

RelaCtrl: Relevance-Guided Efficient Control for Diffusion Transformers

Ke Cao^{1,2,3*} Jing Wang^{3,4*} Ao Ma^{3*} Jiasong Feng³ Zhanjie Zhang^{3,5}
 Xuanhua He^{1,2} Shanyuan Liu³ Bo Cheng³ Dawei Leng³ Yuhui Yin³ Jie Zhang^{1,2}

Abstract

The Diffusion Transformer plays a pivotal role in advancing text-to-image and text-to-video generation, owing primarily to its inherent scalability. However, existing controlled diffusion transformer methods incur significant parameter and computational overheads and suffer from inefficient resource allocation due to their failure to account for the varying relevance of control information across different transformer layers. To address this, we propose the Relevance-Guided Efficient Controllable Generation framework, **RelaCtrl**, enabling efficient and resource-optimized integration of control signals into the Diffusion Transformer. First, we evaluate the relevance of each layer in the Diffusion Transformer to the control information by assessing the “ControlNet Relevance Score”—i.e., the impact of skipping each control layer on both the quality of generation and the control effectiveness during inference. Based on the strength of the relevance, we then tailor the positioning, parameter scale, and modeling capacity of the control layers to reduce unnecessary parameters and redundant computations. Additionally, to further improve efficiency, we replace the self-attention and FFN in the commonly used copy block with the carefully designed Two-Dimensional Shuffle Mixer (**TDSM**), enabling efficient implementation of both the token mixer and channel mixer. Both qualitative and quantitative experimental results demonstrate that our approach achieves superior performance with only 15% of the parameters and computational complexity compared to PixArt- δ . Our project homepage is <https://360cvgroup.github.io/RelaCtrl/>.

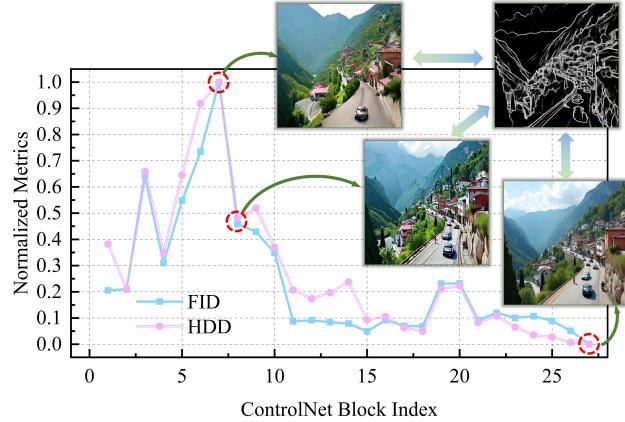


Figure 1. Effect of skipping a specific position within the ControlNet block on the quality of the generated image. Higher FID and HDD indicate a more significant impact of the skipped layer on the quality of the final results, reflecting a stronger correlation with the generated image quality.

1. Introduction

The Diffusion Transformer (DiT) (Peebles & Xie, 2023), with its strong scalability and multi-modal alignment capabilities, has significantly advanced the fields of text-to-image and text-to-video generation (like, PixArt- α (Chen et al., 2023), Flux (BlackForestlabs AI, 2024), Stable Diffusion 3 (Stability AI, 2024), CogVideoX (Yang et al., 2024), Sora (Brooks et al., 2024), HunyuanVideo (Li et al., 2024), and Qihoo-T2X (Wang et al., 2024), etc). By leveraging its robust architecture and scalability, the fidelity of the generated results and consistency with the given textual description are dramatically improved. Recent studies, such as PixArt- δ (Chen et al., 2024) and OminiControl (Tan et al., 2024), focus on controlled text-to-image generation based on the DiT framework, promoting its application in real-world scenarios such as AI-driven content creation and e-commerce shopping.

However, current controlled generation methods for DiT face two main shortcomings. **Firstly**, a significant number of additional parameters and computations are introduced, increasing the burden on training and inference. For example, PixArt- δ directly duplicates the first half of the network’s blocks (i.e., 13 blocks), resulting in a 50% increase in both the number of parameters and computational complexity. Similarly, the control token concatenation in

*Equal contribution ¹University of Science and Technology of China ²Hefei Institutes of Physical Science, Chinese Academy of Sciences ³360 AI Research ⁴Sun Yat-sen University ⁵Zhejiang University. Correspondence to: Dawei Leng <lengdawei@360.cn>, Jie Zhang <zhangjie@iim.ac.cn>.

OminiControl adds only a limited number of parameters but doubles the number of tokens involved in the attention and linear layers, leading to a nearly 70% increase in overall computational complexity. **Secondly**, the varying relevance of control information across different layers of the network is often overlooked, resulting in inefficient allocation of computational resources. Our experiments on “ControlNet Relevance Score”, in which we trained a controlled generative model by copying all blocks and removing control blocks from different layers during inference, revealed that different layers in DiT exhibit varying levels of relevance for control information. As shown in Fig. 1, this relevance follows a trend of increasing and then decreasing, with higher relevance observed in the front-center layers and lower relevance in the deeper layers of the network, resulting in only a slight performance loss (for a more detailed explanation, please refer to Sec. 2.1). Existing methods neglect this variation and apply uniform settings to all layers that introduce control information, resulting in inefficient allocation of parameters and computational resources, such as redundant parameters or computations in layers with low relevance.

To address the above issues, we propose the Relevance-Guided Efficient Controllable Generation framework (i.e., **RelaCtrl**) for diffusion transformer, based on control information relevance analysis. Specifically, to achieve efficient utilization of computational resources, we design relevance-guided allocation and steering strategies. Control blocks are placed at locations with high control information relevance, while locations with weak relevance are left without control blocks. To further reduce the number of parameters and computational complexity introduced by the copy control block operation, we design a lightweight Two-Dimensional Shuffle Mixer (**TDSM**) to replace the self-attention and FFN layers in the copy block. Self-attention plays the role of token mixer and FFN plays the role of channel mixer. Therefore, to efficiently perform the same functions as the token mixer and channel mixer, TDSM first randomly selects a varying number of feature channel groups, then randomly divides the token groups within each channel group, and finally computes the attention within each token-channel group along the token dimension. Theoretical analysis demonstrates that TDSM can overcome the limitations of local grouping and enable non-local modeling in the channel-token dimensions. In addition, we regulate the number of channel division groups in TDSM based on the correlation. In regions with stronger correlation, we reduce the number of channel groups and expand the feature dimensions involved in attention to enhance its modeling capability. The results from multiple controlled experiments demonstrate that our approach achieves superior performance with only a 45M parameter increase (7.4% of PixArt- α) and an additional 46.7 GFLOPs (8.6% of PixArt- α).

The main contribution of this paper can be summarized as

follows:

- We investigate in detail the relevance of control information across different layers of the network, finding that the shallower layers are more sensitive to the control signal, while the deeper layers exhibit weaker relevance to the control effect.
- Based on the relevance analysis, we propose a relevance-guided controlled generation strategy (RelaCtrl), which efficiently allocates the embedding positions of control blocks and the strength of the TDSM modeling capability. This approach minimizes the number of parameters introduced by the control branch and reduces computational complexity without compromising performance.
- We propose a Two-Dimensional Shuffle Mixer (TDSM), which efficiently replaces the self-attention and FFN in the original copy block by calculating attention within randomly divided channel and token groups. Theoretical analysis demonstrates that this design overcomes the limitations of local group modeling, ensuring efficient token-mixing and channel-mixing.
- The experimental results across four different conditional guidance tasks and two text-to-image generation models show that RelaCtrl consistently achieves superior performance while maintaining efficiency, validating the generalization of both the relevance-guided strategy and the proposed TDSM.

