ITM Web of Conferences **80**, 01009 (2025) https://doi.org/10.1051/itmconf/20258001009
ACAAI 2025

Conditional Latent Diffusion for Precision-Controllable Image Generation

Ji Su*

College of Artificial Intelligence and Automation, Hohai University, Changzhou, China

Abstract. Some models based on Latent Diffusion Models (LDMs), like Stable Diffusion, have revolutionized the image generation field in recent years. But LDMs' inherent precision control is often not effective enough to solve practical application problems. This paper reviews and compares five classic or state-of-the-art conditional control mechanisms—ControlNet, T2I-Adapter, Composer, UniControl, and FreeControl—designed to address this limitation. This paper analyze their architectural principles, performance trade-offs (e.g., in average FID score, computational cost, and inference speed), and applicability across different domains. Our comparative analysis demonstrates that while UniControl and Composer excel in dealing with tasks with high-quality requirement for their good performance in fine-grained control, methods like T2I-Adapter and FreeControl offer superior efficiency for mobile deployment due to their low computational demands. As the earliest control mechanism, ControlNet is still an effective mechanism and has certain application value. This overview provides a foundation for selecting appropriate control mechanisms for specific image generation tasks.

1 Introduction

Image synthesis is one of the most popular and developed fields in computer vision. While Generative Adversarial Networks (GANs) are excellent when the data variability is limited, because it is difficult for their adversarial learning procedure to extend to complex, multi-model distributions, the traditional likelihood-based models often need high computational demands, Latent Diffusion Models (LDMs) have made a great breakthrough by dividing the process of image synthesis into perceptual compression and a generative phase [1, 2]. LDMs have outstanding effects in image generation, especially high-resolution image synthesis, and the models based on LDM, such as Stable Diffusion, have become the major tools for image generation in both industrial and academic fields.

However, LDMs face limitations in performing some practical generative tasks. There are several errors that LDMs have: First is coarse-grained control, which means global text prompts often fail at precise spatial editing. The second type of errors is inconsistency errors; the structures of humans and animals generated by LDMs are frequently distorted and unreliable when LDMs face open-domain generation questions. Given that, the controllable

* Corresponding author: 2424020132@hhu.edu.cn

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

generation based on LDM is needed. This article aims to sum up the ways to make conditional latent diffusion.

2 Overview of models

2.1 Latent Diffusion Models (LDMs)

Compared with diffusion denoising probabilistic models (DDPM/ DMs), LDMs learn reverse Markov chains in highly compressed spaces instead of pixel spaces [2, 3]. A pre-trained autoencoder transforms images into latent representations. The decoder is capable of reconstructing the image from the latent space, giving. In order to avoid the latent spaces with high arbitrariness, two kinds of regularization are introduced: The first is KL-regularization, and the other is VQ-regularization.

In the generative phase, a diffusion model is trained directly on the latent space z. This Latent Diffusion Model (LDM) learns a reverse Markov chain of progressively noised latent, optimizing the re-weighted variational bound: where is a time-conditional U-Net [4]. Operating in the compressed space reduces dimensionality, which will slash computational costs while preserving semantic information.

For conditional synthesis, the U-Net with cross-attention layers (equation 1) that map conditioning inputs y (e.g., text prompts) into intermediate features via domain-specific encoders [5].

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)\text{V} \tag{1}$$

The joint objective becomes:

$$L_{LDM} = E_{E(x),y,\epsilon\sim N(\mathbb{0},\mathbb{1}),t}\left[\left|\epsilon-\epsilon_\theta\left(z_t,t,\tau_\theta(y)\right)\right|_2^2\right]. \tag{2}$$

This architecture supports diverse modalities (text, layouts) and enables convolutional high-resolution synthesis.

2.2 Control Mechanisms

2.2.1 ControlNet

ControlNet is a control mechanism that controls LDM by locking original parameters of the pre-trained diffusion models, creating a trainable copy of coding layer, and using zero convolutional layer to connect the copy with the original diffusion models, which can ensure that there is no harmful noise in the early stage of model training, and allows the parameters to grow gradually, efficiently protect pre-training knowledge and support the learning of diversified input conditions [6].

2.2.2 T2I-Adapter

T2I-Adapter was proposed to control the process of LDM by training a low-cost adapter and aligning the internal knowledge of pre-trained text to image diffusion model with external control signals, such as structural sketches and spatial colours [7].

2.2.3 Composer

Composer uses the decomposition-recombination paradigm. Composer first separates images into 8 combinable representations and then trains models using representations to generate reorganized images, which can control image generation precisely [8]. With the help of Bidirectional guidance, decoupling control is easy to realize.

2.2.4 UniControl

UniControl, which supports mixed conditions, extracts specific low-level characteristics of tasks through MOE-style adapters, combines Task-Aware HyperNet to regulate ControlNet weight, and realizes the unified control of 9 types of visual conditions by a single model with lower quantities of parameters than ControlNet [9].

2.2.5 FreeControl

FreeControl was proposed in 2024, which is a lightweight mechanism [10]. With the help of PCA, the semantic base of the seed image diffusion feature is built, and the guiding image feature is projected onto the base to achieve cross-modal structural alignment and the statistical alignment of the appearance of the brother image to solve the appearance leakage problem in training-free control.

3 Comparative analysis

3.1 Trade-off analysis

A direct quantitative comparison of these mechanisms is challenging due to variances in their experimental setups, including the base diffusion models, training datasets, and specific tasks evaluated. The performance metrics reported in Table 1 are primarily sourced from each method's original publication and should be interpreted within their respective contexts. Four levels of training cost are given using V100 GPUs (Low: <1 GPU day, Medium: 1~4 GPU days, High: 4~10 GPU days, Extremely high: >10 GPU days). Inference Speed is measured in iterations per second (it/s) on a standard V100 GPU. Data is sourced from respective original publications and may not be directly comparable due to differing experimental setups.

