IP-Prompter: Training-Free Theme-Specific Image Generation via Dynamic Visual Prompting

YUXIN ZHANG, MINYAN LUO, and WEIMING DONG*, MAIS, Institute of Automation, CAS, China and School of Artificial Intelligence, UCAS, China
XIAO YANG, HAIBIN HUANG, and CHONGYANG MA, ByteDance Inc., China
OLIVER DEUSSEN, University of Konstanz, Germany
TONG-YEE LEE, National Cheng-Kung University, Taiwan
CHANGSHENG XU, MAIS, Institute of Automation, CAS, China and School of Artificial Intelligence, UCAS, China

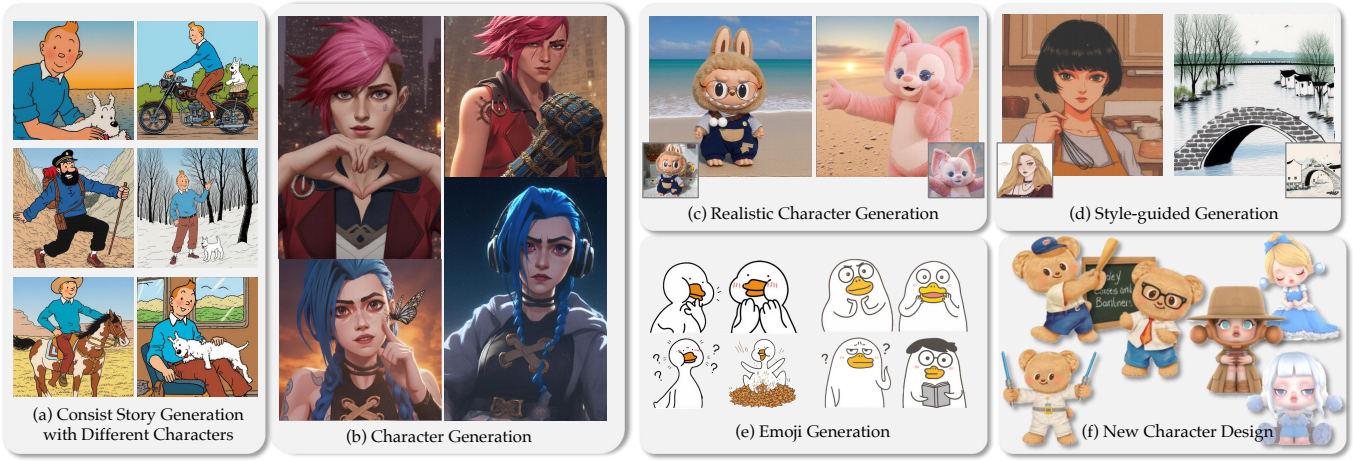


Fig. 1. Training-free generation results of *IP-Prompter*. Each result presented in this paper is generated using a fixed random seed.

The stories and characters that captivate us as we grow up shape unique fantasy worlds, with images serving as the primary medium for visually experiencing these realms. Personalizing generative models through fine-tuning with theme-specific data has become a prevalent approach in text-to-image generation. However, unlike object customization, which focuses on learning specific objects, theme-specific generation encompasses diverse elements such as characters, scenes, and objects. Such diversity also introduces a key challenge: how to adaptively generate multi-character, multi-concept, and continuous theme-specific images (TSI). Moreover, fine-tuning approaches often come with significant computational overhead, time costs, and risks of overfitting. This paper explores a fundamental question: Can image generation models directly leverage images as contextual input, similarly to how large language models use text as context? To address this, we present *IP-Prompter*, a novel training-free TSI generation method. *IP-Prompter* introduces visual prompting, a mechanism that integrates reference images into generative models, allowing users to seamlessly specify the target theme without requiring additional training. To further enhance this process, we propose a Dynamic Visual Prompting (DVP) mechanism, which iteratively optimizes visual prompts to improve the accuracy and quality of generated images. Our approach enables diverse applications, including consistent story generation, character design, realistic character generation, and style-guided image generation. Comparative evaluations against state-of-the-art personalization methods demonstrate that *IP-Prompter* achieves significantly better results and excels in maintaining character identity preserving, style consistency and text alignment, offering a robust and flexible solution for theme-specific image generation. Our project page: https://ip-prompter.github.io/.

CCS Concepts: • **Computing methodologies** → **Image processing**.

Additional Key Words and Phrases: Personalized image generation; Diffusion models; Visual prompting.

ACM Reference Format:
Yuxin Zhang, Minyan Luo, Weiming Dong, Xiao Yang, Haibin Huang, Chongyang Ma, Oliver Deussen, Tong-Yee Lee, and Changsheng Xu. 2025. IP-Prompter: Training-Free Theme-Specific Image Generation via Dynamic Visual Prompting. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25), August*

*Corresponding author: Weiming Dong (weiming.dong@ia.ac.cn)

Authors' Contact Information: Yuxin Zhang; Minyan Luo; Weiming Dong, MAIS, Institute of Automation, CAS, China and School of Artificial Intelligence, UCAS, China, zhangyuxin2020@ia.ac.cn, luominyan21@mails.ucas.ac.cn, weiming.dong@ia.ac.cn; Xiao Yang; Haibin Huang; Chongyang Ma, ByteDance Inc., China, yangxiao.0@bytedance.com, jackiehuanghaibin@gmail.com, chongyangm@gmail.com; Oliver Deussen, University of Konstanz, Germany, oliver.deussen@uni-konstanz.de; Tong-Yee Lee, National Cheng-Kung University, Taiwan, tonylee@ncku.edu.tw; Changsheng Xu, MAIS, Institute of Automation, CAS, China and School of Artificial Intelligence, UCAS, China, csxu@nlpr.ia.ac.cn.

SIGGRAPH Conference Papers '25, Vancouver, BC, Canada

ACM ISBN 979-8-4007-1540-2/2025/08
https://doi.org/10.1145/3721238.3730670

10–14, 2025, Vancouver, BC, Canada. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3721238.3730670

1 Introduction

If a picture is worth a thousand words, then a theme tells an entire story. We define a theme-specific image (TSI) as a visual composition that cohesively integrates characters, objects, and environments within a unified artistic style or narrative framework. TSIs align explicitly with a defined theme or concept, making them essential for thematic communication and audience engagement. These images have applications in areas such as branding, storytelling, and design. Recent advancements in text-to-image generation models have enabled users to synthesize images from text prompts. For generating images of specific concepts, popular methods include personalized techniques such as model fine-tuning [Ruiz et al. 2023], the integration of auxiliary control networks [Mou et al. 2024; Zhang et al. 2023c], and attention exchange mechanisms for concept injection [Chung et al. 2024; Hertz et al. 2024]. However, unlike object customization tasks, which focus on learning specific objects, generating TSIs involves managing a diverse set of elements, including characters, scenes, and objects. For instance, as shown in Fig. 1(a), TSIs of *The Adventures of Tintin* encompass multiple elements such as Tintin, Snowy, and Captain Haddock. This diversity presents a significant challenge: designing methods capable of flexibly and efficiently adapting to multi-character, multi-concept, and continuous generation tasks, all while minimizing costs. Existing fine-tuning approaches struggle with rapid concept switching and introduce high computational and time overhead. Similarly, approaches that rely on auxiliary networks or attention exchange mechanisms often struggle to maintain the identity consistency of characters and objects while necessitating structural modifications to large pre-trained models, thereby introducing additional challenges.