2. Methods

2.1. DiT-ControlNet Relevance Prior

In prior studies, the majority of studies on ControlNet have centered on the U-Net architecture. However, the DiT framework (Chen et al., 2023), constructed by stacking a sequence of transformer blocks without an explicit encoder-decoder structure, poses significant challenges for direct adaptation to achieve effective controllability (Chen et al., 2024). To tackle this issue, PixArt- δ introduced a method that duplicates the first 13 Transformer blocks from the DiT model, integrates the outputs of these copied blocks with their corresponding frozen blocks, and forwards the combined results to subsequent frozen modules for further processing. While this approach has demonstrated qualitatively favorable results, it brings notable drawbacks. Simply duplicating the first half of the frozen modules results in a considerable increase in model parameters and computational overhead, leading to prohibitively high training and inference costs, particularly for high-resolution image generation. Moreover, our observations indicate that different replicated layers in DiT-ControlNet contribute unequally to the overall genera-

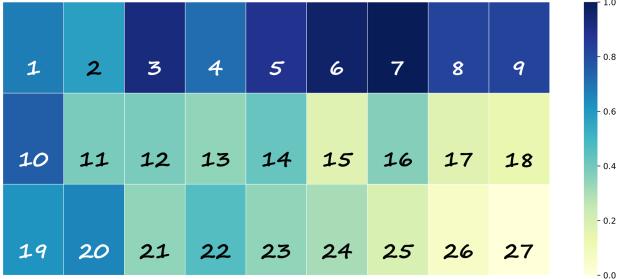


Figure 2. The relevance diagram of different layers in the DiT-ControlNet was calculated based on the FID and HDD ranks. The overall trend shows an initial increase followed by a decrease. The selected placement positions of RelaCtrl in PixArt- α are marked with white numbers.

tion quality and control fidelity. Blindly copying the initial Transformer blocks can introduce unnecessary computational redundancy without proportional performance gains. To systematically evaluate the relevance of individual layers within DiT-ControlNet to generation and controlled quality, we trained a fully controlled PixArt- α network containing 27 replicated modules. During inference, we systematically skip each control block layer and assess its impact on the final generation. For quantitative assessment, we employed the Fréchet Inception Distance (FID) to measure image generation quality and the Hausdorff Distance (HDD) to evaluate control accuracy. These metrics enabled us to analyze the impact of skipping individual blocks from the control branch on overall performance, providing relevant scores for each control block. Finally, we get the ControlNet Relevance Score CRS based on the combination of these two metrics:

$$CRS_i = \frac{1}{2} \left(\frac{F_i - F_{\min}}{F_{\max} - F_{\min}} + \frac{H_i - H_{\min}}{H_{\max} - H_{\min}} \right) \quad (1)$$

Where F and H represent the rank of initial FID and HDD indicators respectively, i shows the index of the control branch block that is removed, and min and max denote the minimum and maximum values within the corresponding rank sequence. If F_i or H_i is higher, it indicates that removing the control block with index i significantly affects the final performance, implying that this module is critically important. Using this approach, we performed single-layer deletions across all replicated blocks in the ControlNet model and derived the regularization metrics and qualitative observations presented in Fig. 1. According to the specified formula 1, the relevance distribution of the ControlNet blocks can be obtained. As illustrated in Fig. 2, the numbers below represent the indices of the control blocks, while the intensity of the shading in each rectangle reflects its relevance to the generation and control performance of the model.

Our findings can be summarized as follows. The most critical modules of DiT-ControlNet are concentrated in the

early-middle layers (e.g., blocks 5, 6, and 7). In contrast, removing the last few modules results in only a minimal decline in performance. Overall, the ControlNet Relevance Score exhibits a trend of initially increasing and then decreasing, which contrasts with observations from prior studies of large language models (Gromov et al., 2024; Zhong et al., 2024; Men et al., 2024) or the main branches of the original DiT architecture (Lee et al., 2024). This indicates that simply increasing or decreasing the number of replicated front transformer blocks in DiT-ControlNet does not offer an effective trade-off between performance and computational cost. Such an approach risks removing control modules essential for maintaining optimal performance. Consequently, we propose dynamically guiding the placement and design of control modules within the network by ranking each layer in the DiT model’s control branch according to its relevance. This strategy ensures a more targeted and efficient utilization of network resources.

2.2. Overall Architecture

Fig. 3 depicts the overall pipeline of our proposed method. Based on the ranking of ControlNet Relevance Score derived in Sec. 2.1 and further validated through ablation studies, we identified and selected the 11 most critical control positions—ranked by relevance from high to low—for integrating the control modules, as is shown in Fig. 2. With this approach, we achieve control performance comparable to PixArt- δ , which utilizes 13 copied modules, while reducing the parameter count by approximately 15%. Although this method effectively decreases the model size and computational overhead, significant redundancy remains in the internal design of the control modules. The transformative power of the Transformer architecture, as emphasized by MetaFormer (Yu et al., 2022), lies in its holistic design, wherein the attention mechanism serves as a token mixer by enabling token-level information mixing, while the remaining components, such as the FFN layers, function as channel mixers to facilitate the integration of channel-wise information. To address the substantial redundancy in the FFN layers within the channel mixer (Pires et al., 2023), we introduce a lightweight module, the Relevance-Guided Lightweight Control Block (RGLC), which unifies token mixing and channel mixing into a single operation. Specifically, we replace the attention and FFN layers in the original PixArt Transformer block with a novel Two-Dimensional Shuffle Mixer (TDSM) design, streamlining the architecture for enhanced efficiency. This method facilitates information interaction and modeling across both the token and channel dimensions, significantly reducing the parameter count and computational demands of the replicated blocks.

