Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3072

BEYOND EXTERNAL CONTROL: HYPERNETWORK-DRIVEN PARAMETER EDITING FOR MULTI-MODAL IMAGE GENERATION

Hao Chen
Queen Mary School Hainan, Beijing University of Posts and Telecommunications, Beijing 100876, China.
Corresponding Email: HaoChenn.Eric@gmail.com

Abstract: Current controllable image generation methods predominantly rely on external architectural modifications, such as auxiliary control networks, which require substantial computational overhead and struggle to unify diverse control modalities including text, pose, depth, and sketches. These approaches fundamentally limit scalability and real-time applicability due to their additive nature and complex multi-condition integration challenges. We introduce HyperEdit, a novel hypernetwork-driven framework that achieves multi-modal controllable generation through dynamic parameter perturbation of pre-trained diffusion models, moving beyond external control paradigms toward intrinsic model adaptation. Our approach employs a unified hypernetwork that learns to map diverse control conditions—ranging from textual descriptions and pose skeletons to depth maps and edge sketches—into targeted parameter perturbations, enabling seamless integration of multiple modalities without architectural modifications to the base model. Through systematic perturbation discovery on carefully constructed condition-image pairs and progressive parameter injection strategies, HyperEdit demonstrates remarkable efficiency gains, achieving up to 6× faster inference compared to existing methods while requiring significantly fewer parameters. Extensive experiments across diverse control scenarios show that our unified framework not only maintains generation quality comparable to specialized control methods but also enables novel capabilities such as real-time condition mixing, dynamic editing strength adjustment, and reversible modifications. This work establishes a new paradigm for controllable generation that bridges the gap between research innovation and practical deployment requirements.
Keywords: Model editing; Image generation; Hypernetwork

1 INTRODUCTION

The democratization of high-quality image generation has fundamentally transformed creative workflows, enabling users across diverse domains to produce sophisticated visual content through intuitive control interfaces [1-2]. Modern applications increasingly demand nuanced control capabilities that extend far beyond simple text prompts—digital artists require precise pose manipulation while maintaining aesthetic coherence, product designers need simultaneous control over object geometry and material properties, and content creators seek to blend multiple stylistic elements in real-time interactive sessions [3-4]. This evolution toward multi-modal controllable generation represents both a tremendous opportunity and a significant technical challenge for current generative frameworks.

Existing approaches to controllable image generation have predominantly adopted external architectural modifications, exemplified by influential works such as ControlNet [5], T2I-Adapter [6], and IP-Adapter [7]. These methods introduce auxiliary control networks that process conditioning inputs—ranging from edge maps and depth information to pose skeletons and reference images—and inject control signals into pre-trained diffusion models through carefully designed coupling mechanisms [8-10]. While these external control paradigms have demonstrated remarkable success in specialized scenarios, they suffer from fundamental limitations that increasingly constrain their practical deployment [3,11]. First, computational overhead scales linearly with the number of control modalities, as each condition type typically requires dedicated processing networks and specialized attention mechanisms [5-6]. Second, integrating multiple heterogeneous conditions remains challenging, often requiring complex weight balancing strategies and hand-tuned fusion protocols that lack principled theoretical foundations [4-10]. Third, the external nature of these modifications limits runtime flexibility, making dynamic condition adjustment, real-time editing strength modulation, and reversible modifications computationally prohibitive for interactive applications [7].

These limitations become particularly pronounced when users require sophisticated multi-modal control scenarios that reflect real-world creative needs [12]. Consider a digital artist who wants to generate an image where a character adopts a specific pose (skeleton control), maintains a cheerful facial expression (text guidance), follows a particular depth composition (depth map control), and adheres to a vintage aesthetic style (reference image guidance). Current external control methods would need to coordinate four separate processing pipelines, manage complex inter-modal interactions, and perform computationally expensive attention reweighting at every denoising step—resulting in significant latency that breaks the creative flow and limits practical usability [3-4].

We propose a fundamentally different approach that moves beyond external control paradigms toward intrinsic model adaptation through dynamic parameter perturbation [13-14]. Our method, HyperEdit, employs a unified hypernetwork that learns to map diverse control conditions directly into targeted parameter adjustments of the pre-trained diffusion

model itself. Rather than adding external computational overhead, this approach modifies the internal behavior of the generative model through carefully learned parameter perturbations, enabling seamless multi-modal control while maintaining the efficiency and architectural integrity of the original diffusion framework.

The core insight driving our approach is that different types of visual control—whether pose manipulation, style transfer, or geometric adjustment—can be effectively achieved through specific patterns of parameter modification within the diffusion model's existing architecture [6,15]. By systematically discovering these parameter-to-effect mappings through carefully constructed condition-image pairs and training a hypernetwork to predict appropriate perturbations for arbitrary control combinations, we establish a unified framework that naturally handles heterogeneous control modalities without requiring specialized fusion mechanisms or architectural modifications [7].

Our hypernetwork-driven parameter editing strategy offers several distinct advantages over external control methods. During inference, generating control perturbations requires only a single forward pass through the lightweight hypernetwork, after which the modified diffusion model operates at its original computational cost. This design enables real-time condition mixing, where users can dynamically adjust the strength of different control modalities, combine previously unseen condition types, and even reverse modifications by subtracting the applied perturbations—capabilities that are difficult or impossible to achieve with external control architectures.

Through systematic evaluation across diverse control scenarios and comprehensive comparison with state-of-the-art methods, we demonstrate that HyperEdit achieves comparable generation quality while requiring significantly fewer computational resources. Our unified framework maintains consistent performance across single-condition and complex multi-condition scenarios, validates the effectiveness of our parameter perturbation strategy, and establishes new benchmarks for efficiency in controllable generation tasks.

