Cross-Resolution Diffusion Models via Network Pruning

Jiaxuan Ren[2*‡] Junhan Zhu[1*] Huan Wang[1†]
[1]Westlake University [2]University of Electronic Science and Technology of China
https://xuan9-9.github.io/CR-Diff/

Figure 1. This paper presents *CR-Diff*, a method to improve the cross-resolution visual consistency of UNet–based diffusion models by masking out some parameters in the model, *i.e.*, *network pruning* – a technique that has been widely used for reducing model size; while here, we novelly repurpose it for generalizing diffusion models to unseen resolutions. The samples above compare the original SDXL [33] model with its counterpart modified by our proposed CR-Diff. The original SDXL is trained at 1024×1024 resolution and can hardly generalize to other resolutions (*e.g.*, 400×560, 480×360), while after CR-Diff prunes some parameters (the kept parameters are unchanged), it manages to generate much more coherent images at these unseen resolutions. This phenomenon suggests some parameters in the UNet-based diffusion models may be like a kind of "impurity"; while pruning, which used to be deemed to *damage* the model's capacity, can actually "purify" the diffusion model, improving its generalizability across resolutions.

Abstract

Diffusion models have demonstrated impressive image synthesis performance, yet many UNet–based models are trained at certain fixed resolutions. Their quality tends to degrade when generating images at out-of-training resolutions. We trace this issue to resolution-dependent parameter behaviors, where weights that function well at the default resolution can become adverse when spatial scales shift, weakening semantic alignment and causing structural instability in the UNet architecture. Based on this analysis, this paper introduces ***CR-Diff****, a novel method that improves the cross-resolution visual consistency by pruning some parameters of the diffusion model. Specifically, CR-Diff has two stages. It first performs block-wise pruning to selectively eliminate adverse weights. Then, a pruned output amplification is conducted to further purify the pruned predictions. Empirically, extensive experiments suggest that CR-Diff can improve perceptual fidelity and semantic coherence across various diffusion backbones and unseen resolutions, while largely preserving the performance at default resolutions. Additionally, CR-Diff supports prompt-specific refinement, enabling quality enhancement on demand.*

*These authors contributed equally to this work.
†Corresponding author: wanghuan@westlake.edu.cn
‡Work done as a visiting research intern at ENCODE Lab, Westlake University.

1. Introduction

Diffusion models [8, 22, 41–44] have achieved remarkable success in text-to-image generation [10, 30, 33, 35, 37, 50], enabling high-quality synthesis across a wide range of visual concepts. However, despite their strong generative capacity, most models are trained at *default resolutions* (e.g., 1024×1024 for SDXL [33]). Although techniques like multi-aspect bucket sampling [31, 33] provide some flexibility by fine-tuning on various aspect ratios, the core problem persists. When applied to *unseen resolutions* outside the training regime, these models tend to exhibit obvious artifacts, reduced semantic alignment, and diminished structural coherence. Recent DiT-based models [2, 10] natively address this limitation through scale-adaptive position encodings. In contrast, foundational UNet-based [36] models [35] lack such inherent robustness, making their generative quality more sensitive to changes in spatial scale.

Network pruning [11, 14, 17, 18, 47, 48] is traditionally used to improve efficiency by reducing computation and memory cost [3, 12, 13, 26, 55]. These approaches primarily aim to compress models while preserving accuracy. Surprisingly, here we observe that pruning in diffusion UNets can play a qualitatively different role. As shown in Figure 2, when applying simple magnitude pruning to SDXL at the unseen resolution of 512×512, we observe a counter-intuitive trend. Instead of degrading performance, moderate sparsity *improves* generation quality. In Figure 2a, metrics such as ImageReward steadily increase as sparsity rises from 0% to 40%, while FID decreases accordingly. This quantitative gain is further reflected in the visual samples in Figure 2b. At 0% sparsity, the dense model fails to produce a coherent object (the "cat" is missing, and the text is incomplete). As sparsity increases to 10–30%, the generated content becomes more semantically aligned. At 40%, both the concept of "a cat holding a sign" and the phrase "hello world" are rendered clearly.

Such phenomena suggest that parameters beneficial at the default resolutions can become *adverse* when applied to unseen resolutions, and pruning mitigates these effects and helps stabilize the generative process. All of these observations lead us to ask: *Can we devise a controllable pruning-based strategy to improve the cross-resolution generability of UNet-based diffusion models?*

To this end, we introduce *CR-Diff*, a two-stage framework that restructures parameter distribution and purifies model outputs of diffusion UNets for improved generation quality at unseen resolutions while maintaining performance at default ones. As shown in Figure 3, CR-Diff first applies *block-wise pruning* to assign differentiated pruning ratios across downsampling, middle, and upsampling blocks, yielding a pruned backbone reflecting the intrinsic importance distribution. Then, a *pruned output amplification* mechanism further purifies predictions by rebalancing dense and pruned outputs, enhancing beneficial signals while suppressing adverse ones. CR-Diff further supports *prompt-specific refinement*, allowing targeted quality enhancement. All are achieved without altering the model architecture, remaining effective across resolutions as shown in Figure 1.

