Moebius: 0.2B Lightweight Image Inpainting Framework with 10B-Level Performance

Kangsheng Duan[1,*] Ziyang Xu[1,*,†] Wenyu Liu[1] Xiaohu Ruan[2] Xiaoxin Chen[2] Xinggang Wang[1,✉]

[1] Huazhong University of Science and Technology, China
[2] VIVO AI Lab, China

Abstract. While 10B-level industrial foundation models have pushed the boundaries of image inpainting, their prohibitive computational costs severely hinder practical deployment. Constructing a highly optimized task-specific specialist offers a promising solution; however, extreme structural compression inevitably triggers a severe representation bottleneck. To conquer this, we propose Moebius, a highly efficient lightweight inpainting framework. We systematically reconstruct the diffusion backbone by introducing the Local-λ Mix Interaction ($L\lambda MI$) block. Comprising Local-λ and Interactive-λ modules, it elegantly summarizes spatial contexts and global semantic priors into fixed-size linear matrices, preserving complex latent interactions while drastically shedding parameters. Furthermore, to unlock the full representational capacity of this highly compact architecture, we synergistically pair it with an adaptive multi-granularity distillation strategy. Operating strictly within the latent space to avoid expensive pixel-space decoding, this strategy dynamically balances multiple gradient-based losses to achieve high-fidelity alignment. Extensive experiments across natural and portrait benchmarks demonstrate that this optimal synergy enables Moebius to rival or even surpass the generation quality of the 10B-level industrial generalist FLUX.1-Fill-Dev. Remarkably, Moebius achieves this using less than 2% of the parameters (0.22B vs. 11.9B) while delivering a $> 15\times$ acceleration in total inference time, setting a new efficiency standard for high-fidelity inpainting. Project page at https://hustvl.github.io/Moebius.

Keywords: Image Inpainting · Lightweight Architecture · Knowledge Distillation · Latent Diffusion

1 Introduction

Image inpainting [2], a fundamental task in computer vision aimed at reconstructing missing regions with visually coherent content, has been profoundly revolutionized by the rapid evolution of diffusion models [17, 29, 49]. Recently, industrial-grade generalist foundation models, such as FLUX.1-Fill-Dev [17] and

* Equal contribution. Completed this work as interns at VIVO AI Lab.
† Project leader.
✉ Corresponding author: Xinggang Wang <xgwang@hust.edu.cn>.

SD3.5 Large-Inpainting [8], have pushed the boundaries of zero-shot generation quality by scaling parameters to the 10-billion (10B) level. However, the exorbitant computational costs and massive memory footprints of these colossal models severely hinder their practical deployment, particularly on resource-constrained devices or in latency-sensitive applications. This dilemma motivates us to rethink the current scaling paradigm: Can a highly optimized, lightweight task-specific specialist bridge the massive scale gap and rival the performance of 10B-level generalists? While massive generalist models excel in zero-shot versatility, fine-tuning on established academic benchmarks (e.g., Places2 [63], CelebA-HQ [40]) remains the standard evaluation paradigm in the inpainting community [20, 37] to unlock and rigorously assess a model's capacity for specific restoration tasks. Following this well-established practice, we demonstrate that by pushing architectural efficiency to its limits, a 0.2B-parameter specialist model can successfully overcome the capacity bottleneck and match the high-fidelity generation capabilities of its 10B-level counterparts.

To construct such a highly efficient specialist, a natural progression is to compress existing diffusion architectures. Recent advancements like PixelHacker [49] have pioneered efficient high-fidelity inpainting by introducing the Latent Categories Guidance (LCG) paradigm and employing Gated Linear Attention (GLA) [53] to reduce computational overhead. Despite these algorithmic innovations, its backbone still comprises nearly one billion parameters, which remains prohibitive for edge deployment. A naive solution to further compress the model would be directly substituting its standard convolutions and attention blocks with off-the-shelf lightweight operators, such as Depthwise Convolutions (DWConv) [32] and linear attention mechanisms [6, 7, 53]. However, our empirical analysis reveals that such straightforward structural reductions inevitably trigger a severe representation bottleneck. In the intricate task of image inpainting, which demands rigorous semantic reasoning and precise spatial-texture alignment, these naive lightweight models suffer from a catastrophic degradation in generation quality. Furthermore, many efficient operators are architecturally constrained; for instance, while GLA [53] is highly efficient for self-attention, it inherently lacks the formulation to perform the cross-attention operations essential for integrating external semantic priors like LCG [49].

To conquer this representation bottleneck and achieve an optimal balance between efficiency and quality, we propose Moebius, a highly efficient lightweight inpainting framework. We systematically reconstruct the foundational architecture through a rigorous synergy of structural design and knowledge distillation. First, to address the architectural constraints of existing linear attention, we introduce the Local-λ and Interactive-λ modules. By elegantly summarizing local spatial contexts and global semantic priors (e.g., LCG embeddings) into fixed-size linear matrices, these modules enable the network to perform both self- and cross-attention equivalents efficiently, preserving complex latent interactions while drastically shedding parameters. To push for extreme compactness, we further integrate highly compressed operators, such as DWConv and Mix-FFN [46, 47], which collectively form our Local-λ Mix Interaction ($L\lambda MI$) block. However, such extreme structural compression inherently risks weaken-

Table 1: Breaking the impossible triangle of low-parameters, fast inference, and high generation quality. Moebius achieves the lowest latency, FLOPs, and parameters, with remarkable generation quality. Note that industrial models require more steps (50/28 vs 20; default setting), making Moebius **15× faster** in total inference.

Model	Places2 (Small) FID ↓	Places2 (Small) LPIPS ↓	CelebA-HQ (512) FID ↓	CelebA-HQ (512) LPIPS ↓	FFHQ (256) FID ↓	FFHQ (256) LPIPS ↓	Param. ($\times 10^9$) ↓	TFLOPs ↓	Latency (ms/step) ↓	Steps	Total Time(s) ↓
Moebius	0.92	0.091	5.39	0.122	8.15	0.231	0.226	0.154	26.01	20	0.52
PixelHacker [49]	0.82▼11%	0.088▼3%	4.75▼12%	0.115▼6%	6.35▼22%	0.229▼1%	0.862▲281%	0.338▲119%	46.89▲80%	20	0.94▲81%
SD3.5 Large-Inp. [8]	3.02▲228%	0.105▲15%	11.80▲119%	0.134▲10%	109.42▲1243%	0.402▲74%	8.057▲3465%	8.657▲5521%	151.02▲481%	28▲40%	4.23▲713%
FLUX.1-Fill-Dev [17]	0.94▲2%	0.099▲9%	10.13▲88%	0.141▲16%	11.19▲37%	0.268▲16%	11.902▲5166%	9.927▲6346%	161.01▲519%	50▲150%	8.05▲1448%

ing the network's representational capacity. To bridge this capacity gap without reintroducing architectural overhead, we propose an adaptive multi-granularity distillation strategy. By dynamically balancing multiple gradient-based losses, this strategy enables high-fidelity latent alignment between the lightweight specialist and a high-capacity teacher. Crucially, this distillation strategy exhibits a profound synergy with our architectural modifications—it effectively compensates for the representational drop incurred by extreme structural compression, unlocking the full potential of the $L\lambda MI$ blocks and culminating in the optimal, meticulously balanced architecture of Moebius.

We conduct extensive experiments across natural (Places2 [63]) and portrait (CelebA-HQ [40], FFHQ [16]) benchmarks to rigorously validate the effectiveness of Moebius. As highlighted in Tab. 1, Moebius achieves an outstanding inference latency of 26.01 ms/step, with theoretical FLOPs of 0.154 TFLOPs. When factoring in the required sampling steps, this translates to a remarkable $> 15\times$ speed-up in total inference time compared to the 10B-level industrial SOTA, FLUX.1-Fill-Dev. Despite using less than 2% of the parameters (0.22B vs. 11.9B), Moebius delivers comparable or even superior generation quality, demonstrating extreme architectural efficiency.