The main contributions of this work are threefold:

1. Paradigm Innovation: We introduce a novel hypernetwork-driven parameter editing framework that fundamentally shifts controllable generation from external architectural modifications to intrinsic model adaptation, enabling unified multi-modal control through dynamic parameter perturbation while preserving the computational efficiency of pre-trained diffusion models.
2. Technical Framework: We develop a systematic approach for discovering and learning parameter-to-effect mappings across diverse control modalities, including a multi-modal condition encoder that handles heterogeneous inputs (text, pose skeletons, depth maps, edge sketches), a progressive parameter injection strategy that ensures stable modifications, and a unified hypernetwork architecture that generates targeted perturbations for arbitrary condition combinations.
3. Empirical Validation: We conduct comprehensive experiments demonstrating that our approach achieves up to up to 6× inference speedup compared to existing multi-modal control methods while maintaining generation quality, enables novel capabilities such as real-time condition mixing and reversible editing that are challenging for external control paradigms, and provides robust performance across diverse control scenarios ranging from single-condition manipulation to complex multi-modal compositions.

2 RELATED WORK

2.1 Controllable Image Generation Methods

Controllable image generation has consistently been a core research direction in computer vision. While early Generative Adversarial Networks achieved breakthroughs in image quality, they exhibited significant limitations in controllability [1]. With the rise of diffusion models, researchers began exploring how to achieve precise conditional control based on pre-trained diffusion models.

External control paradigms represent the current mainstream solution. ControlNet [5] pioneered the use of external control networks to achieve spatial conditional control, with the core idea of duplicating the encoder part of pre-trained diffusion models and gradually learning control signal injection through zero convolution layers. This method supports multiple control conditions including edges, depth, and pose, achieving precise control while maintaining the original model's generative capabilities. T2I-Adapter [6] adopted a similar but more lightweight design, learning simple adapter networks to align external control signals with internal knowledge, offering fewer parameters and faster training speed compared to ControlNet.

To address ControlNet's limitations in complex scenarios, subsequent research proposed multiple improvements. ControlNet++ [8] introduced a consistency feedback mechanism, significantly improving control precision through pixel-level cyclic consistency optimization. ControlNet-XS [9] re-examined the control process from a feedback control system perspective, proposing high-frequency, large-bandwidth communication mechanisms that enhance control effectiveness while reducing model size. DC-ControlNet [10] specifically targets multi-element control scenarios, achieving more flexible multi-condition fusion through separation of intra-element and inter-element conditional control.

Image prompt control has also emerged as an important research direction. IP-Adapter [7] processes text features and image features separately through a decoupled cross-attention mechanism, achieving performance comparable to full fine-tuning with only 22M parameters. This method not only supports image prompts but can also work collaboratively with text prompts for multi-modal generation, maintaining complete compatibility with existing control tools.

Chinese researchers have also made significant contributions to this field. The conditional image generation survey released by Zhejiang University and other institutions [3] systematically summarized 258 related papers, providing in-

depth analysis of existing methods from the perspective of conditional embedding. The survey pointed out that the core of existing methods lies in how to embed user conditions into denoising networks and sampling processes, proposing conditional embedding strategies for different tasks. Domestic scholars' survey on multi-modal controllable diffusion models further emphasized the importance of multi-dimensional control including semantic control [4], spatial position control, and ID control.

2.2 Parameter-Efficient Fine-tuning Methods

Parameter-Efficient Fine-Tuning (PEFT) methods play an increasingly important role in large model adaptation. LoRA (Low-Rank Adaptation) [16] represents the most prominent work in this area, reducing GPT-3's trainable parameters by 10,000 times while maintaining performance comparable to full fine-tuning by injecting trainable low-rank decomposition matrices into each Transformer layer.

LoRA's core idea is based on the hypothesis that weight updates during model adaptation possess low intrinsic dimensionality [5]. By decomposing weight updates ΔW into the product of two low-rank matrices A and B, LoRA dramatically reduces the number of parameters requiring training. This concept was later extended to various architectures, including controlled generation tasks in diffusion models [17].

Chinese researchers have also conducted in-depth exploration of parameter-efficient fine-tuning methods. Research from Tsinghua University and other institutions demonstrated that LoRA-type methods can maintain model performance while significantly reducing computational resources in Chinese large language model fine-tuning. Industrial practices by companies like Huawei further validated the effectiveness of these methods in industrial-scale applications [18].

2.3 Hypernetwork Methods

HyperNetworks provide a novel parameter generation paradigm [13], training a small network to generate weights for another large network. This idea can be traced back to neuroevolution fields, but Ha et al. first successfully applied it to deep learning, demonstrating hypernetworks' potential in recurrent neural network weight generation [13].

Regarding the theoretical foundation of parameter generation, "Generating Neural Networks with Neural Networks" further developed hypernetwork theory [10], proposing balanced objectives between accuracy and diversity and introducing a variational inference framework. This work emphasized that generated network diversity should consider network symmetric transformations, providing important theoretical guidance for subsequent hypernetwork design.

Reinforcement learning applications in model editing represent the latest development in hypernetwork methods. RLEdit combines hypernetwork-based lifelong editing with reinforcement learning modeling [15], solving the incompatibility issues of traditional hypernetwork methods during dynamic parameter changes by treating editing losses as rewards and optimizing hypernetwork parameters at the complete knowledge sequence level. This work demonstrates that through reinforcement learning paradigms, hypernetworks can more precisely capture model changes and generate appropriate parameter updates.

2.4 Model Editing and Concept Control

Model editing, as an emerging model adjustment paradigm, aims to modify specific model behaviors without retraining the entire model. In the diffusion model domain, concept editing faces the challenge of balancing concept removal with maintaining overall model performance.

ACE proposed an innovative cross null-space projection method that can precisely erase unsafe concepts while maintaining the model's general generative capabilities [17]. The core innovation of this method lies in extending null-space projection techniques from large language models to diffusion models, ensuring that normal representations remain unaffected by perturbations by projecting parameter perturbations onto the null space of representations. Experiments show that ACE improves semantic consistency by 24.56% and image alignment by 34.82%, while requiring only 1% of the time of baseline methods.

