An In-Depth Exploration of ControlNet Training Strategy Ablations for Conditional Generation

Anonymous CVPR submission

Paper ID

Abstract

In recent years, diffusion models represented by Stable Diffusion [6] have achieved great success in the field of text-to-image generation. The emergence of ControlNet [8] has greatly enhanced the controllability of the generation process by introducing external structural conditions such as attitude maps and edge maps. However, most of the existing studies follow the standard paradigm of freezing the backbone of pre-trained models, and there is insufficient exploration of the optimal training strategies in different fine-tuning scenarios. This paper aims to systematically study the training strategies of ControlNet, especially the impact of the parameter update (sdlocked) of the pre-trained U-Net backbone and the injection range of the control module (onlymidcontrol) on the model performance. We conducted a series of ablation experiments on the Fill50K dataset under the conditions of pose and edge maps. The experimental results reveal a key finding: Compared with the standard practice of completely freezing the backbone, adopting a strategy of "unlocking the backbone and simplifying the control" (that is, unfreezing the Stable Diffusion backbone while only applying the ControlNet module to the middle layer of U-Net) can achieve a better balance on small datasets. This strategy not only accelerates the "emergent convergence" of the model, improves the quality of the generated images and the accuracy of conditional alignment, but also effectively alleviates the severe overfitting and catastrophic forgetting problems that occur when all parameters are completely thawed. Our research indicates that this fine-tuning method similar to transfer learning offers a new and efficient paradigm for training ControlNet on specific styles or small-scale datasets, providing valuable practical guidance for the design and application in the field of controllable image generation.

1. Introduction

In recent years, diffusion models [2] have achieved significant success in the field of text-to-image generation. Latent diffusion models, represented by Stable Diffusion [6], can generate high-quality images based on natural language prompts. However, relying solely on text prompts makes it difficult to precisely control the spatial layout and structural details of the image; for example, it is challenging to accurately describe the complex poses of characters or the detailed arrangement of a scene through words alone. This often leads to the need for repeated experimentation and adjustment of the prompt to gradually approach the desired composition.

The introduction of ControlNet [8] provides a breakthrough for this challenge. ControlNet is an architecture that adds conditional control to pre-trained text-to-image diffusion models. Without altering the original weights of the Stable Diffusion model, it incorporates external structural information (such as edge detection maps, human keypoint pose diagrams, etc.) as additional conditional inputs, thereby providing fine-grained spatial guidance for the generation process. Specifically, ControlNet duplicates and freezes the weights of the original Stable Diffusion model, using "zero-convolution" layers to connect a new conditional branch to the feature layers of the diffusion model. Since these convolutional layers start with zero initial weights, they have minimal impact on the original model at the beginning of training, gradually learning the conditional control signals from scratch. This design ensures that adding the control branch does not disrupt the original generation capabilities while effectively aligning the generated images with the input conditions. Experiments show that ControlNet can be stably trained under various conditions (regardless of the size of the dataset) and supports flexible combinations of single or multiple conditions such as edges, depth, segmentation, and human poses. By allowing users to provide additional structural images as conditions, ControlNet achieves finer control over image generation compared to pure text, enhancing the match between the generated results and user intentions, and expanding the application prospects of diffusion models in creative design, film and animation, content editing, and more.

Based on this background, we observe that current research and applications of ControlNet mainly focus on controlling the poses of static images of people, with insufficient exploration of its broader potential uses and areas for improvement. For instance, in tasks such as video synthesis, medical imaging, biological feature simulation, and complex scene layout, there is great potential for structural condition control. Similarly, there are possibilities for further optimization in the ControlNet model architecture and training strategies to improve the quality of generation and the fidelity to the conditions. Therefore, this paper takes image generation under the conditions of pose maps (such as human keypoints) and edge maps (such as HED edge detection maps [7]) as a starting point, delving into how to enhance the generation quality and conditional control capabilities of ControlNet. This research is of great significance: it can enrich the application of ControlNet in more fields, making the generation of diffusion models more "what you see is what you get," and also provide new ideas for the design of the next generation of controllable image generation models.

