

Mage-Flow: An Efficient Native-Resolution Foundation Model for Image Generation and Editing

Microsoft Mage Team

Large-scale visual generators are increasingly capable but costly to train, fine-tune, and deploy. We introduce **Mage-Flow**, a compact 4B-scale generative stack for efficient text-to-image generation and instruction-based image editing. The stack is built from two co-designed components: **Mage-VAE**, a lightweight high-fidelity latent tokenizer, and a **Native-Resolution Multimodal Diffusion Transformer** trained with rectified flow matching. *Mage-VAE* uses one-step diffusion-style encoding and decoding with anchor-latent regularization, preserving the reconstruction quality of strong public VAEs while reducing tokenization cost by more than an order of magnitude. Together with native-resolution packing and stack-level CUDA kernel fusion, the stack supports flexible-resolution training and improves end-to-end training throughput by about $2.5\times$. Built on this foundation, we develop a complete model family with Base, RL-aligned, and Turbo variants for both generation and editing. Diffusion-NFT improves prompt following, text rendering, aesthetic quality, and editing fidelity, while few-step distillation with adversarial perceptual guidance produces 4-step Turbo models for low-latency inference. Despite its compact scale, *Mage-Flow* and *Mage-Flow-Edit* achieves competitive performance across standard generation and editing benchmarks. More importantly, the Turbo variants make high-resolution generation and editing practical for interactive use: at 1024^2 resolution on a single NVIDIA A100 GPU, *Mage-Flow-Turbo* generates an image in 0.59s, and *Mage-Flow-Edit-Turbo* edits an image in 1.02s, while maintaining a small memory footprint. These results show that careful tokenizer-backbone-system co-design can deliver strong high-resolution generation and editing within an efficient 4B model family.

Project Page: <https://microsoft.github.io/Mage>

Code: <https://github.com/microsoft/Mage>

Model: <https://huggingface.co/collections/microsoft/mage>

Date: July, 2026

1 Introduction

Recent advances in visual generative modeling have substantially improved both text-to-image generation and instruction-based image editing. Contemporary systems are expected to synthesize high-fidelity images, follow compositional prompts, render multilingual text, preserve spatial layout, and support localized or iterative editing. These capabilities enable a broad range of creation workflows, including poster and document design, product visualization, UI prototyping, scientific diagrams, and interactive image editing.

However, strong visual generation remains difficult to access and study under practical compute budgets. Closed-source systems such as GPT-Image [1], Nano Banana [2], and Seedream [3] provide strong performance but limit transparency and reproducibility. Meanwhile, competitive open-source models increasingly rely on large backbones: recent generators include Z-Image [4] at 6B, Qwen-Image [5] at 20B, FLUX.2 [6] at 32B, and Hunyuan-Image-3.0 [7] at 80B, while strong editing models such as Qwen-Image-Edit [5] and FireRed-Image-Edit [8] also reach 20B scale. While these models achieve impressive absolute quality, their scale increases the cost of



Figure 1 Qualitative showcase of *Mage-Flow*. Native-resolution samples spanning editorial design, photorealism, food photography, bilingual text rendering, portraits, and stylized art. *Mage-Flow* supports flexible generation with height and width from 512 to 2048, including extreme 4:1 aspect ratios such as 512×2048 and 2048×512 . The examples demonstrate coherent layouts, fine visual details, and legible English and Chinese text across diverse resolutions and aspect ratios.

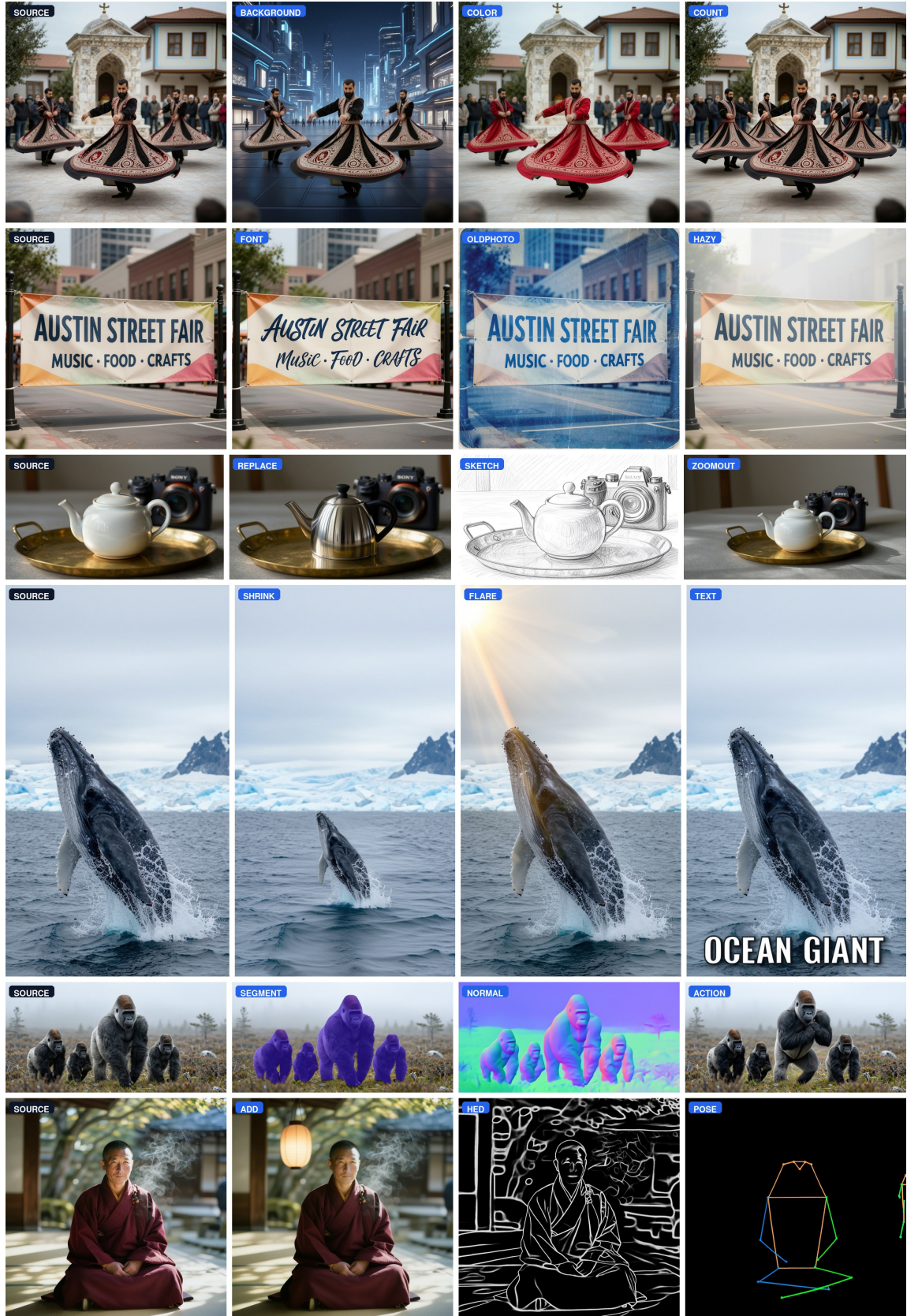


Figure 2 Qualitative showcase of *Mage-Flow-Edit* (I). Source-centric examples spanning background, color, count, font, old-photo and haze effects, subject addition/replacement, sketch rendering, zoom-out, object shrinking, lens-flare synthesis, text addition, segmentation, surface normals, action change, HED edges, and pose extraction. All showcased edit types are supported bidirectionally.

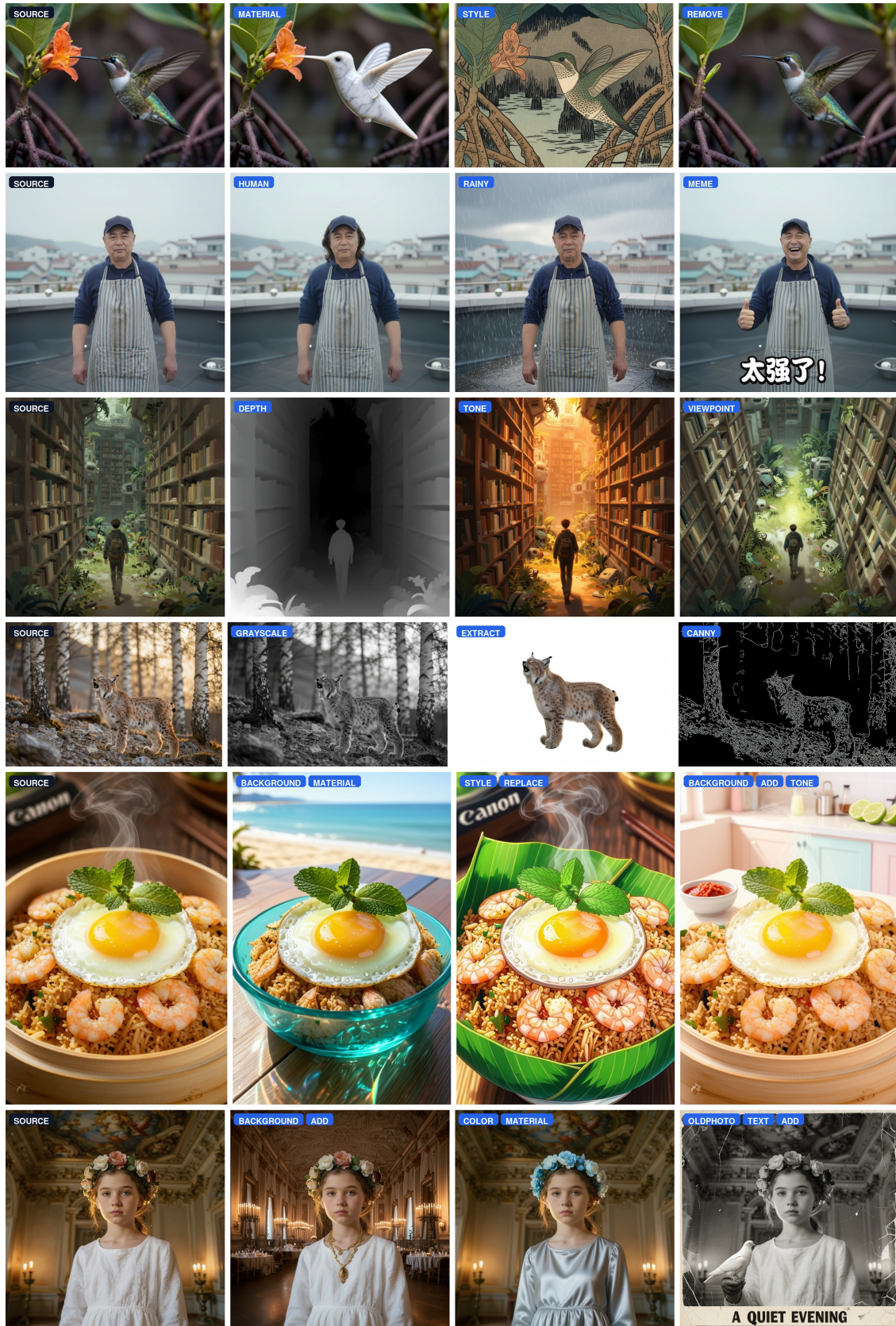


Figure 3 Qualitative showcase of *Mage-Flow-Edit* (II). Additional examples cover material and style changes, subject removal, human retouching, rainy weather, meme creation, depth, tone, viewpoint, grayscale conversion, extraction, Canny edges, and multi-type editing. All showcased edit types are supported bidirectionally.

inference, fine-tuning, controlled ablations, and domain adaptation. As a result, there remains a gap between releasing open weights and making visual generation systems truly practical to study, modify, and deploy under realistic compute budgets.