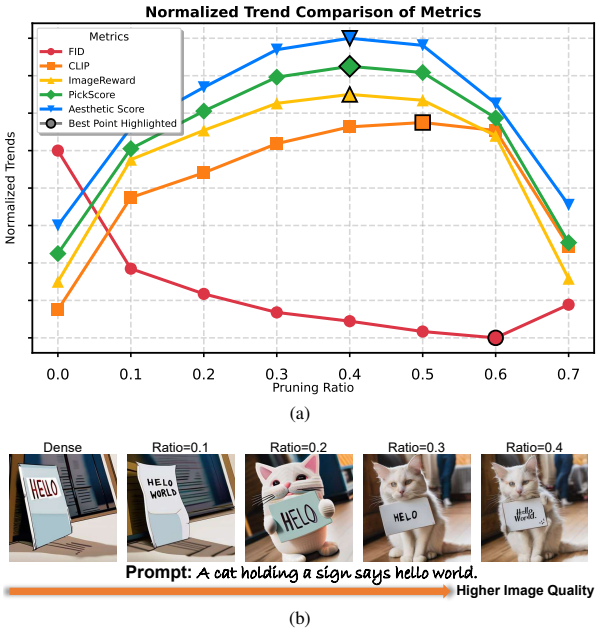


Figure 2. Effects of magnitude-based unstructured pruning on SDXL at unseen resolution 512×512. (a) Quantitative metrics improvement within moderate sparsity. (b) Qualitative illustration of improved semantic alignment as sparsity increases.

Our contributions are summarized as follows:

- We reveal that pruning diffusion UNets can *improve* text-to-image performance, particularly at unseen resolutions where dense models exhibit resolution bias.
- We introduce a *block-wise pruning* and *output amplification* strategy that adapts sparsity across the UNet and refines the pruned subnetwork to improve generation quality and stabilize semantic coherence.
- Experiments show that our method consistently and controllably enhances output quality, improving various metrics across models and resolutions.

2. Related Work

Text-to-Image Diffusion Models. Diffusion models [22, 44] have established themselves as the state-of-the-art for high-fidelity text-to-image synthesis, powering models like the widely-used Stable Diffusion (SD) series [10, 33, 35, 38], DALL-E2 [34], sana [7, 50, 51], Pixart [4–6], and FLUX [2]. However, a significant limitation of traditional UNet architectures, particularly foundational models like SD 1.5, is their limited gener-

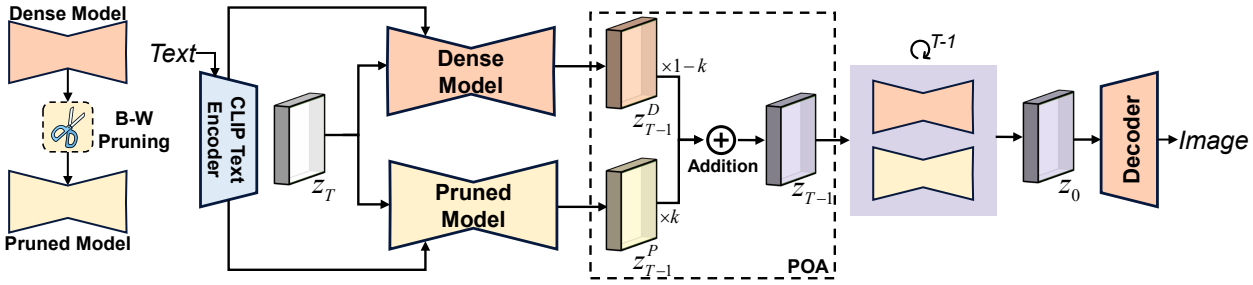


Figure 3. **Overview of CR-Diff.** Most UNet–based diffusion models exhibit resolution-dependent degradation when generating at unseen scales. CR-Diff addresses this issue through a two-stage *pruning and optimizing* process, consisting of a **block-wise (B-W) pruning ratio** strategy and a **pruned output amplification (POA)** mechanism. As shown in Figure 5, *block-wise pruning* adopts a magnitude-based criterion with adaptive ratios across blocks to extract parameter directions that remain stable across resolutions. *Pruned output amplification* refines the model's forward predictions by amplifying the pruned output with an amplification coefficient $k > 1$, which suppresses residual dense model's influences that otherwise introduce artifacts. This leads to cleaner denoising trajectories and higher-quality final images with more stable structure and details.

alization to resolutions and aspect ratios unseen during training. This fragility largely stems from spatially fixed inductive biases such as learned positional encodings in attention layers. Consequently, generating images at novel resolutions directly often leads to obvious degradation in visual coherence and semantic fidelity, such as object duplication or compositional collapse.

To mitigate this, several strategies have been proposed. The most common approach is multi-aspect training [31, 33], where models are explicitly fine-tuned on data "bucketed" into various aspect ratios after pretraining models at a fixed aspect-ratio and resolution, as was done for SDXL [33]. More recently, MMDiT-based [32] architectures like SD3 [10] and FLUX [2] have demonstrated superior flexibility by design. Instead of interpolating fixed embeddings [9], they natively handle variable input dimensions by generating 2D positional grids, which are constructed based on maximum training dimensions and then center-cropped to target resolutions before being frequency embedded. In contrast to these approaches, our work introduces a novel method to improve generation quality at unseen resolutions through a post-hoc, pruning-based strategy.

Neural Network Pruning. Neural network pruning [1, 11, 14, 17–19, 25, 48] is widely used to reduce parameter count and computational cost in deep learning, and has recently seen applications in large language models [16, 28, 45, 49] as well as other large-scale architectures [15, 39, 40, 46]. In diffusion models, pruning has primarily been explored as a compression technique to improve inference efficiency, leading to compact generators such as SnapFusion [26], MobileDiffusion [55], BK-SDM [23], Laptop-Diff [53], and LD-Pruner [3]. Recent general-purpose frameworks, including EcoDiff [54] and OBS-Diff [56], also follow this compression-oriented objective.

