Zero-Shot Multi-Reference Personalization via MLLMs-Guided Layout Planning

Junhao Feng[1], Xinghang Xu[2], Zihao Zhang[1] and Zhonghua Wan[1(✉)]

[1] School of Mechanical Engineering of North University of China, 030051, Taiyuan, China
[2] College of Computer Science and Technology, North University of China, 030051, Taiyuan, China
fjunhao80@gmail.com1, zhwan@nuc.edu.cn✉

Abstract. Though zero-shot adapters excel in image personalization, they often encounter significant challenges in multi-reference personalized generation, specifically failing to precisely adhere to the spatial layouts described in text prompts and suffering from feature leakage between reference images. To address these two challenges, we propose RIG (Regional Image-prompt Generation), a novel training-free framework. For the first challenge, leveraging Multimodal Large Language Models (MLLMs), we introduce a layout planning binder. Leveraging Chain-of-Thought (CoT) reasoning, this module infers and generates precise global layouts from text prompts, while simultaneously binding reference images to their corresponding regions. For the second, we introduce a spatially decoupled diffusion mechanism that isolates feature streams during attention computation. By injecting reference features exclusively into designated regions, this mechanism effectively prevents feature interference between reference images. Extensive experiments demonstrate that RIG significantly outperforms state-of-the-art adapter methods in terms of both personalization fidelity and text-layout alignment.

Keywords: Image Personalization · Diffusion Models · Multimodal LLMs · Zero-Shot Learning.

1 Introduction

Customizing or personalizing image generation with reference images is a significant and practical application of text-to-image diffusion models. In recent years, significant strides have been made in image personalization. Existing approaches generally fall into two paradigms: tuning-based or training-based and training-free methods. Although tuning-based methods (e.g., DreamBooth[11], Custom Diffusion[5]) or training-based methods (e.g., Mix-of-Show[2]) achieve high fidelity and strict prompt adherence via parameter fine-tuning or training, their high training costs and the necessity to manage task-specific parameters hinder rapid deployment in multi-task scenarios.

Fig. 1. A comparison between RIG and previous adapter methods. The yellow and purple text are respectively the layout and background hints in the prompts.

In contrast, training-free adapters (e.g., IP-Adapter[17], OminiControl[13]) achieve plug-and-play personalized generation by injecting reference features into frozen models, demonstrating exceptional performance in single-reference personalization tasks. However, existing models struggle to precisely reason about spatial relationships within text prompt, making it difficult for adapter-based methods to achieve personalized generation that aligns with the described spatial layouts(Fig. 1). The guidance from multiple reference images further overshadows the text prompts, hindering the model's ability to map specific reference images to their corresponding spatial locations as defined by the text. Furthermore, we observe that current adapters tend to broadcast extracted features across the entire generative latent space. Consequently, in multi-reference scenarios, features from a single image are erroneously applied as global guidance, failing to effectively disentangle the semantics of different reference images. This inability leads to conflicts between personalized attributes, manifesting as "feature leakage," where each generated subject exhibits visual characteristics blended from other reference images(Fig. 1).

To address the aforementioned two limitations, we present RIG(Regional Image-prompt Generation), a novel training-free framework for reference-guided image personalization. By translating the Multimodal Large Language Models(MLLMs)' region-level structured generation intents into the underlying control of the diffusion model, RIG explicitly defines the specific spatial region where each reference image should exert its influence. Within these confined regions, we employ spatially decoupled diffusion mechanism to achieve personalized generation.

The main contributions of this paper are as follows.

- A novel training-free framework for personalized generation called RIG is presented. By seamlessly integrating the spatial reasoning capabilities of MLLMs with diffusion models, RIG achieves precise alignment between personalized content and textual spatial layout prompts without requiring training.

- A spatial decoupled diffusion mechanism is devised to enforce strict feature isolation during the attention computation phase, effectively addressing the persistent issues of feature leakage observed in adapter models for multi-reference personalized generation.
- Qualitative and quantitative experiments demonstrate the superiority of RIG. It significantly outperforms current existing advanced zero-shot adapters, particularly in terms of multi-reference and prompt-layout alignment.