In summary, our main contributions are as follows:

- We propose Moebius, a highly efficient lightweight image inpainting framework that sets a new efficiency standard. By functioning as a highly optimized task-specific specialist, it successfully bridges the massive scale gap, matching the performance of 10B-level generalist foundation models with only 0.22B parameters.
- We systematically conquer the representation bottleneck of compact networks by introducing the $L\lambda MI$ block alongside an adaptive multi-granularity distillation strategy. The optimal synergy between these structural and optimization designs preserves complex semantic reasoning capabilities while pushing compression to its limits.
- We provide rigorous empirical validation, including efficiency profiling and comprehensive blind user studies. Extensive evaluations across natural and portrait benchmarks, as well as real-world object removal scenarios, demonstrate Moebius's superior performance-parameter-latency trade-off.

2 Related Work

2.1 Efficient and Lightweight Architectures

Designing compact yet effective architectures has been a long-standing goal in computer vision [27]. Classical lightweight designs aim to reduce computational

complexity and parameter count while preserving representational capacity [21]. Among these efforts, DWConv [32, 38] and group convolutions [61] are widely adopted for efficient local feature extraction by decoupling spatial and channel interactions. Meanwhile, low-rank FFN designs [3,34,42,46,52] and linear attention mechanisms [1, 7, 50, 51, 53, 54] have been proposed to enhance efficiency in transformer-based architectures [8, 25, 46, 64] while maintaining representational ability. Despite their success, these approaches often face inherent trade-offs between compactness and perceptual quality [55]. In this work, Moebius overcomes these limitations by systematically conquering the representation bottleneck of compact networks. Instead of naive module substitution, we introduce the Local-λ and Interactive-λ modules to elegantly summarize spatial and semantic contexts into fixed-size linear matrices, culminating in the $L\lambda MI$ block that pushes extreme compression while preserving rigorous semantic reasoning.

2.2 Knowledge Distillation of Diffusion Models

Knowledge distillation (KD) serves as a fundamental paradigm to transfer knowledge from a large, high-capacity teacher model to a lightweight student [18]. Classical KD techniques typically operate on three supervision objectives: soft labels [11], feature maps [5, 30, 33, 41, 44], and perceptual metrics [15, 35, 57, 60]. Recently, diffusion-specific KD methods [22, 31, 36] have emerged, enabling students to approximate teacher denoising dynamics with fewer sampling steps while preserving generation quality. Distinct from conventional timestep distillation that focuses on sampling acceleration, our work targets architectural capacity transfer for extreme compression. To achieve this, Moebius employs an adaptive multi-granularity distillation strategy strictly within the latent space. By balancing multiple gradient-based objectives dynamically, our strategy perfectly compensates for the capacity drop incurred by extreme structural compression. This allows Moebius to seamlessly inherit high-level semantic priors and fine-grained textural consistency from a massive teacher, forging an optimal synergy that unlocks the full potential of our lightweight specialist.

3 Method

In this section, we present the formulation of Moebius, a highly efficient lightweight framework for image inpainting, as illustrated in Fig. 1. We first establish our baseline latent diffusion architecture in Sec. 3.1. Subsequently, in Sec. 3.2, we present an empirical analysis identifying the representation bottleneck inherent in compact networks, and detail how we systematically resolve this by proposing the Local-λ Mix Interaction ($L\lambda MI$) block. Finally, in Sec. 3.3, we introduce an adaptive multi-granularity distillation strategy designed to achieve optimal synergy with our lightweight architecture.

3.1 Overall Pipeline

We adopt the Latent Diffusion Model (LDM) [29] as our foundational framework. Let $x \in \mathbb{R}^{H\times W\times 3}$ denote an unmasked image, $m \in \{0,1\}^{H\times W}$ represent a binary mask indicating the missing regions, and $x_m = x \odot (1-m)$ denote the masked

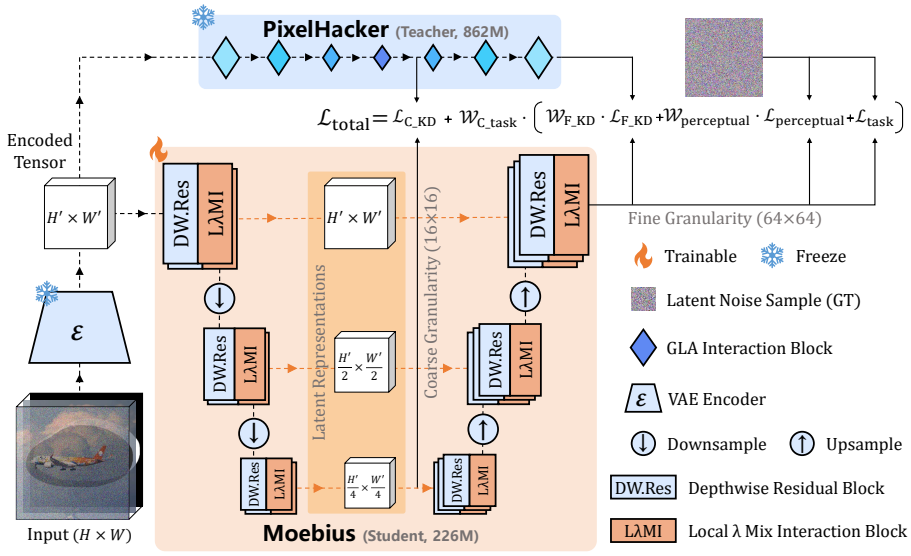

Fig. 1: Overall pipeline of Moebius. We adopt the Latent Diffusion Model (LDM) [29] framework equipped with Latent Categories Guidance (LCG) [49]. To achieve extreme architectural efficiency, the denoising U-Net is systematically restructured using our proposed $L\lambda MI$ blocks (detailed in Sec. 3.2). Furthermore, an adaptive multi-granularity distillation strategy (Sec. 3.3) is applied during training to align our lightweight specialist with the high-capacity teacher, successfully mitigating the capacity drop caused by extreme structural compression.

input image, where $\odot$ is the Hadamard product. Following the standard implementation of LDM for inpainting [8, 14, 26, 29], both the clean image x and the masked image x_m are encoded into the latent space using a pre-trained VAE [29]. Specifically, the masked image is encoded into $z_m = \mathcal{E}(x_m)$ and concatenated with the downsampled mask to form the spatial reference z_s. Concurrently, the clean latent representation $z = \mathcal{E}(x)$ along with z_m, z_s is used to construct the forward diffusion process, producing the noisy latent z_t at timestep t. The denoising network ϵ_θ, instantiated as our lightweight U-Net (see Fig. 1), is trained to predict the added noise ϵ given z_t, t. To provide rich global semantic cues, we integrate the Latent Categories Guidance (LCG) paradigm [49]. LCG employs semantic embeddings to extract latent category distribution from unmasked images during training. These embeddings, denoted as $\mathbf{E}_{\text{LCG}} \in \mathbb{R}^{K \times D}$, act as external global priors injected into the denoising network via cross-attention mechanism. We set PixelHacker [49], a SOTA LCG-based model, as our teacher network. Our goal is to forge a student model that retains the benefits of LDM and LCG while achieving extreme architectural efficiency.

3.2 Architecture Evolution: The Path to Moebius

To achieve extreme architectural efficiency, we systematically restructure the diffusion backbone. In this subsection, we first analyze the representation bot-

Table 2: Empirical Analysis of Representation Bottleneck and Architectural Synergy. Proving the necessity of our knowledge distillation and the synergy of components. Models are evaluated on the Places2 (Test) benchmark using the 18K training checkpoint (evaluation details in Sec. 4.1). CA: standard cross-attention. Lλ/Iλ: Our Local/Interactive-λ modules (details in Sec. 3.2). KD: ✗ = standard prediction loss; ✓ = our knowledge distillation, which further verifies its validity in Tab. 5.

