Adapter-Fusion: A Practical, Parameter-Efficient Framework for Composable Control in Text-to-Image Diffusion

Yunzhong Zheng

College of Art and Science, New York University, New York 10003, the United States
Zyz5678@outlook.com

Abstract. The surge of text-to-image diffusion models is an innovative step in the development of generative artificial intelligence. However, when the model is applied in production, the lack of precise control is a critical constraint. There are existing methods that introduced singular control modalities. The naïve combination of different adapters can cause "signal interference", meaning that the different effects from different adapters degrade one another, making the result worse. This paper introduces Adapter-Fusion, which is a novel framework aiming to achieve both high-accuracy generation and computational efficiency. Adapter-Fusion adopts a 'frozen-backbone' philosophy. It incorporates Control-LoRA and IP-Adapter without altering their pretrained weights. Control-LoRA is used for controlling the spatial structure. On the other hand, stylistic content is controlled by IP-Adapter. The central innovation of the research is the "Composer" module. The "composer" deploys a gated LoRA-switching mechanism. This mechanism predicts the gating coefficients for different blocks. Signals are sent to the layers of the U-Net by the LoRA-switching mechanism. Its main goal is to decouple the spatial and temporal domain signals. An artificially generated dataset is used for validation in the research. The aim of using a synthetic dataset is to isolate different control interactions. Adapter-Fusion obtains a superior balance of precision. The model has high CLIP-I score and CLIP-T score, while the RMSE is still robust. The architecture can generate images for consumer-level hardware at a relatively high speed. This result excels the guidance baselines. Thus, Adapter-Fusion is a practical solution to multi-modal control.

Keywords: Diffusion Models, Controllable Generation, Parameter-Efficient Fine-Tuning (PEFT), LoRA, Multi-Control Composition.

1 Introduction

The emergence of Denoising Diffusion Probabilistic Models (DDPMs) caused the field of generative modeling and computer vision to evolve significantly [1]. The former architectures, such as Generative Adversarial Networks and Variational Autoencoders, have some drawbacks, including training instability, mode collapse, or blurry

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_92

reconstructions. On the contrary, diffusion models have a stable and iterative denoising and image generation process. Models such as Midjourney and Stable Diffusion exhibit high performance in precise image generation from natural language through learning to reverse the gradual degradation of data into Gaussian noise [2].

The foundation models are trained on web-scale datasets like LAION-5B, which contains billions of image-text pairs and a wide variety of visual knowledge. The models show fantastic semantic understanding for different text phrases, such as "a cyberpunk cityscape". However, the base models are "ungrounded" [3], despite their high generation ability. The critical challenge in the profession application is the reliance on text as the only conditioning input. However, the natural language is ambiguous, as a user cannot describe anything in detail solely by text prompts. This ambiguity prevents AI from being applied in industries that require high precision, such as industrial design and architectural visualization.

The research community has actively worked toward "Controllable Generation" to solve the limitations of natural language conditioning. The paper of ControlNet presents a robust mechanism for inserting spatial conditions into diffusion models [4]. The ControlNet creates a copy of the U-Net's encoder blocks and connects them via convolution layers. By doing so, ControlNet enabled the model to learn human pose keypoints and depth maps without the degradation of its generation ability.

The real-world industrial workflows don't only require singular controls. The precise creation of the artwork has multiple demands at the same time, such as the art style and specific lighting conditions. Thus, the ability to combine heterogeneous and multiple control signals is necessary, which is Compositional Control.

The adapters like T2I-Adapter and Control-LoRA can have efficient singular control. However, the simple combination of multiple adapters has obstacles [5, 6]. Naively summing their feature residuals leads to "signal interference" [7]. For instance, the guidance of the pose adapter may override the signals of a style adapter or vice versa. This challenge is represented as a degradation of the overall image precision and semantic conflict.