In this report, we present the ***Mage-Flow* generative stack**, a compact and research-friendly foundation for efficient image generation and editing. The stack consists of ***Mage-VAE***, a high-fidelity lightweight latent tokenizer derived from our prior CoD-Lite design [9], and a 4B **Native-Resolution Multimodal Diffusion Transformer** (NR-MMDiT) trained with rectified flow matching in the *Mage-VAE* latent space. This shared stack, together with native-resolution packing and fused-kernel training infrastructure, supports two model instantiations: ***Mage-Flow*** for text-to-image generation and ***Mage-Flow-Edit*** for instruction-based image editing. Rather than scaling to tens of billions of parameters, our goal is to provide a compact and extensible 4B-scale baseline for visual generation, controllable editing, post-training alignment, and vertical application research.

The *Mage-Flow* stack is designed around system-level efficiency rather than generator scaling alone. In latent generative pipelines, optimizing the diffusion backbone alone is insufficient: the tokenizer is invoked during training, inference, and repeated editing, and its cost grows rapidly with image resolution. We therefore first redesign the tokenizer component of the stack. Most modern public VAEs are optimized for pixel-level reconstruction but retain architectures whose encoder and decoder costs scale unfavorably at high resolution. In few-step high-resolution generation and editing, this cost can approach the diffusion backbone itself in latency and memory, making the tokenizer a practical pipeline bottleneck. To address this issue, we train ***Mage-VAE*** from scratch under three design principles. First, building on our prior CoD-Lite exploration [9], we treat the tokenizer as a learned image codec: the decoder is a fully convolutional pixel-diffusion model pre-trained with a compression-oriented objective and then distilled to a single step, avoiding the global attention blocks and multi-step computation that make conventional VAE decoders expensive at high resolution. Second, motivated by the structural symmetry of auto-encoding, we construct the encoder as the architectural dual of the decoder: a one-step diffusion model that generates latents conditioned on pixels, making encoding as lightweight as decoding. Third, we replace the standard Gaussian-prior KL with an anchor-latent KL that regularizes the posterior toward the latent distribution of a strong public VAE, yielding a generation-ready latent space for downstream diffusion training. As a result, *Mage-VAE* attains reconstruction fidelity on par with FLUX.2-VAE while requiring approximately $12\times$ and $22\times$ fewer encoding and decoding MACs per pixel, respectively. This efficient tokenizer is the first layer of the *Mage-Flow* stack, enabling high-resolution generation and repeated editing with substantially lower tokenization overhead.

On top of this latent space, the stack uses a shared **Native-Resolution Multimodal Diffusion Transformer** as the generative backbone. The 4B NR-MMDiT is trained with rectified flow matching and is used by both text-to-image generation and instruction-based editing. Instead of conventional bucket-based training, where each optimization step is restricted to one predefined resolution and aspect-ratio bucket, we adopt a native-resolution packing scheme [10]. It packs variable-length image sequences with arbitrary resolutions and aspect ratios, together with variable-length text sequences, into a single batch using FlashAttention’s variable-length kernels [11, 12] and per-sample 2D rotary embeddings. This removes the single-bucket restriction, exposes each update to heterogeneous native image sizes, and allows one checkpoint to generalize naturally to flexible output sizes. The same packing mechanism also improves inference efficiency by evaluating the conditional and unconditional classifier-free guidance branches in one packed forward pass.

Beyond architecture, efficient native-resolution training requires stack-level systems optimization. We therefore fuse the dominant memory-bound operator chains in the repeated blocks of *Mage-VAE*, the Qwen3-VL [13] text encoder, and NR-MMDiT into custom CUDA kernels. These fused kernels reduce activation memory traffic and kernel-launch overhead, increasing MFU from approximately 33% to 77% and achieving about a $2.5\times$ end-to-end training speedup (Table 4).

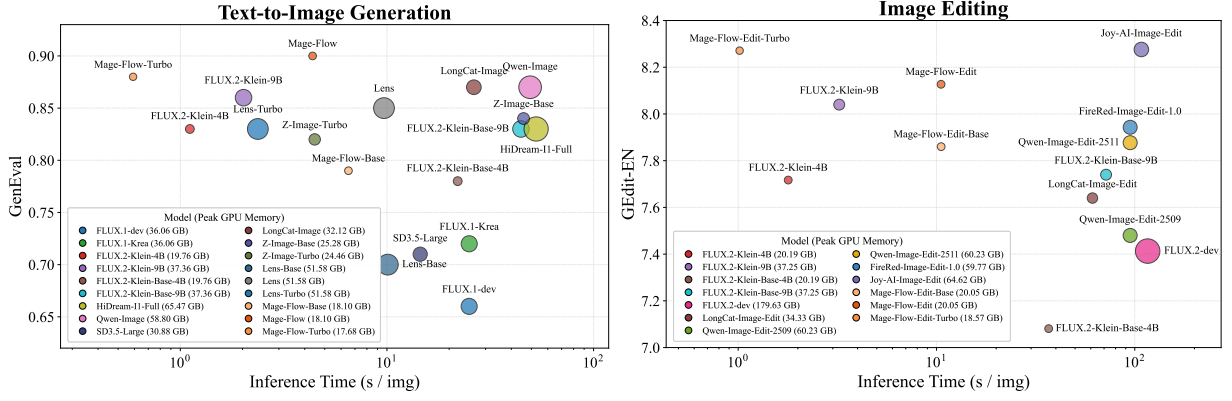


Figure 4 Quality-speed-memory comparison. We compare representative text-to-image generation and image-editing models on NVIDIA A100 GPUs. The left panel reports GenEval performance, and the right panel reports GEdit-Bench-EN performance. The x-axis shows end-to-end inference time per image, the y-axis shows benchmark performance, and the marker area is proportional to peak GPU memory. All models are evaluated on a single A100 GPU except FLUX.2-dev, which is evaluated with two A100 GPUs due to memory requirements; for FLUX.2-dev, the reported time is the two-GPU inference time and the reported memory is the sum of peak memory across both GPUs. *Mage-Flow* and *Mage-Flow-Edit* form a favorable quality-efficiency frontier with strong scores, low latency, and small memory footprint.

Together, the lightweight tokenizer, native-resolution backbone, packed CFG inference, and fused-kernel training infrastructure make efficient 4B-scale native-resolution generation and editing practical under a fixed compute budget.

We instantiate this shared stack into a complete **generation-and-editing model family**. For text-to-image generation, we first train *Mage-Flow-Base* as a 4B native-resolution foundation model, align it into *Mage-Flow* with Diffusion-NFT [14] to improve prompt following, aesthetics, bilingual text rendering, and preference alignment, and distill it into the 4-step *Mage-Flow-Turbo* using Decoupled-DMD guidance [15] with adversarial perceptual guidance [16]. For instruction-based editing, we reuse the same *Mage-VAE* latent space and NR-MMDiT backbone, but change the conditioning format to include editing instructions and source-image latents. This yields *Mage-Flow-Edit-Base*, *Mage-Flow-Edit*, and *Mage-Flow-Edit-Turbo*. Throughout editing training, we mix editing data with generation data, which improves source-conditioned editability while preserving the open-ended generative prior. As shown in Fig. 4, the resulting 4B-scale family achieves a strong quality-efficiency trade-off against representative open-source systems. At 1024² resolution, on a single NVIDIA A100 GPU, *Mage-Flow* generates images in 4.37s, while the 4-step *Mage-Flow-Turbo* reduces latency to 0.59s. For editing, the 30-step *Mage-Flow-Edit* runs at 10.55s, and the 4-step *Mage-Flow-Edit-Turbo* accelerates inference to 1.02s. Across generation and editing, peak GPU memory remains around 18-20 GB, the lowest among the compared models. Combined with competitive or superior benchmark performance against much larger open-source systems such as Qwen-Image [5], Z-Image [4], FLUX.2 [6], and FireRed-Image-Edit [8], these results show that the *Mage-Flow* stack provides a compact, fast, and memory-efficient foundation for local desktop deployment and downstream research. Overall, our main contributions are summarized as follows:

- We propose the ***Mage-Flow* generative stack**, a compact 4B-scale foundation for efficient text-to-image generation and instruction-based image editing. The stack combines a lightweight tokenizer, a native-resolution diffusion backbone, and a fused-kernel training infrastructure, providing an extensible open baseline for visual generation research.
- We design ***Mage-VAE***, a lightweight high-fidelity latent tokenizer based on one-step diffusion-style encoding and decoding with anchor-latent KL regularization. It preserves the reconstruction quality of strong public VAEs while substantially reducing high-resolution encoding

and decoding cost.

- We introduce a **Native-Resolution MMDiT** backbone trained with rectified flow matching. With native-resolution packing, variable-length text/image batching, packed CFG inference, and stack-level CUDA kernel fusion, the backbone enables efficient flexible-resolution training and inference under a fixed compute budget.
- We build a complete **generation-and-editing model family** from the shared stack, including Base, RL-aligned, and Turbo variants for both *Mage-Flow* and *Mage-Flow-Edit*. Across generation and editing settings, our models achieve a strong quality–speed–memory trade-off: they match or surpass much larger open-source systems while delivering faster inference and lower peak GPU memory among the compared models.