However, these methods view pruning solely as a means of model compression. In contrast, we find that pruning can *improve* the generative quality of text-to-image diffusion models, revealing a qualitatively different role for sparsity beyond efficiency.

3. Method: CR-Diff

3.1. Preliminaries

Diffusion models [22, 35] generate images by progressively denoising a latent variable x_t through a learned reverse diffusion process parameterized by a UNet backbone. Given a noisy latent x_t at timestep t, the model predicts the clean signal $\hat{x}_0$ conditioned on a text or image prompt c. The training objective is formulated as:

$$\mathcal{L} = \mathbb{E}_{x_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2\right], \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$. The denoising network ϵ_θ is realized as a hierarchical UNet consisting of convolutional and attention-based modules distributed across multiple feature resolutions.

Block-Specific Contribution Pattern. The UNet architecture in diffusion models can be decomposed into three structural stages, namely the *downsampling blocks*, the *middle blocks*, and the *upsampling blocks*. These stages operate at distinct feature scales and serve complementary purposes in the generative process. These functional asymmetries cause different blocks to contribute unevenly to the denoising process. The ablation results in Table 7a prove that optimal pruning ratios vary accordingly, and *applying differentiated treatment across blocks* leads to improved performance.

Resolution-Sensitive Weight Behavior. Diffusion UNets comprise convolution layers that capture local spatial priors and fine-grained textures, attention layers that establish global semantic relationships and

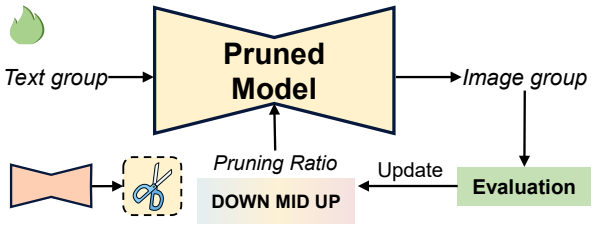


Figure 4. Simulated annealing (SA) search process for determining the optimal block-wise pruning ratios.

text–image alignment, feed-forward layers that reshape intermediate feature representations, and normalization or modulation parameters that encode activation statistics across diffusion steps. Although jointly trained, these components are implicitly adapted to the feature and scale statistics of the trained default resolution.

Consequently, when the model is applied to *unseen resolutions*, the feature distributions shift away from those seen during training. Scale-specific weights no longer align with the altered structure. In such cases, these parameters can be regarded collectively as *adverse weights*, referring to weights that do not align well with the semantic structure required at non-default resolutions, and can lead to degraded visual coherence when generating at unseen resolutions.

3.2. Overall Framework

Building upon the diffusion UNet foundation introduced above, our pruning framework seeks to preserve semantically essential parameters while attenuating adverse ones, thereby improving image generation. The central idea is to apply block-wise sparsification across the UNet hierarchy and subsequently refine the retained subnetwork to mitigate residual degradation. As illustrated in Figure 3, the framework operates in two sequential stages, *pruning* and *optimization*.

In the *pruning* stage, the **block-wise pruning ratio** strategy shown in Figure 5 assigns differentiated pruning ratios to the downsampling, middle, and upsampling blocks, which improves generation quality and yields a pruned backbone that reflects intrinsic importance distributions of weights. In the *optimization* stage, the **pruned output amplification** (POA) mechanism shown in Figure 3 leverages differences between dense and pruned outputs, amplifying pruned prediction while attenuating residual dense signals that introduce artifacts.

After the two-stage refinement, CR-Diff can synthesize images from text prompts with cleaner denoising trajectories and noticeably improved visual quality.

3.3. Block-Wise Pruning Ratio Strategy

As discussed in Section 3.1, heterogeneous weight functions make uniform pruning ratios less effective on diffusion UNets, which inevitably reduces local texture encoding in early downsampling blocks, diminishes global semantic integration in middle blocks, and limits high-frequency detail recovery in upsampling blocks, ultimately compromising visual coherence and fidelity.

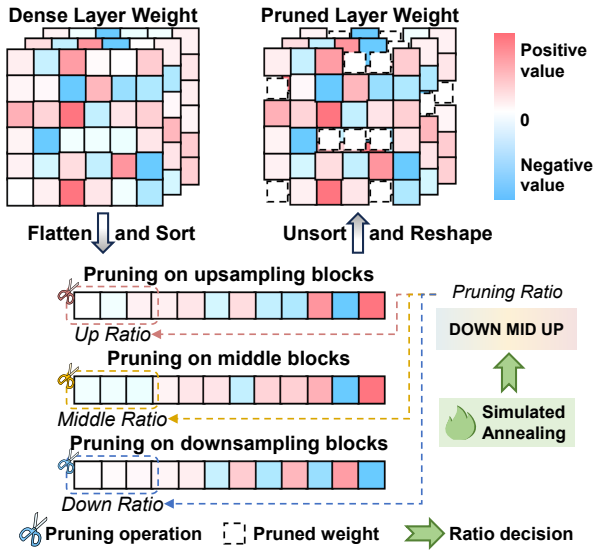


Figure 5. Block-wise (B-W) pruning applies differentiated pruning ratios across blocks to preserve essential structure and provide a pruned backbone for subsequent optimization.