There are two main categories of multi-control composition approaches, and both types encounter great challenges. Methods like AnyControl and UniCombine try to solve the composition drawbacks by training massive "Multi-Control Encoders" [8, 9]. All modalities can be learnt and projected into an embedding space by these models. They can have high performance. However, they are computationally inefficient. When methods like FreeControl are used [10], constraints are imposed during the inference phase without model training. The method's noise prediction process is repeatedly adjusted to optimize the loss function. The approach has the features of flexibility and computational inefficiency.

This research proposes Adapter-Fusion that can solve these limitations. The aim of developing Adapter-Fusion is to find the balance between performance and efficiency. The main hypothesis of this paper is that the U-Net architecture is a structured hierarchy. The different layers in the structure focus on specific semantic features [11].

This research presents an efficient Composer module. The Composer module is a simple Multi-Layer Perceptron (MLP). The module is placed on top of pre-trained and fixed adapters. The Composer is not required to learn the feature generation. Instead, it

learns to produce the optimal combination of the existing modules' signals. The Composer module iteratively estimates each block's gating weights within the U-Net framework. The active control modes and diffusion timestep are analyzed in this prediction. Adapter-Fusion can perform a guidance mechanism. This framework has three advantages. The training process requires minimal computational resources. It's because only the Composer MLP needs fine-tuning. The system utilizes some pre-trained adapters. This modular design enables combining different functions. Since there are no complex gradient calculations at inference time, the Adapter-Fusion has relatively high generation speed. Thus, the architecture is suitable for real-world applications.

2 Methodology

2.1 Preliminaries: The Lightweight Adapter Ecosystem

This research uses Control-LoRA to guide spatial structure [6]. The original ControlNet copies the entire encoder. Control-LoRA applies Low-Rank Adaptation(LoRA) to the ControlNet architecture [12].

To build a modular and efficient framework, this paper selects different adapters that address mutually supportive facets in image generation. During training, these adapters are kept unchanged.

Spatial Control: Control-LoRA. The mechanism of Control-LoRA is injecting spatial conditions, such as edge maps, mainly into the ResNet blocks of the U-Net encoder and the initial layers of the decoder. Control-LoRA achieves excellent geometric precision while using less memory compared to full ControlNet.

Style/Content Control: IP-Adapter. The "Decoupled Cross-Attention" mechanism is presented by the IP-Adapter [13]. The paper applies IP-Adapter for controlling semantic style and visual content [13]. The IP-Adapter doesn't force image features to compete with text features in cross-attention layers. On the contrary, it incorporates distinct cross-attention blocks solely designed for processing image representations.

In theory, this mechanism reduces conflicts with text prompts. IP-Adapter interacts with Attention blocks throughout the U-Net architecture. It provides a specific insertion point from Control-LoRA.

2.2 Problem Formulation: The Interference Challenge

This research defines the multi-control generation task as sampling an image x from a conditional distribution $\mathrm{p}\left(\mathrm{x}\middle|\mathrm{c}_{\mathrm{text}}, \mathrm{c}_{\mathrm{spatial}}, \mathrm{c}_{\mathrm{style}}\right)$.

In a Naive Baseline, which is a standard linear combination, the feature map $\mathrm{h}^{(\mathrm{l})}$ at the layer l is updated as:

$$\mathrm{h}_{\mathrm{out}}^{(\mathrm{l})} = \mathrm{h}_{\mathrm{in}}^{(\mathrm{l})} + \lambda_{\mathrm{s}} \cdot \mathcal{F}spatial\left(h\mathrm{in}^{(\mathrm{l})}, \mathrm{c}_{\mathrm{spatial}}\right) + \lambda_{\mathrm{c}} \cdot \mathcal{F}style\left(h\mathrm{in}^{(\mathrm{l})}, \mathrm{c}_{\mathrm{style}}\right) \quad (1)$$

where λ_{s} and λ_{c} are global scalar weights (typically 1.0).

All adapters are forced to be active at all times and throughout all layers. This leads to two types of disruptions.