2. Objectives

This study aims to systematically enhance the image generation effects based on pose maps and edge maps under conditional control using ControlNet. The specific objectives include:

Expanding Application Scenarios: To explore the applicability of ControlNet in a wider range of tasks, such as character action transfer, medical image synthesis, complex scene layout generation, and video keyframe guidance, thereby validating the value of structured conditions across different fields.

Architectural Innovation and Optimization: To investigate improvements in the architectural design of the ControlNet model, including strategies for inserting conditions at intermediate layers of different diffusion models, optimizing the connection methods between the control branch and the backbone network, and developing efficient mechanisms for fusing conditional features with generated features.

Improving Training Strategies: To develop effective training schemes that enhance the model's control capabilities and stability, such as reasonable freezing/thawing strategies for the main model, methods to improve the expression and alignment of condition signals during training, and regularization techniques to boost training stability.

We aim to significantly improve the quality of images generated under target pose or edge conditions, ensuring that the output results more accurately match the given conditions in terms of details and structure, while maintaining the diversity and visual realism of the images.

2.1. Research Questions

To achieve the research objectives, this paper will focus on exploring the following key issues:

1. Expansion of ControlNet Applications: Besides generating static character images, under the condition of using pose keypoint maps and edge detection maps, in what other generation tasks can ControlNet be expanded? How does it play a role in motion transfer in dynamic videos, synthesis of medical images, generation from scene layout to images, and continuous guidance for video keyframes, and what challenges does it face?
2. Optimization of Architecture Design: Currently, ControlNet introduces control signals by adding a zero-convolution conditional branch to a pre-trained diffusion model. Is there a more optimal network architecture or fusion mechanism that could improve the efficiency of condition utilization? For example, at which levels (shallow vs. deep, encoder vs. decoder) of the diffusion UNet should the conditional features be injected, and what kind of connection and fusion methods (addition, concatenation, or attention mechanisms) would yield the best generation results and precise condition alignment?
3. Improvement of Training Strategies: Without compromising the original generative capabilities, how can effective training strategies be designed to enhance ControlNet's sensitivity and control accuracy towards conditions? For instance, in what situations should the weights of the pre-trained diffusion model be frozen, or gradually unfrozen to adapt to new domains? Can the introduction of additional loss functions (such as cycle consistency loss or intermediate feature alignment) improve the match between conditions and outputs, thereby enhancing training stability and generation quality?

2.2. Related Work

T2I-Adapter: A Lightweight Control Module. Following ControlNet [8], researchers have explored alternative efficient methods for incorporating structural conditions. T2I-Adapter [5] introduces a lightweight, plug-and-play adapter module that injects external control signals without modifying the original Stable Diffusion architecture. This approach trains small convolutional networks for each condition type (e.g., sketch, edge, depth, keypoints) to encode multi-scale features, which are then added to the feature maps at different layers of Stable Diffusion's U-Net encoder. While the base model remains frozen, only the adapter parameters are updated. With substantially fewer parameters (e.g., 77M compared to ControlNet's 567M), T2I-Adapter enables faster inference while achieving comparable or superior structural and textual alignment on datasets like COCO [4]. Additionally, it supports flexible module composition: adapters for different conditions can be weighted and combined during inference to enable multi-condition control without retraining. For instance, merging outputs from sketch and color palette adapters allows simultaneous control over shape and color. The method also employs non-uniform timestep sampling during training to enhance guidance for low-level visual features (e.g., edges, colors). Overall, T2I-Adapter demonstrates an effective pathway for leveraging large models' implicit capabilities through compact modules, offering a valuable complement to the ControlNet framework.

Other Structured Conditional Generation Methods. Beyond the above, several recent architectures have been proposed to enhance controllability by integrating structural information into diffusion models. For example, GLIGEN [3] focuses on controlling scene layout and object placement. By incorporating learnable gated units into Stable Diffusion's cross-attention layers, GLIGEN accepts bounding box coordinates and corresponding object labels as additional inputs, enabling precise object positioning. Similar to other methods, it freezes most of the pre-trained weights and trains only a small set of gating and embedding parameters, thereby endowing the model with grounding capabilities. Experiments show that GLIGEN, built upon pre-trained diffusion models, achieves higher image quality and layout accuracy in layout-to-image tasks compared to models trained from scratch. Another related direction is multi-conditional diffusion, where models are trained to accept composite controls (e.g., text, segmentation, edges) in a single forward pass. However, such approaches typically require full retraining or large-scale fine-tuning, incurring high computational costs and potential limitations in generalizing to unseen condition combinations.