On the other hand, large language models (LLMs) have demonstrated the ability to use user-provided context as knowledge, offering convenience in textual communication. Inspired by this, we extend this paradigm to visual communication in image generation models. Despite the inherent challenges of TSI generation, it offers a unique advantage: the availability of extensive visual context in the form of character, object, and background images, which can be leveraged directly. In this paper, we introduce a visual prompting interaction framework based on image inpainting models. As illustrated in Fig. 2(c), this framework incorporates personalized concepts by directly stitching guiding images as contextual information. Visual prompting eliminates the need for additional networks, modules, or attention mechanisms. This training-free and modification-free approach not only enhances efficiency but also ensures that guidance information remains within the same visual domain as the target output, resulting in precise results.

To address the core challenges of TSI generation, we further propose a Dynamic Visual Prompting (DVP) scheme. DVP matches and arranges visual prompts in real time based on the target theme and user-provided text instructions. Users only need to provide a thematic image collection, while DVP enables precise control over the generation model. The key steps of DVP include analyzing user intentions, matching context information, composing visual prompts, iterative updating, and evaluation. First, DVP automatically extracts key visual elements from the user's input text prompts. Subsequently, it performs visual-textual matching within the theme-specific image collection. The images with the highest matching scores are then composed into visual prompts in a specific arrangement according to an importance assessment. These visual prompts are then updated in self-consistency manner and fed into the generation model. In the last stage, the generated results are evaluated and satisfactory output is returned to the user. By dynamically adjusting reference images according to user instructions, DVP ensures superior flexibility, accuracy, and efficiency. We name our TSI generation method incorporating DVP as *IP-Prompter*.

We also conduct extensive comparisons with a wide range of baseline methods on TSI generation tasks. *IP-Prompter* demonstrates state-of-the-arts performance in theme consistency and text-image alignment. Moreover, *IP-Prompter* can directly assist users in diverse design applications, including consistent story generation, character design, realistic character generation, and style-guided image creation. In summary, our contributions are as follows.

- We introduce a new approach for visual prompting, i.e., interacting with the image generation model in a training-free and modification-free manner through a multi-grid format. We emphasize the importance of optimizing visual prompts in the same way as carefully designing text prompts.
- We propose *IP-Prompter*, a TSI generation method that leverages DVP to dynamically adjust guiding images according to user input. DVP tackles the core challenge of flexibility in TSI generation, and offers exceptional accuracy and efficiency, seamlessly adapting to multi-turn dialogues.
- Extensive experiments demonstrate that *IP-Prompter* achieves state-of-the-art performance in TSI generation tasks. Additionally, *IP-Prompter* supports diverse creative applications, including consistent story generation, character design, realistic character generation, and style-guided image creation.

2 Related Work

Theme-specific and consistency-oriented image generation tasks focus on creating a series of new images that preserve consistent visual and semantic characteristics guided by input directives. Achieving this goal necessitates models capable of reliably maintaining the core identity of personalized content while adapting it to varied contexts. Recent progress in this area has led to the development of diverse approaches, broadly classified into training-based methods and training-free image guidance techniques.

Conventional approaches to personalized generation often depend on fine-tuning models or incorporating supplementary networks to attain precise control over the generated outputs. Techniques such as DreamBooth [Ruiz et al. 2023], LoRA [Hu et al. 2021], and subsequent works [Chen et al. 2024; Jang et al. 2024; Kumari et al. 2023a; Purushwalkam et al. 2024; Shah et al. 2025; Sohn et al. 2024; Xu et al. 2024] focus on fine-tuning the model's attention mechanisms to embed personalized concepts effectively. Another class of methods, including Textual Inversion [Gal et al. 2023] and subsequent works [Avrahami et al. 2023; Huang et al. 2024c; Li et al. 2024a; Vinker et al. 2023; Wei et al. 2023; Zeng et al. 2024b; Zhang

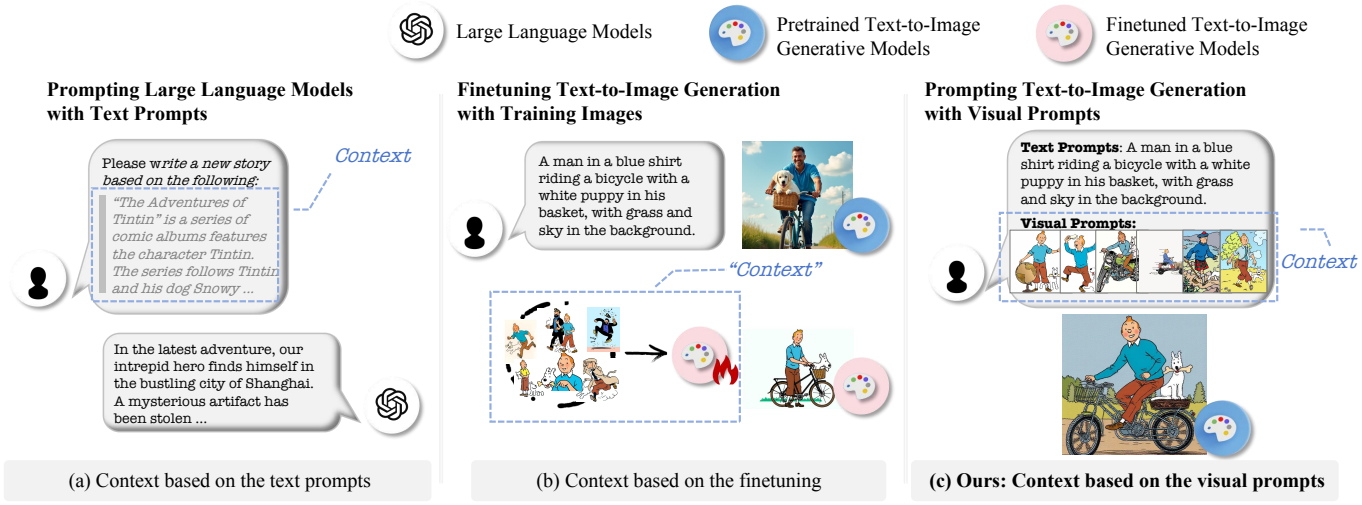


Fig. 2. **Schematic illustration of visual promoting:** (a) Text prompting in LLMs provides context and knowledge for the model to generate target content. (b) Existing personalized methods inject concepts into the model by fine-tuning the model, training a reference network, or altering the model structure to achieve thematic control. (c) Our proposed visual prompting based on inpainting represents a new model interaction paradigm, where visual prompts directly provides contextual information to the model, enabling fast and efficient controllable generation without the need to modify the generative model.

et al. 2023a,b], invert guiding images into the textual space, enabling concept guidance through textual prompts. Frameworks like T2I-Adapter [Mou et al. 2024], ControlNet [Zhang et al. 2023c], and related approaches [Gal et al. 2024; Parmar et al. 2025; Wang et al. 2024b,a; Ye et al. 2023; Zong et al. 2024] incorporate additional reference networks to encode and inject guidance concepts directly into the generation pipeline. In particular, these methods do not require additional fine-tuning during inference. While these approaches facilitate personalized generation, they typically require extensive parameter updates and incur substantial computational overhead, limiting their efficiency in scenarios that demand scalability or rapid adaptation.