2 Related Work

2.1 Image Generation Foundation Models

Text-to-image generators. Text-to-image generation has evolved from pixel-space diffusion and U-Net-based latent diffusion toward large diffusion transformers and rectified-flow models. Stable Diffusion [17] popularized generation in a compressed VAE latent space, greatly reducing the cost of pixel-space diffusion, while SDXL [18] scaled the U-Net design with stronger text conditioning and multi-aspect-ratio training. Diffusion Transformers [19] further shifted the field toward transformer backbones, with SD3 [20] adopting MMDiT-style architectures trained with rectified flow, and SANA [21] improving high-resolution efficiency through linear-attention diffusion transformers. Recent open-source systems continue to scale this paradigm, including Z-Image [4], Qwen-Image [5], FLUX.2 [6], HunyuanImage 3.0 [7], and LongCat-Image [22], while closed-source systems such as Nano Banana [2] and Seedream [3] show strong generation quality but limited transparency and reproducibility. Many of these models operate at 6B–80B scales, creating practical barriers for adaptation and systematic research. In contrast, *Mage-Flow* fixes the generator at a compact 4B scale and co-designs it with a lightweight VAE, aiming to provide an efficient, extensible, and research-friendly foundation for high-quality visual generation.

Instruction-based image editors. Diffusion-based image editing has progressed from inversion- and guidance-based methods to end-to-end instruction-tuned editors. Early methods such as SDEdit [23] enabled global edits through stochastic denoising, while Prompt-to-Prompt [24] and Null-text Inversion [25] improved localized editing by manipulating cross-attention or optimized embeddings, but often required per-image optimization. Another line introduced explicit conditioning, including ControlNet [26] for spatial control and IP-Adapter [27] for image-prompt conditioning. Instruction tuning was popularized by InstructPix2Pix [28], followed by larger-scale supervised editors such as MagicBrush [29], AnyEdit [30], UltraEdit [31], and OmniGen [32]. More recent MMDiT-based editors, including FLUX.1 Kontext [33], Step1X-Edit [34], ICEdit [35], OmniGen2 [36], UniWorld-V1 [37], DreamOmni2 [38], ChronoEdit [39], JoyAI-Image-Edit [40], and FireRed-Image-Edit [8], have improved instruction following, identity preservation, and compositional editing, but often rely on substantially larger backbones. *Mage-Flow-Edit* instead provides a compact 4B instruction-based editor that retains strong editing quality while substantially lowering the cost of adaptation, fine-tuning, and downstream research.

Unified multimodal generation models. A parallel line of work studies unified multimodal models that combine image understanding, generation, and editing in a single backbone [41]. Janus-Pro [42] uses separate visual encoders for understanding and generation while sharing one autoregressive transformer. Transfusion [43] trains a single transformer over interleaved text and image data

with both next-token prediction and diffusion objectives. BAGEL [44] adopts a decoder-only architecture for multimodal understanding and generation, while Emu3.5 [45] further scales native multimodal modeling toward broad generation and editing capabilities. These models pursue general-purpose multimodal unification and are typically larger and architecturally distinct from MMDiT-based generators. *Mage-Flow* focuses on the complementary direction: specializing a 4B MMDiT stack for efficient, high-quality generation and editing, while treating understanding as an orthogonal capability that can be incorporated through external encoders or downstream systems.

2.2 Tokenization for Image Generation

Visual tokenization bridges raw pixels and generative backbones, and largely determines the efficiency frontier of modern image generators. Continuous-latent VAEs, introduced by Latent Diffusion Models [17] and later used in systems such as SD3 [20] and FLUX.1, remain the dominant choice for diffusion-based generation. In parallel, discrete tokenizers such as VQ-VAE [46] and VQ-GAN [47] laid the foundation for autoregressive image generation. However, most open generators inherit public VAEs optimized primarily for pixel-level reconstruction, whose encoder and decoder costs grow rapidly with resolution and can become a major latency and memory bottleneck in high-resolution generation and repeated editing.

A related line of work rethinks tokenization through diffusion-based generative compression. Multi-step diffusion codecs [48–50] can achieve strong perceptual quality at low bitrates, but their iterative sampling makes them too slow for real-time generative pipelines; one-step diffusion codecs [51–53] improve latency, but often rely on heavy DiT- or U-Net-style backbones. CoD-Lite [9] provides an important observation for this regime: at small scales, compression-oriented pre-training transfers better to learned codecs than generation-oriented pre-training, and after distillation and adversarial training, lightweight fully-convolutional decoders are sufficient without global attention. Building on these insights, *Mage-VAE* trains the tokenizer from scratch for downstream MMDiT generation. Its decoder is a fully-convolutional pixel-diffusion model pre-trained with a compression-oriented objective and distilled to a single step, avoiding the global attention blocks and multi-step computation that make conventional VAE decoders expensive at high resolution. Since auto-encoding is symmetric, we construct the encoder as the architectural dual of the decoder: a one-step diffusion model that generates latents conditioned on pixels, making encoding as efficient as decoding. Finally, *Mage-VAE* replaces the standard Gaussian-prior KL with an anchor-latent KL that regularizes the posterior toward a fixed VAE latent distribution, producing a generation-ready latent space with $16\times$ spatial reduction and 128 channels.

2.3 Post-training for Image Generation

Post-training is critical for aligning pre-trained diffusion and flow-matching generators with human preferences, prompt adherence, and aesthetic quality. One major direction extends offline preference optimization to diffusion models. Diffusion-DPO [54] applies the DPO objective over diffusion trajectories, while D3PO [55] fine-tunes directly from preference comparisons without an explicit reward model. These methods are simple and stable because they operate on pre-collected preference data, and have become common components in modern text-to-image post-training pipelines.

A complementary direction uses online reinforcement learning to optimize non-differentiable or black-box rewards. DDPO [56] casts denoising as a sequential decision process and fine-tunes generators with learned or rule-based rewards, often using reward models such as ImageReward [57]. Extending RL to flow matching is more challenging because ODE sampling is deterministic and lacks an explicit action distribution. Flow-GRPO [58] addresses this by converting the ODE sampler into a stochastic SDE with matched marginals, whereas Diffusion-NFT [14] bypasses the reverse process and performs negative-aware fine-tuning on the forward process, requiring neither likelihood estimation nor a specific solver. In *Mage-Flow* and *Mage-Flow-Edit*, we adopt

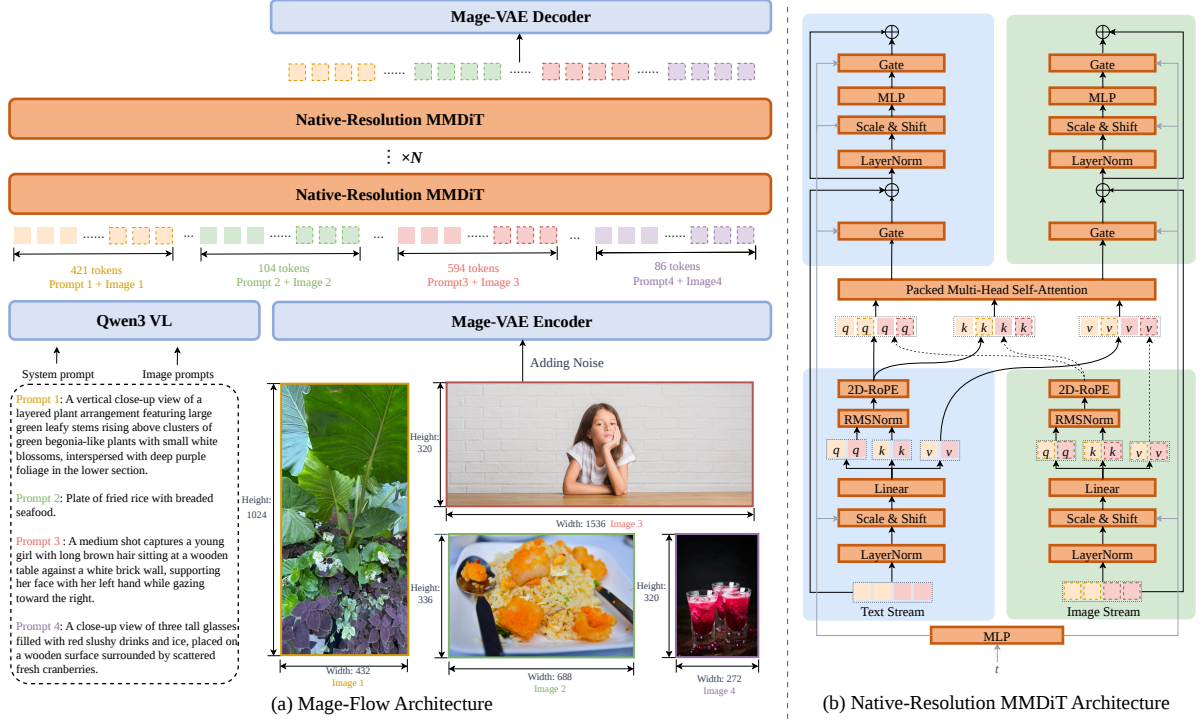


Figure 5 *Mage-Flow* architecture. *Mage-Flow* uses Qwen3-VL to encode text prompts and *Mage-VAE* to map images into compact transformer-ready latents. Images of different resolutions and aspect ratios are flattened into variable-length latent token sequences and packed together with variable-length text tokens in a single batch. The 4B Native-Resolution MMDiT processes these packed sequences with per-sample 2D rotary positional embeddings and FlashAttention variable-length attention, avoiding fixed resolution buckets while preserving each sample’s native spatial layout. The right panel illustrates the MMDiT block, where text and image streams use modality-specific normalization and projection layers but interact through joint self-attention.

Diffusion-NFT as the post-training stage on top of supervised fine-tuning, applying it to curated capability-targeted generation and editing data to improve prompt adherence, text rendering, compositional control, and editing fidelity.

2.4 Distillation for Image Generation

Iterative diffusion sampling is costly at deployment, motivating extensive work on reducing the number of sampling steps. Training-free solvers such as DDIM [59] and higher-order ODE solvers reduce step counts without retraining but often degrade under very small budgets. Distillation methods push this further: progressive distillation [60] iteratively halves teacher trajectories; consistency models [61] and latent consistency models [62] learn few-step mappings through self-consistency along the probability-flow ODE; and InstaFlow [63] improves one-step generation by straightening flow trajectories.