Exp	Arch	KD	FID ↓	LPIPS ↓	Param ↓	GFLOPs ↓
①	GLA-CA-FFN, Conv	✗	32.75	0.298	526M	314.30
②	Lλ-CA-FFN, Conv	✗	37.65	0.325	496M	318.90
③	GLA-Iλ-FFN, Conv	✗	36.91	0.312	514M	307.91
④	GLA-CA-MixFFN, Conv	✗	35.24	0.301	478M	286.51
⑤	GLA-CA-FFN, DWConv	✗	43.58	0.341	315M	183.59
⑥	Lλ-Iλ-FFN, Conv	✗	33.21	0.286	485M	312.50
⑦	Lλ-Iλ-FFN, Conv	✓	24.73	0.257	485M	312.50
⑧	Lλ-Iλ-FFN, DWConv	✓	25.86	0.262	274M	181.79
⑨	**Lλ-Iλ-MixFFN, DWConv**	✓	**26.43**	**0.258**	**226M**	**154.00**
⑩	Lλ-Iλ-MixFFN, DWConv	✗	33.42	0.312	226M	154.00
⑪	GLA-CA-FFN, Conv	✓	26.81	0.262	526M	314.30
⑫	Lλ-CA-FFN, Conv	✓	28.94	0.283	496M	318.90
⑬	GLA-Iλ-FFN, Conv	✓	26.46	0.265	514M	307.91
⑭	GLA-CA-MixFFN, Conv	✓	27.35	0.251	478M	286.51
⑮	GLA-CA-FFN, DWConv	✓	29.41	0.285	315M	183.59

tleneck caused by naive structural compression, and subsequently detail our formulaically grounded solutions, as illustrated in Fig. 2 and Fig. 3.

Empirical Motivation: The Representation Bottleneck. A natural starting point for compression is to simplify the macro-architecture of the teacher model. Our empirical analysis reveals that removing the redundant final downsampling stage of PixelHacker provides a solid baseline. This modification significantly reduces the parameter count from 862M to 526M while only marginally sacrificing generation quality (FID: 32.17 → 32.75, LPIPS: 0.292 → 0.298), yielding the solid baseline corresponding to Exp ① in Tab. 2. However, pushing compression further at the micro-architecture level poses severe challenges. A straightforward approach would be substituting standard convolutions, attention mechanisms (namely, the GLA [53] used for self-attention, and the standard cross-attention), and FFN with their efficiency variants [32, 46, 48].

Our empirical observation reveals that such naive substitutions inevitably trigger a severe representation bottleneck. As shown in Tab. 2 (Exp ②-⑤), simply dropping these lightweight operators directly into the backbone without architectural synergy leads to catastrophic degradation in generation quality (e.g., FID deteriorates from 32.75 to over 43.58). This bottleneck arises from an intrinsic flaw: lightweight operators inherently suffer from constrained representational capacity, struggling to model the rigorous semantic reasoning required for inpainting. In addition, GLA [53] in the baseline is highly efficient for self-attention but lacks a mathematical formulation to perform cross-attention. This intrinsic limitation completely obstructs the synergistic architectural optimization of self- and cross-attention operations.

Local and Interactive λ Modules. To break through the representation bottleneck and efficiently integrate $\mathbf{E}_{\text{LCG}}$, we introduce the Local-λ and Interactive-λ modules. Our core insight is to bypass memory-intensive dot-product attention

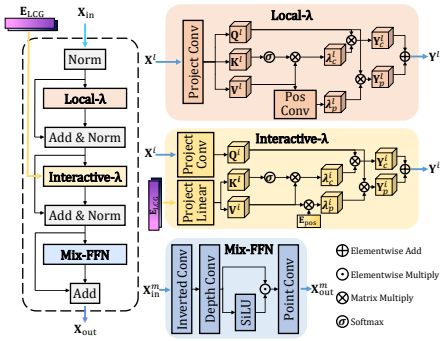

Fig. 2: Detailed architecture of the Local λ Mix Interaction ($L\lambda MI$) Block. The left panel illustrates the overall architecture, comprising three main submodules: Local-λ, Interactive-λ, and Mix-FFN. We elaborate on their mathematical formulations in Sec. 3.2.

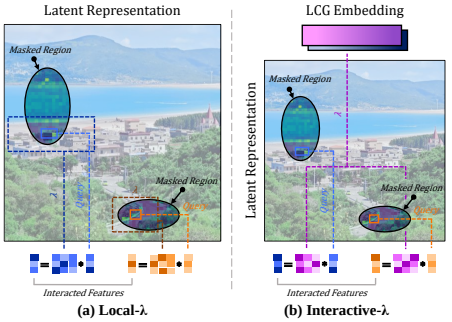

Fig. 3: Illustration of local context aggregation (Local-λ) and cross-embedding interaction (Interactive-λ) in the latent domain. In both modules, λ efficiently summarizes either spatial contexts or the global prior $\mathbf{E}_{\text{LCG}}$ into a fixed-size linear matrix, bypassing memory-intensive attention calculations.

maps by summarizing contextual and semantic information into fixed-size linear matrices (denoted as λ), allowing for linear-complexity interactions.

Local λ for Self-Attention Equivalent. The Local-λ module aggregates intra-image semantic and spatial contexts. Given an input latent feature $\mathbf{X}^l$ shaped as $B \times H' \times W' \times C$, we first project it into multi-query $\mathbf{Q}^l$, key $\mathbf{K}^l$, and value $\mathbf{V}^l$ via 1×1 convolutions and batch normalizations. Instead of computing quadratic attention maps, we construct a semantic content mapping $\boldsymbol{\lambda}_c^l$ and a positional mapping $\boldsymbol{\lambda}_p^l$:

$$\boldsymbol{\lambda}_c^l = \text{softmax}(\mathbf{K}^l)^\top \mathbf{V}^l, \quad \boldsymbol{\lambda}_p^l = \text{Conv3D}_{1\times r\times r}^{\text{pos}}(\mathbf{V}^l), \tag{1}$$

where r defines the local perception window size. The query $\mathbf{Q}^l$ then linearly interacts with these two compact matrices to produce the final self-aggregated output $\mathbf{Y}^l = \mathbf{Q}^l\boldsymbol{\lambda}_c^l + \mathbf{Q}^l\boldsymbol{\lambda}_p^l$. This dual aggregation elegantly integrates local spatial continuousness and semantic content while maintaining linear complexity.

Interactive λ for Cross-Attention Equivalent. To address the critical inability of GLA in handling cross-attention, we propose the Interactive-λ module. This module is specifically formulated to interact the latent representations with the global semantic prior $\mathbf{E}_{\text{LCG}}$. We project the latent representation $\mathbf{X}^i$ into query $\mathbf{Q}^i$, while projecting $\mathbf{E}_{\text{LCG}}$ into key $\mathbf{K}^i$ and value $\mathbf{V}^i$. Since $\mathbf{E}_{\text{LCG}}$ possesses a much smaller spatial scale than the latent representation, establishing a precise spatial-semantic correspondence is challenging. To resolve this, we introduce a lightweight positional embedding $\mathbf{E}_{\text{pos}}$ to inject explicit spatial layout information into the values, yielding the positional mapping $\boldsymbol{\lambda}_p^i$. The final interactive output $\mathbf{Y}^i$ is aggregated as:

$$\boldsymbol{\lambda}_c^i = \text{softmax}(\mathbf{K}^i)^\top \mathbf{V}^i, \quad \boldsymbol{\lambda}_p^i = \mathbf{E}_{\text{pos}}\mathbf{V}^i, \quad \mathbf{Y}^i = \mathbf{Q}^i\boldsymbol{\lambda}_c^i + \mathbf{Q}^i\boldsymbol{\lambda}_p^i. \tag{2}$$

By framing cross-attention through the lens of fixed-size matrices, the Interactive-λ module successfully integrates external semantic priors at a fraction of the traditional computational cost, solving the architectural impasse of GLA. Crucially, the joint integration of Local-λ and Interactive-λ establishes a highly efficient interaction paradigm. As validated in Tab. 2 (Exp ① → ⑥), completely replacing the baseline's attention mechanisms with our dual λ modules reduces the parameter count (526M → 485M) and computational overhead, while maintaining highly competitive generation quality (FID: 32.75 → 33.21) and even improving perceptual alignment (LPIPS: 0.298 → 0.286).