To this end, we employ a *block-wise pruning ratio* strategy, in which downsampling blocks, the middle block, and upsampling blocks of the UNet are each assigned distinct pruning ratios based on magnitude.

To determine the optimal pruning ratio for each block, we adopt a *simulated annealing (SA)* search strategy. Let $\mathbf{r} = r_{\text{down}}, r_{\text{mid}}, r_{\text{up}}$ denote the pruning ratio configuration across the downsampling, middle, and upsampling blocks. Starting from initial configurations, the model generates images for a fixed set of prompts, and their ImageReward is averaged to assess the overall performance of the current ratio setting. SA then perturbs and updates $\mathbf{r}$ iteratively, gradually refining it to maximize generation quality. The search procedure is illustrated in Figure 4, and the full algorithm is provided in the supplementary material due to limited space.

Through this exploration, each block receives a ratio that preserves critical semantic structure while suppressing weights that introduce degradation in generation.

3.4. Pruned Output Amplification

To further refine the generative behavior of the pruned model, we introduce a *pruned output amplification* (POA) mechanism, which operates on the forward denoising trajectory, as illustrated in Figure 3. At each denoising step t, we obtain the predicted output $\mathbf{z}_t^P$ from the pruned model and the corresponding output $\mathbf{z}_t^D$ from

Table 1. Performance comparison on unseen resolutions. Across the evaluated models and resolutions, CR-Diff improves most metrics relative to the dense model, with **bold** values indicating superior performance of our CR-Diff against the dense model.

Model	Resolution	FID ↓		CLIP ↑		ImageReward ↑		PickScore ↑		Aesthetic Score ↑	
		Dense	CR-Diff	Dense	CR-Diff	Dense	CR-Diff	Dense	CR-Diff	Dense	CR-Diff
SDXL	512×512	83.827	**37.918**	0.295	**0.321**	-0.498	**0.735**	20.296	**22.140**	4.335	**5.525**
	400×560	146.984	**36.688**	0.252	**0.311**	-1.734	**0.092**	18.608	**21.074**	3.494	**4.672**
	480×360	211.369	**46.040**	0.225	**0.307**	-2.148	**-0.099**	18.060	**20.956**	3.806	**4.644**
SD1.5	400×560	39.047	39.291	0.309	**0.310**	0.061	**0.151**	21.146	**21.188**	4.736	**4.779**
	480×360	39.797	**37.634**	0.307	**0.307**	-0.068	**-0.026**	20.906	**20.944**	4.710	**4.819**
	768×768	38.832	**38.452**	0.314	**0.315**	-0.050	**0.059**	21.208	**21.232**	5.419	5.385
SD2.1	400×560	48.110	**35.837**	0.296	**0.304**	-0.461	**-0.068**	20.374	**20.540**	4.190	**4.428**
	480×360	73.807	**41.042**	0.278	**0.294**	-0.933	**-0.561**	19.573	**20.177**	3.984	**4.532**
	768×768	37.198	**35.237**	0.317	**0.318**	0.334	**0.419**	21.695	21.451	5.583	5.339

the dense model, and then performs combination:

$$\mathbf{z}_t = k\,\mathbf{z}_t^P + (1-k)\,\mathbf{z}_t^D, \tag{2}$$

where the amplification coefficient k determines the relative contribution of the pruned and dense outputs.

Because $\mathbf{z}_t^P - \mathbf{z}_t^D$ represents the pruning-induced shift that improves generative behavior, choosing $k > 1$ selectively amplifies this beneficial direction while suppressing residual artifact-inducing tendencies inherited from the dense model. This step-by-step refinement stabilizes the denoising trajectory and preserves structural consistency throughout sampling. After applying both block-wise pruning and POA, the resulting model produces higher-quality images directly from text prompts.

4. Experiments

4.1. Settings

Models and Resolutions. To assess the effectiveness of CR-Diff, we apply pruning to three UNet–based diffusion models across both their default training resolutions and a set of unseen resolutions. For SDXL [33], in addition to its default 1024×1024 resolution, we evaluate performance at 512×512, 400×560, and 480×360. For SD1.5 and SD2.1 [35], beyond the default 512×512, we likewise consider 400×560, 480×360 ,and 768×768 as unseen settings. This setup enables us to examine how pruning influences generative robustness when moving away from the resolution regime on which the model was originally trained. In the following experiments, all resolutions are expressed in the format height $\times$ width.

Evaluation Metrics. We evaluate our method on a subset of 5K prompts sampled from the MS-COCO 2014 validation set [27]. Performance is measured along three dimensions: image fidelity, text–image alignment, and aesthetic preference. Specifically, Fréchet Inception Distance (FID) [21] is used to assess image quality, while CLIP Score [20] and ImageReward [52] evaluate semantic alignment between text and image. Plus, PickScore [24] and Aesthetic Score provide assessments of aesthetic appeal and human preference consistency.

4.2. Results of CR-Diff on Unseen Resolutions

As shown in Table 1, CR-Diff demonstrates generally improved performance across multiple diffusion backbones when evaluated at resolutions that deviate from their default training settings.

For SDXL, which is originally optimized for the resolution at 1024×1024, applying CR-Diff at unseen resolutions results in substantial and notable gains across all evaluation metrics. The large magnitude of improvement suggests that scale-mismatched parameters in SDXL strongly contribute to texture degradation and structural inconsistency, and that CR-Diff effectively suppresses these detrimental effects.