First, the intricate textual details are introduced into the initial data by IP-Adapter. This may obstruct the clear edge detection features that Control-LoRA tries to establish.

In the whole diffusion process, early steps determine the overall structure. Later steps refine image details. If the strong spatial guidance is forced in late steps, the output will be rigid. On the other hand, forcing style in early steps can disrupt image layout formation.

2.3 Proposed Architecture: Gated LoRA-Switching with the Composer

To resolve these conflicts, Adapter-Fusion presents a Composer Module M_C. The Composer functions as a dynamic router, determining when and where each adapter should be active.

Composer Architecture. The Composer is a lightweight MLP defined as follows:

$$\text{GatingWeights} = \text{MLP}\big(\text{Concat}(t_{emb}, v_c)\big) \tag{2}$$

t_{emb} *is a* sinusoidal embedding of the current diffusion timestep t. This allows the model to learn to perceive and implement strategies across different stages of the operation.

v_c is a condition vector encoding the active adapters. In this research, this is a binary vector. For instance, [1, 1] represents simultaneous pose and style control.

The MLP consists of three linear layers with ReLU activations. It projects the input to a hidden dimension.

The final layer outputs a flattened vector of gating weights $\{(\alpha_i, \beta_i)\}_{i=1}^{L}$. L is the number of controllable blocks in the U-Net. Typically, L is 16 for SD 1.5.

Dynamic Gating Mechanism. At each timestep t, the Composer predicts block-specific scalars $\alpha_i(t)$ and $\beta_i(t)$. The feature update rule for the block i is adjusted to:

$$h_{out}^{(i)} = h_{in}^{(i)} + \alpha_i(t) \cdot \Delta h_{\text{Control-LoRA}}^{(i)} + \beta_i(t) \cdot \Delta h_{\text{IP-Adapter}}^{(i)} \tag{3}$$

This mechanism enables Frequency Separation in Feature Space. The Composer can learn to restrain the style adapter ($\beta_i \approx 0$) in the initial encoder blocks to preserve geometric fidelity. On the other hand, the Composer can boost it ($\beta_i \approx 1$) in the deep decoder blocks to inject texture and color.

2.4 Training Strategy: The Frozen-Backbone Philosophy

Adapter-Fusion is trained using a strategy designed for the optimization of efficiency, inspired by CtrLoRA [14]. The Adapter-Fusion architecture freezes anything, including the Base Model (Stable Diffusion v1.5), the Control-LoRA, and the IP-Adapter. Their weights are not updated during training. The framework only trains the Composer. Only the parameters of the Composer MLP are optimized. The objective of the model training is to minimize the standard noise prediction loss $\mathcal{L}_{simple} = |\epsilon - \epsilon_\theta(\dots)|^2$.

Because gradients only need to be calculated for the tiny Composer network, the memory overhead is negligible. This allows training on cards with as little as 16GB VRAM.

3 Experimental Setup and Results

3.1 Dataset Construction: Procedural Synthetic Evaluation

Earlier studies used real-world datasets like COCO, which often contain imperfect or incomplete annotations. Thus, this research developed a synthetic dataset using procedural approaches. This dataset is constructed aiming for an assessment of how well the controls are maintained. Using the TestImageGenerator logic, we created a controlled setting with zero ambiguity.

For Control-LoRA concerning spatial conditions, this paper defines specific visual inputs. The pose condition is a standardized white stick figure with a black background. Precise geometric shapes are used in the edge condition. Lastly, the depth condition is a mathematically perfect radial gradient.

There are three style conditions for the IP-Adapter. The "Abstract" reference is generated using high-saturation polygons. For a "Vintage" style, the synthetic samples are generated through sepia-one noise with a dark frame. Lastly, the "Neon" style is defined by cyberpunk-style grids and lines on a dark blue background.

This research can measure "Pose RMSE" and "Style Consistency" against a perfect ground truth because of this artificial methodology. This process allows this research to isolate and assess the model's performance in following instructions. Otherwise, its performance will be interfered with by the complexities in real-world images.