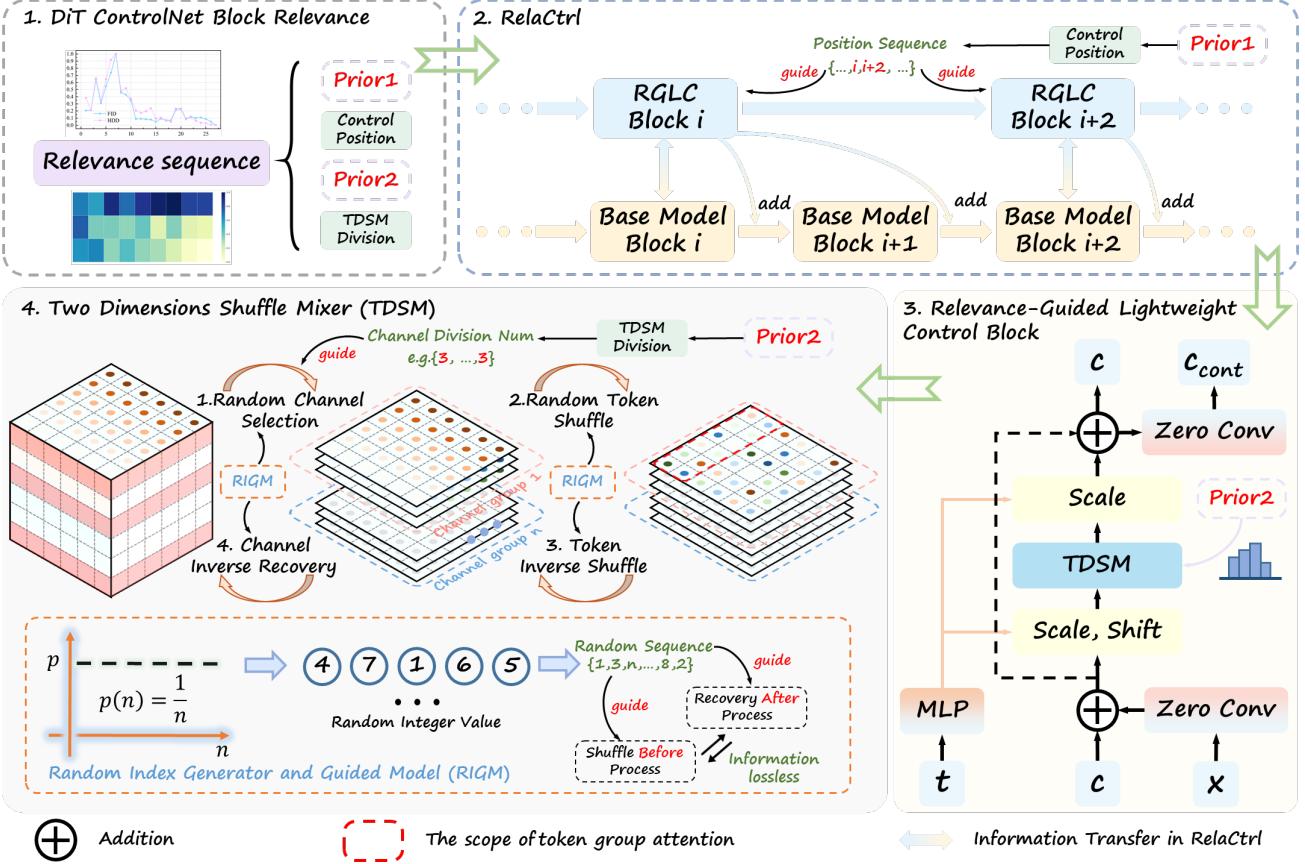


Figure 3. The overall architecture of RelaCtrl. Control block locations are prioritized based on the ControlNet Relevance Score, ranked from highest to lowest. The direct duplication of the main branch in the original ControlNet is replaced with the carefully designed Reference-Guided Lightweight control block. Additionally, the Two-Dimensional Shuffle Mixer effectively reduces model parameters and computational overhead while preserving performance.

2.3. Relevance-Guided Lightweight Control Block

The third part of Fig. 3 illustrates the detailed structure of the RGLC Block. The module takes three inputs: the control condition input c , the diffusion timesteps embedding t used to compute the weights for feature normalization, and the input x from the corresponding frozen block. To enhance information interaction between the control branch and the frozen backbone network, x is passed through a zero convolution layer and added to the conditional input c , producing c_{in} . The resulting c_{in} is then processed by the Two-Dimensions Shuffle Mixer (TDSM). Following the processing, the output is passed through a zero convolution layer, resulting in c_{cond} which is added to the main branch to provide control guidance. This process can be formally expressed as follows:

$$c_{cond} = \text{ZC}(\text{TDSM}(c_{in}) + c_{in}) \quad (2)$$

Where ZC refers to the operation of zero convolution. To address the additional computational overhead introduced by the self-attention mechanism, TDSM employs locally grouped self-attention with shuffle characteristics. This design significantly reduces the computational complexity of

the network while preserving non-local information interactions within the groups. A detailed analysis of this process is provided in Sec. 2.4.

2.4. Two Dimensions Shuffle Mixer

From the perspective of MetaFormer (Yu et al., 2022), the effectiveness of Transformers can be attributed to two key components: the token mixer, implemented via the self-attention mechanism, and the channel mixer, realized through the feed-forward network (FFN) layer. However, studies have revealed that the FFN is often highly redundant despite consuming a significant proportion of the model parameters (Pires et al., 2023). To alleviate the computational burden of the control branch, we propose grouping tokens for computation while employing specific strategies to enhance interaction and modeling capacity across token groups (Huang et al., 2021; Cao et al., 2024). Departing from traditional shuffle methods that operate exclusively on the token dimension, we extend the token mixer in the Transformer architecture to model non-local interactions within local windows by introducing a novel approach that jointly

operates on both the token and channel dimensions. This dual-dimension strategy enables more efficient and effective modeling. Specifically, We begin by performing random channel selection, followed by random shuffling of the input sequence across the 3D dimension space. Afterward, local self-attention calculations are applied. Although the subsequent attention mechanism is confined to a fixed group, the tokens involved may originate outside this group. This operation effectively disrupts the inherent relationships between tokens and introduces the information flow between channels to some extent, thereby breaking the interaction constraints typically imposed by local attention. To provide theoretical validation, we first present the following definition:

Definition 2.1. (Local Partition). For the local group where self-attention calculations are performed, its 3D size $s \times s \times d$ satisfies $s \ll H$, $s \ll W$, and $d \ll D$, where $H \times W \times D$ denotes the dimensions of the module input after the token positions have been arranged.

Definition 2.2. (Random Selection and Shuffle Function). $S : c_{in} \rightarrow c_{rs}^i$, $i \in [1, n]$ denote the random selection and shuffle function. This function randomly divides the input $c_{in} \in \mathbb{R}^{H \times W \times D}$ into n parts along the channel dimension, and then scrambles the elements within each part in the 3D space. As a result, $c_{rs}^i \in \mathbb{R}^{H \times W \times d_i}$, $i \in [1, n]$, where $\sum_{i=1}^n d_i = D$.

Regardless of the initial distances between tokens, the random selection and shuffle operations may place any two tokens within the same window, making it possible to model non-local relationships at both the token and channel levels within local windows.

Definition 2.3. (Interactive Set and Interactive Distance). Consider $c_{rs}^i \in \mathbb{R}^{H \times W \times d_i}$, where we define the set of token pairs within the same local group after the random shuffle operation as $I_S = \{(t_j, t_k)\}$. This set indicates that, after the random shuffle function S , the token indices t_j and t_k are placed within the same local group. At this point, the interactive distance between the two tokens can be defined as d_S :

$$d_S(t_j, t_k) = \begin{cases} \|t_j - t_k\|_2, & \text{if } (t_j, t_k) \in I_S \\ \infty, & \text{if } (t_j, t_k) \notin I_S \end{cases} \quad (3)$$

Here, ∞ indicates that the interaction between the two tokens cannot be captured within the current local group, which does not affect the derivation process of distance modeling. Therefore, the expected interactive distance $d(t_j)$ of token t_j can be defined as follows:

$$d(t_j) = \mathbb{E}_{t_k | (t_j, t_k) \in I_S} [d_S(t_j, t_k)] \quad (4)$$

It can be proven that the lower bound of $d(t_j)$ is $\Omega(\frac{\sqrt{2}}{4}(H + Wd_i))$. The detailed derivation steps can be found in the Appendix B.1.

Theorem 2.4. The lower bound of $d(t_j)$ is:

$$\begin{aligned} d(t_j) \geq & \frac{\sqrt{2}}{4(HWd_i - 1)} \left[Ht_{wj}(t_{wj} + 1) \right. \\ & + Wd_i t_{hj}(t_{hj} + 1) \\ & + H(Wd_i - t_{wj})(Wd_i - t_{wj} - 1) \\ & \left. + Wd_i(H - t_{hj})(H - t_{hj} - 1) \right] \end{aligned} \quad (5)$$

Corollary 2.5. Let $\bar{d}(t_j)$ denote the average interactive distance in c_{rs}^i , and it can be proved that $\bar{d}(t_j) \approx d(t_j)$, with the detailed proof provided in the Appendix B.2.

According to Corollary 2.5, the average distance that can be captured by grouped attention in TDSM is $\Omega(\frac{\sqrt{2}}{4}(H + Wd_i))$, enabling the modeling of non-local interactions. In summary, the random selection in the channel and shuffle operation we introduced disrupts the positional order of tokens. However, since the token order is not explicitly modeled within the self-attention mechanism for visual tasks, this operation does not compromise the effectiveness of the process. Nonetheless, the arrangement of input tokens may affect the semantic information embedded in the latent code during recovery. To resolve this problem, we propose an inverse recovery operation applied to both the token and channel dimensions following the self-attention computation. This overall method with shuffle and recovery is termed the Two Dimensions Shuffle Mixer (TDSM), leverages the capability of this reversible transformation pairs to ensure information preservation during self-attention calculations, thereby enabling efficient non-local information interaction across both the channel and token dimensions.