2 Related Works

Training-free Personalized Image Generation. Personalized generation aims to synthesize new images containing specific subjects or styles derived from a few reference images. To address efficiency bottlenecks, adapter-based methods like IP-Adapter have emerged. By employing a decoupled cross-attention mechanism, they inject reference features into frozen diffusion models. Although efficient, the original IP-Adapter relies on global feature injection, which often leads to attribute mixing and spatial layout error when extended to multi-reference scenarios. Although recent work like Conceptrol[3] attempts to guide attention using text-derived masks, it still fundamentally relies on the base model's weak spatial grounding and fails to guarantee strict feature isolation. In contrast, RIG eliminates global interference by enforcing explicit region-aware feature injection.

Controllable Image Generation. To enhance structural control, methods such as ControlNet[18], T2I-Adapter[8], and GLIGEN[6] have been introduced. These allow users to guide generation using explicit spatial conditions like edge maps, segmentation masks, or bounding boxes. Although powerful, these approaches face two limitations in our context: (1) Dependency on explicit conditions: They require users to provide precise masks or coordinates, failing to support abstract natural language spatial instructions (e.g., "A is to the left of B"). (2) Lack of identity binding: They focus on generic object layout and do not address the challenge of binding specific reference identities to designated regions. RIG solves this by automating the semantic binding process directly from text.

MLLMs as Layout Planners. Leveraging the reasoning capabilities of MLLMs to guide visual generation has become a burgeoning trend. Representative works like LMD[7] and RPG[16] utilize the in-context learning ability of LLMs(Large Language Models) to translate abstract prompts into concrete layout bounding boxes, guiding diffusion models to handle complex compositional scenarios. These methods demonstrate the potential of MLLMs in spatial planning. However, existing MLLM-based approaches are predominantly confined to generic text-to-image (T2I) generation tasks. They lack the mechanism to integrate specific reference image features into the planned regions.

Bridging these gaps, RIG uniquely integrates the "high-level semantic planning" of MLLMs with the "low-level feature injection" of adapters. Unlike previous works, we utilize the layout information parsed by the MLLM to directly modulate the cross-

attention mechanism within the diffusion model. This enables simultaneous precise spatial control and high-fidelity multi-concept personalization without any fine-tuning.

3 Method

3.1 Background

Diffusion Models. Diffusion models are a class of generative models that typically consist of two fundamental processes: a forward diffusion process and a reverse denoising process. In the forward process, a fixed Markov chain is used to gradually perturb the original image data by adding Gaussian noise over t timesteps. In contrast, the reverse process learns a denoising model that progressively transforms Gaussian noise into realistic samples.

Diffusion models can also be conditioned on additional inputs to guide the generation process. For instance, text-to-image diffusion models generate images according to textual prompts, while image prompts have also been widely adopted in image-conditioned generation tasks. Current image personalization methods are commonly built upon such conditional diffusion models. The denoising network is usually parameterized as ϵ_θ and its training objective can be formulated as:

$$\mathcal{L} = \mathbb{E}_{x_0,\epsilon\sim\mathcal{N}(0,\mathbf{I}),p_{\text{text}},p_{\text{img}},t}\left[\left\|\epsilon - \epsilon_\theta(x_t, p_{\text{text}}, p_{\text{img}}, t)\right\|^2\right].$$

Here, x_t denotes the original image, t represents the diffusion timestep, and the noisy latent/image at timestep t is defined as:

$$x_t = \alpha_t x_0 + \sigma_t \epsilon,$$

where α_t and σ_t are predefined coefficients determined by the noise schedule, and p_{text} and p_{img} denote the text and image features, respectively. After training, the model can generate high-quality images by iteratively removing Gaussian noise through the learned denoising process.

Decoupled Cross-Attention Mechanism. The decoupled cross-attention mechanism serves as the foundation of our zero-shot personalized generation method. Compared with directly injecting reference image features into the text feature space, this mechanism provides a more effective way to integrate text and image conditions into the cross-attention layers of the U-Net, thereby guiding the generation process more precisely.

Specifically, an additional cross-attention layer is introduced into each original cross-attention layer of the U-Net to incorporate image features. Given the image feature p_{img}, the output of the image-conditioned attention branch is computed as:

$$Z^{\text{img}} = \text{Attention}\left(Q, K^{\text{img}}, V^{\text{img}}\right) = \text{Softmax}\left(\frac{Q(K^{\text{img}})^\top}{\sqrt{d}}\right)V^{\text{img}}.$$

The query, key, and value matrices are defined as:

$$Q = ZW_q, K^{\text{img}} = p_{\text{img}}W_k^{\text{img}}, V^{\text{img}} = p_{\text{img}}W_v^{\text{img}},$$

where Z denotes the intermediate feature representation in the U-Net, W_q is the query projection matrix, and W_k^{img} and W_v^{img} are the key and value projection matrices for image features, respectively.