Recent trends in personalized generation focus on training-free methods, which achieve concept injection by manipulating features or utilizing the capabilities of pre-trained models. For instance, StyleAligned [Hertz et al. 2024] and concurrent techniques [Alaluf et al. 2024; Chung et al. 2024; Deng et al. 2024] achieve style-consistent image generation by swapping keys and values between reference and generated images in the attention layers. ConsiStory [Tewel et al. 2024] introduces a subject-driven shared attention block and correspondence-based feature injection mechanism. FreeCustom [Ding et al. 2024] introduces a multi-reference self-attention mechanism and a weighted mask strategy to inject concepts. The above methods requires manipulations of model structures and are challenging generation characters of different actions.

Some methods explore visual prompting in the field of computer vision. Bar et al. [2022] propose that various computer vision tasks can be treated as grid inpainting problems. They also highlight the importance of diverse data. Zhang et al. [2023d] conduct an investigation on the impact of in-context examples in computer vision and found that the performance is highly sensitive to the choice of examples. In the field of image generation, some methods [Yeh et al. 2024] utilize image prompting to achieve guidance. Analogist [Gu et al. 2024b] explores the performance of visual in-context learning by proposing self-attention cloning. IC-LoRA [Huang et al. 2024b] maps reference images to generate output via LoRA training. Group diffusion transformers [Huang et al. 2024a] establishes associations using a group-attention block. JeDI [Zeng et al. 2024a] injects reference information through coupled self-attention. OminiControl [Tan et al. 2024] proposes a minimal control framework to achieve conditioning. Diptych Prompting [Shin et al. 2024] leverages the inherent consistency capabilities of pre-trained models and integrates a single reference image via attention re-weighting. In contrast, we introduce prompt engineering to the visual domain by treating the dataset as an independent prompt bank that provides visual elements, rather than embedding dataset-specific information into model parameters.

3 Method

Dynamic Visual Prompting (DVP) is designed to generate novel images that align with a user-provided image set while adhering to the content of a textual prompt. Inspired by human creative processes, theme-specific creation typically involves defining the subject, designing its appearance, iteratively refining the design, and producing high-quality outputs. Therefore, the challenges in theme-specific generation can be distilled into the following key tasks: (1) Clarifying the subject of creation: How can user intentions be effectively decomposed and interpreted to produce meaningful outputs? (2) Establishing textual-visual connections: How can textual prompts be linked to visual elements to ensure generated outputs reflect the desired content and appearance? (3) Optimizing generation results: How can visual prompts be organized and selected to maximize output quality? Furthermore, how can the generated results be systematically evaluated?

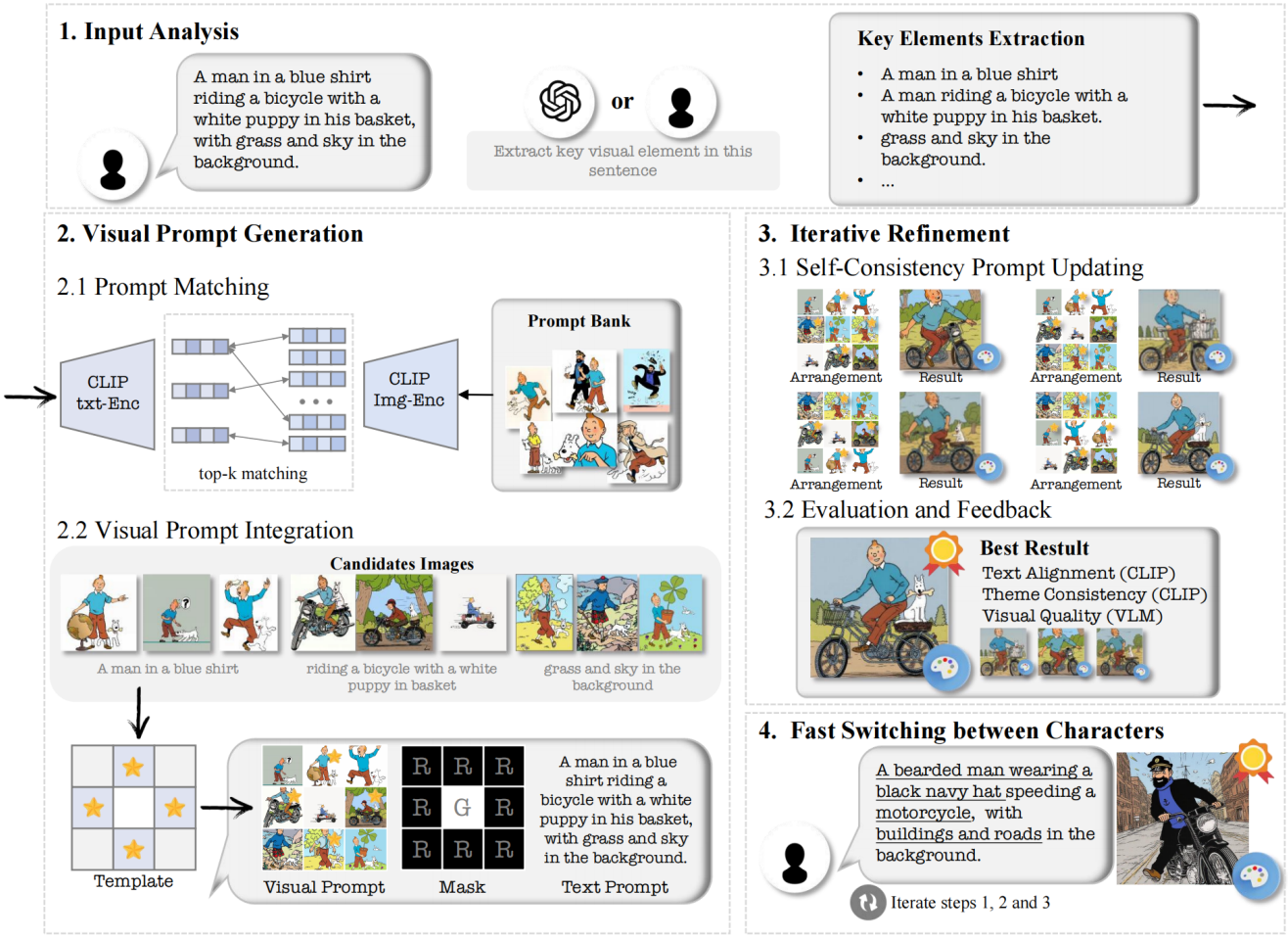


Fig. 3. **Pipeline of *IP-Prompter*:** Dynamic visual prompting (DVP) includes three key stages: (1) Comprehending user intent and extracting key elements; (2) Matching and generating visual prompts; and (3) Updating and evaluating prompts through self-consistency. This way (4) DVP enables effortless transition between diverse creative subjects, thereby enhancing the flexibility and efficiency of content generation.

To address these challenges, we propose DVP (see Fig.3), a framework composed of three steps: (1) user intent understanding and key element extraction; (2) visual prompt matching and generation; and (3) self-consistent prompt updating and evaluation. DVP is designed to (4) seamlessly switch between different creative subjects, providing enhanced flexibility and efficiency in content generation.