3. Experiment

3.1. Experiment Setup

Evaluation Metrics. To comprehensively assess the quality of the generated images, we employed multiple evaluation metrics. The Fréchet Inception Distance (FID) (Heusel et al., 2017) and CLIP-Aesthetics score (Schuhmann et al., 2022) measure the visual fidelity of the generated images, while the CLIP Score (Hessel et al., 2021) evaluates text consistency. The control fidelity is explicitly assessed through the distance between the generated and target images. For images guided by HED and Canny conditions, we utilized the Hausdorff Distance (HDD) (Huttenlocher et al., 1993) for evaluation. For depth conditions, we used MSE-depth, and for segmentation map control, we employed the mean Intersection over Union (mIoU).

Table 1. Quantitative comparisons of different methods on the COCO validation set (Lin et al., 2014). The best results are highlighted in **bold**, while the second-best results are marked with underline.

Model	Method	Controllability		Quality		Text Consistency		Controllability		Quality		Text Consistency	
		Canny				Hed				Hed			
		HDD↓	FID↓	CLIP-Ae↑	CLIP-Score↑	HDD↓	FID↓	CLIP-Ae↑	CLIP-Score↑	HDD↓	FID↓	CLIP-Ae↑	CLIP-Score↑
SD1.5	Uni-ControlNet	<u>95.40</u>	33.81	5.207	0.259	<u>98.78</u>	59.72	5.086	0.252	100.52	91.94	4.819	0.251
	Uni-Control	97.90	91.29	4.965	0.249								
SDXL	ControlNet-XS	101.34	21.57	5.134	0.286	-	-	-	-	-	-	-	-
	ControlNext	117.59	49.32	4.816	0.275	-	-	-	-				
PixArt- α	PixArt- δ	96.26	<u>21.38</u>	<u>5.508</u>	0.279	98.91	<u>29.22</u>	<u>5.243</u>	0.275	<u>98.91</u>	<u>27.73</u>	5.451	<u>0.276</u>
	RelaCtrl	94.04	20.34	5.584	<u>0.282</u>	96.11	27.73	5.451	<u>0.276</u>				
Model	Method	Depth				Segmentation							
		MSE-d↓	FID↓	CLIP-Ae↑	CLIP-Score↑	mIoU↑	FID↓	CLIP-Ae↑	CLIP-Score↑	mIoU↑	FID↓	CLIP-Ae↑	CLIP-Score↑
	Uni-ControlNet	102.75	43.17	5.230	0.250	0.316	40.83	5.270	0.255	102.46	43.17	5.230	0.258
SDXL	ControlNet-XS	<u>99.20</u>	<u>34.38</u>	5.235	0.281	-	-	-	-	101.63	73.26	4.919	0.253
	ControlNext	101.63	73.26	4.919	0.253	-	-	-	-				
PixArt- α	PixArt- δ	99.69	35.21	<u>5.723</u>	<u>0.283</u>	0.379	<u>35.50</u>	<u>5.668</u>	0.282	<u>99.11</u>	<u>33.93</u>	5.887	<u>0.285</u>
	RelaCtrl	99.11	33.93	5.887	0.285	0.405	33.76	5.702	0.287				

Baseline. We conducted a comparative analysis of RelaCtrl against state-of-the-art (SOTA) control techniques, including Uni-ControlNet (Zhao et al., 2024), UniControl (Qin et al., 2023), ControlNet-XS (Zavadski et al., 2025), ControlNext (Peng et al., 2024), and PixArt- δ (Chen et al., 2024). For the first four methods, we utilized their officially released pre-trained weights. It is worth noting that ControlNet-XS and ControlNext only provided weights for Canny and Depth conditions, with the Canny weights for ControlNext being trained on an animation dataset. To further highlight the performance of our proposed method, we trained the control branch for both PixArt- δ and our method entirely from scratch under identical experimental settings, enabling a rigorous quantitative and qualitative comparison.

3.2. Compared with SOTA methods

As shown in Table 1, we comprehensively evaluated the proposed method against existing controllable generation techniques across four conditional control tasks. Our method consistently outperforms alternatives in control accuracy, as demonstrated by the superior performance on control indicators for various conditions, highlighting its precision in generating controlled images. Furthermore, our approach achieves consistently better results in terms of FID and CLIP aesthetic scores, reflecting enhanced image quality. In text similarity evaluations, the CLIP Score confirms that our method achieves superior text-image consistency across diverse tasks, demonstrating improved semantic alignment while maintaining high control accuracy and visual fidelity.

Fig. 4 presents a visual comparison of RelaCtrl and other

methods under Canny and Depth conditions. Our method significantly reduces computational complexity and parameter usage while maintaining generation quality comparable to PixArt-ControlNet, outperforming other approaches based on SD1.5 and SDXL. Additional controllable generation results of RelaCtrl are shown in Fig. 5, further illustrating its effectiveness in extracting and injecting information from conditional images during the generation process. This enables the production of results that align seamlessly with conditional controls.

3.3. Algorithm efficiency analysis

We conducted a comprehensive analysis of RelaCtrl and ControlNet with 13 replicated modules (PixArt- δ) under consistent experimental settings. Specifically, the input data was in bfloat16 format with a resolution of 512, the test batch size was set to 1, and all experiments were performed on an NVIDIA A100 GPU. Table 2 presents a comparative evaluation of the two methods, with each indicator expressed as a percentage relative to the original base model. Notably, memory consumption excludes usage by the CLIP and T5 encoders. Compared to the original PixArt- α , RelaCtrl incurred a modest 7.38% increase in parameters and an 8.61% increase in computational complexity. These increments are significantly lower than those of the ControlNet method, which demonstrates nearly a 50% increase in both parameters and complexity. Additionally, RelaCtrl showed some advantages in memory usage and inference time, although it is worth noting that inference time is predominantly influenced by the speed of the main network rather than the

RelaCtrl: Relevance-Guided Efficient Control for Diffusion Transformers

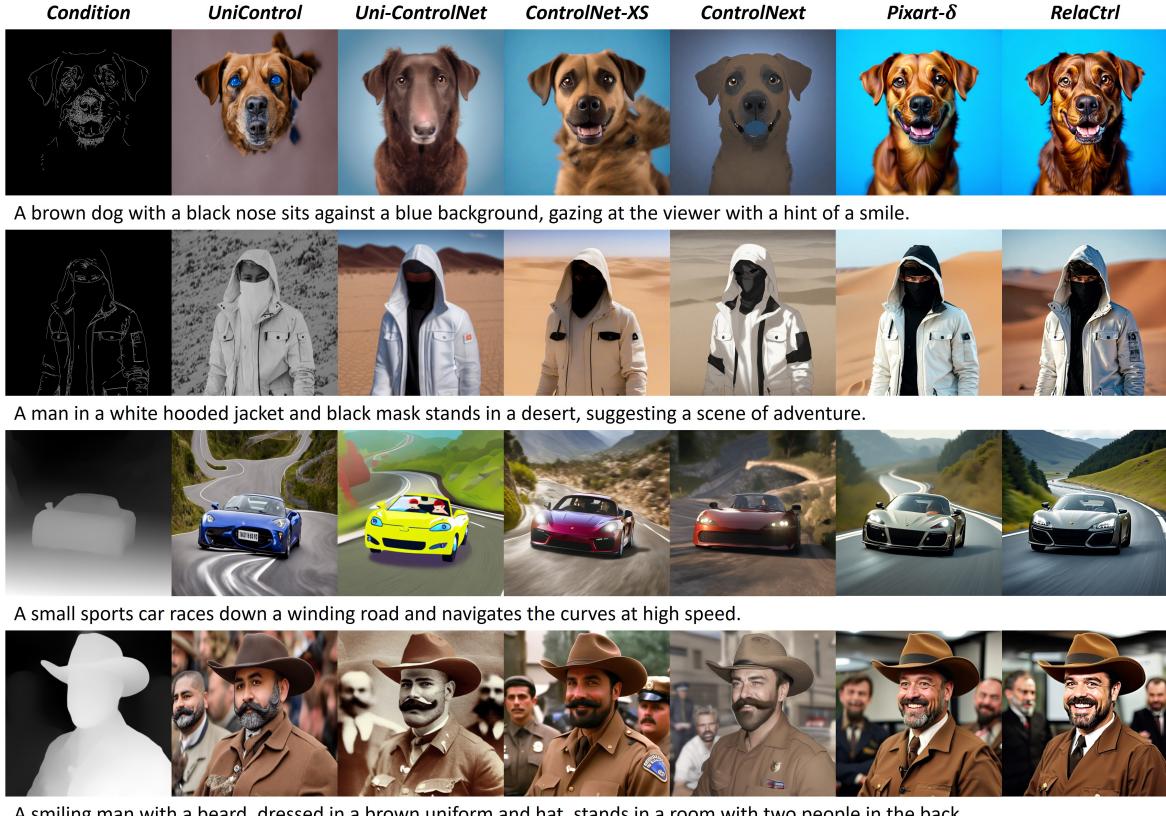


Figure 4. Qualitative comparison of different methods. Please zoom in for better details.