Similarly, given the text feature p_{text}, the text-conditioned cross-attention branch is computed as:

$$Z^{\text{text}} = \text{Attention}(Q, K^{\text{text}}, V^{\text{text}}) = \text{Softmax}\left(\frac{Q(K^{\text{text}})^{\top}}{\sqrt{d}}\right)V^{\text{text}}.$$

Where:

$$Q = ZW_q, K^{\text{text}} = p_{\text{text}}W_k^{\text{text}}, V^{\text{text}} = p_{\text{text}}W_v^{\text{text}}.$$

Here, $W_k^{text}$and W_v^{img} denote the key and value projection matrices for text features, respectively. It is worth noting that the image-conditioned and text-conditioned cross-attention branches share the same query Q, while their keys and values are projected from different conditioning features. The final output of the decoupled cross-attention mechanism is obtained by summing the outputs of the two branches:

$$Z^{\text{final}} = Z^{\text{img}} + Z^{\text{text}}.$$

3.2 Framework Overview and Conception

To address the challenges of inaccurate spatial layout and feature leakage in multi-reference zero-shot generation, we propose RIG, a training-free framework that requires no fine-tuning of any pretrained models. Instead of directly fusing multiple reference features in a global or weakly constrained manner, RIG explicitly decomposes the generation process into two stages: multimodal layout planning and binding and spatially-decoupled diffusion.

As shown in Fig. 2, the multimodal layout planning and binding module takes a text prompt and multiple reference images as input. It leverages the reasoning capability of multimodal large language models to identify key subjects, infer their spatial arrangement, and establish explicit bindings between each reference subject and its target region. This structured layout serves as a high-level generation plan, specifying where each reference object should appear and how different objects are spatially related.

Based on the planned layout, the spatially-decoupled diffusion module, illustrated in Fig. 3, performs region-aware conditional generation during the denoising process. Specifically, the target image space is partitioned into regions associated with different reference subjects, and each region only activates the reference features bound to it. This design restricts the propagation of reference features across unrelated regions, thereby reducing cross-reference interference and feature leakage. Meanwhile, the pretrained diffusion model is leveraged to naturally synthesize backgrounds, boundaries, and interactions between objects, ensuring global visual coherence.

Overall, RIG enables precise reference-to-region control, improved identity and attribute preservation, and effective feature isolation in multi-reference scenarios. Since the entire framework operates at inference time without model fine-tuning, it provides a plug-and-play solution for controllable zero-shot multi-subject generation.

3.3 Multimodal Layout Planning and Binding

This module executes two primary tasks: Hierarchical layout decomposition and visual-semantic binding.

Given a global text prompt P containing multiple entities and their complex relationships, along with a set of reference images $IP = \{IP_1, \dots, IP_j\}$. Traditional diffusion models struggle to process the intricate spatial relationships within P. We utilize an MLLM to deconstruct the image space $H * W$ at a semantic level. The planner infers and outputs a Generation Blueprint Φ formally defined as:

$$\Phi = (S, T, B) = \text{MLLM}(P, IP),$$

This blueprint consists of three key components:

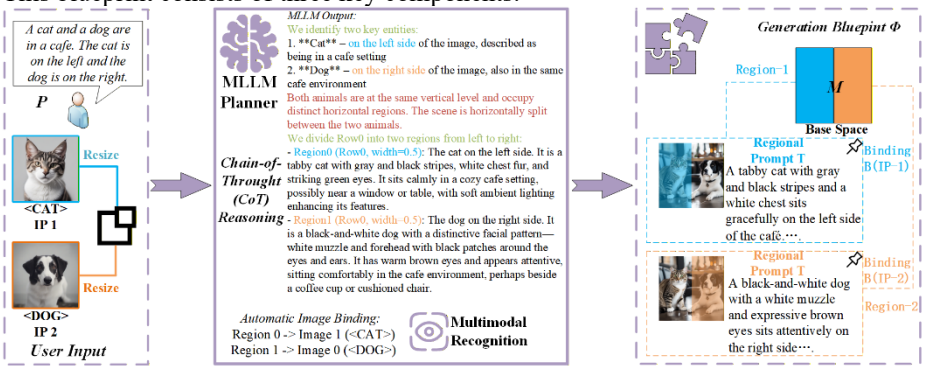

Fig. 2. Given a global prompt P and reference images IP_j, the MLLM performs chain-of-thought reasoning to decompose the scene. It outputs a generation blueprint Φ containing: (1) Spatial partition descriptor (S) represented by non-overlapping sub-regions M; (2) Regional text prompts (T) composed of fine-grained descriptions t_k; and (3) Visual-semantic bindings (B). The purple dashed lines illustrate the automated binding process, where specific regions are explicitly linked to reference images (e.g., the blue region binds to the CAT image IP_1).