User Intent Understanding and Key Element Extraction. The first step in DVP is to interpret the user's creative intentions and identify the central subject of the creation. To accomplish this, DVP employs visual element extraction based on the user's input textual prompt. This process can be performed either automatically, using LLMs, or manually, through user input. Specifically, we utilize a structured text command structured as follows: "*Please extract N key visual elements from this paragraph.*" Here, N is either set to a default value (e.g., $N = 3$) or specified by the user. This process yields a set of key elements $\{element_0, \ldots, element_N\}$. Identifying these elements is critical, as failing to do so may result in visual prompts lacking the specific semantic details necessary for accurate and meaningful content generation.

Visual Prompt Matching and Generation. The second step establishes a connection between the textual and visual prompts by leveraging the CLIP model [Radford et al. 2021] to map those elements into a shared embedding space. The CLIP text encoder and image encoder are used to transform the extracted key elements and images into text embeddings $\{E_0, \ldots, E_N\}$, and image embeddings $\{I_0, \ldots, I_M\}$, where M corresponds to the number of user-provided images. The similarity between each text embedding and image embedding is then computed as: $similarity_{i,j} = \frac{E_i \cdot I_j}{\|E_i\| \|I_j\|}$. For each visual element, the top-K images with the highest similarity scores are selected, resulting in $N \times K$ image candidates. We set $N = 3$ and $K = 3$, and this process yields a 3×3 image template (see Fig. 3).

The arrangement of visual prompts within the mask-filling generative model plays a crucial role in the generation process. To analyze this influence, we conduct attention visualization experiments. For

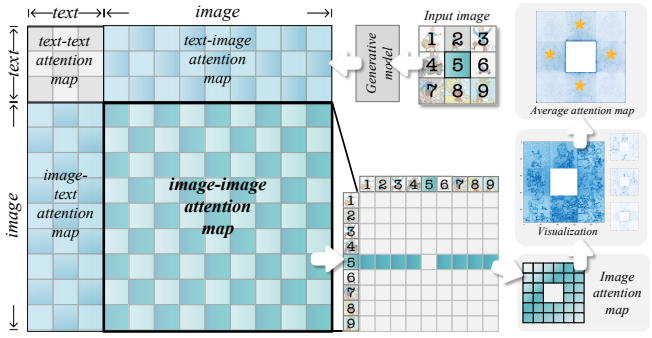


Fig. 4. **Attention maps computed during model inference**. Here, a lots of visual prompts are combined in various arrangements. As shown in the upper right corner, areas with deeper colors are allocated more attention. These are referred to as significant regions and are marked with star symbols.

the same group of reference images, inference is performed with different grid arrangement. We calculate the average attention maps (annular) of the reference image to the generated image obtained during all inferences, and the results are shown in the upper right corner of Fig. 4. This reveals that specific starred regions exhibit higher average attention intensities (deeper colors) compared to other areas, indicating their significance within the image grid. Consequently, the most important images are placed in these high-attention regions, and are complemented by a central mask. These inputs are then fed into an image inpainting model, which synthesizes the target image within the central mask, guided by the user-provided textual input. We use six arrangements in all experiments. Three visual elements are fully permuted across rows.

Self-consistent Prompt Updating and Evaluation. For images within the starred/unstarred region, variations in images arrangement can still influence the generated images. Similar to the behavior observed in LLMs, where semantically identical prompts may yield divergent outputs, a self-consistent strategy [Wang et al. 2022] was adopted to address this variability. In LLMs, self-consistency typically involves evaluating outputs generated from different prompt arrangements and selecting the most plausible result through a voting mechanism. Inspired by dynamic prompting [Yang et al. 2023] and self-consistency in LLMs, we introduced a self-consistent prompt updating approach. As shown in the iterative refinement stage of Fig. 3, the model iteratively rearranges a set of visual prompts to produce diverse outputs. The best match to the user's requirements is then selected using quantitative metrics, including text-image consistency (CLIP text score), thematic consistency (CLIP image score), and visual quality (evaluated using visual-language models). In this way, DVP can help users obtain the most suitable combination of reference images and reduce their burden on selection.

4 Experiments

Evaluation baselines. We compared our method with state-of-the-arts personalization methods, including FLUX 1.0 [Labs 2023], Textual Inversion (TI) [Gal et al. 2023] with SDXL [Podell et al. 2023], DreamBooth+LoRA [Ruiz et al. 2023] with FLUX, Kolors [Team 2024] Character with FLUX, IP-Adapter [Ye et al. 2023] with FLUX, EasyRef [Zong et al. 2024], FreeCustom [Ding et al. 2024], and ConsiStory [Tewel et al. 2024].

Implementation details. We used FLUX-Fill 1.0 [Labs 2023] with the default hyper-parameters in all our experiments. The guidance scale is 30 and the number of inference steps is 50. *IP-Prompter* generates strong results with 15 diverse images per character and excels with 30. The matching and evaluation process take about 10 seconds. The synthesis process takes about 30 seconds for a 512×512 image, which is comparable with baseline training-free methods and faster than minute-level finetuning-based methods and day-level adapter-based methods. Training-based methods like LoRA [Hu et al. 2021] offer higher adaptability and precision at the cost of additional setup, whereas training-free methods prioritize convenience and general applicability.

4.1 Qualitative Evaluation

We compare our method with eight state-of-the-arts personalization methods. As illustrated in Fig. 5, we categorize the evaluated methods into three groups: training-free methods, methods that require training an additional control network, and methods that involve separate learning for each concept. Notably, except for FLUX, none of the other methods explicitly leverages thematic information.

ConsiStory [Tewel et al. 2024] demonstrates the capability to generate a series of consistent images in text-to-image scenarios but struggles to achieve precise control in image-guided generation tasks. FreeCustom [Ding et al. 2024] is able to generate consistent objects, but has difficulty with actions and background changes EasyRef [Zong et al. 2024] utilizes the representation capabilities of multi-modal models and produces favorable results in certain scenarios, as shown in the 4^{th} row of Fig. 5. However, it fails to preserve identity information when dealing with non-human domains, such as cartoons and animals. Kolors-Character [Team 2024] and IP-Adapter [Ye et al. 2023] are based on single-image guidance, effectively captures basic image characteristics, such as cartoon styles, and exhibits strong text-image consistency. However, they struggle with precise character control. For example, Kolors fails to preserve character identity in the 1^{st}, 2^{nd}, 5^{th}, and 6^{th} rows, while IP-Adapter fails in the 1^{st}, 3^{rd}, 5^{th}, and 6^{th} rows. TI [Gal et al. 2023] achieves commendable generation quality and text-image consistency, but struggles to capture intricate character appearances, particularly when character actions vary significantly. DreamBooth+LoRA [Ruiz et al. 2023] delivers the best results among the baseline methods. However, slight compromises in overall style and character identity are noticeable in the 1^{st} and 2^{nd} rows of Fig. 5. In contrast, as demonstrated in the 2^{nd} column of Fig. 5, our method achieves superior thematic and text-image consistency. *IP-Prompter* effectively controls diverse character types, enabling them to perform dynamic actions and appear in novel scenes, while preserving fine-grained details, such as clothing logos. Notably, the results of *IP-Prompter* are difficult to distinguish from the real ones. For fair comparison, all results are generated using the a single random seed.