Another family distills at the distribution level. Distribution Matching Distillation (DMD) [64] matches the student distribution to a teacher using an auxiliary score model, while DMD2 [65] removes regression loss, introduces adversarial training, and supports multi-step student sampling. Adversarial diffusion distillation [66] and Hyper-SD [67] further combine score- or consistency-based objectives with adversarial losses to preserve visual detail at small step counts. More recently, Decoupled DMD [15] separates CFG augmentation from distribution matching and allows independent noise schedules, while SenseFlow [16] introduces adversarial guidance based on frozen Vision Foundation Models (VFMs) for flow-based generators. Specifically, frozen VFMs such as CLIP [68] and DINOv2 [69] are used to extract semantic features from both generated and

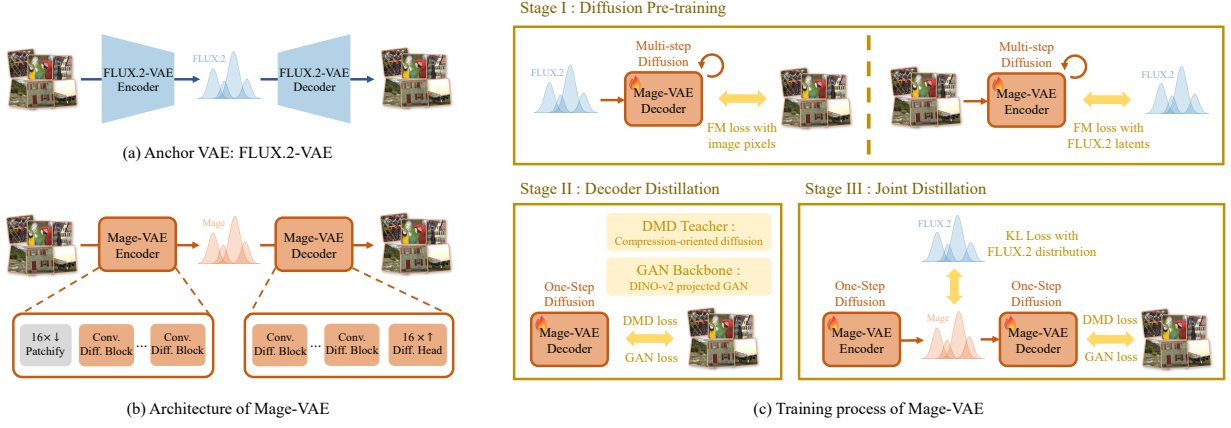


Figure 6 *Mage-VAE* is a lightweight pixel-diffusion-based VAE distilled from the FLUX.2-VAE latent space. FLUX.2-VAE provides the anchor latent distribution used for diffusion pre-training and anchor-latent regularization. The *Mage-VAE* decoder is distilled to reconstruct images from FLUX.2-aligned latents, while the *Mage-VAE* encoder is trained to produce latents compatible with the same anchor space. This design yields an efficient one-step encoder-decoder that preserves a generation-ready latent structure and can be interchanged with FLUX.2-VAE in downstream generators.

real images, and a lightweight feature discriminator trained on these representations provides adversarial gradients to improve realism and semantic alignment. *Mage-Flow* and *Mage-Flow-Edit* use Decoupled DMD as the main distillation backbone and add this VFM-based adversarial guidance as a complementary signal, producing 4-step Turbo variants for both generation and editing.

3 Mage-Flow Stack

As shown in Fig. 5(a), the *Mage-Flow* stack is built around two core model components: *Mage-VAE*, a lightweight high-fidelity latent tokenizer, and a 4B Native-Resolution Multimodal Diffusion Transformer (NR-MMDiT) trained with rectified flow matching. *Mage-VAE* provides a compact generation-ready latent space, while NR-MMDiT models packed latent sequences with flexible resolutions and aspect ratios. The stack is supported by a training and inference infrastructure that makes native-resolution 4B-scale learning practical. We describe *Mage-VAE* in Section 3.1, NR-MMDiT in Section 3.2, and the training infrastructure in Section 3.3.

3.1 Mage-VAE: Lightweight Image Tokenizer

Public VAEs released alongside SD3 [20] and the FLUX family [6] largely inherit the VQGAN [70] and LDM [71] autoencoder architectures, which were originally designed for low-resolution 256×256 images. When scaled to modern 1K and 2K image generation, inference latency and memory consumption grow unfavorably due to architectural components such as global attention and computationally expensive high-resolution blocks. As a result, VAE encoding and decoding can become a significant source of latency, approaching the diffusion Transformer itself in few-step high-resolution generation. For example, during the 4-step generation of 1K-resolution images using FLUX.2-Klein-4B, VAE decoding accounts for 14% of the total generation time. *Mage-VAE* is designed to remove this bottleneck while preserving reconstruction fidelity. It replaces heavy high-resolution autoencoder components with lightweight one-step diffusion-style encoding and decoding, and regularizes the latent space with an anchor-latent KL toward a strong public VAE. Its architecture and training pipeline are summarized in Fig. 6.

Table 1 Evaluation results of *Mage-VAE*. The model parameters and computational complexity (measured by kMACs/pixel) of both encoding and decoding are reported. Reconstruction quality is measured on CLIC 2020 testset at native resolution and FFHQ val 10k at 1024×1024 resolution.

Model	Params (M)		kMACs/px		CLIC 2020 Test			FFHQ (val 10k)		
	Enc.	Dec.	Enc.	Dec.	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
SD-VAE [71]	34	49	2130	4796	30.10	0.8156	0.0565	33.00	0.8662	0.0461
SD-3.5-VAE [20]	34	50	2132	4797	32.79	0.8896	0.0271	36.23	0.9330	0.0189
FLUX-VAE [33]	34	50	2132	4797	34.80	0.9264	0.0188	38.43	0.9587	0.0129
FLUX.2-VAE [6]	34	50	2134	4798	36.88	0.9447	0.0139	40.47	0.9682	0.0102
Qwen-Image-VAE [5]	54	73	3378	5423	34.95	0.9129	0.0571	38.75	0.9515	0.0422
HunyuanVideo-VAE [74]	100	146	6198	14096	36.14	0.9305	0.0250	39.86	0.9610	0.0205
HunyuanImage-3.0-VAE [7]	389	871	28053	49836	33.80	0.8864	0.0546	36.84	0.9259	0.0468
Mage-VAE	49	52	173	215	36.61	0.9450	0.0148	40.67	0.9708	0.0107

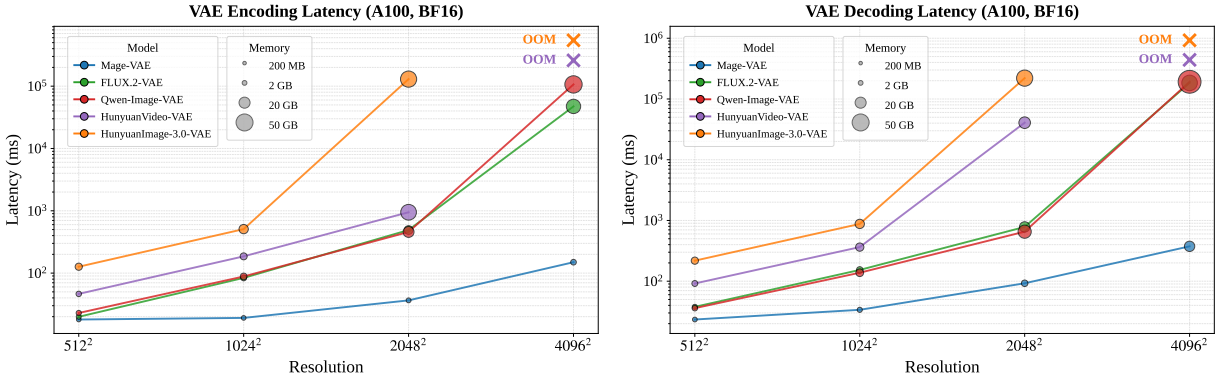


Figure 7 Inference cost of *Mage-VAE*. Encoding and decoding latency and memory is tested across resolutions on 80GB A100 GPU. SD-VAE, SD-3.5-VAE, and FLUX-VAE are not shown, as their model structures are almost identical to FLUX.2-VAE, causing their inference costs to completely overlap.

3.1.1 Architecture

Lite codec-style decoder. Our decoder design is inspired by CoD-Lite [9], a lightweight diffusion-based image codec for high-fidelity perceptual reconstruction. CoD-Lite shows that compression-oriented diffusion pre-training, followed by one-step distillation, can yield strong reconstruction quality without heavy DiT- or U-Net-style decoders. Following this principle, we formulate the *Mage-VAE* decoder as a one-step pixel diffusion model. It uses stacked convolutional diffusion blocks [72] and a decoupled pixel diffusion head [73] to reconstruct RGB pixels directly from latents. This fully convolutional design avoids global attention and keeps decoding cost nearly linear in image resolution.

Lite symmetric encoder. Unlike image codecs, where decoding efficiency is often the primary concern, generative models also require efficient encoding: training needs to prepare image latents at scale, and editing models must repeatedly encode reference or source images. We therefore design the encoder as the architectural dual of the decoder. Since a decoder can be viewed as a pixel generator conditioned on latents, we view the encoder as a latent generator conditioned on pixels. Concretely, the encoder is also implemented as a one-step diffusion model, consisting of a patch embedding layer followed by stacked convolutional diffusion blocks. This symmetric design makes latent extraction as lightweight as pixel reconstruction.

Table 2 Anchor-latent tokenizer compatibility. We evaluate generation and editing backbones with different latent tokenizers. The comparison swaps between *Mage*-VAE and FLUX.2-VAE while keeping the downstream backbone fixed. Higher is better for all metrics.

(a) Text-to-image generation.										
Backbone	Tokenizer	GenEval	DPG-Bench	TIIF-Short	TIIF-Long	CVTG-2K	OneIG-EN	OneIG-CN	LongText-EN	LongText-CN
<i>Mage-Flow-Turbo</i>	<i>Mage</i> -VAE	0.88	85.48	83.58	84.16	0.873	0.523	0.491	0.911	0.801
	FLUX.2-VAE	0.88	85.49	81.18	81.09	0.869	0.523	0.491	0.909	0.800
FLUX.2-Klein-4B [6]	FLUX.2-VAE	0.83	85.53	78.91	79.04	0.628	0.500	0.364	0.649	0.068
	<i>Mage</i> -VAE	0.83	85.52	77.40	79.43	0.623	0.502	0.361	0.654	0.068

(b) Instruction-based editing.						
Backbone	Tokenizer	ImgEdit	GEEdit-EN	GEEdit-CN	TextEdit-Syn	TextEdit-Real
<i>Mage-Flow-Edit-Turbo</i>	<i>Mage</i> -VAE	4.38	8.271	8.264	12.77	15.41
	FLUX.2-VAE	4.34	8.098	8.112	12.85	15.26
FLUX.2-Klein-4B [6]	FLUX.2-VAE	4.01	7.717	7.750	11.84	14.46
	<i>Mage</i> -VAE	3.95	7.734	7.672	11.65	14.46

3.1.2 KL regularization with an anchor latent distribution

CoD-Lite constrains its latent space with a bitrate objective. Under the commonly used uniform-noise posterior approximation, this bitrate objective can be interpreted as a KL-style regularization term in latent space. From this perspective, a VAE can be viewed as a learned image codec with a Gaussian latent posterior, suggesting that codec architectures can be transferred to VAE design by replacing bitrate control with variational regularization.