Table l. Mechanisms used in LDMs (average FID score ↓, training cost, inference speed).

Control Mechanism	Average FID Score	Training Cost	Inference Speed(it/s)
FreeControl	N/A	No training required	25.96
T2I-Adapter	26.3	Low	27.96
ControlNet	26.2	Medium	18.78
UniControl	24.0	High	N/A
Composer	9.2	Extremely high	N/A

To make the trade-off comparison visualization, the Fig. 1 is provided.

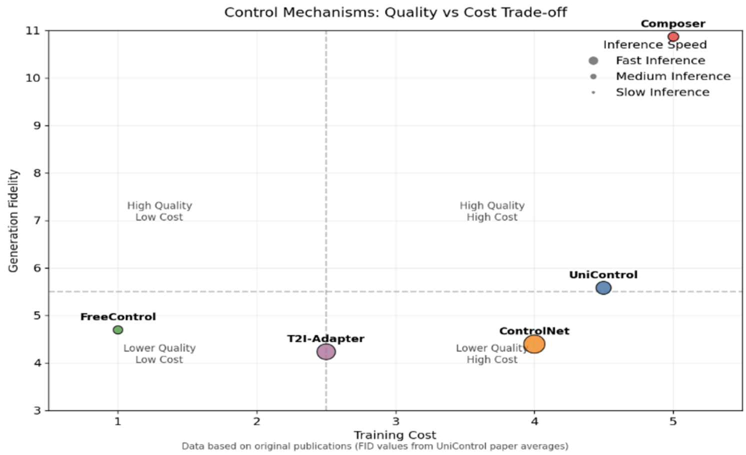

Fig. 1. Trade-off analysis of quality and cost of 5 control mechanisms

3.2 Analysis for the application

The five control mechanisms mentioned above are representative approaches, which are relatively mature or have great potential. Different mechanisms have different advantages and drawbacks, so they are applicable to various fields.

ControlNet is the earliest attempt to control the process of LDM. Compared with the other four mechanisms, ControlNet is good at fine-grained control (ControlNet has relatively low values of Fréchet Inception Distance (FID)) by introducing a new convolutional layer, but it also brings some negative effects to ControlNet, like high computational costs. [6, 11]

T2I-Adapter adopts a lightweight adapter architecture, which has an obvious advantage in generation efficiency compared with other mechanisms. [7] However, the T2I-Adapter has lower fineness with relatively high values of FID. So, this mechanism is suitable for mobile devices.

Composer supports multi-modal input combination, and it works excellently in the fields of style changing and decoupling controlling of text-to-image generation. [8] Composer has an advantage in fidelity and flexibility indicators, but the training complexity is relatively high, and the computing efficiency is medium. Composer can be utilized in advertising design and media art creation.

UniControl balances the abilities of generalization and fineness, and has a relatively higher efficiency by improving the architecture of ControlNet. [9] So, UniControl can cope with some problems with high requirements, such as game screen design, personalized image generation, and medical image synthesis with high-fidelity requirements.

For the reason that FreeControl does not need training, its advantages are prominently reflected in the easy deployment and high efficiency, but its abilities of fineness and generalization are weak. [10] So, like T2I-Adapter, FreeControl is suitable for use in mobile devices and low-resource deployment.

4 Conclusion

This paper has presented a systematic overview of five advanced control mechanisms for enhancing the precision controllability of Latent Diffusion Models. Our analysis confirms that a trade-off exists between generation fidelity (e.g., FID score) and computational efficiency. Composer and UniControl are well-suited for applications demanding high precision, such as medical imaging and industrial design, whereas T2I-Adapter and FreeControl provide practical solutions for resource-constrained environments like mobile devices. Despite these advancements, challenges remain in achieving real-time, high-fidelity control under diverse conditions. Future work should focus on developing more efficient architectures that mitigate this trade-off, exploring unified frameworks that can handle multimodal conditions seamlessly, and improving the robustness of open-domain generation.

Reference

1. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., & Bengio, Y. Generative adversarial networks. Commun. ACM 63, 139-144 (2020). https://doi.org/10.1145/3422622(https://doi.org/10.1145/3422622)
2. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. High-resolution image synthesis with latent diffusion models. Proc. IEEE 112, 465-524 (2024). https://doi.org/10.1109/JPROC.2024.3361356(https://doi.org/10.1109/JPROC.2024.3361356
3. Ho, J., Jain, A., & Abbeel, P. Denoising diffusion probabilistic models. Commun. ACM 67, 72-81 (2024). https://doi.org/10.1145/3652902(https://doi.org/10.1145/3652902)
4. Ronneberger, O., Fischer, P., & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Med. Image Anal. 26, 1-11 (2015). https://doi.org/10.1016/j.media.2015.02.009(https://doi.org/10.1016/j.media.2015.02.009)
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. Attention is all you need. In Adv. Neural Inf. Process. Syst. 30, 5998-6008 (2017).
6. Zhang, L., Rao, A., & Agrawala, M. Adding conditional control to text-to-image diffusion models. In Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 3839-3848 (2023).
7. 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. In Proc. AAAI Conf. Artif. Intell. 38, 4296-4304 (2024).
8. Huang, L., Chen, D., Liu, Y., & Shen, Y. Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023).
9. Qin, C., Zhang, S., Yu, N., Huang, Y., Yang, K., & Tong, Y. UniControl: A unified diffusion model for controllable visual generation in the wild. arXiv preprint arXiv:2305.11147 (2023).
10. Mo S, Mu F, Lin K H, et al. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 7465-7475.
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In Adv. Neural Inf. Process. Syst. **30**, 6626-6637 (2017).