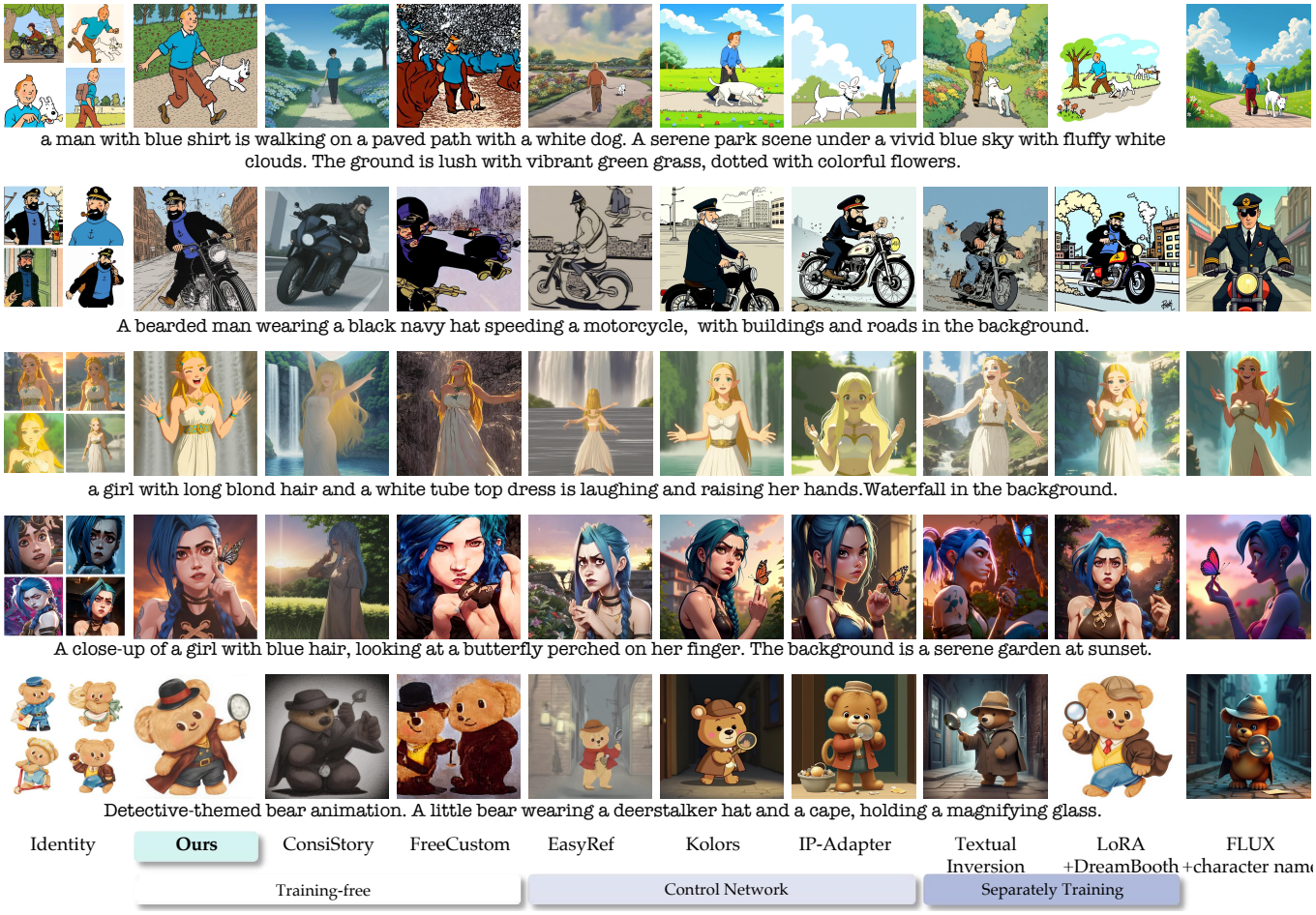


Fig. 5. **Qualitative Results.** We compare *IP-Prompter* with the SOTA personalization methods, including FLUX 1.0, Textual Inversion (TI) with SDXL, DreamBooth+LoRA with FLUX, Kolors Character with FLUX, IP-Adapter with FLUX, EasyRef, FreeCustom and ConsiStory.

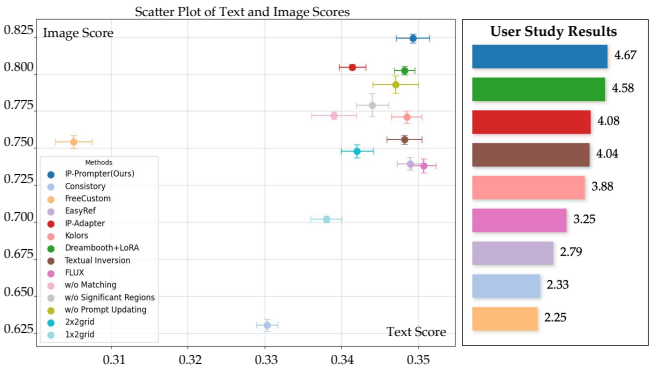


Fig. 6. **Quantitative evaluation and user study results.** *IP-Prompter* achieves comparable scores to the fine-tuned FLUX model.

4.2 Quantitative Evaluation

We employ two metrics for quantitative evaluation. We select ten familiar themes. For each theme, we use five prompts, generating ten images per prompt with different random seeds, resulting in a total of 500 images for each method. As shown in Fig. 6, the vertical axis represents the *image similarity*, measured as the pairwise CLIP cosine similarity between the reference images and the generated images. The higher the better thematic fidelity. The horizon axis represents the *text similarity*, measured as the CLIP similarity between all generated images and their textual conditions. The higher the better editability. Our method achieves the highest thematic consistency while maintaining a text instruction-following capability comparable to FLUX. *IP-Prompter* achieves performance comparable to the fine-tuned FLUX model using the DreamBooth+LoRA approach. For user study, we invited 50 participants to rate the results generated by each method for 10 common themes on a scale of 0 to 5, and the results are shown in the right part of Fig. 6. *IP-Prompter* received the highest user preference.

(a) Ablation results of different components in DVP

(b) Ablation results of different number of references

Fig. 7. Ablation study results.

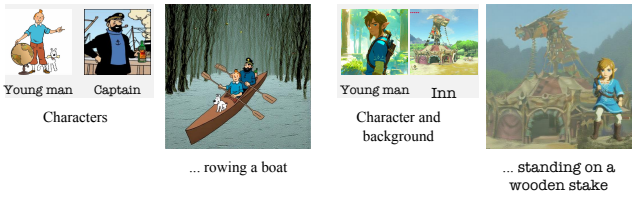
4.3 Ablation Study

As shown in Figs. 6 and 7(a), we conduct an ablation study on different steps in DVP. The ablation of each component contain 100 generated results. Without the self-consistency prompt updating process, it is difficult for the model to match the most suitable image in the most important position, resulting in a loss of detail, such as the absence of the red coat in the 2nd row. Without attention-based arrangement, this detail loss is exacerbated, for instance, the identity in the 2nd row loses. Without prompt matching, the guiding image and the target content are misaligned, which causes inconsistencies in the characters. Finally, without any DVP mechanism, it is difficult to generate images with consistent style and characters, demonstrating the capability of the visual prompting mechanism.

As shown in Fig. 7(b), we conducted an ablation study on different numbers and formats of visual prompting. The black regions in the mask indicate the position of the reference image, while the white regions represent the canvas generated by the model. From left to right, the configurations correspond to nine, four, and two reference images, respectively. It is evident that as the number of reference images decreases, the generated results suffer from style degradation and even character identity loss. A reduced number of reference images fails to provide sufficient context for the model, and this lack of diversity causes the model to ignore the influence of the reference images when generating new content. The quantitative ablation study results in Fig. 6 demonstrate the importance of the number of reference images in maintaining character consistency. Reducing the grid size results in a significant decrease in image scores.