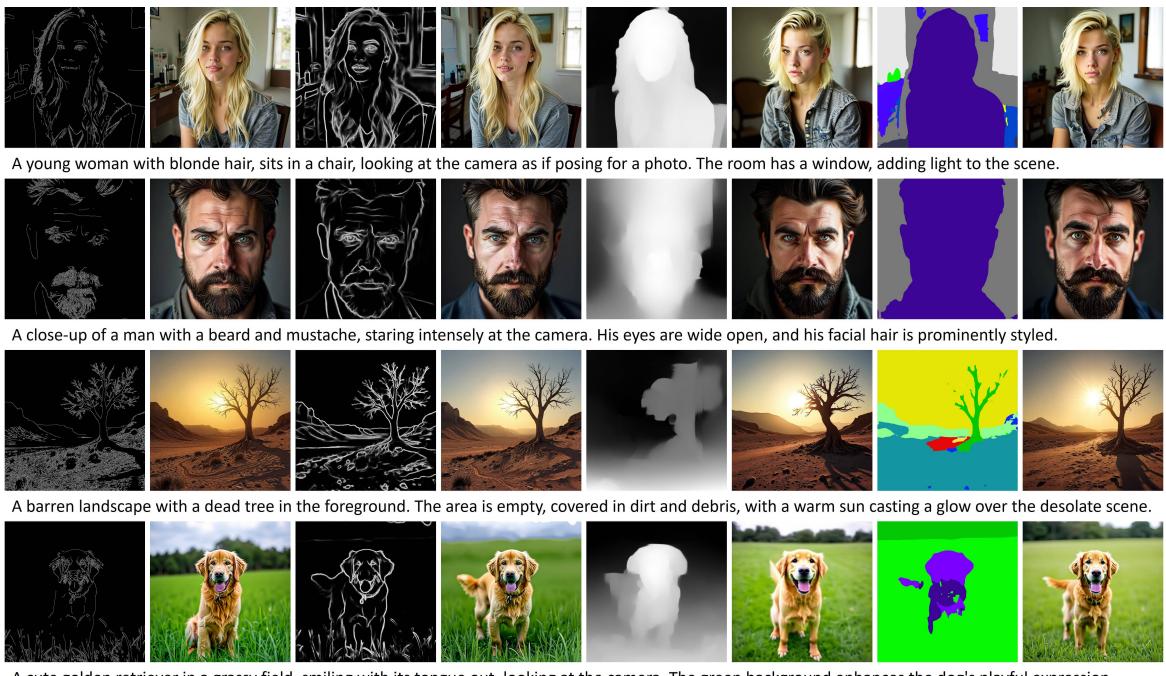


Figure 5. Generation effects of RelaCtrl under varying control conditions.

control branch, leading to similar small increases in this metric for both methods. Overall, RelaCtrl outperforms the original ControlNet method applied to the DiT model by achieving comparable or superior metrics and visual results while significantly reducing computational resource consumption. This demonstrates the effectiveness and efficiency of the proposed RelaCtrl framework.

Table 2. Evaluation of the proposed method’s effectiveness, with the following units for the four metrics: Parameters (M), Complexity (GFLOPs), Inference Time (s), and Memory Usage (MiB)

Method	Parameters	Complexity	Inference	Memory
PixArt- α	611.15	542.56	3.81	2114
w/ ControlNet	+294.34	+270.57	+0.51	+1694
	(+48.16%)	(+49.87%)	(+13.39%)	(+80.13%)
w/ RelaCtrl	+45.15	+46.71	+0.24	+395
	(+7.38%)	(+8.61%)	(+6.30%)	(+18.70%)

3.4. Ablation study

We conducted quantitative experiments to evaluate the efficacy of guiding control block placement based on DiT-ControlNet Relevance. Utilizing relevance scores, we ranked the control block positions from highest to lowest and placed copy blocks at the top-ranked positions, evaluating configurations with 13, 12, 11, and 10 blocks. The results are summarized in Table 3. As anticipated, reducing the number of control blocks resulted in a gradual decline in FID and HDD metrics. Notably, under the guidance of relevance scores, using only 11 control blocks achieved performance comparable to the original ControlNet with 13 replicated blocks. When further reduced to 10 blocks, the quality of the generated results slightly declined relative to the 13-block setup. Consequently, we selected the top 11 positions ranked by relevance scores for control placement, as illustrated by the white numbers in Fig. 2.

To further validate the significance of the RGLC block and the incorporation of Prior 2 based on revelation, we conducted additional quantitative experiments which are shown in Table 4. Employing the ControlNet with 13 copy blocks as the baseline and the RelaCtrl network as the foundation, we evaluated the impact of removing the RGLC block and relevance prior 2, respectively. The results demonstrate that the absence of either component leads to a decline in image quality and control performance. Replacing the RGLC block with the original copy block not only significantly increases the parameter scale but also results in performance degradation, highlighting the efficacy of the RGLC block. In the experiment (w/o Prior 2), we applied a uniform TDSM channel division across all RGLC blocks and eliminated our correlation-specific settings. The poorer HDD and FID metrics observed indicate that allocating more parameters and computational resources to control positions with higher

correlation is crucial for achieving high-quality controllable generation. These results underscore the importance of both the RGLC block and relevance prior 2 in enhancing generation quality and control precision.

Table 3. The impact of control block placement guided by DiT-ControlNet Relevance. ControlNet-top13, which directly replicates the first 13 blocks of the main branch, serves as the baseline for parameter volume comparison.

Setting	HDD \downarrow	FID \downarrow	Para Ratio
ControlNet-top13	96.26	21.38	100%
Relevance-top13	94.57	20.31	100%
Relevance-top12	95.88	20.79	92.5%
Relevance-top11	95.57	21.28	84.6%
Relevance-top10	96.36	22.24	76.9%

Table 4. The impact of the RGLC block and the number of TDSM partitions within it on generation performance. The PixArt- δ with 13 copied blocks serves as the baseline for parameter comparison.