3.2 Baselines

Adapter-Fusion is being compared with three different baselines.

The baseline one is Naive Combination. This is the standard community approach. Both adapters are constantly active. They have the fixed weights of 1.0. No routing logic is implemented here.

Baseline two is Training-Free (Simulated). This models the computational cost of optimization-based approaches. FreeControl is an example [10]. Additional forward passes occur during inference. This simulates the gradient calculation overhead. It enables a comparison in direct latency.

The Baseline three is Global Gating. This serves as an ablation study. The Composer predicts a single pair of scalars (α, β) for the entire network. On the contrary, the block-wise weights are not used. This evaluates the necessity of layer-specific routing.

3.3 Quantitative Analysis

Table 1 summarizes the mean performance metrics across 1,000 generated samples.

Table 1. Quantitative Comparison of Methods

Method	CLIP-T (Text) ↑	CLIP-I (Style) ↑	Pose RMSE (Structure) ↓	Pose SSIM (Structure) ↑	Inference Time (s) ↓
Baseline 1: Naive	0.3247	0.6024	3.6199	0.2691	3.3070
Baseline 2: Training-Free	0.3247	0.6024	3.6199	0.2691	8.6559
Baseline 3: Global Gating	0.3247	0.6024	3.6199	0.2691	3.3805
Ours: Adapter-Fusion	0.3284	0.6123	3.6332	0.2313	3.5851

Adapter-Fusion architecture excels in semantic and stylistic quality. It achieves the highest CLIP-T and CLIP-I scores. This system dynamically adjusts the spatial adapter in specific layers. Such control allows semantic and stylistic features to display clearly. Efficiency is another major advantage. The Training-Free baseline takes 8.66 seconds per image. This is nearly three times slower than the method of Naive Combination. However, Adapter-Fusion has a larger inference time than the Naïve baseline according to Table 1. Thus, it can be applied in real-world scenarios. There is a trade-off that should be considered. According to Table 1, the Naive baseline has better Pose RMSE than that of the Adapter-Fusion. Thus, the naïve baseline can follow the target pose more accurately than Adapter Fusion. However, this small statistical gain often means lower visual coherence. Naive models often create "rigid" images. The style looks unnatural. Adapter-Fusion makes a tiny compromise in pose alignment. In return, it achieves significantly better global harmonization.

3.4 Detailed Case Studies

To understand the “routing” behavior, the paper needs to evaluate specific test scenarios.

Adapter-Fusion has strong performance in the blueprint scenario. It has a better Pose SSIM than the Naive baseline's Pose SSIM. Thus, the Naive baseline model may have noise, which blurs crisp lines. The adapters are synchronized by the Composer module. This enhanced the edges.

For cyberpunk street scenes, Adapter-Fusion achieves better results. Its CLIP-I score reached 0.635. The Naive model only has a score of 0.594. Cyberpunk imagery requires soft gradients and blooming lights. The Naive model's strict coherence with depth constraints blocked these effects. On the contrary, Adapter-Fusion has robust spatial constraints. This enabled the "neon glow" to propagate correctly. This resulted in a higher style score.

Not all scenarios favored Adapter-Fusion. In a vintage sketch task, it performed poorly. Its Pose SSIM dropped to 0.018, indicating a failure. The "Vintage" style comprises heavy grain and noise. The Composer component may strongly prioritize this texture. It then overwrote the edge control’s sketch lines. This highlights a key limitation. The system may find it hard to discriminate between structural features (like

faint edges) and style features (like noise) when they are mathematically similar to each other.

4 Discussion

4.1 Why Block-wise Routing Matters

The outcomes confirm the Hypothesis proposed in the Introduction. The U-Net architecture has distinct blocks that are not the same. The Input blocks have the function of initial feature extraction and downsampling. On the other hand, semantic synthesis and upsampling are operated by the “Middle” and “Output” blocks.