For SD1.5 and SD2.1, which are natively trained at 512×512, CR-Diff also provides consistent gains when evaluated at unseen resolutions. Improvements are reflected in enhanced semantic alignment as measured by CLIP and ImageReward, as well as better visual preference captured by PickScore and Aesthetic Score.

Compared with SDXL, however, the improvements appear more moderate. This is due to the intrinsic resolution characteristics of SD1.5 and SD2.1. Their training data encourages coarser semantic representation, with objects occupying larger spatial regions and containing relatively low detail density. As a result, reducing resolution does not heavily disrupt global structure because the models are designed to perform well under limited texture complexity.

4.3. Results of Prompt-Specific Optimization

CR-Diff already provides substantial gains over the dense model, with improvements observed on over 85% of evaluated prompts under global refinement. This demonstrates that the two-stage framework is broadly

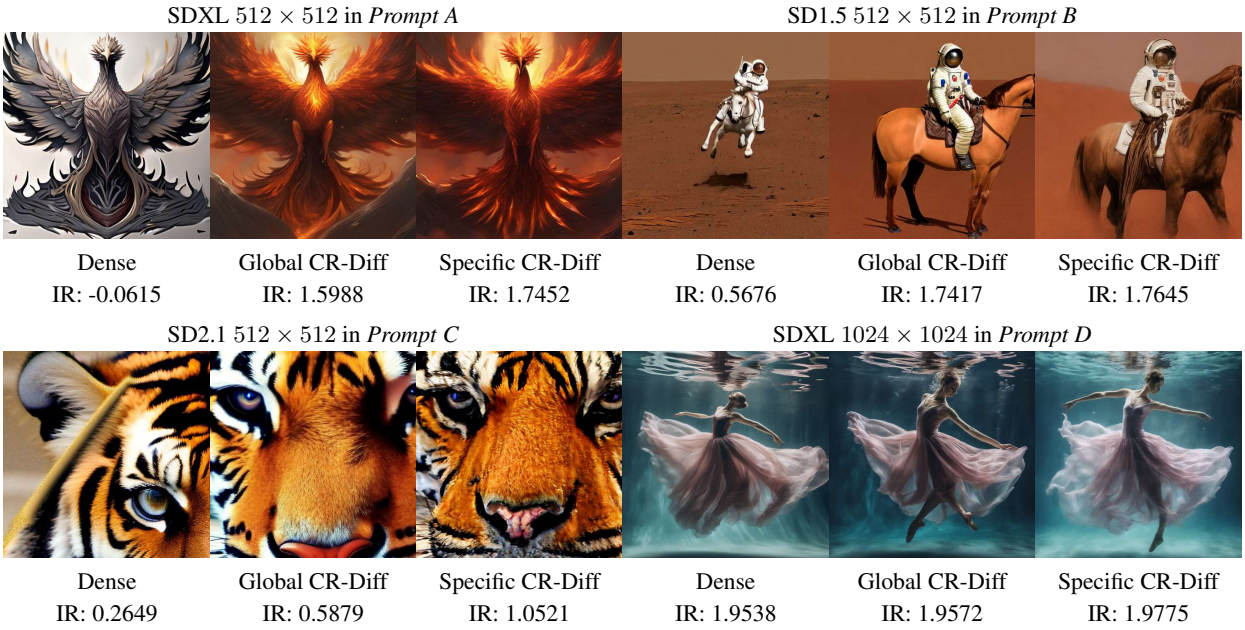


Figure 6. Visual comparison across three generation settings. *Dense* denotes the original unpruned model. *Global CR-Diff* applies a pruning ratio optimized for a specific model and resolution, shared across all prompts. *Specific CR-Diff* further refines this ratio for the given prompt, enabling prompt-specific optimization. Each group corresponds to a specific prompt, and the ImageReward (IR) scores are shown below each image. Global CR-Diff improves generative fidelity in a prompt-agnostic manner, while Specific CR-Diff further enhances semantic alignment and visual coherence for the specific prompt.

Table 2. Performance comparison on default resolutions. Across the evaluated models, CR-Diff consistently improves or maintains performance, with **bold** values indicating gains over the dense model.

Model	Resolution	FID ↓		CLIP ↑		ImageReward↑		PickScore ↑		Aesthetic Score ↑	
		Dense	CR-Diff	Dense	CR-Diff	Dense	CR-Diff	Dense	CR-Diff	Dense	CR-Diff
SDXL	1024 × 1024	33.186	33.562	0.322	**0.322**	0.788	**0.946**	22.512	**22.639**	6.123	6.106
SD1.5	512 × 512	38.368	**37.773**	0.315	0.314	0.239	0.203	21.539	21.377	5.205	**5.233**
SD2.1	512 × 512	45.583	**36.792**	0.308	**0.309**	-0.100	**-0.052**	20.943	**20.960**	4.728	**5.082**

effective in enhancing overall fidelity and semantic consistency across diverse scenes. Nevertheless, some prompts involve particularly fine-grained textures, rare materials, or compositionally intricate structures that can benefit from more specialized treatment than what global refinement alone can supply. For such cases, CR-Diff provides prompt-specific optimization that tailors pruning configurations to individual prompts, searching for locally optimal patterns that preserve finer visual details and offer more precise prompt-dependent control.

