.-K.; Tang, H.; Wang, Y.; and Xu, C. 2025. StoryImager: A Unified and Efficient Framework for Coherent Story Visualization and Completion. In *European Conference on Computer Vision*, 479–495. Springer.

Tao, M.; Bao, B.-K.; Tang, H.; and Xu, C. 2023. Galip: Generative adversarial clips for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14214–14223.

Xu, M.; Wang, J.; Tao, M.; Bao, B.-K.; and Xu, C. 2024. CookGALIP: Recipe Controllable Generative Adversarial CLIPs With Sequential Ingredient Prompts for Food Image Generation. *IEEE Transactions on Multimedia*.

Xue, Z.; Song, G.; Guo, Q.; Liu, B.; Zong, Z.; Liu, Y.; and Luo, P. 2023. RAPHAEL: Text-to-Image Generation via Large Mixture of Diffusion Paths. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 41693–41706. Curran Associates, Inc.