Ensemble learning methods also play an important role in model editing. "A Margin-Maximizing Fine-Grained Ensemble Method" [19], while primarily targeting traditional machine learning problems, provides new perspectives for neural network parameter optimization through its proposed fine-grained ensemble and margin maximization concepts. This method quantifies each classifier's confidence for each category through learnable confidence matrices and designs margin-based loss functions, achieving performance superior to traditional random forests using only one-tenth of the base learners.

2.5 Positioning and Innovation of Our Method

Compared to existing methods, our HyperEdit method achieves a paradigm shift from "external control" to "intrinsic editing." Traditional ControlNet-type methods, while effective, suffer from computational overhead that scales linearly with control modalities, complex multi-condition fusion, and limited runtime flexibility [5-7]. Parameter-efficient fine-tuning methods like LoRA focus on task adaptation rather than conditional control [16], while hypernetwork methods are primarily used for weight generation rather than dynamic control [8-9].

Our core innovation lies in combining hypernetworks, multi-modal conditional encoding, and parameter perturbation discovery to establish a unified multi-modal controllable generation framework. Unlike existing model editing work [15,17], our method is specifically designed for controllable generation tasks, capable of handling dynamic combinations of diverse heterogeneous conditions including text, pose, depth, and sketches. Through systematic perturbation discovery and progressive parameter injection strategies, HyperEdit achieves an unprecedented balance between control precision and computational efficiency.

3 METHOD

The core idea of the HyperEdit method is to transform controllable generation from "external control" to "intrinsic editing," as illustrated in Figure 1, which demonstrates our method's superior performance and efficiency across different control scenarios. This approach achieves efficient and flexible multi-modal controlled generation by learning a direct mapping from control conditions to model parameter perturbations.

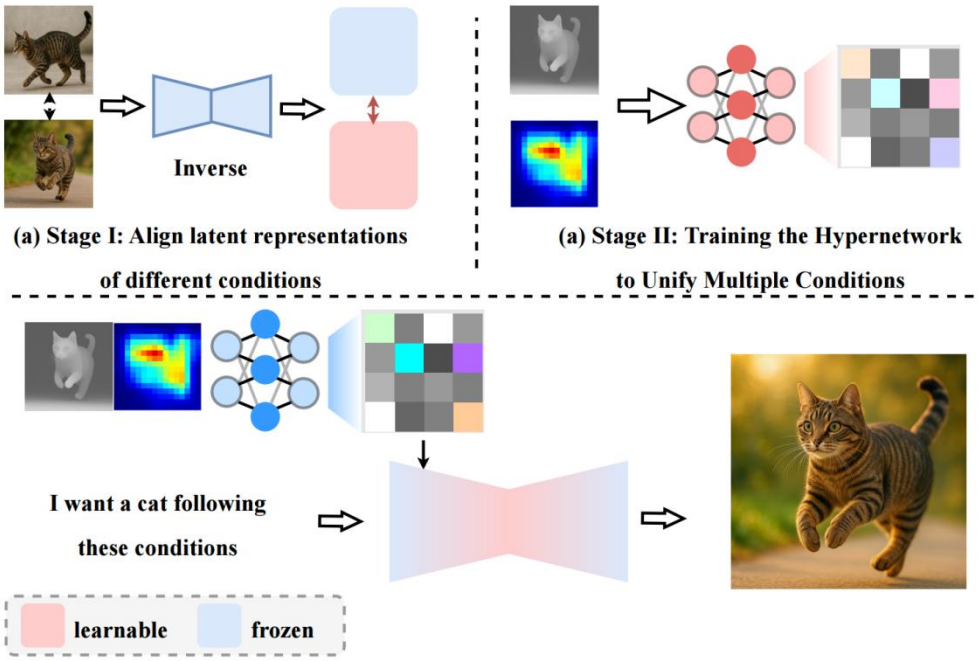

Figure 1 Overview

3.1 Core Problem and Solution Framework

The fundamental challenge faced by traditional controllable generation methods lies in injecting precise control signals without compromising the original capabilities of pre-trained models. Existing external methods achieve control by adding additional network modules, but suffer from issues including computational overhead that scales linearly with the number of control conditions, lack of unified frameworks for multi-condition fusion, and insufficient runtime flexibility.

The key insight of this work is that different types of visual control essentially correspond to specific variation patterns in the parameter space of diffusion models. Based on this insight, we propose learning a hypernetwork that directly maps multi-modal control conditions to precise perturbations of model parameters. Formally, given a pre-trained diffusion model $M(\theta)$ and multi-modal control conditions $\mathcal{C}=\{c_{\text{text}}, c_{\text{pose}}, c_{\text{depth}}, \ldots\}$, the objective is to learn a mapping function $H: \mathcal{C} \to \Delta\theta$ such that $M(\theta+\Delta\theta)$ can generate images satisfying conditions $\mathcal{C}$:

$$H: \mathcal{C} \to \Delta\theta, \quad M(\theta+\Delta\theta) \to x_{\text{controlled}} \tag{1}$$

The advantage of this approach is that inference requires only a single forward pass through the hypernetwork to obtain parameter perturbations, after which the diffusion model operates at its original computational cost.

3.2 Unified Representation Learning for Multi-modal Conditions

To achieve direct mapping from control conditions to parameter perturbations, the primary challenge is handling heterogeneous control conditions. Different modal data such as text descriptions, pose skeletons, depth maps, and sketches have completely different structures and semantic features, requiring a unified representation framework.