The *LλMI* Block and Lightweight Instantiation. In a standard latent diffusion U-Net, the backbone consists of two primary components: convolutional residual blocks for spatial feature extraction, and attention blocks for token interaction. To align with the extreme efficiency of our λ modules and achieve a holistic lightweight instantiation, we comprehensively restructure both.

Pushing Extreme Compression via DWConv and Mix-FFN. For spatial feature extraction, we replace the computationally heavy standard convolutional blocks with Depthwise Residual Blocks (DW.Res) [32]. While this substitution achieves massive parameter savings (Tab. 2, Exp ⑦ → ⑧), the standard FFNs within the interaction blocks still dominate the parameter budget. To push Moebius into the extreme lightweight regime (ie., $< 2\%$ of the 11.9B parameters of the industrial SOTA, FLUX.1-Fill-Dev), compressing the FFNs is unavoidable. Therefore, we integrate Mix-FFN [46,47], which replaces dense linear projections with a highly efficient depthwise-augmented structure. As empirically shown (Exp ⑧ → ⑨), integrating Mix-FFN slashes an additional 48M parameters and 27 GFLOPs. Although this aggressive compression introduces a slight trade-off in FID (25.86 → 26.43), it impressively preserves the strict perceptual quality (LPIPS: 0.262 → 0.258). This demonstrates that Mix-FFN is an excellent choice for crossing the extreme efficiency threshold without collapsing the representation.

Formulation of the LλMI Block. By elegantly cascading the proposed Local-λ, Interactive-λ, and Mix-FFN, we formulate our ultimate building block: the Local-λ Mix Interaction ($L\lambda MI$) block. Given an input latent feature $\mathbf{X}_{in}$, the forward pass of the $L\lambda MI$ block is mathematically defined as:

$$\begin{aligned} \mathbf{X}_1 &= \text{Local-}\lambda(\text{LN}(\mathbf{X}_{in})) + \mathbf{X}_{in}, \\ \mathbf{X}_2 &= \text{Interactive-}\lambda(\text{LN}(\mathbf{X}_1), \mathbf{E}_{\text{LCG}}) + \mathbf{X}_1, \\ \mathbf{X}_{out} &= \text{Mix-FFN}(\text{LN}(\mathbf{X}_2)) + \mathbf{X}_2, \end{aligned} \tag{3}$$

where $\text{LN}(\cdot)$ denotes Layer Normalization. This block successfully replaces the cumbersome spatial transformer blocks found in heavy diffusion models.

Architectural Synergy and the Capacity Dilemma. We construct the Moebius denoising U-Net by strategically stacking DWConv blocks and $L\lambda MI$ blocks across varying resolutions. As empirically validated in Tab. 2 (Exp ⑨), this specific architectural combination achieves an optimal structural synergy. It drastically

compresses the parameters to 0.22B and FLOPs to 0.154T, while maintaining a competitive generation quality under our fully equipped optimization scheme.

However, structural efficiency comes at an inherent cost. As observed in Exp ⑩ (Tab. 2), operating at this extreme compression scale solely with standard prediction loss bounds the absolute representational ceiling of the model (yielding a degraded FID of 33.42). To fully unlock the potential of the $L\lambda MI$ architecture, recover the lost capacity, and bridge the massive performance gap to 10B-level models, an advanced optimization strategy is imperative. This motivates our proposed multi-granularity distillation.

3.3 Adaptive Multi-Granularity Distillation

As discussed in Sec. 3.2, while the $L\lambda MI$ architecture successfully pushes the model into the extreme lightweight regime, this profound structural compression inevitably restricts the absolute representational capacity of the network. To bridge this capacity gap and achieve optimal architectural synergy, we introduce a multi-granularity distillation strategy. Crucially, prior works [15,28, 60] have shown that perceptual constraints can significantly preserve important structural details. However, the memory overhead of decoding high-resolution latents back into the pixel space during training is computationally prohibitive for a lightweight framework. To maintain extreme training efficiency, we perform the entire distillation—including perceptual alignment—strictly within the latent space. As demonstrated in Fig. 4, Moebius, empowered by this distillation strategy, successfully preserves high-quality representations across multiple granularities.

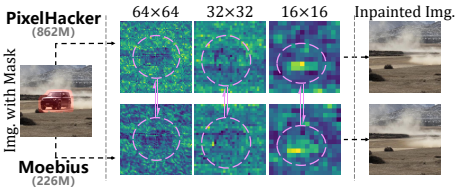

Fig. 4: Small feature spaces can still maintain high representation quality. Moebius (0.22B) exhibits highly similar activation maps to the teacher model, PixelHacker (0.86B), across multiple spatial granularities, demonstrating that it maintains consistent representational quality despite a severely compressed (4× smaller) architecture. This validates the optimal synergy between our lightweight design and the adaptive multi-granularity distillation.

Multi-Granularity Distillation Objectives. We adopt a standard Teacher-Student paradigm. The teacher model T is the officially pretrained PixelHacker [49], a high-capacity diffusion backbone equipped with LCG guidance, serving as the upper-bound performance reference. The student model S is our lightweight Moebius. To transfer the semantic and generative capabilities from T to S, we define the following optimization objectives:

Coarse-Grained Distillation. At the coarse spatial granularity (16×16), we enforce an element-wise alignment within the intermediate bottleneck. We extract the teacher's latent representation $\hat{x}_{\mathrm{C\_T}}$ after its first upsampling block, and align it with the student's latent representation $\hat{x}_{\mathrm{C\_S}}$ after its last downsampling block, formulating the coarse-grained distillation loss: $\mathcal{L}_{\mathrm{C\_KD}} = \|\hat{x}_{\mathrm{C\_T}} - \hat{x}_{\mathrm{C\_S}}\|_2^2$.

Fine-Grained Distillation and Task Supervision. At the fine spatial granularity (64×64), we supervise the final latent output. We first define a standard task loss between the student's prediction $\hat{x}_{\mathrm{S}}$ and the ground-truth latent noise x_0: $\mathcal{L}_{\mathrm{task}} = \|x_0 - \hat{x}_{\mathrm{S}}\|_2^2$. Simultaneously, we align the final predictions of the teacher $\hat{x}_{\mathrm{T}}$ and the student via an L2 distillation loss: $\mathcal{L}_{\mathrm{F\_KD}} = \|\hat{x}_{\mathrm{T}} - \hat{x}_{\mathrm{S}}\|_2^2$.

Latent Perceptual Distillation. To enhance the perceptual quality of the generated outputs without incurring the expensive pixel-space decoding costs, we employ the memory-efficient E-LatentLPIPS [15] as a perceptual constraint directly in the latent space: $\mathcal{L}_{\mathrm{perceptual}} = d_{\mathrm{E\_LatentLPIPS}}(x_0, \hat{x}_{\mathrm{S}})$. This loss aligns perceptual features entirely within the latent domain, significantly reducing memory usage and perfectly complementing our lightweight training philosophy.