As shown in Figure 6, the prompt-specific optimization consistently enhances both semantic fidelity and visual coherence compared with dense models and globally optimized CR-Diff. Taking the SDXL 512 × 512 case under Prompt A as an illustrative example, the dense model on the left fails to express phoenix fire or molten lava and instead resembles a cold carved bird, so the semantic intent is largely lost. The global CR-Diff result in the middle restores the fiery theme and atmosphere, but the molten quality remains limited. The prompt-specific optimized result on the right most accurately conveys both the burning phoenix and the flowing rebirth from molten lava, achieving the clearest and most consistent expression of the prompt.

4.4. Generalization of CR-Diff

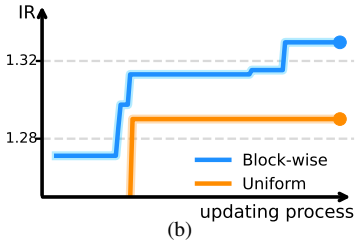
Results of CR-Diff on Default Resolutions. Although CR-Diff is primarily designed to address degradation at unseen resolutions, it also preserves or even improves model performance at the default resolutions. As shown in Table 2, across SDXL, SD1.5, and SD2.1, the two-stage framework maintains generative fidelity while often improving metrics. On SDXL at 1024 × 1024, for instance, CR-Diff preserves image fidelity and text–image alignment comparable to the dense model with ImageReward increasing and FID remaining.

Table 3. Performance comparison on DiTs. **Bold** values indicate our CR-Diff is better than the dense model. The results show that CR-Diff preserves or even improves the generation quality, demonstrating the generalizability of our method beyond UNets.

Model	Resolution	FID ↓		CLIP ↑		ImageReward ↑		PickScore ↑		Aesthetic Score ↑	
		Dense	CR-Diff	Dense	CR-Diff	Dense	CR-Diff	Dense	CR-Diff	Dense	CR-Diff
SD3Medium	512 × 512	40.453	**38.901**	0.317	**0.317**	0.972	**1.038**	22.187	22.121	4.886	4.884
SD3Medium	1024 × 1024	37.841	**37.026**	0.320	**0.320**	1.081	**1.128**	22.609	22.543	5.513	5.455
FLUX.dev	1024 × 1024	35.799	**35.708**	0.311	**0.312**	0.945	0.935	22.793	22.775	6.295	6.263

Model_Resolution	Uniform		Block-wise	
	Ratio	IR↑	Ratio	IR↑
SDXL_1024 × 1024	0.124	0.921	0.295 / 0.194 / 0.236	**0.946**
SDXL_512 × 512	0.288	0.688	0.397 / 0.434 / 0.387	**0.735**
SD2.1_480 × 360	0.369	-0.663	0.651 / 0.138 / 0.271	**-0.561**

(a)



(b)

Figure 7. Ablation results of block-wise pruning. (a) Performance comparison under uniform and block-wise pruning strategies across different models and resolutions. For block-wise pruning, the ratios are listed in the order *down-sampling / middle / up-sampling*. ImageReward (IR) of the better-performing strategy is highlighted in **bold**, showing that differentiated ratios improve image quality. (b) Conceptual illustration of ImageReward trends during the updating process for the optimal pruning configuration, showing generally higher and more stable values under block-wise pruning compared to uniform pruning.

Results of CR-Diff Applied to DiT. Furthermore, while primarily designed for diffusion UNets, CR-Diff can also be safely applied to Diffusion Transformer (DiT) without causing performance degradation. We evaluate CR-Diff on representative DiT models, including SD3Medium [10] and Flux.dev [2]. As shown in Table 3, the framework preserves generative fidelity at default resolutions, and in some cases even improves certain metrics such as ImageReward, FID, and CLIP scores. For instance, on SD3Medium at 1024 × 1024, ImageReward increases from 1.081 to 1.128 while FID decreases from 37.841 to 37.026, indicating that CR-Diff's pruning and optimization stages generalize beyond UNet architectures.

4.5. Ablation Study

Block-Wise Pruning Ratio. Table 7a presents representative examples comparing uniform pruning with the proposed block-wise pruning strategy. Across the shown models and resolutions, block-wise pruning yields higher ImageReward scores than uniform pruning. For instance, on SDXL at 1024 × 1024, IR improves from 0.921 to 0.946, and on SDXL at 512 × 512, IR improves from 0.688 to 0.735. These results reflect the advantage of allocating differentiated pruning ratios that match the functional roles of the corresponding blocks. Full best pruning ratio configurations for all resolution settings are listed in the supplementary material, where substantial differences across downsampling, middle, and upsampling blocks can be observed.

The evolution of ImageReward during the optimal pruning config updating process on SDXL 512 × 512 is illustrated in Figure 7b. Uniform pruning applies the same ratio across all blocks and therefore tends to reduce capacity in regions where parameters are more functionally critical, resulting in a lower and flatter performance plateau during optimization. In contrast, block-wise pruning preserves information flow more effectively, particularly in the middle and upsampling stages that contribute strongly to global structure and fine-grained detail. This leads to a more favorable optimization trajectory and a higher final quality level, as reflected by the consistently stronger ImageReward scores.

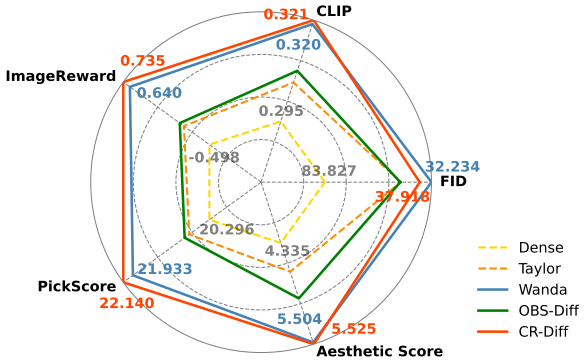


Figure 8. Radar comparison across pruning strategies on SDXL at 512 × 512. Five metrics are normalized with direction alignment so that larger radial values denote better performance. CR-Diff achieves the strongest overall results, highlighting its superior perceptual and semantic quality.