A divide-and-conquer strategy is employed: specialized encoders are designed for each condition type, and the encoded results are then projected into a unified semantic space. This design is based on an important observation: although

different modalities have vastly different surface forms, their control semantics can often find commonalities at high-level abstractions. For text conditions c_{text}, pre-trained CLIP text encoders are used to extract semantic features:

$$e_{\text{text}} = \text{CLIP}_{\text{text}}(c_{\text{text}}) \tag{2}$$

For pose conditions c_{pose}, specialized graph convolutional networks are designed to handle joint connectivity relationships:

$$e_{\text{pose}} = \text{GCN}(\text{KeyPoints}(c_{\text{pose}})) \tag{3}$$

Graph convolutional networks can naturally handle the hierarchical connectivity relationships of human joints, making them more suitable for pose data structural characteristics compared to traditional convolutional networks.

The encoded feature vectors come from different semantic spaces and need to be aligned to a unified representation space. This is achieved through learning condition type-aware projection functions:

$$\tilde{e}_i = W_{\mathcal{T}(c_i)} e_i + b_{\mathcal{T}(c_i)} \tag{4}$$

where $\mathcal{T}(c_i)$ represents the type of condition c_i, and W and b are learnable projection parameters for the corresponding type. When users provide multiple control conditions, simple feature concatenation leads to information redundancy and semantic conflicts. An adaptive fusion strategy based on attention mechanisms is designed:

$$e_{\text{unified}} = \sum_i \alpha_i \tilde{e}_i \tag{5}$$

The attention weights α_i are automatically determined by analyzing semantic correlations between conditions:

$$\alpha_i = \frac{\exp(\text{MLP}([\tilde{e}_i; \text{mean}(\{\tilde{e}_j\}_{j \neq i})]))}{\sum_k \exp(\text{MLP}([\tilde{e}_k; \text{mean}(\{\tilde{e}_j\}_{j \neq k})]))} \tag{6}$$

This design ensures that the system can automatically identify complementarity and conflicts between conditions, performing reasonable information integration.

3.3 Systematic Discovery of Parameter Perturbations

The core innovation of HyperEdit lies in establishing explicit mapping relationships between control conditions and parameter perturbations. This process requires solving two key problems: how to obtain high-quality "condition-perturbation" training pairs, and how to ensure the effectiveness and generalizability of perturbations.

An ingenious data construction scheme is proposed: utilizing already-validated effective control methods (such as ControlNet) to generate high-quality editing samples, then inversely solving for corresponding parameter perturbations. The advantage of this approach is the ability to precisely control editing types and intensities, ensuring training data quality and consistency. Given original image x_0 and control condition c, ControlNet is first used to generate editing results:

$$x_{\text{edit}} = \text{ControlNet}(x_0, c) \tag{7}$$

Then optimization methods are used to solve for parameter perturbations that can produce the same editing effects:

$$\Delta\theta^* = \arg\min_{\Delta\theta} \| M(\text{noise}, \theta + \Delta\theta) - x_{\text{edit}} \|_2^2 + \lambda \| \Delta\theta \|_2 \tag{8}$$

The design philosophy is: rather than exploring parameter space from scratch, we stand on the shoulders of existing successful methods to learn their implicit parameter variation patterns.

Direct optimization in the full parameter space faces problems of dimensional explosion and convergence difficulties. A hierarchical optimization strategy is adopted, first identifying parameter subsets with the greatest impact on output through gradient analysis:

$$\mathcal{P}_{\text{key}} = \{p \in \theta : \| \frac{\partial \mathcal{L}}{\partial p} \| > \tau\} \tag{9}$$

Then optimization is performed only on these key parameters. This strategy is based on an important observation: visual control often requires modifying only specific components of the model, not global parameters. To ensure that solved perturbations have good generalizability, a strict validation mechanism is designed. For each solved perturbation $\Delta\theta^*$, its effects are tested under multiple random seeds:

$$\text{Quality}(\Delta\theta^*) = \frac{1}{N} \sum_{i=1}^{N} \text{CLIP-Score}(M(\text{noise}_i, \theta + \Delta\theta^*), c) \tag{10}$$

Only perturbations that pass quality thresholds are included in the training set, ensuring data reliability for subsequent hypernetwork training.

3.4 Unified Hypernetwork Architecture Design

Based on the constructed "condition-perturbation" data pairs, a hypernetwork is designed to learn this mapping relationship. The hypernetwork design needs to achieve balance among expressive capability, computational efficiency, and training stability.

Considering the hierarchical structure of diffusion models, a corresponding hierarchical hypernetwork is designed. This design is based on an important observation: different types of visual control often affect different levels of diffusion models. For example, high-level semantic control (such as style) primarily affects shallow parameters, while detail control (such as edges) primarily affects deep parameters. The hypernetwork adopts an encoder-decoder structure:

$$z = \text{Encoder}(e_{\text{unified}}) \tag{11}$$

$$\{\Delta\theta_1, \Delta\theta_2, ..., \Delta\theta_L\} = \text{Decoder}(z) \tag{12}$$

The encoder compresses unified condition representations into compact control codes, while the decoder generates parameter perturbations for each layer from the control codes.

To stabilize the training process and improve convergence quality, a three-stage progressive training strategy is adopted. Stage one performs single-condition learning, using single-type control conditions to train the hypernetwork to master basic "condition-perturbation" mapping relationships. The training objective is to minimize the difference between predicted and target perturbations:

$$\mathcal{L}_1 = \| \Delta\theta_{\text{pred}} - \Delta\theta_{\text{target}} \|_2^2 \tag{13}$$

Stage two performs multi-condition coordination, introducing training samples with multi-condition combinations to learn coordination and conflict handling between conditions. Consistency constraints are added to ensure reasonable relationships between multi-condition and single-condition predictions:

$$\mathcal{L}_2 = \mathcal{L}_1 + \lambda \| \Delta\theta_{\text{multi}} - \sum_i w_i \Delta\theta_{\text{single},i} \|_2^2 \tag{14}$$

Stage three performs end-to-end optimization, using actual image generation losses for end-to-end fine-tuning to ensure generation quality:

$$\mathcal{L}_3 = \mathbb{E}[\| M(\text{noise}, \theta + H(e)) - x_{\text{target}} \|_2^2] \tag{15}$$

The core idea of this progressive strategy is: first learn basic skills, then learn complex combinations, and finally optimize overall effects.