Following this view, we convert the CoD-Lite-style codec into a generation-ready VAE by replacing the bitrate constraint with a KL regularizer. Unlike conventional VAEs that match the posterior to a standard Gaussian prior, we regularize the posterior toward an anchor latent distribution induced by the FLUX.2-VAE [6]. Since FLUX.2-VAE uses 32 latent channels with an $8\times$ spatial reduction and these latents are typically $2\times$ patchified before entering the diffusion Transformer, *Mage*-VAE internalizes this patchification by directly producing $16\times$ -downsampled latents with 128 channels. We use the patchified FLUX.2 latents as anchor targets for both encoder pre-training and KL regularization, allowing *Mage*-VAE to inherit a generation-friendly latent structure while producing Transformer-ready latents for downstream diffusion training.

3.1.3 Training

The training pipeline contains three stages. In **Stage I**, we pre-train the encoder and decoder separately as multi-step diffusion models using a standard flow-matching objective in the $\mathcal{X}$ -prediction parameterization. The decoder learns to reconstruct image pixels, while the encoder learns to predict the patchified FLUX.2-VAE anchor latents from pixels. In **Stage II**, we distill the decoder into a one-step model using reconstruction loss, a DMD loss [64] with a compression-oriented diffusion teacher [75], together with a DINOv2-projected GAN loss [76] to improve perceptual fidelity. In **Stage III**, we jointly fine-tune the one-step encoder and decoder with the anchor-latent KL regularizer, producing a compact VAE whose latent space is ready for *Mage-Flow* training. The detailed training process can be found in Appendix A.

3.1.4 Evaluation

Reconstruction. In Table 1, we evaluate the reconstruction performance of *Mage*-VAE using two complementary settings: the high-quality general-domain CLIC 2020 test set [77] at its native resolution ($\sim 2K$), and the portrait-domain FFHQ validation set [78] at 1024×1024 . *Mage*-VAE achieves reconstruction fidelity comparable to the strongest FLUX.2-VAE baseline on CLIC 2020,

Table 3 Inference efficiency of packed CFG evaluation. By packing conditional and unconditional branches into a single batch, *Mage-Flow* reduces the classifier-free guidance overhead while preserving the original denoising trajectory. All configurations are benchmarked on a single NVIDIA A100 GPU.

Model	Steps	Separate CFG	Packed CFG	Speedup
<i>Mage-Flow-Base</i>	30	7.5089s	6.5159s	1.15×
<i>Mage-Flow</i>	20	5.0076s	4.3680s	1.15×
<i>Mage-Flow-Edit-Base</i>	30	11.5463s	10.5582s	1.09×
<i>Mage-Flow-Edit</i>	30	11.5985s	10.5475s	1.10×

while obtaining the best or near-best reconstruction quality on FFHQ among the compared VAEs. This shows that the proposed lightweight design preserves high visual fidelity across both general-domain and face-centric images.

Inference efficiency. Table 1 and Fig. 7 further evaluate the efficiency. *Mage-VAE* reduces the computational complexity of FLUX.2-VAE by about 12.3× for encoding and 22.3× for decoding. This translates into consistently lower latency and memory usage across all tested resolutions. The advantage becomes more pronounced at high resolution, where baseline VAEs become extremely slow or run out of memory at 4096×4096 , while *Mage-VAE* remains efficient for both encoding and decoding. These results demonstrate that *Mage-VAE* provides a substantially better quality–efficiency trade-off, making it suitable for high-resolution generation, repeated editing, and interactive deployment.

Generation and editing. Beyond pixel-level reconstruction and efficiency, we further evaluate whether *Mage-VAE* preserves the generation-ready latent structure of FLUX.2-VAE. Since *Mage-VAE* is distilled from and regularized toward the latent distribution of the high-capacity but computationally expensive FLUX.2-VAE, the two tokenizers should remain compatible when used by downstream generators. Table 2 reports a cross-tokenizer ablation that swaps *Mage-VAE* and FLUX.2-VAE while keeping the generation or editing backbone fixed. Replacing *Mage-VAE* with FLUX.2-VAE in our Turbo models yields similar benchmark performance, and replacing FLUX.2-VAE with *Mage-VAE* in FLUX.2-Klein-4B [6] also maintains comparable results. This suggests that *Mage-VAE* not only reconstructs images well, but also preserves the latent geometry required by FLUX.2-style diffusion generators, allowing it to serve as an efficient substitute for the original FLUX.2-VAE in downstream generation and editing models.

Finding 1: Anchor-latent supervision enables efficient VAE distillation without breaking the generation-ready latent space: by aligning a lightweight one-step encoder–decoder to the latent distribution of a strong but computationally expensive VAE, the tokenizer preserves reconstruction quality and cross-generator compatibility while substantially reducing encoding and decoding cost.

3.2 Native-Resolution MMDiT

3.2.1 Model Architecture and Native Packing

The generative backbone of *Mage-Flow* is a 4B-parameter Native-Resolution Multimodal Diffusion Transformer, as illustrated in Fig. 5(b). It follows the MMDiT block design introduced by SD3 [20], where text and image tokens are concatenated and processed by joint self-attention. Each block uses modality-specific normalization and projection layers to preserve the distinct statistics of

text and visual streams, while cross-modal interaction is performed through self-attention over the combined sequence.

A key difference from standard MMDiT training is that NR-MMDiT does not restrict each optimization step to a predefined resolution bucket. In conventional bucket-based training, images are assigned to a finite set of fixed resolution and aspect-ratio buckets, and each step draws samples from only one bucket so that all visual token grids share the same spatial shape. This simplifies batching, but it discretizes the native resolution distribution, introduces bucket-quantization mismatch, and limits the diversity of aspect ratios observed within each update. It also makes extremely wide or tall outputs difficult to support unless corresponding buckets are explicitly added. Inspired by NiT [10], we instead train directly on native-resolution sequences. Images with different resolutions and aspect ratios are encoded by *Mage-VAE*, flattened into variable-length latent token sequences, and packed into a single contiguous batch under a fixed token budget. We apply the same packing principle to text conditions: prompt embeddings of different lengths are packed rather than padded to a common maximum length. With FlashAttention’s variable-length kernels [11, 12], the packed text and image sequences are processed using per-sample cumulative offsets, which restrict attention within each sample without constructing explicit block-diagonal masks.

This native-resolution formulation has several practical advantages. First, samples are packed under a fixed token budget, so training can naturally balance heterogeneous image sizes together with variable-length text conditions. This improves batching flexibility and avoids unnecessary padding on the text side, while allowing the image branch to preserve each sample’s native latent grid. Second, it removes the single-bucket restriction during training: each optimization step can contain images with different native resolutions and aspect ratios, rather than drawing all samples from one predefined resolution bucket. This exposes the model to a richer and less discretized resolution distribution within every update, and avoids the bucket-quantization mismatch introduced by mapping continuous image sizes to a finite set of buckets. Therefore, a single checkpoint generalizes naturally to flexible output sizes at inference time. As shown in Fig. 1, both height and width can range from 512 to 2048, enabling not only standard square, portrait, and landscape outputs, but also extreme aspect ratios such as 512×2048 and 2048×512 .

The same packing mechanism also improves inference efficiency. For classifier-free guidance, the conditional and unconditional branches can be packed together and evaluated in a single forward pass, avoiding the redundant computation introduced by separate CFG evaluation while preserving the original denoising trajectory. As shown in Table 3, packed CFG consistently accelerates inference by $1.09\times$ – $1.15\times$ across Mage-Flow and Mage-Edit variants. Therefore, native-resolution packing serves as both a training strategy for heterogeneous-resolution learning and an inference optimization for efficient CFG evaluation.

Finding 2: Native-resolution packing turns resolution diversity into a training signal: removing the single-bucket restriction allows each update to mix heterogeneous image sizes and aspect ratios, improving resolution flexibility while also enabling efficient packed CFG inference.

3.2.2 Conditioning and rectified-flow training

Text conditioning is provided by a frozen Qwen3-VL-4B-Instruct [13] text encoder, which maps each prompt into contextual embeddings τ . Given an image x , *Mage-VAE* encodes it into latents $z = \text{Mage-VAE}(x)$ at a $16\times$ spatial reduction with 128 channels. These latents are flattened into visual tokens and linearly projected to the NR-MMDiT hidden dimension. The model is trained

Table 4 Training-efficiency ablation. All configurations are benchmarked on a single 8-GPU NVIDIA B200 node with FlashAttention-4 [12]. The global batch size is 8, with one packed sample per GPU, and each sample contains a fixed packed sequence length of 50,000 tokens. We progressively replace FLUX.2-VAE with *Mage-VAE* and enable fused CUDA kernels for *Mage-VAE*, the Qwen3-VL text encoder, and NR-MMDiT. Peak per-GPU memory, model FLOP utilization (MFU), per-step wall-clock time, and relative speedup are reported.

Tokenizer	VAE Fuse	Text Fuse	DiT Fuse	Mem. (GB)	MFU	Time (s)	Speedup
FLUX.2-VAE [6]	–	–	–	175.45	33.20%	1.9259	1.00×
<i>Mage-VAE</i>	–	–	–	175.47	45.63%	1.3634	1.41×
<i>Mage-VAE</i>	✓	–	–	175.47	45.85%	1.3586	1.42×
<i>Mage-VAE</i>	✓	✓	–	175.47	47.20%	1.3196	1.46×
<i>Mage-VAE</i>	✓	✓	✓	141.44	77.26%	0.7748	2.49×

in the *Mage-VAE* latent space with the rectified flow-matching objective [20],

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,\tau),t,\epsilon} \left[\|v_\theta(z_t, t, \tau) - (z - \epsilon)\|_2^2 \right],$$

where

$$z_t = (1 - t)z + t\epsilon, \quad \epsilon \sim \mathcal{N}(0, I).$$

The same rectified-flow objective and NR-MMDiT backbone are used for both generation and editing, but with different conditioning inputs. For *Mage-Flow*, Qwen3-VL encodes the text prompt into τ , and NR-MMDiT denoises the target latent tokens conditioned on τ . For *Mage-Flow-Edit*, Qwen3-VL encodes the editing instruction together with the source image into multimodal conditioning embeddings τ , while *Mage-VAE* encodes the source and target images into z^{src} and z^{tgt} . Thus, the NR-MMDiT input sequence concatenates τ , z^{src} , and the noisy target latent tokens.

To distinguish source and target visual tokens, *Mage-Flow-Edit* extends the 2D rotary positional embedding (RoPE) used in *Mage-Flow* with an additional frame dimension. Each visual token is assigned a 3D position index (h, w, f) , where (h, w) denotes its native spatial location and f denotes the image index, covering all source images and the target image. This frame-aware RoPE preserves source-target spatial correspondence while indicating token roles inside the shared attention sequence. The loss is computed only on target tokens, so *Mage-Flow-Edit* can be initialized directly from *Mage-Flow* without adding separate editing modules.