1.Spatial Partition Descriptor (S): To ensure the generated spatial layout adheres strictly to the textual description, we employ a custom template to enforce a Chain-of-Thought (CoT) reasoning process within the MLLM. The MLLM outputs a set of structured segmentation instructions (e.g., specifying the row and column proportions for each part), which we convert into a set of mutually exclusive binary regions: $M = \{M_0, \dots, M_k\}$ where $M_k \in \{0,1\}^{H \times W}$ defines the spatial extent of the k-th region.

2.Regional Text Prompts (T): User prompts are typically global, mixing attributes and relationships of various entities. For more effective regional guidance, the global prompt is rewritten into a set of fine-grained sub-prompts:

$$T = \text{Redescribe}t_0, \ldots, t_k.$$

Where t_k focuses on describing the local visual content of the k-th region. These descriptions are also derived by the MLLM based on our preset template.

3.Visual-Semantic Binding (B): Existing layout control methods (e.g., GLIGEN[6]) often require users to manually specify image-location correspondences, limiting practicality. A core innovation of RIG is the introduction of Automated Visual-Reasoning Binding. The MLLM jointly analyzes the visual features of the reference images IP_j and the semantic content of the regional prompts t_k to construct a binding mapping function $B: \{0, \ldots, k\} \rightarrow \{1, 1, \ldots, N\}$. If $(B(k) = j\ (j > 0))$, it indicates that the k-th region is explicitly bound to the reference image (e.g., binding the reference image "cat" to the left region). Additionally, we introduce a text-only control identifier (denoted as "-1"). If selected, this indicates that the region is controlled solely by text (e.g., background regions).

This mechanism (Fig. 2) ensures that the correspondence between reference images and fixed layout regions is logically locked prior to the commencement of the generation process.

A chain-of-thought-inspired structured reasoning template. To mitigate hallucinations and unstable outputs in region prompt generation, we design a structured chain-of-thought template to explicitly constrain the reasoning path of the multimodal large language model. Although multimodal large language models exhibit strong cross-modal reasoning capabilities in image understanding and text generation tasks, their outputs can still be affected by visual perception bias, language priors, and stochastic decoding. When the model fails to sufficiently align visual evidence with textual descriptions, it may generate objects, attributes, or spatial relationships that are not actually present in the image. Moreover, given the same input image and textual prompt, the model may produce region allocation ratios and region-level prompts with inconsistent formats or unstable semantics, which can affect the parsing and generation performance of subsequent modules.

To address this issue, we decompose the multimodal reasoning process into five stages: example-guided format learning, key visual evidence extraction, cross-modal semantic alignment, candidate conclusion verification, and final response generation. Guided by predefined examples, the model first learns the desired output format and constraints. It then extracts task-relevant visual evidence from the image and aligns it with the textual condition. Before generating the final response, the model further verifies whether the candidate region allocation and region-level prompts are supported by the image content. This staged reasoning mechanism reduces the model's over-reliance on language priors, suppresses hallucinated content, and ensures that the outputs are parseable in format, stable in semantics, and consistent with the visual evidence. Our design paradigm is presented in Table 1.

Table 1. Region-aware structured Chain-of-Throught template for parseable regional prompt generation.

You are a master of composition who excels at extracting key objects and their attributes from input text and supplementing the original text with more detailed imagination, creating layouts that conform to human aesthetics. Your task is described as follows:
Input: Text prompt + reference images
1. Entity Extraction: Identify key objects, attributes, and spatial relations from the text prompt.
2. Layout Planning: Split the canvas into rows and regions according to the spatial arrangement of entities.
3. Region Assignment: Assign each region to one primary entity or background content.
4. Regional Prompt Generation: Generate a detailed local prompt for each region while preserving the original semantics.
5. Image-Region Binding: Match each region to the most relevant reference image; assign -1 to background or content absent from all reference images.
Output: Split ratio + regional prompts + image bindings.