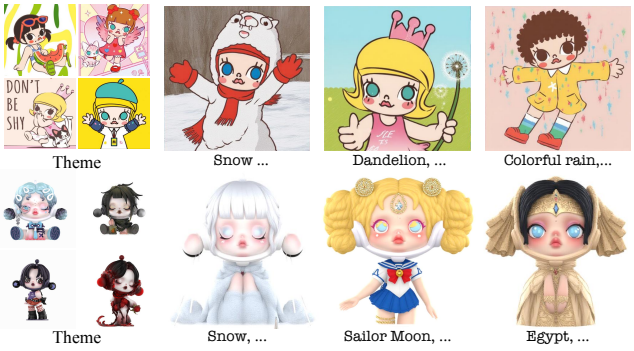
5 Applications

Multi-Concept Generation within a Single Theme. One core challenge of multi-concept generation is synthesis of images that feature multiple elements, including character interactions, backgrounds,

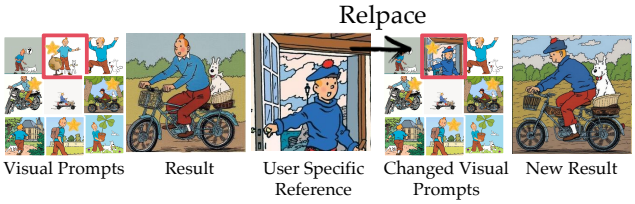


(a) Multi-concept-guided generation results.

(b) Realistic character generation results.



(c) New character generation results.



(d) User-specific refinement results.

Fig. 8. Our results of diverse applications.

and objects, while maintaining consistency and meaningful interaction between these elements. Some methods have explored multi-concept generation [Avrahami et al. 2023; Gu et al. 2024a; Jiang et al. 2024; Jin et al. 2024; Kong et al. 2024; Kumari et al. 2023b; Liu et al. 2023; Xu et al. 2023; Yeh et al. 2024; Zhang et al. 2024], but these methods require additional training or modification to the network structure. Fig. 8(a) shows the ability of *IP-Prompter* to generate complex combinations of characters and backgrounds. *IP-Prompter* cohesively integrates multiple concepts and facilitates interactions

(a) Story generation results.

(b) Emoji sticker generation in story generation setting.

Fig. 9. Generation results for comic stories and emoji stickers.

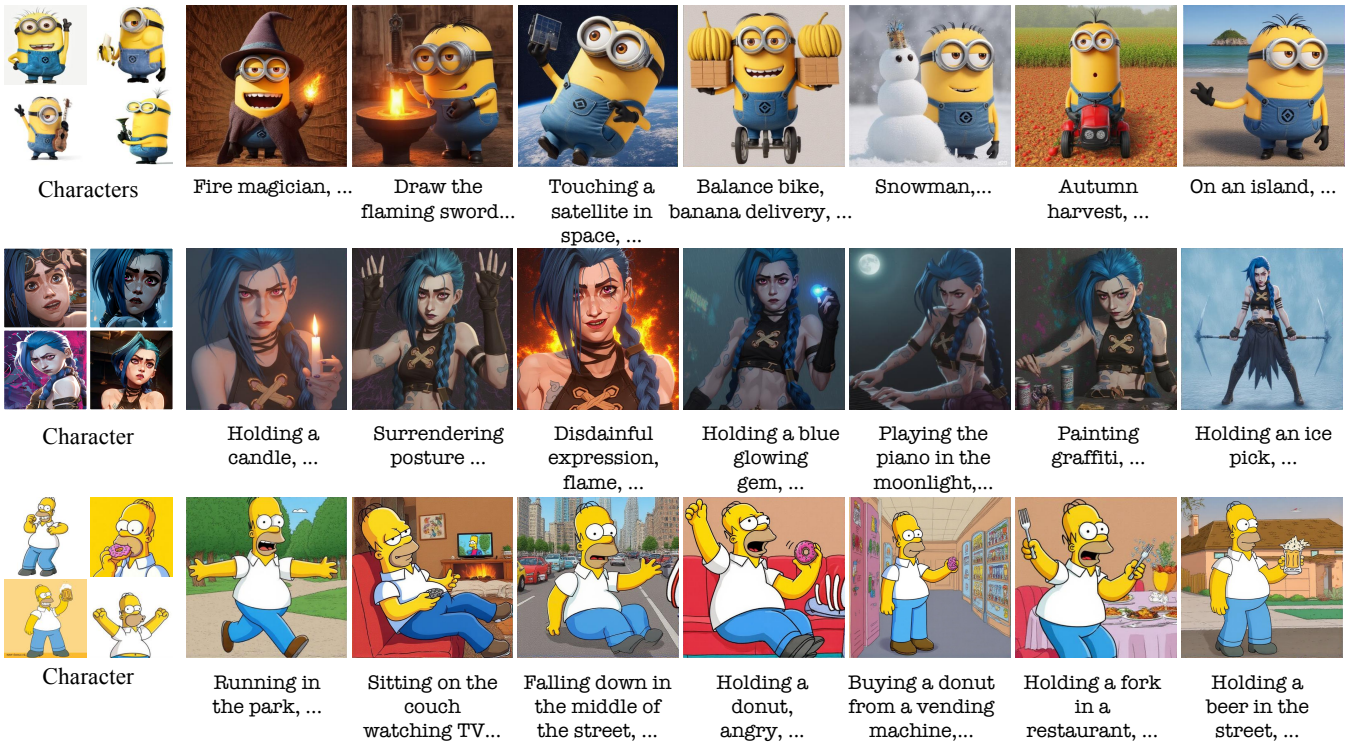


Fig. 10. More results of story generation.

between them, unlocking possibilities for intricate storytelling and character-driven narratives.

Realistic Character Generation in Photographic Contexts. *IP-Prompter* enables the generation of realistic characters seamlessly integrated into photographic backgrounds. The primary challenge of this task lies in achieving consistent character identity while maintaining editability. Comparisons with state-of-the-arts methods highlight *IP-Prompter* 's superior performance in maintaining consistency, as demonstrated in Fig. 8(b). These advancements hold considerable potential for applications such as advertisement generation.

New Character Design. The generation of new characters is less explored [Richardson et al. 2024]. *IP-Prompter* facilitates character design by generating images that adhere to the stylistic and character-specific features of a target theme, while introducing novel concepts. *IP-Prompter*'s visual prompting template design and DVP facilitates diverse content generation. As shown in Fig. 8(c), experiments validate *IP-Prompter's* ability to maintain the style and character appearances. *IP-Prompter* provides both professional designers and hobbyists a versatile tool for sparking creativity.

User-Specific Refinement. As shown in Fig. 8(d), DVP allows users to inject personalized images at specified locations in the visual prompts, achieving more refined concept customization.