3.3 Training Infrastructure

The training efficiency of the *Mage-Flow* stack is determined by three repeatedly executed modules: the convolutional diffusion blocks in *Mage-VAE*, the Transformer blocks in the frozen Qwen3-VL text encoder, and the 4B NR-MMDiT blocks. Although their main arithmetic comes from convolutions, matrix multiplications, and attention, each block also contains many memory-bound operator chains, including normalization, adaptive modulation, RoPE application, gating, activation, and residual addition. When launched as separate CUDA kernels, these operators repeatedly read and write large activation tensors, causing substantial memory traffic and kernel-launch overhead.

To reduce this overhead, we fuse the dominant operator chains inside the repeated blocks of the full stack. In *Mage-VAE*, we fuse normalization–activation–residual chains in the convolutional diffusion blocks. In the Qwen3-VL text encoder and NR-MMDiT, we fuse common Transformer-side operations such as adaptive normalization, rotary embedding application, and gated residual updates. These fused kernels keep intermediate values in on-chip memory and write back only the final outputs, reducing both activation memory movement and kernel launches. Because these

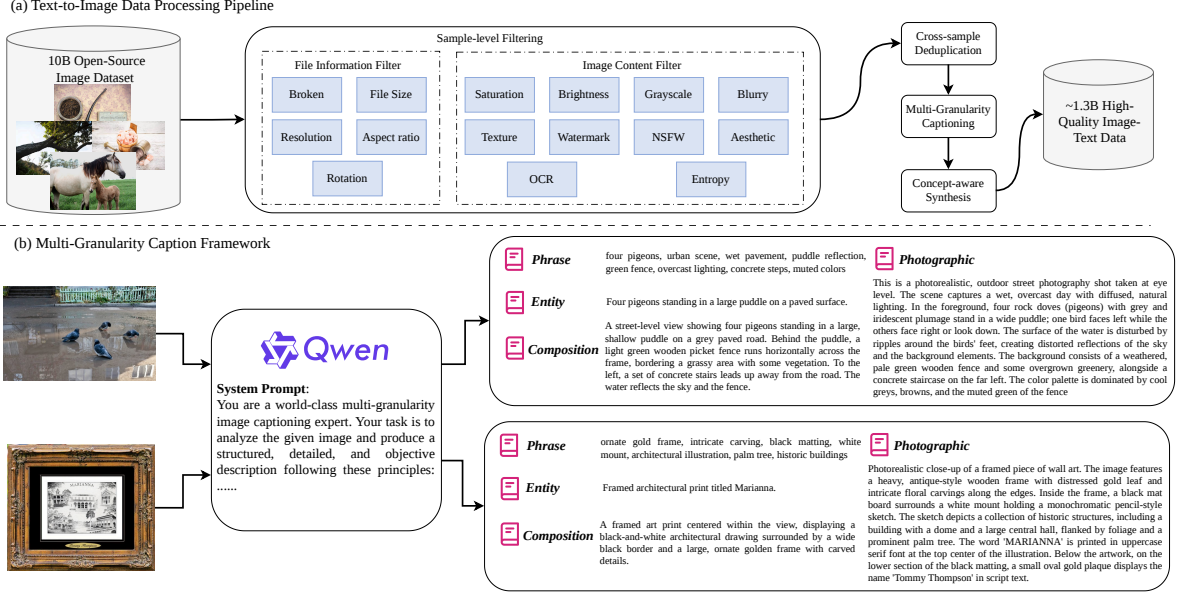


Figure 8 Text-to-image data processing and captioning pipeline. (a) Raw web image-text pairs are processed by sample-level filtering, cross-sample deduplication, multi-granularity captioning, and concept-aware synthesis, resulting in a high-quality image-text corpus. (b) The multi-granularity captioning framework uses Qwen3-VL to produce phrase-level, entity-level, composition-level, and photographic captions, providing training prompts with different levels of semantic detail and visual specificity.

blocks are executed many times during each forward and backward pass, local operator fusion translates into a substantial stack-level throughput gain.

As reported in Table 4, replacing FLUX.2-VAE with *Mage-VAE* already reduces per-step time from 1.9259 s to 1.3634 s, giving a $1.41\times$ speedup and showing the importance of a lightweight tokenizer. Fusing *Mage-VAE* and Qwen3-VL brings additional but moderate gains, while fusing the repeated NR-MMDiT blocks provides the largest improvement because the 4B diffusion backbone dominates the training step. Overall, the full system increases MFU from 33.20% to 77.26%, reduces peak per-GPU memory from 175.45 GB to 141.44 GB, and improves per-step training speed by $2.49\times$. These results show that efficient native-resolution training requires not only an efficient tokenizer and architecture, but also stack-level kernel optimization that removes memory-bound overhead from the repeated blocks.

Finding 3: Efficient native-resolution generation requires co-designing the model and the training system: lightweight tokenization reduces the arithmetic cost, while stack-level kernel fusion removes memory-bound overhead that would otherwise dominate repeated VAE, text-encoder, and MMDiT blocks.

4 Data Collection and Curation

The *Mage-Flow* stack relies on two complementary data pipelines: one for text-to-image generation and one for instruction-based image editing. The generation pipeline curates large-scale image-text pairs into high-quality and balanced prompt-image supervision, while the editing pipeline constructs and filters source-image, instruction, and target-image triples. Both pipelines are designed to improve visual quality, safety, semantic alignment, diversity, and coverage of capability-specific skills that are under-represented in raw web data.

Table 5 Sample-level filtering thresholds across pre-training and SFT stages. Thresholds are progressively tightened from the 256² stage to the 1024² and SFT stages. Early stages retain broad visual coverage, while later stages emphasize higher resolution, stronger aesthetics, lower watermark probability, and cleaner image content.

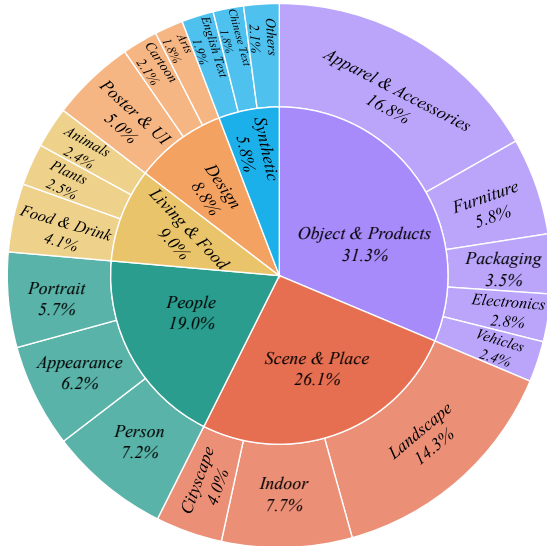
	256 ²	512 ²	1024 ²	SFT
Pixel count $h \times w$	$\geq 256^2$	$\geq 512^2$	$\geq 1024^2$	$\geq 1024^2$
$\min(h, w)$	≥ 128	≥ 256	≥ 512	≥ 512
Aspect ratio	[0.1, 10.0]	[0.1, 10.0]	[0.1, 10.0]	[0.1, 10.0]
File size	≥ 1 KB	≥ 1 KB	≥ 1 KB	≥ 1 KB
NSFW score	≤ 0.1	≤ 0.1	≤ 0.1	≤ 0.1
Watermark score	< 0.5	< 0.3	< 0.1	< 0.05
Aesthetic-V2.5 score	≥ 4.5	≥ 5.5	≥ 6.0	≥ 6.5
OCR text-area ratio	≤ 0.3	≤ 0.3	≤ 0.3	≤ 0.3
OCR num. regions	≤ 5	≤ 5	≤ 5	≤ 5

4.1 Text-to-Image Data Collection and Filtering

The *Mage-Flow* generation corpus is built from roughly 10B raw image-text pairs collected from large-scale open-source datasets. As shown in Fig. 8, the curation pipeline contains four major stages: sample-level filtering, cross-sample deduplication, multi-granularity captioning, and concept-aware synthesis. The sample-level filters remove corrupted, low-quality, unsafe, or visually unsuitable images; deduplication suppresses near-duplicate visual modes; multi-granularity captioning standardizes textual supervision; and concept-aware synthesis supplements long-tail images. After curation, approximately 1.3B high-quality image-text pairs are retained, from which the stage-wise pre-training subsets are sampled.

Sample-level filtering. We first filter each image-caption pair independently using the file-information and image-content filters shown in Fig. 8(a). File-information filters remove corrupted or near-empty files, images with insufficient resolution or pixel count, extreme aspect ratios, and samples with incorrect orientation metadata. Image-content filters then score decoded images for visual quality and safety: brightness and saturation filters remove over-exposed, under-exposed, or unnaturally saturated images; grayscale and blurry filters remove near-monochrome or low-sharpness samples; entropy and texture filters suppress near-empty images and texture-like non-semantic patterns; and watermark, aesthetic, OCR, and NSFW filters remove watermarked, low-quality, document-like, or unsafe images. Many filters are threshold-based rather than binary. As shown in Table 5, the thresholds are progressively tightened across the 256², 512², 1024², and SFT stages so that early training preserves broad coverage, while later stages focus on cleaner and higher-quality data.

Cross-sample deduplication. Web-scale data is highly redundant both within individual sources and across datasets. After sample-level filtering, we deduplicate images at two levels. Each retained image is encoded into an SSCD copy-detection descriptor [79], which is robust to re-encoding, cropping, resizing, and light edits. All descriptors are indexed with FAISS [80] for efficient nearest-neighbor search. Within each dataset, image pairs with cosine similarity above 0.9 are grouped as duplicates, and only the highest-quality representative is retained. Very large clusters are capped to suppress repeated web templates such as stock photos, banners, and product layouts. Across datasets, we maintain a persistent descriptor index of already accepted images and reject new samples that reproduce existing ones above the same similarity threshold. We also match against a held-out benchmark index to reduce evaluation contamination. Together, these steps improve diversity and prevent duplicated sources from dominating the training distribution.



(a) Generation pre-training data.



(b) Editing pre-training data.

Figure 9 Data composition for generation and editing pre-training. (a) Concept distribution of the curated generation corpus after merging filtered web data with targeted supplemental data. (b) Final composition of the editing pre-training mixture after adjusting sampling rates across constituent editing datasets.

Multi-granularity captioning. To standardize textual supervision, we caption retained images with Qwen3-VL-32B-Instruct [13]. Following [22], each image is assigned captions at multiple levels: a phrase-level description for concept statistics, an entity-level caption describing major objects and attributes, a composition-level caption describing spatial layout and relations, and a photographic caption describing style, lighting, viewpoint, atmosphere, and fine visual details. During training, the model samples from these descriptive caption channels so that it learns to follow prompts with different lengths and specificity. For text-rich images, the captioner is explicitly prompted to recognize visible text and convert it into rendering instructions.