Training Strategies and Enhanced Control Performance. As structural conditional generation methods evolve, improving condition adherence and output quality has become a key research focus. Some studies introduce additional constraint losses during training to strengthen condition-image consistency. For example, a pixel-level cycle consistency loss can be used: pre-trained discriminative models (e.g., edge detectors, segmentation networks) are applied to generated images to extract condition signals, which are then compared with the original input conditions to directly optimize alignment. To mitigate the high computational cost of full diffusion sampling during loss calculation, an efficient single-step perturbation strategy can approximate the generated output for consistency evaluation.

Another line of work explores intermediate feature alignment training. At each denoising step, lightweight convolutional probes are trained to reconstruct input condition maps (e.g., edges or depth) from the U-Net's intermediate features. During training, a consistency loss is computed between the predicted "pseudo-condition" from noisy latents and the ground-truth condition, encouraging the model to maintain condition awareness throughout the diffusion process. This strategy embeds control signals more deeply into generation, enhancing structural fidelity and control precision.

In summary, recent advances—spanning model architectures (e.g., ControlNet, T2I-Adapter, GLIGEN) and training strategies (e.g., frozen fine-tuning, adapter composition, consistency constraints)—have continuously advanced the field of structurally conditioned controllable generation. Building upon these works, this study further extends applicable scenarios and proposes improvements for generation tasks under pose and edge map conditions, contributing to enhanced quality and precision in controllable image synthesis.

3. Methods

3.1. Merging Conditional Control with the Stable Diffusion Backbone

ControlNet [8] augments a pretrained text-to-image diffusion model by introducing a secondary, learnable pathway that injects spatial conditioning signals into the denoising network while preserving the capabilities of the original backbone (Stable Diffusion [6]). Let $F(\cdot;\Theta)$ denote a pretrained U-Net block that maps a feature map $x \in \mathbb{R}^{h\times w\times c}$ to an output $y = F(x;\Theta)$. To incorporate structural conditions such as edges, poses, or depth maps, ControlNet constructs a *dual-path* architecture: the original block is kept intact as a frozen backbone, and a trainable copy $F(\cdot;\Theta_c)$ is created to process condition-aware features. These two paths are fused through lightweight 1×1 convolutions, denoted by $Z(\cdot;\Theta_z)$, which align feature dimensions and modulate the backbone activations with task-specific information.

Given a conditioning feature map c, the ControlNet block computes

$$y_c = F(x;\Theta) + Z\big(F(x + Z(c;\Theta_{z1});\Theta_c);\Theta_{z2}\big). \tag{1}$$

This formulation allows the trainable copy to interpret the conditioning signal and contribute corrective residuals, while the frozen backbone ensures that the pretrained generative behavior is preserved. The connector convolutions $Z(\cdot;\Theta_{z1})$ and $Z(\cdot;\Theta_{z2})$ serve as projection layers that map the conditioning features into the appropriate channel space and then inject the processed representation back into the main feature stream.

The conditioning feature map c itself is derived from a raw conditioning image $c_i \in \mathbb{R}^{512\times512\times3}$ through a small encoder $E(\cdot)$. This encoder consists of a few strided convolution layers, where the first convolution primarily aligns the spatial resolution with the latent space of Stable Diffusion [6] (typically 64×64). The goal of $E(\cdot)$ is not to perform deep semantic extraction but to convert the conditioning image into a feature map compatible with the U-Net resolution at which ControlNet operates.

3.2. Training Considerations: Frozen Backbone and Zero-Initialized Connectors

To stabilize learning and prevent catastrophic forgetting, all pretrained parameters Θ of Stable Diffusion [6] are frozen throughout training. Since conditional datasets are often much smaller than the large-scale corpora used to train the original model, freezing the backbone prevents over-specialization and ensures that the core generative abilities remain intact. Only the parameters of the trainable copy Θ_c, the conditioning encoder $E(\cdot)$, and the connector convolutions Θ_{z1}, Θ_{z2} are updated.