Consistent Story Generation. Here we aim to create sequences of images that consistently maintain certain character identities, depict diverse narratives, and facilitate transitions between multiple characters. These are also challenges in the story generation task. Some methods have explored story generation using text-to-image models [Ahn et al. 2023; Gong et al. 2023; He et al. 2023; Liu et al. 2024; Maharana et al. 2022; Mao et al. 2024; Tao et al. 2024; Wang et al. 2024c; Yang et al. 2024; Zhou et al. 2024b,a], but these methods require additional training to maintain character identities. In contrast, *IP-Prompter* leverages visual prompting to effectively maintain identity within the image domain. The DVP mechanism enables adaptive matching of visual prompts to accommodate varying plots and characters. As shown in Figs. 9 and 10, *IP-Prompter* not only creates novel scenes beyond the scope of the original theme, such as characters "riding in space", but also seamlessly incorporates iconic visual elements from the source theme, like the "orange astronaut suit". These capabilities highlight the potentiality of *IP-Prompter* in picture book creation, emoji sticker creation and educational storytelling.

Consistent Style Image Generation. Artistic style presents distinct challenges in image generation, particularly in maintaining stylistic coherence while introducing diverse content. Some methods have explored style-guided generation using text-to-image models [Chung et al. 2024; Hertz et al. 2024; Jeong et al. 2024; Junyao et al. 2024; Li et al. 2024b; Sohn et al. 2024; Wang et al. 2024b; Zhang et al. 2023b], but these methods often require additional training or modification of the network structure. As demonstrated in Fig. 11, experiments

Fig. 11. Style-guided generation results.

Fig. 12. Failure cases due to data limitation.

conducted on two distinct artistic styles validate *IP-Prompter* 's ability to achieve consistent style generation.

6 Limitations

IP-Prompter can generate diverse actions and backgrounds that do not exist in the dataset, but it may be limited by overly homogeneous or insufficient data. We recommend that users provide images of the target characters in multiple poses and varied backgrounds. As shown in Fig. 12, if only facial images of a character are provided and the model is tasked with generating a full-body image, the results may appear reasonable but lack consistency with the actual target. Beyond collecting richer data, this limitation could be addressed through data augmentation leveraging generative models.

7 Conclusion

This paper proposes *IP-Prompter*, a novel training-free, modification-free method of theme-specific image generation with high flexibility and accuracy. We introduce visual prompting, a form of interaction with generative models, which can provide more accurate and direct guidance for models in the visual domain. Our dynamic visual prompting pipeline leverages the capabilities of multi-modal models and LLMs while utilizing data-driven advantages to meet the demanding requirements of flexible theme-specific generation. In comparisons with several state-of-the-art baseline methods, *IP-Prompter* achieves superior results in both qualitative and quantitative evaluations. Following a training-free technical paradigm, *IP-Prompter* delivers performance comparable to fine-tuned models, with the added benefit of producing highly realistic outputs. Experimental results demonstrate the feasibility and effectiveness of *IP-Prompter* across a variety of applications, and its low cost lowers the barrier to use. Additionally, the DVP working mode highlights the potential of inpainting-based visual models in enhancing image generation processes. We believe that more targeted training and fine-tuning approaches will further unlock the potential of more efficient, controllable generation in future applications.

Acknowledgments

This work was supported in part by the Beijing Natural Science Foundation under nos. L221013 and QY24384, in part by the Beijing Training Program of Innovation and Entrepreneurship for Undergraduates under no. S202414430024, in part by the National Science and Technology Council under no. 111-2221-E-006-112-MY3, Taiwan, in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy-EXC 2117-422037984.

References

Daechul Ahn, Daneul Kim, Gwangmo Song, Seung Hwan Kim, Honglak Lee, Dongyeop Kang, and Jonghyun Choi. 2023. Story Visualization by Online Text Augmentation with Context Memory. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 3125–3135.

Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2024. Cross-Image Attention for Zero-Shot Appearance Transfer. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA). Article 132, 12 pages.

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. 2023. Break-A-Scene: Extracting Multiple Concepts from a Single Image. In *SIGGRAPH Asia 2023 Conference Papers* (Sydney, NSW, Australia). Association for Computing Machinery, New York, NY, USA, Article 96, 12 pages.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. 2022. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems* 35 (2022), 25005–25017.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. 2024. AnyDoor: Zero-shot object-level image customization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6593–6602.

Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. 2024. Style Injection in Diffusion: A Training-free Approach for Adapting Large-scale Diffusion Models for Style Transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8795–8805.

Yingying Deng, Xiangyu He, Fan Tang, and Weiming Dong. 2024. Z*: Zero-shot Style Transfer via Attention Reweighting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6934–6944.

Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. 2024. FreeCustom: Tuning-Free Customized Image Generation for Multi-Concept Composition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9089–9098.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *International Conference on Learning Representations (ICLR)*.

Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2024. Lcm-lookahead for encoder-based text-to-image personalization. In *European Conference on Computer Vision*. Springer, 322–340.

Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. 2023. Interactive Story Visualization with Multiple Characters. In *SIGGRAPH Asia 2023 Conference Papers* (Sydney, NSW, Australia). Article 101, 10 pages.

Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. 2024a. Mix-of-Show: Decentralized Low-Rank Adaptation for Multi-Concept Customization of Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36.

Zheng Gu, Shiyuan Yang, Jing Liao, Jing Huo, and Yang Gao. 2024b. Analogist: Out-of-the-box visual in-context learning with image diffusion model. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–15.

Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. 2023. Animate-A-Story: Storytelling with Retrieval-Augmented Video Generation. *arXiv preprint arXiv:2307.06940* (2023).

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. 2024. Style Aligned Image Generation via Shared Attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4775–4785.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Huanzhang Dou, Yupeng Shi, Yutong Feng, Chen Liang, Yu Liu, and Jingren Zhou. 2024a. Group diffusion transformers are unsupervised multitask learners. (2024).

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. 2024b. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775* (2024).

Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin C.K. Chan, and Ziwei Liu. 2024c. ReVersion: Diffusion-Based Relation Inversion from Images. In *SIGGRAPH Asia 2024 Conference Papers* (Tokyo, Japan) *(SA '24)*. Association for Computing Machinery, New York, NY, USA, Article 4, 11 pages.

Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. 2024. Identity decoupling for multi-subject personalization of text-to-image models. *arXiv preprint arXiv:2404.04243* (2024).

Jaeseok Jeong, Junho Kim, Yunjey Choi, Gayoung Lee, and Youngjung Uh. 2024. Visual Style Prompting with Swapping Self-Attention. *arXiv preprint arXiv:2402.12974* (2024).

Jiaxiu Jiang, Yabo Zhang, Kailai Feng, Xiaohe Wu, and Wangmeng Zuo. 2024. MC^2: Multi-concept Guidance for Customized Multi-concept Generation. *arXiv preprint arXiv:2404.05268* (2024).

Chen Jin, Ryutaro Tanno, Amrutha Saseendran, Tom Diethe, and Philip Alexander Teare. 2024. An Image is Worth Multiple Words: Discovering Object Level Concepts using Multi-Concept Prompt Learning. In *International Conference on Machine Learning (ICML)*.

Gao Junyao, Liu Yanchen, Sun Yanan, Tang Yinhao, Zeng Yanhong, Chen Kai, and Zhao Cairong. 2024. StyleShot: A Snapshot on Any Style. *arXiv preprint arxiv:2407.01414* (2024).

Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. 2024. OMG: Occlusion-friendly Personalized Multi-concept Generation in Diffusion Models. In *European Conference on Computer Vision (ECCV)*. Springer, 253–270.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023a. Multi-concept customization of text-to-image diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1931–1941.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023b. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1931–1941.