Concept-aware synthesis and balancing. Although large-scale web data provides broad visual coverage, it remains sparse in several capability-critical areas, including long-text rendering, rare objects, uncommon attributes, structured layouts, and under-represented styles. Thus, we construct targeted supplemental data for these long-tail concepts, including synthetic text-rendering samples with diverse fonts, layouts, languages, colors, and backgrounds, as well as additional image-text pairs covering rare concepts and compositional cases. All supplemental samples are passed through the same filtering and quality-control pipeline before being merged into the curated pre-training corpus. Then, we use phrase-level captions to estimate the concept distribution of the combined corpus, as shown in Fig. 9(a). The resulting distribution remains long-tailed: *Object & Products* and *Scene & Place* form the two largest coarse domains, while *Design* and *Synthetic* provide important coverage for layout, product-style, poster-style, and text-rendering capabilities. During training, we apply concept-aware sampling to reduce the dominance of frequent objects, natural scenes, and common photorealistic styles. Across training stages, we progressively tighten filtering thresholds and strengthen reweighting, moving from broad visual-prior learning in early stages to cleaner and more capability-focused learning in later stages.

4.2 Image-Editing Data Collection and Filtering

The *Mage-Flow-Edit* corpus consists of (source image, edit instruction, target image) triples. As shown in Fig. 10, the raw pool contains roughly 90M triples from two complementary sources: about 50M triples aggregated from open-source instruction-based editing datasets and about

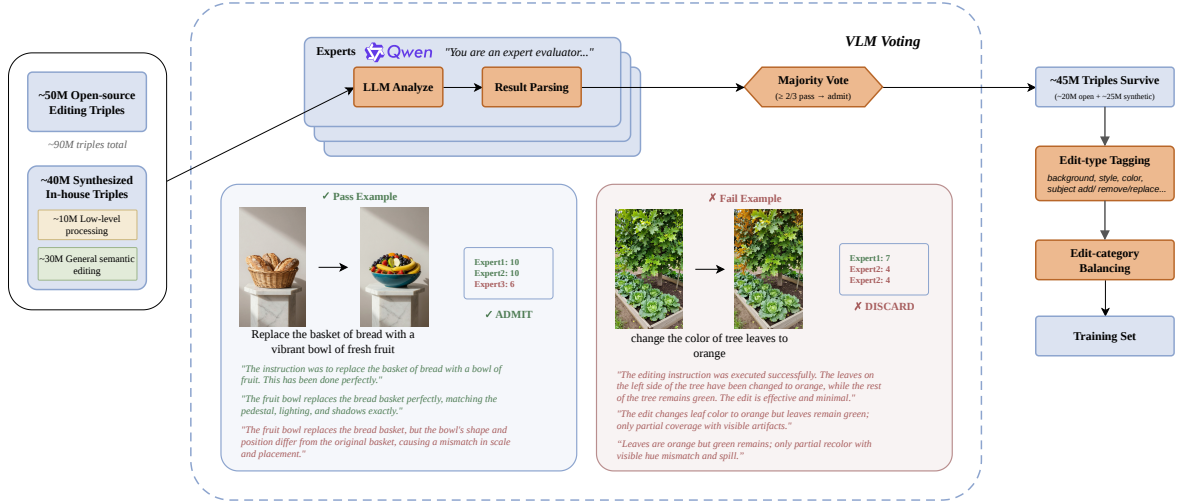


Figure 10 Editing data filtering pipeline. Raw editing triples are collected from open-source editing datasets and in-house synthesis. Each triple is evaluated by three Qwen3.5-9B experts with different system prompts and partially overlapping rubrics. Each expert analyzes the source image, target image, and edit instruction, and its reasoning output is parsed into a pass/fail decision. A triple is retained only if it receives a majority vote from the three experts. The surviving 45M triples are then assigned to a manually defined edit-type taxonomy and reweighted across categories to form the final editing training set.

40M triples synthesized in-house. The synthesized split contains approximately 10M low-level image-processing triples and 30M general semantic-editing triples. Then, we apply a VLM-based voting filter, followed by edit-type tagging and category balancing, to construct the final editing training set.

Editing data synthesis. The in-house semantic-editing subset is constructed to cover a broad range of user-facing editing skills, including background replacement, color and material modification, tone and style transfer, subject addition, removal, and replacement, object-count change, motion change, viewpoint change, text editing, portrait retouching, old-photo restoration, and global adjustment. Each edit type is generated with a type-specific pipeline that combines off-the-shelf generation, inpainting, segmentation, image processing, and template-based instruction generation. This provides explicit source–target pairings and clear editing instructions, while supplementing edit categories that are sparse or unreliable in open-source datasets.

VLM-based dataset filtering. Raw editing triples often contain instruction-inconsistent or visually degraded samples. For example, the target image may fail to apply the requested edit, modify irrelevant regions, change the source identity or layout unnecessarily, or introduce visible artifacts. Therefore, we evaluate each triple with three independent Qwen3.5-9B experts [81]. Each expert is configured with a different system prompt and evaluates a partially overlapping set of criteria, so that the three judgments are complementary rather than identical. Given the source image, target image, and edit instruction, each expert checks whether the requested edit is correctly executed, whether unrelated regions are preserved, and whether the edited result remains visually plausible. The reasoning output is parsed into a criterion-level assessment and then converted into a pass/fail decision using a predefined threshold. We retain a triple only when at least two of the three experts vote to pass. After filtering, roughly 20M open-source triples and 25M synthesized triples survive, forming a 45M-triple retained pool for editing training.

Edit-type tagging and balancing. We manually define a taxonomy of 19 edit categories that covers all retained editing data. Instead of tagging each sample independently with a VLM, we

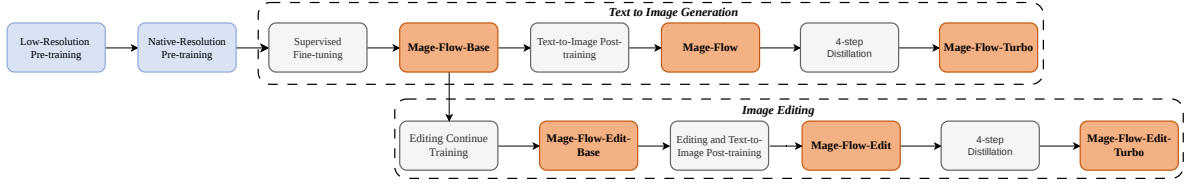


Figure 11 Overview of the *Mage-Flow* training pipeline. The shared backbone is first trained with progressive text-to-image pre-training, summarized as low-resolution pre-training followed by native-resolution pre-training, and then supervised fine-tuning produces the base generation checkpoint *Mage-Flow-Base*. From this checkpoint, the text-to-image branch applies Diffusion-NFT post-training to obtain *Mage-Flow* and 4-step distillation to obtain *Mage-Flow-Turbo*. The editing branch forks from *Mage-Flow-Base*: continued source-conditioned editing training produces *Mage-Flow-Edit-Base*, mixed editing-and-generation post-training produces *Mage-Flow-Edit*, and 4-step distillation produces *Mage-Flow-Edit-Turbo*.

determine the annotation unit from the organization of each data source. If a dataset contains a consistent editing operation, the whole dataset is treated as one unit. If a dataset is organized into semantically consistent sub-datasets, each sub-dataset is treated as a separate unit. If an explicit edit-type field is available, samples sharing the same field value form one unit. We then manually map every unit to one of the 19 edit categories, ensuring that each retained sample is assigned to the unified taxonomy. During balancing, we adjust the sampling rate of each constituent dataset and edit category so that the aggregate training mixture maintains broad and well-proportioned coverage across the taxonomy. This prevents frequent operations from dominating the gradient signal while keeping rare but important edit types sufficiently represented. The resulting editing pre-training composition is shown in Fig. 9(b).

5 Training

We train the *Mage-Flow* model family with a unified recipe shared by text-to-image generation and instruction-based editing. Both tasks operate in the same *Mage-VAE* latent space and use the same 4B Native-Resolution MMDiT backbone with a rectified-flow objective. The main differences are the conditioning format and data mixture: text-to-image generation is conditioned on prompts, while editing is conditioned on editing instructions and source image(s). As shown in Fig. 11, we organize the recipe by training stage: progressive pre-training and supervised fine-tuning produce the base checkpoint, Diffusion-NFT post-training produces the aligned models, and few-step distillation produces the Turbo variants.

5.1 Pre-training and Supervised Fine-tuning

Text-to-image generation. The generation model is trained with a progressive pre-training and SFT curriculum. As shown in Table 5, pre-training contains three stages. First, we train on 1.2B filtered and recaptioned image-text pairs at a fixed 256×256 resolution to learn broad visual-language alignment at low computational cost. Second, we move to a higher-quality 600M subset under a 512-pixel native-aspect-ratio regime, where each sample keeps an approximately 512×512 pixel budget while preserving its original aspect ratio. Third, we train on an even cleaner 300M subset under the 1024-pixel native-aspect-ratio regime to improve fine details, layout fidelity, text rendering, and aesthetic quality. All stages use the same rectified-flow objective in the *Mage-VAE* latent space, while progressively increasing resolution, data quality, and concept-aware reweighting strength. We then perform supervised fine-tuning on a curated 150M high-quality subset at the 1024-pixel native-aspect-ratio regime, using stricter aesthetic, alignment, watermark, OCR, and duplication filters together with increased weights for capability-targeted data. This produces the text-to-image base checkpoint, *Mage-Flow-Base*.