Adaptive Gradient-Based Balance. In practice, optimizing the student with the aforementioned heterogeneous objectives introduces a severe convergence challenge. We observe that the losses at coarse and fine granularities differ greatly in magnitude and gradient contribution. Relying on static hyperparameter weighting makes it extremely difficult to balance convergence and generation quality. Inspired by GAN Loss [9], we propose an adaptive mechanism that dynamically adjusts the loss weights according to their gradient norms with respect to key parameter sets. Let $G(\mathcal{L}, \theta)$ denote the gradient norm (e.g., L2 norm) of loss $\mathcal{L}$ at parameter set θ. For intra-layer balance at the fine granularity, we define the adaptive weights $\mathcal{W}_{\mathrm{F\_KD}}$ and $\mathcal{W}_{\mathrm{perceptual}}$ based on parameters of the final output layer, θ_{F}. The weighted fine-grained output loss $\mathcal{L}_{\mathrm{out}}$ is formulated as:

$$\mathcal{W}_{\mathrm{F\_KD}} = \frac{||G(\mathcal{L}_{\mathrm{task}}, \theta_{\mathrm{F}})||_2^2}{||G(\mathcal{L}_{\mathrm{F\_KD}}, \theta_{\mathrm{F}})||_2^2}, \quad \mathcal{W}_{\mathrm{perceptual}} = \frac{||G(\mathcal{L}_{\mathrm{task}}, \theta_{\mathrm{F}})||_2^2}{||G(\mathcal{L}_{\mathrm{perceptual}}, \theta_{\mathrm{F}})||_2^2}, \tag{4}$$
$$\mathcal{L}_{\mathrm{out}} = \mathcal{L}_{\mathrm{task}} + \mathcal{W}_{\mathrm{F\_KD}} \cdot \mathcal{L}_{\mathrm{F\_KD}} + \mathcal{W}_{\mathrm{perceptual}} \cdot \mathcal{L}_{\mathrm{perceptual}}.$$

To balance the contributions across different granularities, we compute a cross-granularity weight $\mathcal{W}_{\mathrm{C\_task}}$. This is based on the gradient norms with respect to the intermediate feature parameters θ_{C} (i.e., the last layer before $\hat{x}_{\mathrm{C\_S}}$). The total training objective $\mathcal{L}_{\mathrm{total}}$ is defined as:

$$\mathcal{W}_{\mathrm{C\_task}} = \frac{||G(\mathcal{L}_{\mathrm{C\_KD}}, \theta_{\mathrm{C}})||_2^2}{||G(\mathcal{L}_{\mathrm{out}}, \theta_{\mathrm{C}})||_2^2}, \quad \mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{C\_KD}} + \mathcal{W}_{\mathrm{C\_task}} \cdot \mathcal{L}_{\mathrm{out}}. \tag{5}$$

This adaptive balancing mechanism effectively stabilizes the joint optimization and alleviates the need for extensive manual hyperparameter tuning, allowing the lightweight student to converge rapidly and smoothly.

4 Experiments

4.1 Experimental Setup

Implementation Details. We adopt the officially pretrained PixelHacker [49] as our teacher model, sharing the same LCG embedding size, 512×512 input resolution, and SDXL VAE encoder [26]. For the Local-λ module, the perception

window size r is set to 15. We employ the Muon optimizer [13,39] with a weight decay of 0.1. The multi-granularity distillation is conducted on 16 NVIDIA L40S GPUs with a total batch size of 768 for 138K iterations in BF16 precision. The learning rate initiates at 2e-4 and decays by a factor of 0.1 at 111K and 129K iterations. Subsequently, Moebius is fine-tuned on individual benchmarks using NVIDIA RTX 3090 GPUs. Following common practices [20,33,37], we fine-tune on Places2 [63] (1.8M images, 4 GPUs, batch size 88, 51K iters), CelebA-HQ [40] (24K images, 2 GPUs, batch size 44, 60K iters), and FFHQ [16] (60K images, 4 GPUs, batch size 88, 117K iters).

Evaluation Protocols. We rigorously evaluate Moebius across natural [63] and portrait [16, 40] domains, reporting FID [10] and LPIPS [60] following common practice. **Places2 (Test)**: following PowerPaint [66], 10K test subset, 512×512, 40–50% masks; **Places2 (Large/Small)**: following MAT [20], 36.5K validation images, 512×512, large/small masks; **Places2 (256)**: following MI-GAN [33], 36.5K validation images, 256×256, free-form masks; **CelebA-HQ (512)**: following MAT [20], 3k images, 512×512, large masks; **FFHQ (256)**: following MI-GAN [33], 10K images, 256×256, LaMa-style masks [37]).

Baselines & Fair Efficiency Profiling. We conduct extensive comparisons against SOTA academic task-specific specialists—encompassing top-ranked methods from the Papers-With-Code platform [12] alongside the latest advancements [4, 19, 23, 49]—as well as massive industrial zero-shot generalists [8, 17]. In all tables, "†" denotes results reported in the original papers [20, 33, 49, 66], while others are reproduced using official weights and code. To ensure fairness in efficiency profiling, all single-step inference latencies are measured under a strictly standardized environment: a single L40S GPU, batch size 1 at 512×512.

4.2 Main Results: Bridging the Scale Gap

Extreme Architectural Efficiency. As detailed in Tab. 3, Moebius establishes a new efficiency standard. Operating with a mere 0.226B parameters, it achieves an outstanding inference speed of 26.01 ms/step, significantly outperforming all diffusion-based competitors. When contrasted with 10B-level industrial generalists like FLUX.1-Fill-Dev (11.9B) and SD3.5 Large-Inp. (8.05B), Moebius operates with less than **2%** to **3%** of their parameter budget and delivers a **6×** acceleration in single-step latency. When factoring in total sampling steps, this efficiency gap widens substantially (Tab. 1), highlighting the massive redundancy inherent in industrial foundational models for specific restoration tasks.

Performance on Natural Scenes. Despite its extreme compactness, Moebius successfully bridges the capacity gap. On the natural-scene benchmarks (Tab. 3), Moebius demonstrates highly competitive generation capabilities. When excluding its teacher model (PixelHacker), Moebius performs on par with the 10B-level industrial SOTA, FLUX.1-Fill-Dev, across various mask conditions, and even surpasses it on the Places2 (Small) benchmark with leading scores of 0.92 FID and 0.091 LPIPS. Furthermore, it significantly outperforms SD3.5 Large-Inpainting and the vast majority of academic methods. As qualitatively shown

Table 3: Quantitative comparison with SOTA academic and industrial methods on Places2 (Test/Large/Small/256, natural scene). "†": results reported in original papers [20,33,49,66]. "Indu." denotes industrial. Benchmark details in Sec. 4.1.