Setting	HDD \downarrow	FID \downarrow	Para Ratio
RelaCtrl	94.04	20.34	15.3%
w/o RGLC	95.57	21.28	84.6%
w/o Prior2	97.30	22.47	17.1%
Baseline	96.26	21.38	100%

3.5. RelaCtrl in Flux

To further validate the effectiveness and generalizability of the proposed method, we conducted additional experiments on Flux.1-dev, with the results provided in the Appendix D.2

4. Conclusion

In this paper, we explore the control information relevance across different layers in the diffusion transformer by the “ControlNet Relevance Score” experiments. We discover that layers with strong relevance to control information are located in the shallow to middle layers, while the deeper layers exhibit weaker relevance. Then, we propose a relevance-guided strategy (RelaCtrl) for allocating control block insertion positions and adjusting the block’s modeling capabilities, enhancing the efficiency of control information integration. Additionally, we design TDSM, which efficiently replaces the original self-attention and FFN through a randomized channel-token grouping attention mechanism. Experimental results show that RelaCtrl achieves superior performance across two T2I models and four conditional guidance generation tasks, while also exhibiting significant efficiency advantages. We hope RelaCtrl provides valuable insights and references for controlled generation research based on diffusion transformers.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

- BlackForestlabs AI Flux. <https://blackforestlabs.ai/#get-flux>, 2024. Accessed: 2024-09-03.
- Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., and Ramesh, A. Video generation models as world simulators. 2024. URL <https://openai.com/research/video-generation-models-as-world-simulators>.
- Cao, K., He, X., Hu, T., Xie, C., Zhang, J., Zhou, M., and Hong, D. Shuffle mamba: State space models with random shuffle for multi-modal image fusion. *arXiv preprint arXiv:2409.01728*, 2024.
- Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.
- Chen, J., Wu, Y., Luo, S., Xie, E., Paul, S., Luo, P., Zhao, H., and Li, Z. Pixart- δ : Fast and controllable image generation with latent consistency models. *arXiv preprint arXiv:2401.05252*, 2024.
- Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Feng, J., Ma, A., Wang, J., Cheng, B., Liang, X., Leng, D., and Yin, Y. Fancyvideo: Towards dynamic and consistent video generation via cross-frame textual guidance. *arXiv preprint arXiv:2408.08189*, 2024.
- Gromov, A., Tirumala, K., Shapourian, H., Glorioso, P., and Roberts, D. A. The unreasonable ineffectiveness of the deeper layers. *arXiv preprint arXiv:2403.17887*, 2024.
- Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., and Dai, B. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.
- He, X., Liu, Q., Qian, S., Wang, X., Hu, T., Cao, K., Yan, K., and Zhang, J. Id-animator: Zero-shot identity-preserving human video generation. *arXiv preprint arXiv:2404.15275*, 2024.
- Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- Huang, Z., Ben, Y., Luo, G., Cheng, P., Yu, G., and Fu, B. Shuffle transformer: Rethinking spatial shuffle for vision transformer. *arXiv preprint arXiv:2106.03650*, 2021.
- Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993.
- Lee, Y., Lee, Y.-J., and Hwang, S. J. Dit-pruner: Pruning diffusion transformer models for text-to-image synthesis using human preference scores. In *European Conference on Computer Vision (ECCV) 2024*, pp. 1–9, 2024.
- Li, Z., Zhang, J., Lin, Q., Xiong, J., Long, Y., Deng, X., Zhang, Y., Liu, X., Huang, M., Xiao, Z., et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.
- Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- Men, X., Xu, M., Zhang, Q., Wang, B., Lin, H., Lu, Y., Han, X., and Chen, W. Shortgpt: Layers in large language models are more redundant than you expect. *arXiv preprint arXiv:2403.03853*, 2024.
- Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.
- Peng, B., Wang, J., Zhang, Y., Li, W., Yang, M.-C., and Jia, J. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.
- Pires, T. P., Lopes, A. V., Assogba, Y., and Setiawan, H. One wide feedforward is all you need. *arXiv preprint arXiv:2309.01826*, 2023.

- Qin, C., Zhang, S., Yu, N., Feng, Y., Yang, X., Zhou, Y., Wang, H., Niebles, J. C., Xiong, C., Savarese, S., et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294, 2022.
- Shuai, X., Ding, H., Ma, X., Tu, R., Jiang, Y.-G., and Tao, D. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*, 2024.
- Stability AI. Stablediffusion3. <https://stability.ai/news/stable-diffusion-3>, 2024. Accessed: 2024-09-03.
- Tan, Z., Liu, S., Yang, X., Xue, Q., and Wang, X. Omnicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024.
- Wang, J., Ma, A., Feng, J., Leng, D., Yin, Y., and Liang, X. Qihoo-t2x: An efficient proxy-tokenized diffusion transformer for text-to-any-task. *arXiv preprint arXiv:2409.04005*, 2024.
- Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.
- Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., and Yan, S. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10819–10829, 2022.
- Zavadski, D., Feiden, J.-F., and Rother, C. Controlnet-xs: Rethinking the control of text-to-image diffusion models as feedback-control systems. In *European Conference on Computer Vision*, pp. 343–362. Springer, 2025.
- Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.
- Zhang, Z., Zhang, Q., Xing, W., Li, G., Zhao, L., Sun, J., Lan, Z., Luan, J., Huang, Y., and Lin, H. Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7396–7404, 2024.
- Zhao, S., Chen, D., Chen, Y.-C., Bao, J., Hao, S., Yuan, L., and Wong, K.-Y. K. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.
- Zhong, L., Wan, F., Chen, R., Quan, X., and Li, L. Block-pruner: Fine-grained pruning for large language models. *arXiv preprint arXiv:2406.10594*, 2024.

A. Related works

A.1. Diffusion-based models

In recent years, diffusion-based methods have garnered significant success in the field of generation (He et al., 2024; Feng et al., 2024; Zhang et al., 2024), particularly in text-to-image (T2I) generation (Guo et al., 2023; Shuai et al., 2024). These methods utilize text embeddings derived from pre-trained language encoders such as CLIP (Radford et al., 2021), Bert (Devlin, 2018) and T5 (Raffel et al., 2020) to generate images with high fidelity and diversity through an iterative denoising process. Recently, the introduction of the latent diffusion model (Rombach et al., 2022) has marked a significant advancement in this field, enhancing the quality and efficiency of generated content. In pursuit of greater scalability and enhanced generation quality, models like DiT (Peebles & Xie, 2023; Chen et al., 2023) integrate large-scale Transformer architectures into the diffusion framework, pushing the boundaries of generative performance. Building on this foundation, Flux (BlackForestlabs AI, 2024) synthesizes flow-matching (Lipman et al., 2022) and Transformer-based architecture to achieve state-of-the-art performance.

A.2. Controllable generation with diffusion models

Controllable generation has emerged as a prominent area of research in diffusion models (Zhang et al., 2023; Qin et al., 2023; Zhao et al., 2024; Chen et al., 2024; Peng et al., 2024). Currently, there are two main approaches to incorporating controllable conditions into image generation: (1) training a large diffusion model from scratch to enable control under multiple conditions, and (2) fine-tuning a lightweight structure while keeping the original pre-trained model frozen. However, the first approach demands significant computational resources, which limits its accessibility for broader dissemination and personal use. In contrast, recent studies have explored the addition of supplementary network structures to pre-trained diffusion models, allowing for control over the generated outputs without the need to retrain the entire model.

ControlNet (Zhang et al., 2023) enables image generation that aligns with control information by reproducing specific layers within the network and connecting them to the original layers using zero convolution. Building on this foundation, Uni-Control (Qin et al., 2023) introduces a Mixture-of-Experts (MoE) framework, which unifies control across multiple spatial conditions. ControlNet-XS (Zavadski et al., 2025) further improves the interaction bandwidth and frequency between the control and main branches within the ControlNet architecture, drawing inspiration from principles of feedback control systems. Nonetheless, these approaches are primarily based on the U-Net structure and may not yield the desired results when directly applied to Diffusion Transformers (DiT) without modification (Chen et al., 2024). PixArt- δ (Chen et al., 2024) proposed a design methodology specifically tailored for DiT, but directly copying the first half of the network results in a 50% increase in both parameter count and computational complexity, resulting in high computational cost and inconvenient for community research and practical deployment.