Black Forest Labs. 2023. FLUX. https://github.com/black-forest-labs/flux.

Wen Li, Muyuan Fang, Cheng Zou, Biao Gong, Ruobing Zheng, Meng Wang, Jingdong Chen, and Ming Yang. 2024b. StyleTokenizer: Defining Image Style by a Single Instance for Controlling Diffusion Models. In *European Conference on Computer Vision (ECCV)*. Springer, 110–126.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. 2024a. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8640–8650.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. 2024. Intelligent Grimm - Open-ended Visual Storytelling via Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6190–6200.

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. 2023. Cones: Concept Neurons in Diffusion Models for Customized Generation. In *International Conference on Machine Learning (ICML)*.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2022. StoryDALL-E: Adapting Pretrained Text-to-Image Transformers for Story Continuation. In *European Conference on Computer Vision (ECCV)*. Springer, 70–87.

Jiawei Mao, Xiaoke Huang, Yunfei Xie, Yuanqi Chang, Mude Hui, Bingjie Xu, and Yuyin Zhou. 2024. Story-Adapter: A Training-free Iterative Framework for Long Story Visualization. *arXiv preprint arXiv:2410.06244* (2024).

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 38. 4296–4304.

Gaurav Parmar, Or Patashnik, Kuan-Chieh Wang, Daniil Ostashev, Srinivasa Narasimhan, Jun-Yan Zhu, Daniel Cohen-Or, and Kfir Aberman. 2025. Object-level Visual Prompts for Compositional Image Generation. *arXiv preprint arXiv:2501.01424* (2025).

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *International Conference on Learning Representations (ICLR)*.

Senthil Purushwalkam, Akash Gokul, Shafiq Joty, and Nikhil Naik. 2024. Bootpig: Bootstrapping zero-shot personalized image generation capabilities in pretrained diffusion models. *arXiv preprint arXiv:2401.13974* (2024).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*. 8748–8763.

Elad Richardson, Kfir Goldberg, Yuval Alaluf, and Daniel Cohen-Or. 2024. ConceptLab: Creative Concept Generation using VLM-Guided Diffusion Prior Constraints. *ACM Transactions on Graphics* 43, 3, Article 34 (June 2024), 14 pages.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22500–22510.

Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. 2025. ZipLoRA: Any Subject in Any Style by Effectively Merging LoRAs. In *European Conference on Computer Vision (ECCV)*. Springer, 422–438.

Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. 2024. Large-Scale Text-to-Image Model with Inpainting is a Zero-Shot Subject-Driven Image Generator. *arXiv preprint arXiv:2411.15466* (2024).

Kihyuk Sohn, Lu Jiang, Jarred Barber, Kimin Lee, Nataniel Ruiz, Dilip Krishnan, Huiwen Chang, Yuanzhen Li, Irfan Essa, Michael Rubinstein, and Dilip Krishnan. 2024. StyleDrop: Text-to-Image Synthesis of Any Style. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. 2024. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098* 3 (2024).

Ming Tao, Bing-Kun Bao, Hao Tang, Yaowei Wang, and Changsheng Xu. 2024. StoryImager: A Unified and Efficient Framework for Coherent Story Visualization and Completion. In *European Conference on Computer Vision (ECCV)* (Milan, Italy). Springer-Verlag, Berlin, Heidelberg, 479–495.

Kolors Team. 2024. Kolors-Character. https://huggingface.co/spaces/Kwai-Kolors/Kolors-Character-With-Flux.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. 2024. Training-Free Consistent Text-to-Image Generation. *ACM Transactions on Graphics* 43, 4, Article 52 (July 2024), 18 pages.

Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. 2023. Concept Decomposition for Visual Exploration and Inspiration. *ACM Transactions on Graphics* 42, 6, Article 241 (Dec. 2023), 13 pages.

Haofan Wang, Matteo Spinelli, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. 2024b. InstantStyle: Free Lunch towards Style-Preserving in Text-to-Image Generation. *arXiv preprint arXiv:2404.02733* (2024).

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. 2024a. InstantID: Zero-shot Identity-Preserving Generation in Seconds. *arXiv preprint arXiv:2401.07519* (2024).

Wen Wang, Canyu Zhao, Hao Chen, Zhekai Chen, Kecheng Zheng, and Chunhua Shen. 2024c. AutoStory: Generating Diverse Storytelling Images with Minimal Human Effort. *International Journal of Computer Vision* (2024), 1–22.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 15943–15953.

Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. 2023. Versatile Diffusion: Text, Images and Variations All in One Diffusion Model. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 7754–7765.

Yu Xu, Fan Tang, Juan Cao, Yuxin Zhang, Oliver Deussen, Weiming Dong, Jintao Li, and Tong-Yee Lee. 2024. Break-for-Make: Modular Low-Rank Adaptations for Composable Content-Style Customization. *arXiv preprint arXiv:2403.19456* (2024).

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. 2024. SEED-Story: Multimodal Long Story Generation with Large Language Model. *arXiv preprint arXiv:2407.08683* (2024). https://arxiv.org/abs/2407.08683

Xianjun Yang, Wei Cheng, Xujiang Zhao, Wenchao Yu, Linda Petzold, and Haifeng Chen. 2023. Dynamic prompting: A unified framework for prompt tuning. *arXiv preprint arXiv:2303.02909* (2023).

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *arXiv preprint arxiv:2308.06721* (2023).

Chun-Hsiao Yeh, Ta-Ying Cheng, He-Yen Hsieh, Chuan-En Lin, Yi Ma, Andrew Markham, Niki Trigoni, Hsiang-Tsung Kung, and Yubei Chen. 2024. Gen4Gen: Generative Data Pipeline for Generative Multi-Concept Composition. *arXiv preprint arXiv:2402.15504* (2024).

Weili Zeng, Yichao Yan, Qi Zhu, Zhuo Chen, Pengzhi Chu, Weiming Zhao, and Xiaokang Yang. 2024b. Infusion: Preventing customized text-to-image diffusion from overfitting. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 3568–3577.

Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. 2024a. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6786–6795.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023c. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 3836–3847.

Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. 2023a. ProSpect: Prompt Spectrum for Attribute-Aware Personalization of Diffusion Models. *ACM Transactions on Graphics* 42, 6, Article 244 (dec 2023), 14 pages.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023b. Inversion-Based Style Transfer with Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10146–10156.

Yang Zhang, Rui Zhang, Xuecheng Nie, Haochen Li, Jikun Chen, Yifan Hao, Xin Zhang, Luoqi Liu, and Ling Li. 2024. SPDiffusion: Semantic Protection Diffusion for Multi-concept Text-to-image Generation. *arXiv preprint arXiv:2409.01327* (2024).

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2023d. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems* 36 (2023), 17773–17794.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. 2024b. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhengguang Zhou, Jing Li, Huaxia Li, Nemo Chen, and Xu Tang. 2024a. StoryMaker: Towards Holistic Consistent Characters in Text-to-image Generation. *arXiv preprint arXiv:2409.12576* (2024).

Zhuofan Zong, Dongzhi Jiang, Bingqi Ma, Guanglu Song, Hao Shao, Dazhong Shen, Yu Liu, and Hongsheng Li. 2024. EasyRef: Omni-Generalized Group Image Reference for Diffusion Models via Multimodal LLM. *arXiv preprint arXiv:2412.09618* (2024).