A second key design choice is the zero-initialization of the connector convolutions. At initialization, both $Z(c; \Theta_{z1})$ and $Z(\cdot; \Theta_{z2})$ output zero for any input, causing the entire ControlNet block to reduce to

$$y_c = F(x; \Theta), \tag{2}$$

which means the model behaves identically to the pretrained Stable Diffusion at the start of training. This avoids injecting untrained, potentially harmful noise into the backbone and allows the influence of the conditional branch to "grow in" smoothly as training progresses. Empirically, this strategy leads to a stable optimization process and a characteristic "sudden convergence" effect: after a period in which the model behaves like the vanilla backbone, the connector weights learn meaningful transformations that allow the network to abruptly acquire strong conditioning fidelity.

Together, the dual-path structure, frozen backbone, and zero-initialized connector convolutions form a robust and data-efficient architecture. ControlNet is thus able to learn spatially localized, task-specific controls while preserving the expressive power and visual quality of the underlying diffusion model.

3.3. Conditioning Generation

In our experiments, we evaluated the *Human Pose* conditioning. The human pose condition is generated using an OpenPose [1]-based keypoint detector, which produces a structured skeletal representation that captures the global configuration and articulation of the subject. This conditioning type is widely used in controllable generation tasks because pose encodes coarse geometry and provides a salient spatial scaffold for the generative model to follow.

For the condition-type experiments, we adopted a publicly available pretrained ControlNet model from HuggingFace. This model was originally trained on large-scale general-domain datasets and has demonstrated strong generalization to diverse image conditions. In our replication, the human pose conditioning yielded high-quality generations, and the model consistently interpreted the skeletal signals provided by the pose skeletons. The results indicate that the pretrained ControlNet weights possess robust representational capacity and that the conditioning mechanism is sufficiently expressive to guide the generation process in a semantically meaningful way.

3.4. Training Procedure

We train our models using the Fill50K dataset released with the ControlNet training pipeline [8], a collection of 50,000 image–mask pairs used for conditioning-guided finetuning. The dataset contains diverse structural patterns and object boundaries, providing meaningful supervision for conditioning-guided generation even though it is significantly smaller than the large-scale datasets used for training Stable Diffusion. As our generative backbone, we adopt Stable Diffusion v1.5 [6], a latent diffusion model trained on LAION-5B that produces high-quality natural images through a U-Net denoiser operating in a compressed latent space. Its strong visual prior makes it an ideal foundation for studying conditional finetuning behavior.

To understand the effect of architectural choices during finetuning, we performed an ablation study following the configuration conventions in the ControlNet training documentation.[1] Specifically, we varied two key flags: `sd_locked` and `only_mid_control`. These options determine which portions of the Stable Diffusion U-Net are updated during training.

Definition of `sd_locked`. When the configuration `sd_locked = True`, all parameters of the original Stable Diffusion model are frozen. Only the ControlNet branch (trainable copy, connector convolutions, and condition encoder) receives gradient updates. This configuration preserves the pretrained generative behavior and prevents the model from drifting away from the domain of natural images. When `sd_locked = False`, the Stable Diffusion U-Net becomes trainable, greatly increasing the number of parameters updated at each iteration. This allows strong adaptation to the conditioning task, but also increases the risk of overfitting and destabilizing the pretrained backbone.

Definition of `only_mid_control`. When the configuration `only_mid_control = True`, ControlNet attaches trainable modules only to the middle block of the U-Net, rather than to all encoder blocks. This substantially reduces the number of trainable parameters, as the skip-connected encoder pathways remain unchanged. When

[1] https://github.com/lllyasviel/ControlNet/blob/main/docs/train.md

`only_mid_control = False`, ControlNet is attached to every encoder block, maximizing control capacity but also enlarging the effective finetuning footprint.