B. Proof to Theorem

B.1. Proof to Theorem 3.4

For token, $t_j = (t_{hj}, t_{wj})$, each remaining token in the input sequence has an equal probability of being shuffled into the same group as t_j . Under this condition, the expected interactive distance between tokens can be calculated as follows:

$$\begin{aligned} d(t_j) &= \mathbb{E}_{t_k | (t_j, t_k) \in I_S} [d_S(t_j, t_k)] \\ &= \frac{1}{HWd_i - 1} \sum_{h=0}^{H-1} \sum_{w=0}^{Wd_i-1} \sqrt{(h - t_{hj})^2 + (w - t_{wj})^2} \end{aligned} \tag{6}$$

Using the following mean inequality

$$\sqrt{\frac{x^2 + y^2}{2}} \geq \frac{x + y}{2} \tag{7}$$

we can calculate a lower bound on $d(t_j)$:

$$\begin{aligned} d(t_j) &= \frac{1}{HWd_i - 1} \sum_{h=0}^{H-1} \sum_{w=0}^{Wd_i-1} \sqrt{(h - t_{hj})^2 + (w - t_{wj})^2} \\ &\geq \frac{1}{HWd_i - 1} \sum_{h=0}^{H-1} \sum_{w=0}^{Wd_i-1} \frac{\sqrt{2}}{2}(|h - t_{hj}| + |w - t_{wj}|) \end{aligned} \quad (8)$$

Due to the following formulas:

$$\begin{aligned} \sum_{h=0}^{H-1} |h - t_{hj}| &= \sum_{h=0}^{t_{hj}} |t_{hj} - h| + \sum_{h=t_{hj}}^{H-1} |h - t_{hj}| \\ &= \frac{t_{hj}(t_{hj} + 1)}{2} + \frac{(H - t_{hj})(H - t_{hj} - 1)}{2} \end{aligned} \quad (9)$$

and

$$\begin{aligned} \sum_{w=0}^{Wd_i-1} |w - t_{wj}| &= \sum_{w=0}^{t_{wj}} |t_{wj} - w| + \sum_{w=t_{wj}}^{Wd_i-1} |w - t_{wj}| \\ &= \frac{t_{wj}(t_{wj} + 1)}{2} + \frac{(Wd_i - t_{wj})(Wd_i - t_{wj} - 1)}{2} \end{aligned} \quad (10)$$

we can obtain

$$\begin{aligned} d(t_j) &\geq \frac{\sqrt{2}}{4(HWd_i - 1)} [Hm_{wj}(m_{wj} + 1) + Wd_im_{hj}(m_{hj} + 1) \\ &\quad + H(Wd_i - m_{wj})(Wd_i - m_{wj} - 1) + Wd_i(H - m_{hj})(H - m_{hj} - 1)] \end{aligned} \quad (11)$$

thus proving the original theorem.

B.2. Proof to Corollary 3.5

For tokens within the same local group, we calculate their average interactive distance:

$$\bar{d}(t_j) = \frac{1}{s^2 - 1} \sum_{j=1}^{s^2-1} d(t_j | HWd_i - j) \quad (12)$$

Among these, $d(t_j | HWd_i - j)$ represents the expected interactive distance when the number of remaining tokens is $HWd_i - j$. Based on Assumption 2.1, we know $j \in [2, s^2 - 1]$, and we can derive the following approximation:

$$d(t_j | HWd_i - j) \approx d(t_j | HWd_i - 1) = d(t_j) \quad (13)$$

Therefore, the value of $\bar{d}(t_j)$ can be estimated as follows:

$$\begin{aligned} \bar{d}(t_j) &= \frac{1}{s^2 - 1} \sum_{j=1}^{s^2-1} d(t_j | HWd_i - j) \\ &\approx \frac{1}{s^2 - 1} \sum_{j=1}^{s^2-1} d(t_j | HWd_i - 1) \\ &= \frac{1}{s^2 - 1} \sum_{j=1}^{s^2-1} d(t_j) \\ &= d(t_j) \end{aligned} \quad (14)$$

C. Training Details

Training and Testing Datasets. We curated a dataset containing 1.73 million high-resolution, high-quality images, each with an aesthetic score of 5.5 or higher. Following the production methodology outlined in ControlNet (Zhang et al., 2023), we generated various types of conditional images, including HED, Canny, Depth, and Segment. For the quantitative experiments, the model was trained at a resolution of 512 and evaluated on the COCO validation dataset (Lin et al., 2014), consisting of 5000 images. For qualitative visual evaluation, the model was trained at a resolution of 1024 and tested on a high-quality, high-resolution test set of 1000 images. To ensure a fair evaluation, all experiments were conducted with a training of 5 epochs on 16 NVIDIA A100 GPUs, using identical settings across all trials.

D. Additional experimental results

D.1. More exploration on Relevance

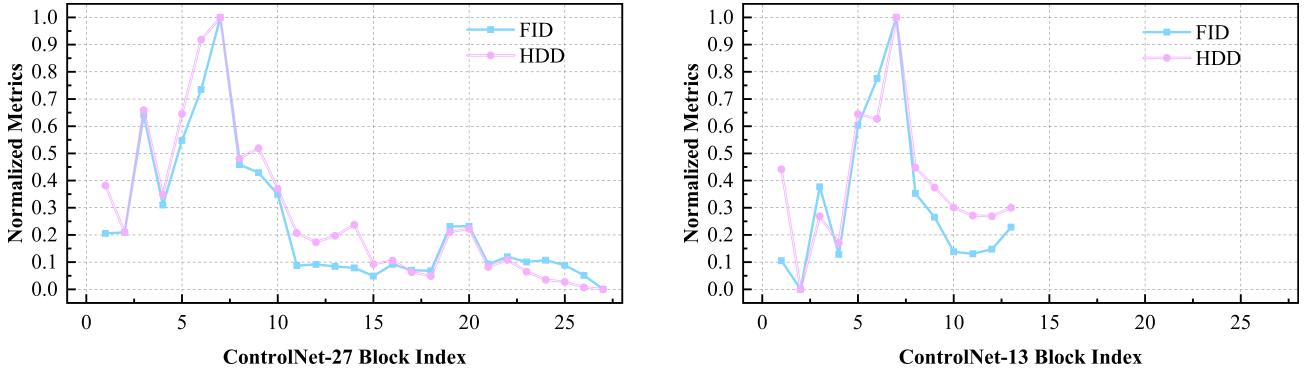


Figure 6. Impact of deleting specific locations on the generated metrics in ControlNet with 27 and 13 blocks.

In the main text, we conducted a study of DiT-ControlNet Relevance based on the PixArt- δ with 27 control blocks and obtained the following insights: the most critical modules of DiT-ControlNet are concentrated in the early-middle layers (e.g., blocks 5, 6, and 7). Overall, the ControlNet Relevance Score exhibits a trend of initially increasing and then decreasing.

To further validate the generalizability of this observation, we also conducted a similar study on a PixArt model with only the first 13 control blocks. Fig. 6 presents the normalized FID and HDD metrics of the generated images after removing the control module at a specific position. It is evident that the experimental results align with our previous findings: the most influential blocks in the control branch remain blocks 5, 6, and 7, which follow the same trend of increasing and then decreasing in relevance. This demonstrates the robustness of our observations, which can provide valuable guidance for subsequent specific design choices.

In Fig. 7, we present additional visual demonstrations of skipping specific ControlNet layers, specifically layers 7, 9, and 27, which correspond to the highest, moderate, and lowest impact on the generated image. The results illustrate that removing layer 7 significantly degrades both the quality of the generated image and the control accuracy. In contrast, skipping the later layers, such as layer 27, has minimal negative effects on the overall performance.

D.2. Experiment on Flux

In this section, we apply RelaCtrl to the latest Flux.1-dev with 12 billion parameters for experimentation. The comparison methods include OminiControl (Tan et al., 2024) and Flux ControlNet which is implemented based on the Diffusers library. Flux ControlNet uses the standard configuration of a double-stream layer with the first four replication layers, alongside a single-stream layer with the first ten replication layers.