3.4 Spatially Decoupled Diffusion

To precisely translate the generated blueprint into high-fidelity images, we propose the spatially decoupled diffusion mechanism built upon pretrained text-to-image models. Leveraging the planned region M and binding mapping B, our method integrates into the Cross-Attention layers of the U-Net, employing a feature routing strategy to achieve disentangled control over feature generation(Fig. 3).

At the $(t-1)$-th denoising step of the diffusion process, let $z_{(t-1)} \in \mathbb{R}^{\mathbb{h}\times\mathbb{w}\times\mathbb{c}}$denote the current latent feature map. To prevent feature interference between different subjects, we first perform spatial decoupling on the Query vectors Q based on the downsampled regions M_k Specifically, we compute the region-specific query $Q^{(k)}$ as:

$$Q^{(k)} = Q \odot M_k.$$

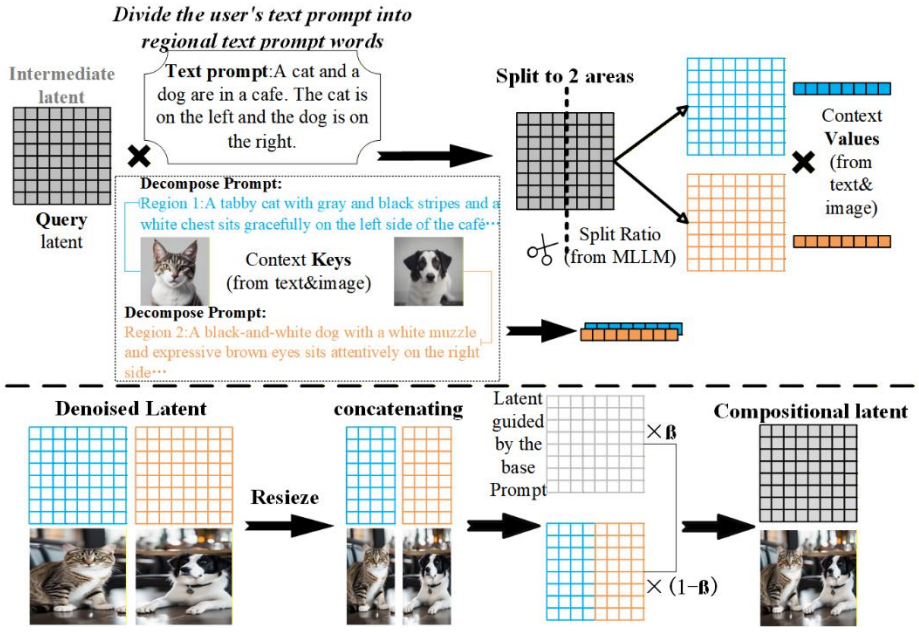

Fig. 3. Overview of the spatially decoupled diffusion architecture. These regional features are processed in parallel within isolated attention blocks to prevent feature leakage and are subsequently fused with the global base content to ensure overall coherence.

This operation spatially isolates the query features for the k-th region, ensuring they interact exclusively with the authorized conditional information, thereby fundamentally blocking feature confusion.

We did not consider each reference image as a global feature, We construct exclusive Key-Value pairs for each region k. The textual condition is derived from the local promptt_k, while the visual condition comes from the bound reference image $IP_{B(k)}$ The attention output for the region is calculated as:

$$O^{(k)} = \underbrace{\text{Attn}(Q^{(k)}, K_{\text{txt}}, V_{\text{txt}})}_{\text{Textual Control}} + \underbrace{\lambda \cdot IP_{[B(k)\neq -1]} \cdot \text{Attn}(Q^{(k)}, K_{\text{img}}, V_{\text{img}})}_{\text{Visual Control}}.$$

where $K_{txt}^{(k)}$ and $V_{txt}^{(k)}$ are projected from the CLIP[10] embeddings of the region prompt t_k, and $K_{txt}^{(k)}$ and $V_{txt}^{(k)}$ are projected from the image encoder features of the reference image $IP_{B(k)}$. $I_{[\cdot]}$ is an indicator function that activates the visual injection branch only when a valid reference image is bound to the region. λ serves as a scaling factor to modulate the influence weight of the reference image.