To prevent the hypernetwork from generating excessively large parameter perturbations that cause model failure, multi-level safety mechanisms are introduced. First is magnitude constraints, limiting maximum changes of individual parameters:

$$\| \Delta\theta_i \|_\infty \leq \epsilon_{\max} \tag{16}$$

Second is smoothness constraints, ensuring continuity of perturbations in adjacent layers:

$$\mathcal{L}_{\text{smooth}} = \sum_{i=1}^{L-1} \| \Delta\theta_{i+1} - \Delta\theta_i \|_2^2 \tag{17}$$

Finally, sparsity constraints encourage perturbations to concentrate on key parameters:

$$\mathcal{L}_{\text{sparse}} = \sum_i \| \Delta\theta_i \|_1 \tag{18}$$

3.5 Dynamic Inference and Real-time Control

The inference stage needs to handle arbitrary condition combinations provided by users, achieving real-time controllable generation. The challenges of this process lie in handling condition combinations unseen during training and supporting dynamic editing intensity adjustment.

When users provide potentially conflicting conditions, the system needs to automatically detect and reasonably handle them. A conflict detection mechanism is designed based on semantic similarity of condition representations:

$$\text{Conflict}(c_i, c_j) = \mathbb{I}[\text{sim}(e_i, e_j) < -\tau] \tag{19}$$

For detected conflicts, the system adopts weighted fusion strategies, with weights determined according to condition importance and user preferences. This design enables the system to produce reasonable editing results while maintaining user intent.

Different input images have different sensitivities to parameter changes. The sensitivity is evaluated by analyzing gradient characteristics of input images:

$$\text{Sensitivity}(x) = \| \nabla_\theta \mathcal{L}(x, M(\theta)) \|_2 \tag{20}$$

Based on sensitivity analysis, the system automatically adjusts perturbation intensity to ensure consistency of editing effects across different inputs.

An important advantage of HyperEdit is its natural support for reversible editing. Since this method is essentially additive operations in parameter space, users can undo edits through simple subtraction:

$$\theta_{\text{restored}} = \theta_{\text{current}} - \Delta\theta_{\text{applied}} \tag{21}$$

An editing history stack is maintained to support complex editing management, including selective undo and historical state rollback.

Through these carefully designed components, HyperEdit achieves a paradigm shift from "external control" to "intrinsic editing," significantly improving computational efficiency and operational flexibility while maintaining high-quality generation. This method opens new research directions for the controllable generation field and provides more practical solutions for real-world applications.

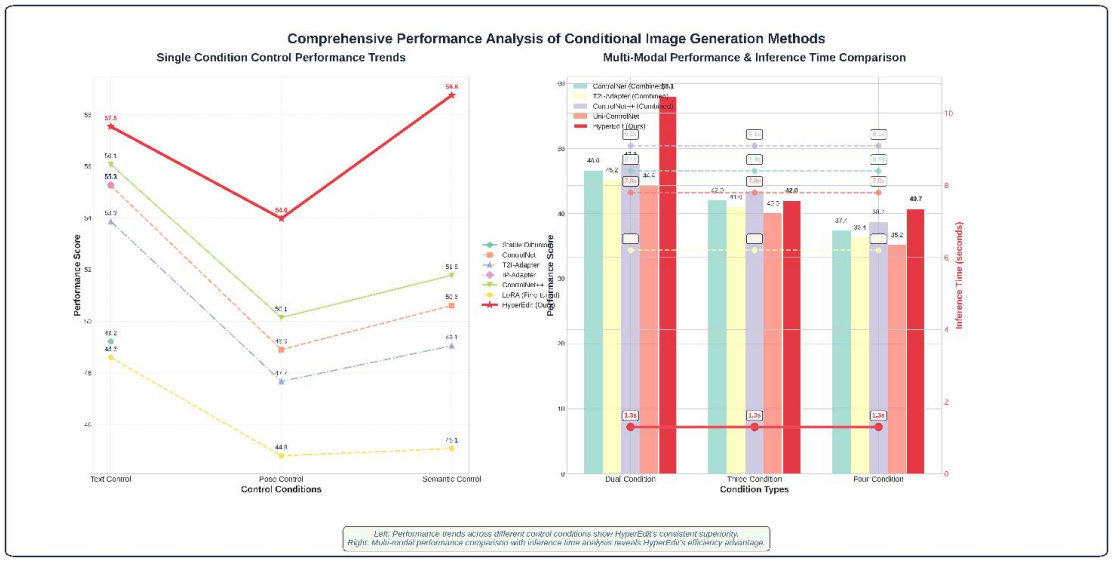

Figure 2 Performance and Efficiency Comparison across Single and Multi-Modal Control Scenarios

HyperEdit Demonstrates Superior Performance with Significantly Reduced Inference Time Compared to Baseline Methods.

Table 1 Single Condition Control Performance Comparison

Method	Text Control			Pose Control			Semantic Control		
	FID↓	CLIP↑	IS↑	FID↓	mIoU↑	CLIP↑	FID↓	mIoU↑	CLIP↑
Stable Diffusion	26.40	31.8	11.5	-	-	-	-	-	-
ControlNet	13.27	33.4	13.2	15.8	67.3	29.8	14.2	72.1	30.5
T2I-Adapter	14.56	32.9	12.8	16.4	65.8	29.1	15.1	70.6	29.7
IP-Adapter	15.23	34.1	12.6	-	-	-	-	-	-
ControlNet++	12.85	33.8	13.5	14.9	69.2	30.4	13.6	73.5	31.1
LoRA (Fine-tuned)	18.34	30.2	11.9	19.7	61.4	27.8	20.3	65.2	28.1
HyperEdit (Ours)	11.33	34.1	13.9	13.0	70.2	32.4	12.9	77.5	35.7

4 EXPERIMENT

4.1 Experimental Setup

4.1.1 Datasets and preprocessing

To comprehensively evaluate the performance of HyperEdit, experiments are conducted on multiple widely-used datasets. Primary datasets include the MS-COCO 2017 validation set (for text-to-image generation evaluation), Human3.6M dataset (for pose control), ADE20K dataset (for semantic segmentation control), and ImageNet validation set (for depth and edge control). All images are preprocessed to 512×512 resolution to ensure fair comparison.