When applying RelaCtrl to Flux, we aim to leverage as many insights and designs from the main text as possible. Accordingly, we retain the control positions of the double-stream layer unchanged and select seven positions out of ten on the single-stream layer for control. Table 5 presents the efficiency comparison between the methods, with GFLOPs calculated at a resolution of 512, and inference time measured at 512 resolution with 30 DDIM sample steps. Although OminiControl introduces

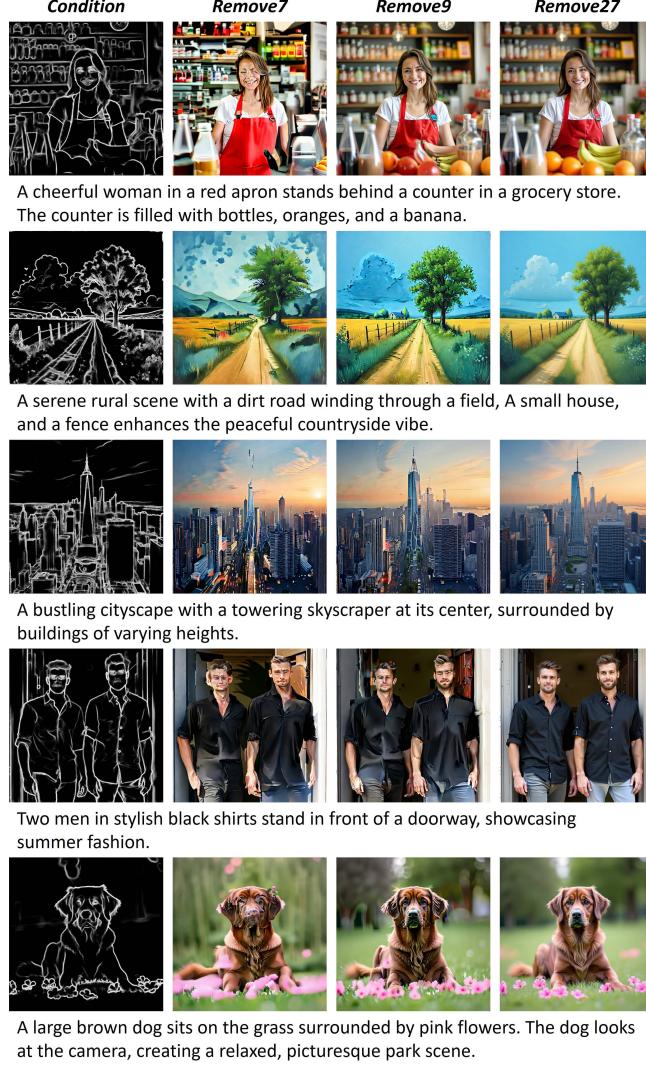


Figure 7. Additional visual results of skipping specific ControlNet layers (7, 9, and 27), correspond to the highest, moderate, and lowest impact on the generated image.

only a modest increase in parameters, it leads to a significant rise in computational load and reasoning time, severely limiting its practical efficiency. Flux ControlNet, on the other hand, shows moderate increases across several efficiency metrics. In contrast, RelaCtrl for Flux significantly reduces both computational complexity and reasoning time, achieving an optimal balance between efficiency and parameter volume. Fig. 8 illustrates the visual results of various control methods on Flux.1-dev. Among them, RelaCtrl not only achieves precise and detailed control effects but also offers significant advantages in computational efficiency, demonstrating the applicability and generalization of the proposed method across different DiT architectures.

Table 5. Effectiveness comparison experiment conducted on Flux.1-dev model.

Setting	Parameters (M)	Complexity (GFLOPs)	Inference (s)
Flux.1-dev	11901.39	9925.78	4.78
+OminiControl	+14.49	+6637.41	+3.11
+ControlNet	+2952.65	+2578.33	+0.79
+RelaCtrl	+549.19	+495.03	+0.34

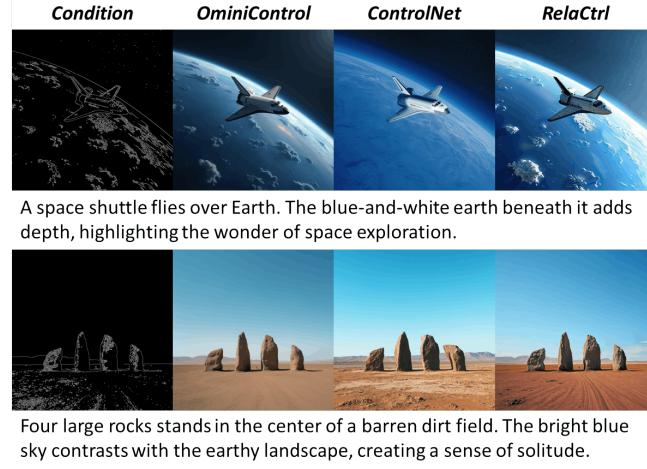


Figure 8. Visual comparison of different control methods on Flux.1-dev.

D.3. Inference with Community Models

We perform inference using the PixArt weights which were fine-tuned with Lora. Although RelaCtrl has not previously trained in these weights, it can still utilize them effectively. Fig. 9 showcases the model’s generated paint, oil, gufeng, and pixel-style images under the specified conditions.

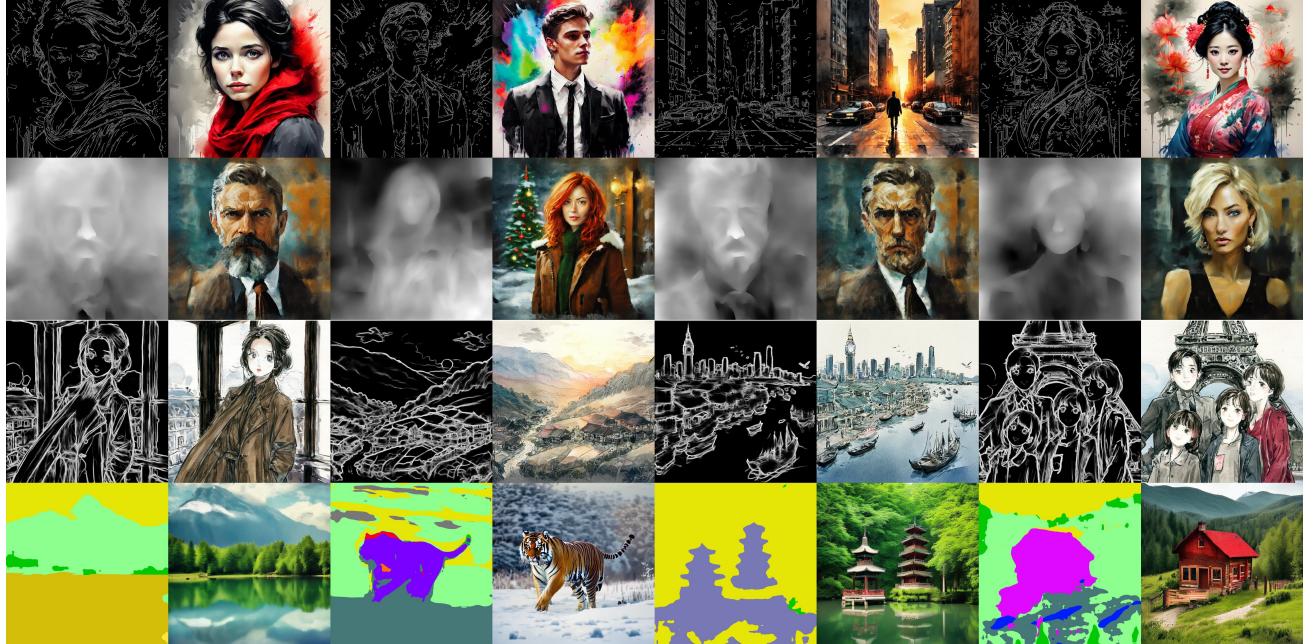


Figure 9. The control effect of RelaCtrl on the fine-tuned PixArt model. The upper and lower rows show the four transitions: (1) Canny to paint, (2) Depth to oil, (3) HED to gufeng, and (4) Segmentation to pixel.