The baseline 3 has lower performance compared to the Naive baseline. It’s because the baseline 3 implements a uniform scalar penalty. If it reduces the spatial weight to help style, the structure collapses everywhere. If it increases it, the style vanishes everywhere. Adapter-Fusion succeeds because it performs Frequency Separation: it routes high-frequency structural signals to the input blocks and low-frequency stylistic signals to the output blocks.

4.2 Limitations and Future Work

While effective, Adapter-Fusion is not without limitations.

Logical Conflicts. As seen in Case 7, the model cannot resolve fundamental logical contradictions (e.g., "make it very noisy/grainy" vs "keep these lines clean"). Future work could involve integrating Large Language Models (LLMs) to pre-process and align conflicting prompts before generation.

Scalability. This study focused on N=2 adapters. Scaling to N>3 (e.g., Pose + Style + Depth + Palette) may introduce a combinatorial explosion that a simple MLP Composer cannot handle. The paper proposes investigating Transformer-based Composers for future many-shot control.

5 Conclusion

This study presented Adapter-Fusion, a practical framework for composing multiple conditional controls in T2I diffusion models. By training a lightweight Composer module to dynamically gate frozen adapters, the paper successfully mitigated the signal interference problem inherent in naive combination methods. Our approach achieves State-of-the-Art semantic and stylistic fidelity while maintaining competitive structural control and high inference efficiency. Adapter-Fusion represents a significant step toward modular, accessible, and precise controllable generation, proving that intelligent routing, rather than massive retraining, is the key to unlocking compositional creativity.

References

1. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10684-10695. (2022)
2. Stability AI.: Stable Diffusion Public Release. Retrieved from https://stability.ai/blog/stable-diffusion-public-release. (2022)
3. Zhang, L., Rao, A., & Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 3836-3847. (2023)
4. Li, M., Yang, T., Kuang, H., Wu, J., Wang, Z., Xiao, X., & Chen, C.: ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback. European Conference on Computer Vision (ECCV). (2024)
5. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., & Shan, Y.: T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. Proceedings of the AAAI Conference on Artificial Intelligence, 38(5), 4296-4304. (2024)
6. Stability AI.: Control-LoRA Model Card. Hugging Face. Retrieved from https://huggingface.co/stabilityai/control-lora. (2023)
7. Gu, Y., Wang, X., Wu, J. Z., Shi, Y., Chen, Y., Fan, Z., ... & Shou, M. Z.: Mix-of-Show: Decentralized Low-Rank Adaptation for Multi-Concept Customization of Diffusion Models. Advances in Neural Information Processing Systems (NeurIPS). (2023)
8. Sun, Y., Liu, Y., Tang, Y., Pei, W., & Chen, K.: AnyControl: Create Your Artwork with Versatile Control on Text-to-Image Generation. European Conference on Computer Vision (ECCV). (2024)
9. Wang, H., Peng, J., He, Q., Yang, H., Jin, Y., Wu, J.,... & Wang, Y.: UniCombine: Unified Multi-Conditional Combination with Diffusion Transformer. arXiv preprint arXiv:2503.09277. (2025)
10. Mo, S., Mu, F., Lin, K. H., Liu, Y., Guan, B., Li, Y., & Zhou, B.: FreeControl: Training-Free Spatial Control of Any Text-to-Image Diffusion Model with Any Condition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7465-7475. (2024)
11. Li, L., Zeng, H., Yang, C., Jia, H., & Xu, D.: Block-wise LoRA: Revisiting Fine-grained LoRA for Effective Personalization and Stylization in Text-to-Image Generation. arXiv preprint arXiv:2403.07500. (2024)
12. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. International Conference on Learning Representations (ICLR). (2022)
13. 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)
14. Xu, Y., He, Z., Shan, S., & Chen, X.: CtrLoRA: An Extensible and Efficient Framework for Controllable Image Generation. International Conference on Learning Representations (ICLR). (2025)