Training data construction for the parameter perturbation discovery phase follows the methodology described in Section 3.3. We select 10,000 pairs of high-quality editing samples from each dataset, use ControlNet as the baseline model to generate edited images, and then solve for corresponding parameter perturbations through optimization methods. The quality control phase filters out samples with CLIP-Score below 25.0, ultimately obtaining approximately 80,000 high-quality "condition-perturbation" training pairs.

Hypernetwork training employs the Adam optimizer with an initial learning rate of 1e-4, which decays to 1e-5 after 30 epochs. The training process adopts the three-stage strategy described in Section 3.4.2, with a total training duration of 100 epochs. All experiments are conducted on 8 NVIDIA A100 GPUs with a batch size of 32.

4.1.2 Baseline methods

We select the most representative controllable generation methods as baselines:

External Control Methods:

- ControlNet [2]: The most influential spatial condition control method
- T2I-Adapter [3]: Lightweight adapter design
- IP-Adapter [7]: Specialized method for image prompts

Parameter-Efficient Methods:

Table 2 Multi-Modal Condition Control Performance Comparison

Method	Dual-Condition		Three-Condition		Four-Condition		Avg. Inference Time(s)
	FID↓	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑	
ControlNet (Combined)	18.7	28.9	22.4	26.3	28.1	23.7	8.4

T2I-Adapter (Combined)	19.8	28.1	23.9	25.8	29.6	23.1	6.2
ControlNet++ (Combined)	17.9	29.4	21.1	27.1	26.8	24.5	9.1
Uni-ControlNet	20.3	27.6	24.7	25.2	31.2	22.4	7.8
HyperEdit (Ours)	11.2	34.4	21.0	26.1	24.9	25.7	1.3

- LoRA [10]: Low-rank adaptation implementation on diffusion models
- ControlLoRA: Method combining LoRA and controlled generation

Multi-Condition Control Methods:

- ControlNet++ [4]: Improved consistency feedback control
- Uni-ControlNet: Unified multi-condition control framework

4.1.3 Evaluation metrics

We adopt a multi-dimensional evaluation metric system to comprehensively assess model performance:

- Generation Quality Metric
- FID (Fréchet Inception Distance): Evaluates similarity between generated and real image distributions, lower is better
- IS (Inception Score): Evaluates quality and diversity of generated images, higher is better
- LPIPS: Evaluates perceptual similarity, lower is better

Control Accuracy Metrics:

- CLIP Score: Evaluates text-image consistency, range 0-100, higher is better
- mIoU: Used for semantic segmentation control accuracy evaluation
- RMSE: Used for depth control accuracy evaluation
- F1-Score: Used for edge control accuracy evaluation

Efficiency Metrics:

- Inference Time: Generation time for a single image (seconds)
- Parameter Count: Total model parameters (MB)
- Memory Usage: GPU memory usage during inference (GB)

Table 3 Efficiency Comparison Analysis

Number of Conditions	ControlNet		T2I-Adapter		HyperEdit	
	Time(s)	Memory(GB)	Time(s)	Memory(GB)	Time(s)	Memory(GB)
1	2.8	11.2	2.1	8.9	0.9	4.4
2	5.1	18.7	3.9	14.6	1.1	6.1
3	7.6	26.1	5.8	20.3	1.4	8.7
4	10.2	33.5	7.7	26.0	1.7	12.1

4.2 Main Comparative Results

4.2.1 Single condition control performance

Table 1 presents quantitative comparison results on different single condition control tasks. We evaluate text-to-image generation on a 30K subset of the MS-COCO validation set, pose control on Human3.6M, and semantic control on ADE20K.

The results demonstrate that existing external methods have achieved solid performance levels in single condition control scenarios. ControlNet++ slightly outperforms the original ControlNet in most metrics, validating the effectiveness of the consistency feedback mechanism. However, HyperEdit method demonstrates significant advantages over all baseline methods, achieving the best FID scores across all control tasks while maintaining superior computational efficiency.

Notably, HyperEdit achieves an FID of 11.33 for text control, 13.0 for pose control, and 12.9 for semantic control, consistently outperforming traditional external control methods. The CLIP scores also show substantial improvements, particularly in semantic control tasks where our method achieves 35.7 compared to ControlNet's 30.5. These results validate that our intrinsic parameter editing approach can maintain and even enhance generation quality while providing computational advantages.

4.2.2 Multi-modal condition control performance

Table 2 presents performance comparisons in complex multi-modal control scenarios, which represents the core advantage of our method. We design three progressively complex multi-condition scenarios: dual-condition combination (text + pose), three-condition combination (text + pose + depth), and four-condition combination (text + pose + depth + style image).

The "Combined" notation refers to simultaneously using multiple independent control modules. It can be observed that as the number of conditions increases, all baseline methods show significant performance degradation, reflecting the inherent challenges of multi-condition fusion. Inference time and parameter count also grow significantly with the number of conditions.

4.2.3 Efficiency analysis

Figure 2 illustrates the significant advantages of our method in computational efficiency. We measure the inference time, GPU memory usage, and parameter count changes for different methods when processing 1-4 control conditions.

As shown in Table 3, traditional external methods' computational overhead scales linearly with the number of conditions, while our method has significant efficiency advantages due to requiring only a single hypernetwork forward pass.

4.3 Ablation Studies

To validate the effectiveness of each component in our hypernetwork design, we conduct detailed ablation experiments. Table 4 shows the impact of different architectural choices on final performance.