Comparison of Different Pruning Criteria. Unlike traditional pruning methods that mainly aim for efficiency and seek to retain performance comparable to the dense model, CR-Diff is designed to surpass the dense model in generative quality. The magnitude-based prun-

Table 4. Ablation results of the pruned output amplification component. Values in the table denote the performance difference between models with and without POA, where consistent positive values demonstrate the effectiveness of POA.

Resolution	ImageReward Improvement		
	SDXL	SD1.5	SD2.1
512×512	+0.205	+0.144	+0.236
400×560	+0.364	+0.155	+0.266
480×360	+0.417	+0.164	+0.261

ing in CR-Diff no longer treats gradient magnitude as a sufficient indicator of parameter importance, since a large gradient only reflects strong influence on the output, not whether that influence is beneficial.

Here we compare our CR-Diff with other representative pruning criteria: Taylor [29], Wanda [45], and OBS-Diff [56], on SDXL [33] at 512×512. Figure 8 presents the evaluation results of images generated under each pruning strategy alongside the dense model in the form of a radar plot. Although other pruning methods can yield moderate improvements over the dense model, CR-Diff consistently delivers the strongest overall performance across the five metrics. Notably, CR-Diff achieves the best results in four of the five metrics.

While Wanda achieves performance relatively close to ours, it requires a full Hessian-free weight importance estimation that takes approximately 420.70s per pruning pass, whereas our magnitude-based block-wise pruning completes in only 0.38s.

Pruned Output Amplification. Table 4 highlights the effect of the *pruned output amplification* (POA) mechanism on ImageReward. Across all tested models and resolutions, POA yields consistently positive ImageReward gains over the corresponding pruned baselines. Notably, the improvement becomes even more pronounced under resolution shifts. For instance, on SD1.5, POA yields a +0.144 ImageReward gain at the default 512×512 resolution, and this improvement further increases to +0.155 and +0.164 at unseen resolutions. This consistent upward trend suggests that POA serves as an effective component for enhancing pruned diffusion models' performance under cross-resolution conditions. Full-resolution results and corresponding metrics are deferred to our supplementary material due to limited space, where the aggregated evaluations consistently confirm better performance across all resolutions.

Effect of the Amplification Coefficient k. To examine the influence of the amplification coefficient k used in pruned output amplification, we conduct an ablation study with $k \in 1.5, 2.0, 2.5$ and evaluate the resulting generative performance. As shown in Table 5, $k = 1.5$ yields the most consistent improvements across all metrics, indicating that a moderate amplification effectively strengthens the beneficial deviation introduced by pruning while maintaining coherent semantic structure.

Table 5. Ablation study on the amplification coefficient k in POA, on SDXL (512×512). A moderate amplification ($k = 1.5$) yields the most stable performance gains across evaluation metrics. Best per-metric values are shown in **bold**.

Metric	K = 1.5	K = 2.0	K = 2.5
Best IR % ↑	**54.31**	29.89	15.80
FID ↓	**37.918**	43.08	61.71
CLIP ↑	**0.321**	0.305	0.290
ImageReward↑	**0.735**	-0.003	-0.557
PickScore ↑	**22.140**	21.14	20.34
Aesthetic Score ↑	**5.525**	5.040	4.630

In contrast, increasing k to 2.0 or 2.5 leads to clear degradation. Excessive amplification suppresses meaningful residual signals from the dense output, resulting in weakened semantic alignment and reduced perceptual fidelity. For instance, ImageReward drops from 0.514 to -0.557, and FID rises from 37.21 to 61.71 when k increases from 1.5 to 2.5. This highlights that the improvement brought by POA arises from balancing the contributions of the pruned and dense outputs, rather than replacing the latter entirely.

Overall, $k = 1.5$ achieves a stable compromise between preserving semantic faithfulness and enhancing visual quality. Accordingly, we adopt $k = 1.5$ as the default setting in all main experiments.

5. Conclusion

This work introduces *CR-Diff*, a pruning-based approach to improve cross-resolution consistency in UNet–based text-to-image diffusion models. CR-Diff operates in two stages. First, a *block-wise pruning* strategy allocates differentiated pruning ratios to the downsampling, middle, and upsampling blocks, preserving resolution-stable structure while removing redundant parameters. Second, a *pruned output amplification* mechanism refines the forward denoising trajectory by amplifying the beneficial output tendencies introduced by pruning and suppressing residual artifact-related signals inherited from the dense model. Unlike existing pruning works that typically pose pruning as an efficiency-improving technique to reduce model size, here we expand its role, *for the first time*, to improving cross-resolution generation quality of diffusion models. Experiments on SDXL, SD1.5, and SD2.1 demonstrate that CR-Diff enhances perceptual fidelity and semantic coherence at unseen resolutions while preserving performance at default resolutions and on DiT models. CR-Diff also supports optional prompt-specific optimization for adaptive, on-demand enhancement.

Acknowledgment

This paper is supported by Young Scientists Fund of the National Natural Science Foundation of China (NSFC) (No. 62506305), and Scientific Research Project of Westlake University (No. WU2025WF003).