These configurations produce four experimental variants. To ensure column compatibility, we present the results in a compact table:

ExpID	sd_locked	mid_only	Sudden Conv.	Behavior
A	True	False	2700	Medium quality
B	True	True	N/A	Low quality
C	False	False	1500	Overfitting, forgetting
D	False	True	2400	High quality, mild overfit

Table 1. Ablation results of training configurations.

Across all configurations, we consistently observe the "sudden convergence" phenomenon: for several training steps, the model behaves similarly to the pretrained backbone, and then at a specific iteration, it abruptly begins to follow conditioning signals and generate structurally aligned images. However, the step at which this convergence occurs varies significantly between training settings, as shown in Table 1.

When both the Stable Diffusion model and all ControlNet modules are trainable (`sd_locked = False`, `only_mid_control = False`), the network receives gradients across a very large parameter space. This accelerates the sudden convergence (1500 steps) and yields high-quality generations early in training. However, this flexibility also leads to severe overfitting: as training continues, images develop blurry or irregular boundaries, and catastrophic forgetting may occur, with some samples collapsing into disordered patterns.

When the backbone is frozen but all ControlNet blocks are active (`sd_locked = True`, `only_mid_control = False`), the training process becomes more stable. This is the recommended configuration in the ControlNet paper [8]. The model maintains reasonable image quality and avoids overfitting even at high training iterations. Notably, the generated images tend to inherit aesthetic properties of Stable Diffusion v1.5 itself—such as characteristic textures and shading—indicating that the frozen backbone strongly shapes the model output.

Restricting ControlNet to the middle block while keeping the backbone frozen (`sd_locked = True`, `only_mid_control = True`) significantly limits the effective gradient pathways. As a result, we observe no sudden convergence even after 12k steps, and the model fails to learn meaningful structural patterns from the dataset. The limited backpropagation route through the single mid-block likely reduces optimization efficiency.

Interestingly, when the backbone is unfrozen but ControlNet is attached only to the middle block (`sd_locked = False`, `only_mid_control = True`), the model becomes more stable than in the fully trainable case. Overfitting still appears but is substantially mitigated, and image quality improves. We hypothesize two contributing factors: (1) the number of trainable parameters under this configuration becomes comparable to the standard setting (`sd_locked = True`, `only_mid_control = False`), whereas the fully trainable configuration involves significantly more parameters; and (2) because only a single decoding pathway is controllable, backpropagation becomes more direct, enabling more effective learning of dataset-specific features without destabilizing the entire U-Net.

Taken together, these results highlight the trade-offs between model stability, learning efficiency, and overfitting. Allowing too many components to update accelerates convergence but risks rapid degradation, whereas restricting updates can improve robustness but slow down or even prevent meaningful learning. Based on our observations, it may be advantageous to unfreeze the backbone while attaching ControlNet only to the middle block (`sd_locked = False`, `only_mid_control = True`) when finetuning on a small dataset with a strong and distinctive style, which behaves more like a transfer-learning procedure. In contrast, when the pretrained generative properties of Stable Diffusion v1.5 are crucial—such as maintaining realism or preserving its characteristic aesthetic—and the primary goal of finetuning is to teach the model how to interpret a new conditioning modality rather than to change its visual style, the configuration (`sd_locked = True`, `only_mid_control = False`) is preferable.

4. Members and Contributions

Hao Zheng: In charge of model training and experimental work, setting up test platforms for different tasks (such as action transfer, medical synthesis, etc.), comparing and analyzing the performance differences between Baseline (original ControlNet) and improved methods, ensuring the rigor and reliability of the experiments. **Zhiyi Chen**: Responsible for preparing and processing training/ testing data and conducting quantitative and qualitative analysis of experimental results, including metric evaluation, visualization result comparison, identifying the strengths and weaknesses of the methods and suggesting improvement directions. **Rongpei Li:** Mainly responsible for proposing and implementing architectural improvement plans, including new ControlNet intermediate layer connection strategies, conditional feature fusion module designs, etc., and collaborating on the formulation of training schemes.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, 2017.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffu-

sion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[3] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22511–22521, 2023.

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, 2014.

[5] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. 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] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022.

[7] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[8] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3813–3824, 2023.