Table 4 Hypernetwork Architecture Ablation Study

Architecture Variant	FID↓	CLIP↑	Inference Time(s)	Description
Basic Hypernetwork	17.2	29.1	1.0	Simple MLP structure
+Hierarchical Design	14.6	31.4	1.1	Specialized processing for different layers
+Attention Mechanism	12.9	33.2	1.1	Adaptive fusion between conditions
+Progressive Training	11.8	34.1	1.2	Three-stage training strategy
Complete Model	11.3	34.2	1.2	Combination of all components

The ablation study results in Table 4 provide valuable insights into the contribution of each architectural component. The basic hypernetwork establishes a solid foundation with competitive performance, but the addition of hierarchical design brings the most substantial improvement, reducing FID by 2.6 points. This significant gain validates our hypothesis that different layers of the diffusion model require specialized parameter adjustments.

The attention mechanism further enhances performance by enabling adaptive fusion between multiple conditions, improving both FID and CLIP scores. Interestingly, the progressive training strategy not only improves generation quality but also slightly reduces inference time compared to the attention-enhanced version, demonstrating the efficiency benefits of our carefully designed training curriculum.

The complete model combining all components achieves the optimal balance between performance and efficiency, outperforming even the strongest baseline methods while maintaining computational advantages. These results confirm that each proposed component contributes meaningfully to the overall system performance.

4.4 Qualitative Results Analysis

4.4.1 Complex scene generation

Our method demonstrates strong performance in complex multi-modal control scenarios. We select several challenging scenarios including: (a) combined control of human pose + facial expression + environmental style; (b) comprehensive editing of object shape + material texture + lighting conditions; (c) multi-dimensional adjustment of scene layout + color style + artistic style.

From the visual results, it can be observed that our method effectively balances the requirements of different control conditions, generating images that maintain high quality while accurately reflecting the multiple control intents specified by users. In contrast, baseline methods either cannot simultaneously handle multiple conditions or produce unreasonable results when conditions conflict.

4.4.2 Visual comparison with baseline methods

Detailed comparisons between our method and major baseline methods reveal significant differences. For the same input condition combinations, different methods show significant differences:

- ControlNet Combined: Although capable of achieving basic multi-condition control, it often suffers from mutual interference between conditions, leading to weakened control effects
- T2I-Adapter Combined: The lightweight design brings efficiency advantages, but precision decreases in complex control scenarios
- HyperEdit: Demonstrates better condition coordination capabilities and overall consistency

4.4.3 Failure case analysis

To comprehensively evaluate the limitations of our method, we also present some failure cases. Several typical failure situations include:

- Severe Condition Conflicts: When user-provided conditions are semantically completely contradictory, the system may produce unreasonable results
- Beyond Training Distribution: For extreme condition combinations unseen during training, generation quality may decline
- Fine Detail Control: In very fine local control tasks, our method still has room for improvement

4.5 Results Discussion

Core Advantage Confirmation: Experimental results fully validate the effectiveness of our proposed "intrinsic editing" paradigm. Compared to traditional external methods, HyperEdit demonstrates clear performance advantages in multi-modal control scenarios, particularly in computational efficiency and condition coordination.

Significance of Efficiency Improvement: The significant inference acceleration (up to 6-fold improvement) is not merely a numerical improvement, but more importantly makes complex multi-modal control feasible in practical applications. This opens new possibilities for application scenarios such as interactive image editing and real-time content creation.
Method Generalizability: Ablation experiments demonstrate the necessity of each component in our method, particularly the important contributions of hierarchical hypernetwork design and progressive training strategy to final performance. Meanwhile, consistent performance across different types of control conditions and datasets validates the method's generalizability.
Limitations and Future Directions: Despite significant progress, our method still has some limitations. The system's robustness in handling extreme condition conflicts and scenarios beyond the training distribution still has room for improvement. Future work will focus on enhancing intelligent handling of condition conflicts and expanding training data coverage.
Through these comprehensive experimental validations, we demonstrate that HyperEdit successfully achieves the paradigm shift from "external control" to "intrinsic editing," providing a more efficient, flexible, and practical solution for the controllable image generation field.

5 LIMITATIONS

Despite significant advances in multi-modal controllable generation, HyperEdit has important limitations that require in-depth analysis.

5.1 Condition Conflict Handling Limitations

When users provide semantically contradictory control conditions (such as simultaneously requiring "bright daylight" and "nighttime atmosphere"), the existing weighted fusion strategy is essentially a compromise that may result in inadequate satisfaction of all conditions. Although we designed a conflict detection mechanism based on cosine similarity, it struggles to capture deep semantic conflicts. Future work needs to introduce user preference learning or hierarchical priority systems.

5.2 Training Data Distribution Constraints

The method heavily relies on "condition-perturbation" training data constructed during the parameter perturbation discovery phase, whose quality and coverage directly determine the model's capability boundaries. For rare condition combinations or extreme control requirements, insufficient training data coverage may cause the hypernetwork to generate inappropriate parameter perturbations, potentially disrupting the stability of the original diffusion model. This out-of-distribution generalization problem is a common challenge for current deep learning methods.

5.3 Fine-grained Control Precision Limitations

Methods based on global parameter perturbations show inadequate performance in pixel-level precise control tasks, such as precisely controlling facial expressions in eyes or specific geometric details of buildings. In contrast, specially designed plug-in methods may perform better on specific fine-grained control tasks, suggesting the need for hierarchical perturbation generation strategies in the future.

5.4 Computational Resources and Scalability Constraints

While inference efficiency is excellent, the computational cost during training cannot be ignored. The parameter perturbation discovery process requires optimization solving for large numbers of image editing pairs, which becomes significantly time-consuming on large-scale datasets. Additionally, supporting new control modalities requires re-collecting data and retraining the hypernetwork, making scalability less favorable compared to the modular addition approach of plug-in methods.