References

[1] Alireza Aghasi, Afshin Abdi, Nam Nguyen, and Justin Romberg. Net-trim: Convex pruning of deep neural networks with performance guarantee. In *NeurIPS*, 2017. 3

[2] Black Forest Labs. Flux. https://blackforestlabs.ai/, 2024. Accessed: 2025-09-25. 2, 3, 7

[3] Thibault Castells, Hyoung-Kyu Song, Bo-Kyeong Kim, and Shinkook Choi. Ld-pruner: Efficient pruning of latent diffusion models using task-agnostic insights. In *CVPR*, 2024. 2, 3

[4] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *ECCV*, 2024. 2

[5] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *ECCV*. Springer, 2024.

[6] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James T Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024. 2

[7] Junsong Chen, Shuchen Xue, Yuyang Zhao, Jincheng Yu, Sayak Paul, Junyu Chen, Han Cai, Song Han, and Enze Xie. Sana-sprint: One-step diffusion with continuous-time consistency distillation. In *ICCV*, 2025. 2

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 2

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 3

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 2, 3, 7

[11] Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. Depgraph: Towards any structural pruning. In *CVPR*, 2023. 2, 3

[12] Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. In *NeurIPS*, 2023. 2

[13] Gongfan Fang, Kunjun Li, Xinyin Ma, and Xinchao Wang. Tinyfusion: Diffusion transformers learned shallow. In *CVPR*, 2025. 2

[14] Sicheng Feng, Keda Tao, and Huan Wang. Is oracle pruning the true oracle? *arXiv preprint arXiv:2412.00143*, 2024. 2, 3

[15] Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *NeurIPS*, 2022. 3

[16] Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *ICML*, 2023. 3

[17] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *NeurIPS*, 2015. 2, 3

[18] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In *ICLR*, 2016. 2

[19] Babak Hassibi, David G Stork, and Gregory J Wolff. Optimal brain surgeon and general network pruning. In *NeurIPS*, 1992. 3

[20] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*. Association for Computational Linguistics, 2021. 5

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 5

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3

[23] Bo-Kyeong Kim, Hyoung-Kyu Song, Thibault Castells, and Shinkook Choi. Bk-sdm: A lightweight, fast, and cheap version of stable diffusion. In *ECCV*, 2024. 3

[24] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, 2023. 5

[25] Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. In *NeurIPS*, 1989. 3

[26] Yanyu Li, Huan Wang, Qing Jin, Ju Hu, Pavlo Chemerys, Yun Fu, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. In *NeurIPS*, 2023. 2, 3

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5, 2, 4, 6

[28] Gui Ling, Ziyang Wang, and Qingwen Liu. Slimgpt: Layer-wise structured pruning for large language models. In *NeurIPS*, 2024. 3

[29] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *CVPR*, 2019. 8

[30] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *ICML*, 2022. 2

[31] NovelAI. NovelAI improvements on Stable Diffusion. https://blog.novelai.net/novelai-improvements-on-stable-diffusion-e10d38db82ac, 2022. 2, 3

[32] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 3

[33] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 1, 2, 3, 5, 8

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 5

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[38] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 2

[39] Xuan Shen, Hangyu Zheng, Yifan Gong, Zhenglun Kong, Changdi Yang, Zheng Zhan, Yushu Wu, Xue Lin, Yanzhi Wang, Pu Zhao, et al. Sparse learning for state space models on mobile. In *ICLR*, 2025. 3

[40] Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. Efficient unstructured pruning of mamba state-space models for resource-constrained environments. In *EMNLP*, 2025. 3

[41] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

[43] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019.

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2

[45] Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *ICLR*, 2024. 3, 8

[46] Kaiwen Tuo and Huan Wang. Sparsessm: Efficient selective structured state space models can be pruned in one-shot. *arXiv preprint arXiv:2506.09613*, 2025. 3

[47] Huan Wang and Yun Fu. Trainability preserving neural pruning. In *ICLR*, 2023. 2

[48] Huan Wang, Can Qin, Yulun Zhang, and Yun Fu. Neural pruning via growing regularization. In *ICLR*, 2021. 2, 3

[49] Jiateng Wei, Quan Lu, Ning Jiang, Siqi Li, Jingyang Xiang, Jun Chen, and Yong Liu. Structured optimal brain pruning for large language models. In *EMNLP*, 2024. 3

[50] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. In *ICLR*, 2025. 2

[51] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng YU, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. In *ICML*, 2025. 2

[52] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *NeurIPS*, 2023. 5

[53] Dingkun Zhang, Sijia Li, Chen Chen, Qingsong Xie, and Haonan Lu. Laptop-diff: Layer pruning and normalized distillation for compressing diffusion models. *arXiv preprint arXiv:2404.11098*, 2024. 3

[54] Yang Zhang, Er Jin, Yanfei Dong, Ashkan Khakzar, Philip Torr, Johannes Stegmaier, and Kenji Kawaguchi. Effortless efficiency: Low-cost pruning of diffusion models. *arXiv preprint arXiv:2412.02852*, 2024. 3

[55] Yang Zhao, Yanwu Xu, Zhisheng Xiao, Haolin Jia, and Tingbo Hou. Mobilediffusion: Instant text-to-image generation on mobile devices. In *ECCV*, 2024. 2, 3

[56] Junhan Zhu, Hesong Wang, Mingluo Su, Zefang Wang, and Huan Wang. Obs-diff: Accurate pruning for diffusion models in one-shot. *arXiv preprint arXiv:2510.06751*, 2025. 3, 8