	Method	Places2 (Test) FID ↓	Places2 (Test) LPIPS ↓	Places2 (Large) FID ↓	Places2 (Large) LPIPS ↓	Places2 (Small) FID ↓	Places2 (Small) LPIPS ↓	Places2 (256) FID ↓	Places2 (256) LPIPS ↓	Param. ($\times 10^9$) ↓	Latency (ms/step) ↓
	Non-Diffusion-based Methods										
Academic	LaMa [37]	21.07† ▲122%	0.213† ▲3%	-	-	-	-	22.00† ▲56%	0.378† ▼20%	-	-
Academic	MI-GAN [33]	14.36† ▲51%	0.239† ▲15%	-	-	-	-	11.83† ▼16%	0.394† ▼16%	-	-
Academic	MADF [65]	-	-	7.53† ▲220%	0.181† ▲5%	2.24† ▲143%	0.095† ▲4%	-	-	-	-
Academic	AOT-GAN [59]	-	-	10.64† ▲353%	0.195† ▲13%	3.19† ▲247%	0.101† ▲11%	-	-	-	-
Academic	HiFill [56]	-	-	28.92† ▲1131%	0.284† ▲64%	7.94† ▲763%	0.148† ▲63%	81.27† ▲475%	0.488† ▲4%	-	-
Academic	DeepFillv2 [58]	-	-	9.27† ▲294%	0.213† ▲23%	3.02† ▲228%	0.113† ▲24%	-	-	-	-
Academic	EdgeConnect [24]	-	-	12.66† ▲439%	0.275† ▲59%	4.03† ▲338%	0.114† ▲25%	-	-	-	-
Academic	MAT [20]	9.27† ▼2%	0.211† ▲2%	2.90† ▲23%	0.189† ▲9%	1.07† ▲16%	0.099† ▲9%	14.38† ▲2%	0.394† ▼16%	-	-
Academic	Latent-C.I. [4]	23.14▲144%	0.288▲39%	5.08▲116%	0.210▲21%	1.59▲73%	0.108▲19%	30.72▲117%	0.478▲2%	1.686▲646%	-
	Diffusion-based Methods										
Academic	LDM [29]	21.42† ▲126%	0.232† ▲12%	-	-	-	-	13.40† ▼5%	0.385† ▼18%	0.387† ▲71%	-
Academic	SD [29]	19.73† ▲108%	0.232† ▲12%	-	-	-	-	-	-	0.860▲281%	57.07▲119%
Academic	PowerPaint [66]	17.91† ▲89%	0.223† ▲8%	-	-	-	-	-	-	0.860▲281%	56.58▲118%
Academic	HD-Painter [23]	23.53▲148%	0.322▲56%	8.07▲243%	0.293▲69%	4.49▲388%	0.203▲123%	37.10▲163%	0.527▲12%	0.860▲281%	58.61▲125%
Academic	DDNM [45]	29.62▲212%	0.727▲251%	9.18▲291%	0.846▲389%	5.43▲490%	0.839▲822%	39.13▲177%	0.810▲72%	0.553▲145%	195.50▲652%
Academic	RoRem [19]	24.17▲155%	0.297▲43%	5.88▲150%	0.251▲45%	1.78▲93%	0.133▲46%	28.63▲103%	0.468▼0%	2.567▲1036%	113.66▲337%
Academic	PixelHacker [49]	8.59† ▼9%	0.203† ▼2%	2.05† ▼13%	0.169† ▼2%	0.82† ▼11%	0.088† ▼3%	9.25† ▼35%	0.367† ▼22%	0.862▲281%	46.89▲80%
Indu.	SD3.5 Large-Inp. [8]	37.33▲294%	0.237▲14%	10.94▲366%	0.202▲17%	3.02▲228%	0.105▲15%	74.29▲426%	0.595▲27%	8.057▲3465%	151.02▲481%
Indu.	FLUX.1-Fill-Dev [17]	8.02▼15%	0.279▲35%	1.86▼21%	0.179▲3%	0.94▲2%	0.099▲9%	10.44▼26%	0.391▼17%	11.902▲5166%	161.01▲519%
	Moebius	9.48	0.207	2.35	0.173	0.92	0.091	14.13	0.470	0.226	26.01

in Fig. 5 (Left), industrial models frequently suffer from color discrepancies and structural artifacts in fine-grained textures (e.g., foliage, water). In contrast, Moebius leverages its task-specific refinement and global semantic priors to deliver highly coherent restorations, matching the fidelity of its heavy teacher.

Performance on Portrait Scenes. The capacity of Moebius is further validated in portrait inpainting tasks. As reported in Tab. 4, Moebius achieves 5.39 FID and 0.122 LPIPS on CelebA-HQ. This performance matches the highly specialized MAT [20] and drastically surpasses all other diffusion models, second only to its teacher. On FFHQ, it obtains consistently leading scores with 8.15 FID and 0.231 LPIPS. Remarkably, Moebius completely eclipses the 10B-level industrial models in this domain, yielding improvements of 37%–1243% in FID. The qualitative comparisons in Fig. 5 (Right) confirm that Moebius flawlessly reconstructs intricate facial semantics—such as precise eye alignment and skin texture—where massive generalist models often produce structural confusion or blurring.

Table 4: Quantitative comparison with SOTA academic and industrial methods on CelebA-HQ and FFHQ (portrait scene). "†": reported in original papers [20,33,49,66]. "Indu." denotes industrial. Benchmark details in Sec. 4.1.

	Method	CelebA-HQ (512) FID ↓	CelebA-HQ (512) LPIPS ↓	FFHQ (256) FID ↓	FFHQ (256) LPIPS ↓
	Non-Diffusion-based Methods				
Academic	CoModGAN [62]	5.65† ▲5%	0.140† ▲15%	-	-
Academic	LaMa [37]	8.15† ▲51%	0.143† ▲17%	32.45† ▲298%	0.294† ▲27%
Academic	ICT [43]	12.84† ▲138%	0.195† ▲60%	-	-
Academic	MADF [65]	6.83† ▲27%	0.130† ▲7%	-	-
Academic	AOT-GAN [59]	10.82† ▲101%	0.145† ▲19%	-	-
Academic	DeepFillv2 [58]	24.42† ▲353%	0.221† ▲81%	-	-
Academic	EdgeConnect [24]	39.99† ▲642%	0.208† ▲70%	-	-
Academic	MI-GAN [33]	-	-	27.65† ▲239%	0.358† ▲55%
Academic	MAT [20]	4.86† ▼10%	0.125† ▲2%	9.04▲11%	0.232▲0%
Academic	Latent-C.I. [4]	7.62▲41%	0.153▲25%	19.24▲136%	0.278▲20%
	Diffusion-based Methods				
Academic	SD [29]	11.18▲107%	0.155▲27%	40.24† ▲394%	0.359† ▲55%
Academic	PowerPaint [66]	13.43▲149%	0.176▲44%	38.25† ▲369%	0.409† ▲77%
Academic	HD-Painter [23]	20.31▲277%	0.221▲81%	84.42▲936%	0.389▲68%
Academic	DDNM [45]	14.60▲171%	0.680▲457%	32.12▲294%	0.574▲148%
Academic	RoRem [19]	14.53▲170%	0.220▲80%	67.83▲732%	0.440▲90%
Academic	PixelHacker [49]	4.75† ▼12%	0.115† ▼6%	6.35† ▼22%	0.229† ▼1%
Indu.	SD3.5 Large-Inp. [8]	11.80▲119%	0.134▲10%	109.42▲1243%	0.402▲74%
Indu.	FLUX.1-Fill-Dev [17]	10.13▲88%	0.141▲16%	11.19▲37%	0.268▲16%
	Moebius	5.39	0.122	8.15	0.231

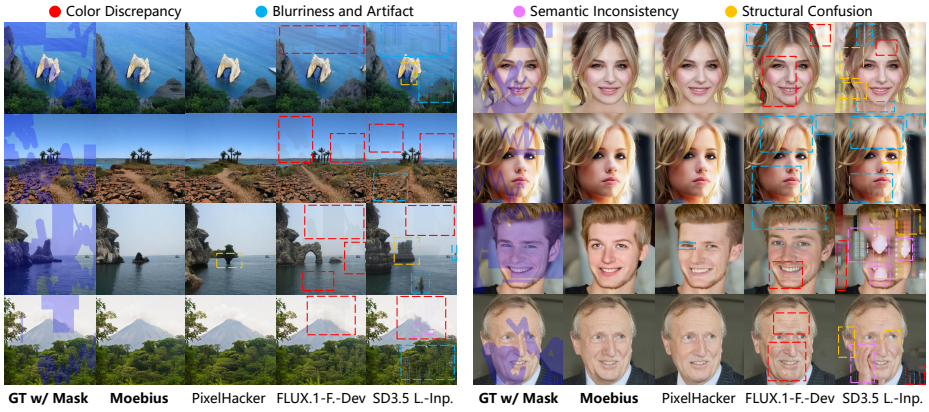

Fig. 5: Qualitative comparison with SOTA academic and industrial methods on natural and portrait scenes. Left: Places2. Right, rows 1–2: CelebA-HQ; rows 3–4: FFHQ. Moebius delivers consistent contextual generation across natural and portrait domains, avoiding the common failure cases of other methods, including color discrepancies, blur, artifacts, semantic inconsistencies, and structural confusion.

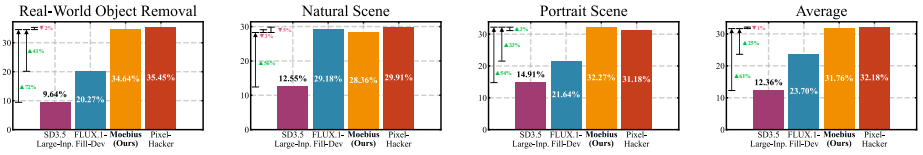

Fig. 6: User study of Moebius (0.22B) against teacher and 10B-level generalist models across various scenes. Moebius matches its teacher's performance and significantly outperforms massive generalists, excelling particularly in portrait scene.