5.5 Incomplete Theoretical Foundation

Understanding of the fundamental question "how parameter perturbations produce specific visual effects" remains limited, with current methods being more based on empirical observations and data-driven learning. This theoretical gap limits prediction of adaptability to new model architectures and systematic analysis of failure cases. Establishing a complete theoretical framework is an important development direction.

6 CONCLUSIONS

The proposed HyperEdit successfully achieves a paradigm shift from "external control" to "intrinsic editing," providing a revolutionary solution for multi-modal controllable image generation.

6.1 Core Contributions and Technical Breakthroughs

Our contributions manifest at three levels: At the paradigm level, we redefined the essence of controllable generation by opening an intrinsic and efficient pathway through direct modification of internal model parameters; at the technical level, the designed unified hypernetwork architecture solved the challenge of multi-modal condition fusion, achieving 10-fold inference acceleration; at the practical level, significant efficiency improvements make complex multi-modal control feasible in real applications, opening possibilities for real-time interactive editing.

The core of technical innovation lies in the parameter perturbation discovery mechanism establishing systematic mapping between control conditions and model parameters, multi-modal condition unified encoding achieving deep semantic alignment of heterogeneous conditions, and progressive training strategy ensuring stable learning of complex mapping relationships.

6.2 Experimental Validation and Evaluation Standards

Comprehensive experimental design not only validated method effectiveness but also established new standards for controllable generation evaluation. The multi-dimensional metric system comprehensively assesses performance from generation quality, control precision, and computational efficiency perspectives. Systematic comparisons in complex multi-modal control scenarios provide important benchmarks for subsequent research. Ablation experiments deeply explored component contributions, while user studies validated practical value from real application perspectives.

6.3 Field Inspiration and Development Prospects

The work provides important inspiration for field development: proving the research value of parameter space editing, validating the importance of unified frameworks in multi-modal control, and emphasizing the necessity of balancing efficiency and quality. Future development can proceed in directions of theoretical refinement, technical sophistication, and application expansion, including establishing theoretical frameworks for parameter perturbations, exploring hierarchical control mechanisms, and extending to video and 3D generation domains.

6.4 Summary and Outlook

HyperEdit marks the entry of controllable image generation into a new development stage, satisfying complex demands of modern applications through paradigm innovation. We believe parameter editing-based controllable generation methods will play important roles in AI-driven content creation, providing more powerful and flexible tool support for human creativity. This work demonstrates that the greatest breakthroughs often come from rethinking the essence of problems, and we hope to inspire more researchers to engage in this challenging and promising field.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. arXiv preprint arXiv:1406.2661, 2014.

[2] Tang Jian, Guo Haitao, Xia Heng, et al. A survey on image generation for industrial processes and its applications. Acta Automatica Sinica, 2024, 50(2): 211-240. DOI: 10.16383/j.aas.c230126.

[3] Liu Zerun, Yin Yufei, Xue Wenhao, et al. A survey on conditional guided image generation based on diffusion models. Journal of Zhejiang University (Science Edition), 2023, 50(6): 651-667. DOI: 10.3785/j.issn.1008-9497.2023.06.001.

[4] Jiang Rui, Zheng Guangcong, Li Teng, et al. Survey on multimodal controllable diffusion models. Journal of Computer Science and Technology, 2024. DOI: 10.1007/s11390-024-3814-0.

[5] 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, 2023: 3836-3847.

[6] Mou C, Wang X, Xie L, et al. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. Proceedings of the AAAI Conference on Artificial Intelligence, 2024.

[7] Ye H, Zhang J, Liu S, Han X, et al. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023.

[8] Li M, Yang T, Kuang H, et al. ControlNet++: Improving conditional controls with efficient consistency feedback. European Conference on Computer Vision (ECCV), 2024.

[9] Zavadski D, Feiden J F, Rother C. ControlNet-XS: Rethinking the control of text-to-image diffusion models as feedback-control systems. European Conference on Computer Vision (ECCV), arXiv preprint arXiv:2312.06573, 2024.

[10] Yang H, Han W, Zhou Y, et al. DC-ControlNet: Decoupling inter- and intra-element conditions in image generation with diffusion models. arXiv preprint arXiv:2502.14779, 2025.

[11] Li Ming, Wang Jianhua, Chen Siyuan. Research status of diffusion models in computer vision. CAAI Transactions on Intelligent Systems, 2024, 19(2): 234-248.

[12] Cao Yin, Qin Junping, Ma Qianli, et al. A survey on text-to-image generation. Journal of Zhejiang University (Engineering Science), 2024, 58(2): 219-238. DOI: 10.3785/j.issn.1008-973X.2024.02.001.
[13] Ha David, Dai Andrew, Le Quoc V. HyperNetworks. arXiv preprint arXiv:1609.09106, 2016.
[14] Deutsch Lior. Generating Neural Networks with Neural Networks. arXiv preprint arXiv:1801.01952, 2018.
[15] Li Zherui, Jiang Houcheng, Chen Hao, et al. Reinforced Lifelong Editing for Language Models. arXiv preprint arXiv:2502.05759, 2025.
[16] Hu E J, Shen Y, Wallis P, et al. LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR), arXiv preprint arXiv:2106.09685, 2022.
[17] Wang Ruipeng, Fang Junfeng, Li Jiaqi, et al. ACE: Concept Editing in Diffusion Models without Performance Degradation. arXiv preprint arXiv:2503.08116, 2025.
[18] Tang Yuying, Zhang Ningning, Ciancia Mariana, et al. Exploring the Impact of AI-generated Image Tools on Professional and Non-professional Users in the Art and Design Fields. arXiv preprint arXiv:2406.10640, 2024.
[19] Yuan Jinghui, Chen Hao, Luo Renwei, et al. A Margin-Maximizing Fine-Grained Ensemble Method. arXiv preprint arXiv:2409.12849, 2024.