After computing the local attention outputs for all regions in parallel, we recompose them in the latent space. Since the regions are spatially mutually exclusive, the aggregated output O^{out} is obtained by summation:

$$O^{\text{out}} = \sum_{k=0}^{N} O^{(k)}.$$

Global Consistency Reconstruction. However, direct concatenation of regional contents can lead to artificial segmentation boundaries or unnatural transitions. To mitigate this we fuse the aggregated output O^{out} with the global content O^{base}:

$$O_{\text{final}} = \beta \cdot O^{\text{base}} + (1 - \beta) \cdot O^{\text{out}}.$$

This fusion mechanism, in conjunction with the original Self-Attention layers of the U-Net, ensuring consistency in background and style across different regions for seamless image generation.

4 Experiment

4.1 Experimental Details and Comparisons

We implement our scalable RIG framework using Qwen3-max[15] for spatial planning and Stable Diffusion XL[9] as the backbone (50 inference steps).

In major personalized generation scenarios, we compare our framework against previous zero-shot adapters and several commercial closed-source methods. The comparison primarily focuses on the following two aspects: 1. Personalized generation scenarios involving abstract spatial layouts and relational prompts. 2. Personalized generation scenarios with multi-reference image objects.

As shown in the figure(Fig. 4), the RIG framework significantly outperforms current zero-shot adapter methods on generation tasks with abstract spatial cues. In particular, it exhibits good visual and textual alignment in terms of the layout of generated content following textual representations, achieving similar performance to existing closed-source models that require finetuning and training. At the same time, in the scenario of multiple reference images generation, RIG well isolates the features between each reference image, and solves the challenge of reference image feature leakage in the scenario of multiple reference images personalized generation by the previous zero-shot adapter.

In Table 2. We re-implemented it according to the code and weights it released (the original evaluation datasets have not been made public). To assess the spatial-semantic alignment, we utilize the comprehensive T2I-CompBench++[4] benchmark, specifically focusing on 2D spatial relationships (UniDet). Regarding personalized generation, we evaluate identity fidelity using CLIP-I[10] and text alignment using CLIP-T[10] (scaled by a factor of 2.5). Furthermore, we report the overall time (Load the model and generate the first image with RTX 4090D*1, 24G, Round off) to demonstrate the deployment efficiency of our framework.

Table 2. Quantitative comparison on our collected Ref-Bench. We evaluate 2D-Spatial, identity fidelity and text alignment. Bold: best results. (In the table, FT represents the method based on fine-tuning, and no-FT represents the method without fine-tuning)

Method	UniDet ↑	CLIP-T(*2.5) ↑	CLIP-I ↑	Overall time (s) ↓
Textual Inversion (FT)[1]	0.157	0.693	0.673	>350
DreamBooth-LoRA (FT)[11]	<u>0.164</u>	<u>0.718</u>	**0.692**	>350
ELITE (no-FT)[14]	0.128	0.657	0.633	**24**
IP-Adapter (Baseline/no-FT)[17]	0.134	0.645	0.629	<u>33</u>
Ours (Zero-shot)	**0.183**	**0.733**	<u>0.688</u>	39

Fig. 4. Qualitatively comparisons of our method with the previous zero-shot adapter work and closed-source methods.

4.2 Ablation Experiment

Ablation Test for Spatial Layout Planning. Spatial layout is essential for enforcing layout constraints. Its removal forces the model to rely exclusively on text prompts, often resulting in ambiguous compositions. By comparing our method with the base model[9] (Fig. 5), we demonstrate that this module provides the explicit guidance for the base model to achieve accurate and coherent image composition.

Fig. 5. A test of the impact of the MLLM planning on the composition of base model.

Spatial Decoupling Diffusion Ablation Test. Conventional zero-shot adapters rely on global spatial decoupling, which inherently causes feature leakage in multi-subject scenarios due to the lack of instance-specific isolation. We conducted an ablation study using the cat-and-dog benchmark by retaining regional binding but removing our proposed spatial decoupling diffusion mechanism. As illustrated in Fig. 6, the absence of this mechanism resulted in severe global feature entanglement. This manifested as identity confusion or the blending of distinct subjects into a single hybrid entity, confirming the necessity of our approach for precise multi-subject generation.

Fig. 6. The dissolution of the spatial decoupling diffusion mechanism.