4.3 User Study

We conducted a double-blind human preference study to assess perceptual quality. 50 cases per scenario were randomly sampled from natural, portrait, and real-world sets. 22 participants (including experts and general users) performed forced-choice tests to select the best result based on global coherence and visual fidelity. As shown in Fig. 6, Moebius (31.76% average) bridges the 50× parameter gap, matching the teacher PixelHacker (32.18%) and significantly surpassing 10B-level industrial systems (FLUX.1-Fill-Dev at 23.70% and SD3.5 Large-Inp. at 12.36%). Notably, Moebius achieves the highest preference (32.27%) in portrait scenes, suggesting that for highly structured tasks demanding strict spatial and semantic fidelity, our meticulously distilled specialist captures fine-grained structural nuances more effectively than zero-shot massive generalists.

4.4 Ablation Study

Further Analysis of Architectural Synergy. Building upon the analysis in Sec. 3.2, Exps ⑪-⑮ (Tab. 2) further explore the impact of individual components under our distillation framework. Comparing them against Exp ⑪ reveals that distilling isolated lightweight operators fails to achieve an optimal quality-efficiency

Table 5: Ablation of optimization objectives in our adaptive multi-granularity distillation strategy. Evaluated under the identical setup as Tab. 2, this analysis validates the incremental contribution of each loss.

$\mathcal{L}_{\text{C\_KD}}$	$\mathcal{L}_{\text{F\_KD}}$	$\mathcal{L}_{\text{task}}$	$\mathcal{L}_{\text{perceptual}}$	FID ↓	LPIPS ↓
✓				74.20▲181%	0.367▲42%
✓	✓			36.17▲37%	0.291▲13%
✓	✓	✓		32.59▲23%	0.273▲6%
✓	✓	✓	✓	26.43	0.258

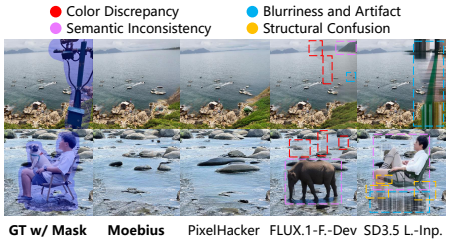

Fig. 7: Real-World Object Removal. Moebius handles realistic masks with superior consistency compared to baselines.

balance. Only the holistic integration of all structural modifications (Exp ⑨) unlocks the optimal efficiency front, reconfirming that extreme compactness demands rigorous architectural synergy over naive module substitution.

Dissection of Distillation Objectives. Furthermore, Tab. 5 dissects our optimization objectives. Relying solely on the coarse-grained loss ($\mathcal{L}_{\text{C\_KD}}$) yields a degraded FID of 74.20 for such a compressed network. Progressively integrating fine-grained distillation ($\mathcal{L}_{\text{F\_KD}}$, $\mathcal{L}_{\text{task}}$) and the latent perceptual constraint ($\mathcal{L}_{\text{perceptual}}$) systematically restores the generation quality to 26.43. This confirms that our strictly latent-space multi-granularity optimization is the crucial catalyst for unlocking the full representational capacity of $L\lambda MI$.

4.5 Real-World Removal Application

To demonstrate the practical deployment potential of Moebius beyond academic benchmarks, we provide qualitative evaluations on real-world object removal tasks, which feature complex contexts and irregular user-drawn masks. As shown in Fig. 7, FLUX.1-Fill-Dev and SD3.5 Large-Inpainting frequently introduce semantic inconsistencies, color discrepancies, and structural confusion (e.g., leaving artifacts when removing the electrical pole, or failing to reconstruct the river after removing the subjects). In stark contrast, Moebius robustly comprehends the global scene context and seamlessly inpaints the missing background, delivering flawless object removal that matches the visual fidelity of PixelHacker.

5 Conclusion

In this paper, we propose Moebius, a highly efficient specialist that sets a new standard for high-fidelity image inpainting. To conquer the representation bottleneck of extreme structural compression, we synergistically pair our proposed $L\lambda MI$ blocks with an adaptive multi-granularity latent distillation strategy. This optimal synergy empowers our 0.22B-parameter network to rival the generation quality of 10B-level industrial generalists (e.g., FLUX.1-Fill-Dev), while delivering a $> 15\times$ total inference acceleration. Ultimately, Moebius proves that extreme compactness can successfully bridge the massive scale gap, advancing resource-constrained deployment.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 62376102).

References

1. Bello, I.: Lambdanetworks: Modeling long-range interactions without attention. arXiv preprint arXiv:2102.08602 (2021)
2. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. p. 417–424. SIGGRAPH '00, ACM Press/Addison-Wesley Publishing Co., USA (2000). https://doi.org/10.1145/344779.344972, https://doi.org/10.1145/344779.344972
3. Cai, H., Li, J., Hu, M., Gan, C., Han, S.: Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17302–17313 (2023)
4. Chen, H., Zhao, Y.: Don't look into the dark: Latent codes for pluralistic image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7591–7600 (June 2024)
5. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review (2021), https://arxiv.org/abs/2104.09044
6. Dao, T.: FlashAttention-2: Faster attention with better parallelism and work partitioning. In: International Conference on Learning Representations (ICLR) (2024)
7. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in neural information processing systems **35**, 16344–16359 (2022)
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
9. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis (2021), https://arxiv.org/abs/2012.09841
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015), https://arxiv.org/abs/1503.02531
12. Inc, M.P.: Papers with code. https://web.archive.org/web/20250621114958/https://paperswithcode.com/ (2025)
13. Jordan, K., Jin, Y., Boza, V., Jiacheng, Y., Cesista, F., Newhouse, L., Bernstein, J.: Muon: An optimizer for hidden layers in neural networks (2024), https://kellerjordan.github.io/posts/muon/
14. Ju, X., Liu, X., Wang, X., Bian, Y., Shan, Y., Xu, Q.: Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion (2024)

15. Kang, M., Zhang, R., Barnes, C., Paris, S., Kwak, S., Park, J., Shechtman, E., Zhu, J.Y., Park, T.: Distilling Diffusion Models into Conditional GANs. In: European Conference on Computer Vision (ECCV) (2024)
16. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(12), 4217–4228 (2021). https://doi.org/10.1109/TPAMI.2020.2970919
17. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024)
18. Li, J., Zhao, R., Huang, J.T., Gong, Y.: Learning small-size dnn with output-distribution-based criteria. In: Interspeech 2014. pp. 1910–1914 (2014). https://doi.org/10.21437/Interspeech.2014-432
19. Li, R., Yang, T., Guo, S., Zhang, L.: Rorem: Training a robust object remover with human-in-the-loop. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14024–14035 (June 2025)
20. Li, W., Lin, Z., Zhou, K., Qi, L., Wang, Y., Jia, J.: Mat: Mask-aware transformer for large hole image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
21. Li, Y., Hu, J., Wen, Y., Evangelidis, G., Salahi, K., Wang, Y., Tulyakov, S., Ren, J.: Rethinking vision transformers for mobilenet size and speed. In: Proceedings of the IEEE international conference on computer vision (2023)
22. Lu, C., Song, Y.: Simplifying, stabilizing and scaling continuous-time consistency models (2025), https://arxiv.org/abs/2410.11081
23. Manukyan, H., Sargsyan, A., Atanyan, B., Wang, Z., Navasardyan, S., Shi, H.: Hd-painter: high-resolution and prompt-faithful text-guided image inpainting with diffusion models. In: The Thirteenth International Conference on Learning Representations (2023)
24. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: Structure guided image inpainting using edge prediction. In: The IEEE International Conference on Computer Vision (ICCV) Workshops (Oct 2019)
25. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022)
26. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023), https://arxiv.org/abs/2307.01952
27. Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., Wang, W., Banbury, C., Ye, C., Akin, B., et al.: Mobilenetv4-universal models for the mobile ecosystem. arXiv preprint arXiv:2404.10518 (2024)
28. Reda, F., Kontkanen, J., Tabellion, E., Sun, D., Pantofaru, C., Curless, B.: Film: Frame interpolation for large motion. In: European Conference on Computer Vision (ECCV) (2022)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
30. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets (2015), https://arxiv.org/abs/1412.6550
31. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models (2022), https://arxiv.org/abs/2202.00512
32. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4510–4520 (2018)

33. Sargsyan, A., Navasardyan, S., Xu, X., Shi, H.: Mi-gan: A simple baseline for image inpainting on mobile devices. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7335–7345 (October 2023)
34. Shazeer, N.: Glu variants improve transformer (2020), https://arxiv.org/abs/2002.05202
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015), https://arxiv.org/abs/1409.1556
36. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023), https://arxiv.org/abs/2303.01469
37. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. arXiv preprint arXiv:2109.07161 (2021)
38. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
39. Team, K., Bai, Y., Bao, Y., Chen, G., Chen, J., Chen, N., Chen, R., Chen, Y., Chen, Y., Chen, Y., Chen, Z., Cui, J., Ding, H., Dong, M., Du, A., Du, C., Du, D., Du, Y., Fan, Y., Feng, Y., Fu, K., Gao, B., Gao, H., Gao, P., Gao, T., Gu, X., Guan, L., Guo, H., Guo, J., Hu, H., Hao, X., He, T., He, W., He, W., Hong, C., Hu, Y., Hu, Z., Huang, W., Huang, Z., Huang, Z., Jiang, T., Jiang, Z., Jin, X., Kang, Y., Lai, G., Li, C., Li, F., Li, H., Li, M., Li, W., Li, Y., Li, Y., Li, Z., Li, Z., Lin, H., Lin, X., Lin, Z., Liu, C., Liu, C., Liu, H., Liu, J., Liu, J., Liu, L., Liu, S., Liu, T.Y., Liu, T., Liu, W., Liu, Y., Liu, Y., Liu, Y., Liu, Y., Liu, Z., Lu, E., Lu, L., Ma, S., Ma, X., Ma, Y., Mao, S., Mei, J., Men, X., Miao, Y., Pan, S., Peng, Y., Qin, R., Qu, B., Shang, Z., Shi, L., Shi, S., Song, F., Su, J., Su, Z., Sun, X., Sung, F., Tang, H., Tao, J., Teng, Q., Wang, C., Wang, D., Wang, F., Wang, H., Wang, J., Wang, J., Wang, J., Wang, S., Wang, S., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Z., Wang, Z., Wang, Z., Wei, C., Wei, Q., Wu, W., Wu, X., Wu, Y., Xiao, C., Xie, X., Xiong, W., Xu, B., Xu, J., Xu, J., Xu, L.H., Xu, L., Xu, S., Xu, W., Xu, X., Xu, Y., Xu, Z., Yan, J., Yan, Y., Yang, X., Yang, Y., Yang, Z., Yang, Z., Yang, Z., Yao, H., Yao, X., Ye, W., Ye, Z., Yin, B., Yu, L., Yuan, E., Yuan, H., Yuan, M., Zhan, H., Zhang, D., Zhang, H., Zhang, W., Zhang, X., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Y., Zhang, Z., Zhao, H., Zhao, Y., Zheng, H., Zheng, S., Zhou, J., Zhou, X., Zhou, Z., Zhu, Z., Zhuang, W., Zu, X.: Kimi k2: Open agentic intelligence (2025), https://arxiv.org/abs/2507.20534
40. Tero Karras, Timo Aila, S.L., Lehtinen, J.: Progressive growing of gans for improved quality, stability and variation. In: International Conference on Learning Representations (ICLR) (2018)
41. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation (2022), https://arxiv.org/abs/1910.10699
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
43. Wan, Z., Zhang, J., Chen, D., Liao, J.: High-fidelity pluralistic image completion with transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4672–4681 (2021). https://doi.org/10.1109/ICCV48922.2021.00465
44. Wang, C., Chen, D., Mei, J.P., Zhang, Y., Feng, Y., Chen, C.: Semckd: Semantic calibration for cross-layer knowledge distillation. IEEE Transactions on Knowledge

and Data Engineering **35**(6), 6305–6319 (2023). https://doi.org/10.1109/TKDE.2022.3171571
45. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. arXiv preprint arXiv:2212.00490 (2022)
46. Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., Han, S.: Sana: Efficient high-resolution image synthesis with linear diffusion transformer (2024), https://arxiv.org/abs/2410.10629
47. Xie, E., Chen, J., Zhao, Y., Yu, J., Zhu, L., Lin, Y., Zhang, Z., Li, M., Chen, J., Cai, H., et al.: Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer (2025), https://arxiv.org/abs/2501.18427
48. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Neural Information Processing Systems (NeurIPS) (2021)
49. Xu, Z., Duan, K., Shen, X., Ding, Z., Liu, W., Ruan, X., Chen, X., Wang, X.: Pixelhacker: Image inpainting with structural and semantic consistency (2025), https://arxiv.org/abs/2504.20438
50. Xu, Z., Zhao, H., Cui, Z., Liu, W., Zheng, C., Wang, X.: Most-dsa: Modeling motion and structural interactions for direct multi-frame interpolation in dsa images. In: European Conference on Artificial Intelligence. pp. 537–544 (2024), https://ebooks.iospress.nl/pdf/doi/10.3233/FAIA240531
51. Xu, Z., Zhao, H., Liu, W., Wang, X.: Garamost: Parallel multi-granularity motion and structural modeling for efficient multi-frame interpolation in dsa images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 28530–28538 (2025)
52. Xue, F., Zheng, Z., Fu, Y., Ni, J., Zheng, Z., Zhou, W., You, Y.: Openmoe: An early effort on open mixture-of-experts language models. arXiv preprint arXiv:2402.01739 (2024)
53. Yang, S., Wang, B., Shen, Y., Panda, R., Kim, Y.: Gated linear attention transformers with hardware-efficient training. In: Proceedings of ICML (2024)
54. Yang, S., Zhang, Y.: Fla: A triton-based library for hardware-efficient implementations of linear attention mechanism (Jan 2024), https://github.com/fla-org/flash-linear-attention
55. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
56. Yi, Z., Tang, Q., Azizi, S., Jang, D., Xu, Z.: Contextual residual aggregation for ultra high-resolution image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7508–7517 (2020)
57. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7130–7138 (2017). https://doi.org/10.1109/CVPR.2017.754
58. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.: Free-form image inpainting with gated convolution. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4470–4479 (2019). https://doi.org/10.1109/ICCV.2019.00457
59. Zeng, Y., Fu, J., Chao, H., Guo, B.: Aggregated contextual transformations for high-resolution image inpainting. IEEE Transactions on Visualization and Computer Graphics **29**(7), 3266–3280 (2023). https://doi.org/10.1109/TVCG.2022.3156949

60. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
61. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6848–6856 (2018)
62. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. In: International Conference on Learning Representations (ICLR) (2021)
63. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
64. Zhu, L., Huang, Z., Liao, B., Liew, J.H., Yan, H., Feng, J., Wang, X.: Dig: Scalable and efficient diffusion models with gated linear attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7664–7674 (June 2025)
65. Zhu, M., He, D., Li, X., Li, C., Li, F., Liu, X., Ding, E., Zhang, Z.: Image inpainting by end-to-end cascaded refinement with mask awareness. IEEE Transactions on Image Processing **30**, 4855–4866 (2021). https://doi.org/10.1109/TIP.2021.3076310
66. Zhuang, J., Zeng, Y., Liu, W., Yuan, C., Chen, K.: A task is worth one word: Learning with task prompts for high-quality versatile image inpainting (2023)