Verification of the Necessity of Global Consistency Reconstruction. Directly stitching generated sub-regions compromises image coherence. We address this via global consistency reorganization that fuses decoupled features with the global base latent. The ablation study (Fig. 7) highlights the trade-off controlled by the base ratio: a high ratio enhances coherence at the cost of personalization, while a zero ratio results in marked fragmentation. Notably, this fragmentation validates the strict spatial isolation of our RIG framework, demonstrating that it successfully eliminates feature leakage but requires base latent injection to ensure seamless global composition.

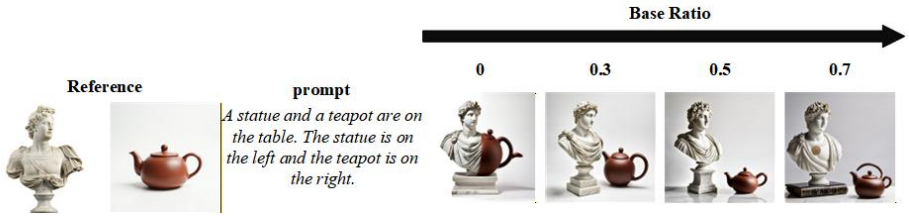

Fig. 7. The overall impact of different base radios on the image.

5 Conclusion

In this paper, we introduce RIG - a framework that requires no training, which significantly enhances the zero-shot adapter method's compliance with spatial layout prompts in text and feature preservation between reference images in the task of image personalization generation. Our framework is based on observations of previous attention mechanisms and base pre-trained models: (1) The previous cross-attention mechanism decouples the global injection of image features. (2) The ambiguity of the native model for spatial layout prompts. These are the key factors that make it difficult to follow the spatial layout and feature leakage during the personalized generation process. Our RIG precisely starts from these perspectives, leveraging the powerful capabilities of multimodal large language models and our unique spatial decoupling diffusion mechanism to place reference images in the correct positions and isolate them for decoupling, achieving personalized tasks on the basis of aligning with text prompt words. This framework, which requires no training at all, offers a new paradigm for image personalization methods. Undoubtedly, it can be extended to more models and technologies, promoting the in-depth application of personalization generation tasks.

6 Discussions

While RIG demonstrates promising improvements over existing methods, the current comparisons may not fully characterize its generalization ability in highly complex spatial planning scenarios. Most evaluated cases involve a moderate number of reference subjects and relatively unambiguous spatial configurations, where the correspondence between reference inputs and target regions can be reasonably inferred. As a result, the reported gains may partially reflect performance under favorable settings rather than robustness under more challenging compositional layouts. In scenarios involving crowded object arrangements, severe occlusions, hierarchical spatial relations, fine-grained positional constraints, or long-range interactions among multiple subjects, maintaining reliable reference-to-region binding and preventing cross-reference feature leakage may become substantially more difficult. Future work should therefore investigate more rigorous stress-test protocols and dedicated benchmarks for multi-reference spatial reasoning, so as to better assess the robustness and scalability of training-free generation frameworks.

7 Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62401520

References

[1] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)

[2] Gu, Y., Wang, X., Wu, J.Z., Shi, Y., Chen, Y., Fan, Z., Xiao, W., Zhao, R., Chang, S., Wu, W., et al.: Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. Advances in Neural Information Processing Systems **36**, 15890-15902 (2023)

[3] He, Q., Yao, A.: Conceptrol: Concept control of zero-shot personalized image generation. arXiv preprint arXiv:2503.06568 (2025)

[4] Huang, K., Duan, C., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-combench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)

[5] Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1931-1941 (2023)

[6] Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22511-22521 (2023)

[7] Lian, L., Li, B., Yala, A., Darrell, T.: Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. arXiv preprint arXiv:2305.13655 (2023)

[8] Mou, C., Wang, X., Xie, L., Wu, Z., Zhang, J., Qi, Z., Shan, Y.: 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. vol. 38, pp. 4296-4304 (2024)

[9] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

[10] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748-8763. PMLR (2021)

[11] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500-22510 (2023)

[12] Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, H., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)

[13] Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: Ominicontrol: Minimal and universal control for diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14940-14950 (2025)

[14] Wei, Y., He, Y., Chen, L., Liu, X., Wang, R., Li, H.: Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12345-12354 (2023)
[15] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
[16] Yang, L., Yu, Z., Meng, C., Xu, M., Ermon, S., Cui, B.: Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In: Forty-first International Conference on Machine Learning (2024)
[17] Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
[18] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836-3847 (2023)