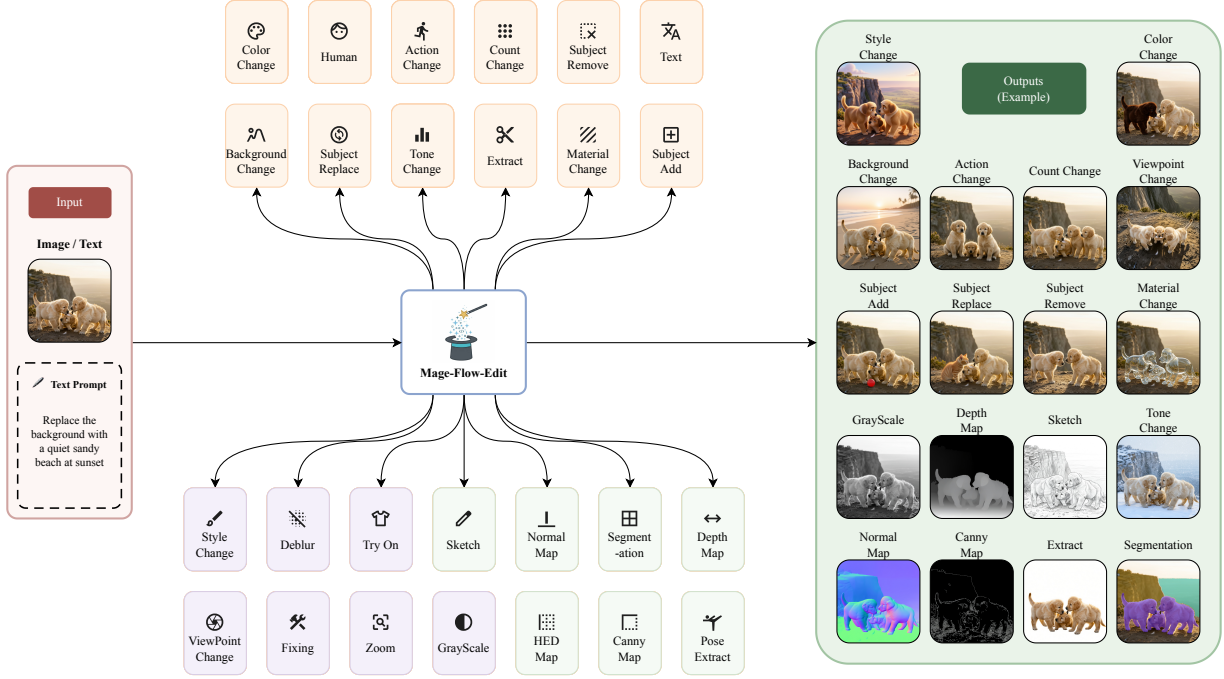


Figure 12 Unified image-editing capabilities and representative outputs. *Mage-Flow-Edit* supports semantic content editing, appearance transformation, image restoration, and structure-aware outputs within a unified image-and-text-conditioned model. The displayed prompt illustrates background replacement, while the remaining outputs correspond to task-specific instructions or output requests.

Instruction-based editing. The editing model is initialized from *Mage-Flow-Base* and trained with a two-stage editing adaptation recipe. In the first stage, we train on a balanced mixture of 35M editing triples and 35M generation pairs, adapting the model to source-conditioned instruction following while preserving the visual prior and open-ended synthesis capability inherited from *Mage-Flow-Base*. In the second stage, we further train on a higher-quality mixture of 20M editing triples and 10M generation pairs to improve editing fidelity and robustness. In both stages, multi-image editing examples account for no more than 0.5% of the editing data, while the majority are single-image edits. This two-stage adaptation produces the editing base checkpoint, ***Mage-Flow-Edit-Base***. Fig. 12 illustrates the breadth of editing capabilities learned through this adaptation recipe. Given the same source image, *Mage-Flow-Edit* can follow diverse text instructions to produce semantic edits, appearance transformations, restoration results, and structure-aware outputs, demonstrating that the editing checkpoint supports a unified one-to-many editing interface rather than a collection of task-specific models.

5.2 Diffusion-NFT Post-training

Starting from the Base checkpoints, we apply Diffusion-NFT [14] as the post-training stage for both generation and editing. Diffusion-NFT operates directly on the forward process of flow-matching generators and performs negative-aware fine-tuning with online samples, requiring no likelihood estimation and remaining compatible with arbitrary black-box samplers. We use the same Diffusion-NFT objective for *Mage-Flow* and *Mage-Flow-Edit*, while adapting the condition format, data mixture, and reward models to each task.

Text-to-image generation. For *Mage-Flow*, we curate a compact RL prompt pool of approximately 20K prompts spanning three capability groups: about 10K text-rendering prompts, 4K aesthetic-quality prompts, and 6K semantic-understanding prompts. Each prompt is assigned a capability tag that determines both its evaluation rubric and its reward evaluator. Text-rendering prompts are

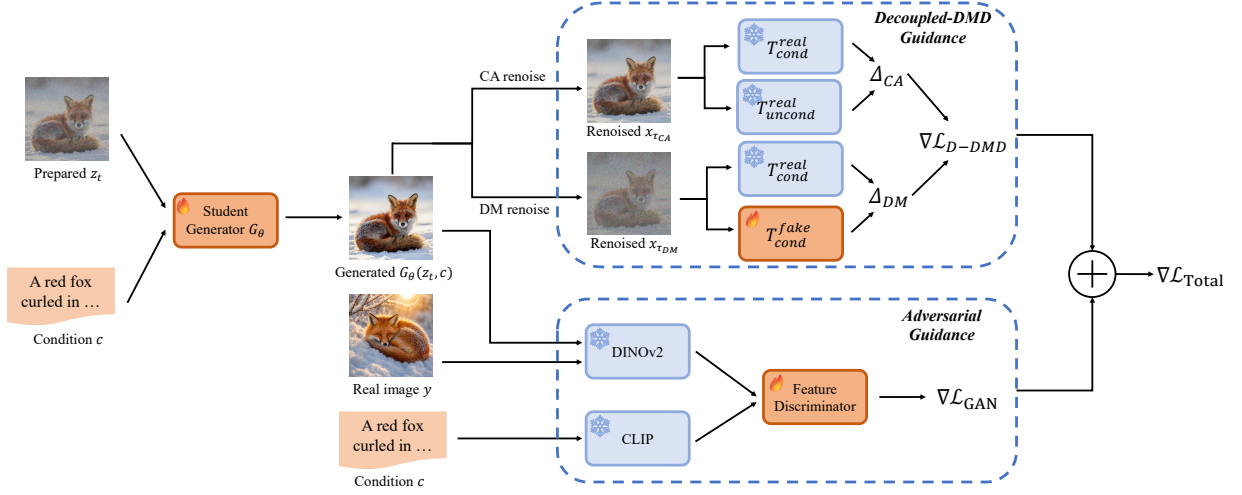


Figure 13 Few-step distillation framework. The student maps a prepared noised latent z_t and condition c to a predicted clean sample. Decoupled DMD re-noises the student output at two independent noise levels: the classifier-free augmentation branch queries the frozen teacher with and without the condition to form Δ_{CA} , while the distribution-matching branch contrasts the teacher with a trainable fake-score model to form Δ_{DM} . In parallel, generated and real images are encoded by frozen vision foundation models such as DINOv2 and CLIP, and a lightweight feature discriminator provides the adversarial gradient $\nabla \mathcal{L}_{GAN}$. The Decoupled DMD and adversarial gradients are combined to train the 4-step Turbo student.

scored by PaddleOCR-VL-1.5 [82], which reads the generated image and compares the recognized strings against the target text to measure OCR fidelity, scene text, typography, and text-object interactions. Aesthetic-quality and semantic-understanding prompts are scored by Qwen3.5-72B [81] with two different rubric system prompts. The aesthetic rubric evaluates photographic quality, lighting, composition, color harmony, texture, and artistic style, while the semantic rubric decomposes prompts into checkable questions covering compositional reasoning, object attributes, spatial relations, counting, actions, and multi-concept scenes. Details of the reward evaluators, rubrics, and OCR scoring formula are provided in Appendix C.

Optimization is performed with online rollout groups. We use a global batch size of 48, and each optimizer step contains prompts from all three capability groups. Although a step mixes heterogeneous prompts, each prompt is routed to exactly one evaluator according to its capability tag; its reward is computed only by that evaluator and is never summed or averaged across capabilities. For each prompt, we sample a group of candidates from the current generator using 10 denoising steps with guidance scale 5.0, and score the candidates with the assigned evaluator. Since different evaluators have different score distributions, advantages are normalized separately within each reward type: candidates scored by one evaluator are normalized only against candidates from the same evaluator. This per-type normalization produces the optimality probabilities $r_i^{(s)} \in [0, 1]$ used by the Diffusion-NFT loss.

We use a two-stage schedule that shifts the capability mixture over time while keeping all three groups present throughout. Stage 1 runs for 140 optimizer steps on a balanced mixture $\mathcal{P}_{\text{aes}}:\mathcal{P}_{\text{text}}:\mathcal{P}_{\text{sem}} = 1:1:1$, with text-rendering prompts that emphasize stable OCR cases including single words, short phrases, simple signs, logos, and short scene text, so that the model sharpens character- and word-level rendering while jointly improving visual quality and semantic alignment. Stage 2 then continues from the best Stage-1 checkpoint for a further 60 steps, up-weighting the harder text-rendering prompts to $\mathcal{P}_{\text{aes}}:\mathcal{P}_{\text{text}}:\mathcal{P}_{\text{sem}} = 2:4:1$ and targeting complete sentences, dense captions, multi-line layouts, punctuation-rich text, and text embedded in complex scenes; retaining aesthetic and semantic prompts throughout prevents over-specialization to OCR and preserves the general generation quality built up in the first stage. Together, the two stages produce the aligned generation checkpoint, **Mage-Flow**.

Table 6 Effect of adversarial perceptual guidance in 4-step distillation. Adversarial guidance consistently improves generation and benefits text editing under the highly compressed four-step trajectory, while changes on general editing benchmarks are mixed. Higher scores indicate better performance.

(a) Text-to-image generation.

Model	GenEval	DPG-Bench	TIIF-Short	TIIF-Long	CVTG-2K	OneIG-EN	OneIG-CN	LongText-EN	LongText-CN
<i>Mage-Flow-Turbo</i>	0.88	85.48	83.58	84.16	0.873	0.523	0.491	0.911	0.801
<i>Mage-Flow-Turbo</i> w/o Adv.	0.89	85.37	80.99	82.03	0.847	0.518	0.486	0.882	0.783

(b) Instruction-based editing.

Model	ImgEdit	GEEdit-EN	GEEdit-CN	TextEdit-Syn	TextEdit-Real
<i>Mage-Flow-Edit-Turbo</i>	4.38	8.271	8.264	12.77	15.41
<i>Mage-Flow-Edit-Turbo</i> w/o Adv.	4.29	8.003	8.025	11.64	14.77

Instruction-based editing. For *Mage-Flow-Edit*, we apply Diffusion-NFT on top of *Mage-Flow-Edit-Base* using a joint stream of editing and generation data. This sharpens source-conditioned instruction following while preserving the generation prior. The two streams are interleaved at a 4:1 ratio of four editing updates per generation update, and share the same Diffusion-NFT objective, global batch size, and optimizer as the text-to-image run.

The generation stream reuses the text-to-image RL prompt pool and its capability-routed rewards. This stream helps preserve the text-rendering, compositional, and open-ended generation ability of *Mage-Flow-Edit* during editing post-training. The editing stream draws instructions from a curated editing RL pool of approximately 30K prompts, uniformly sampled across edit tasks in the editing pre-training corpus, including object replacement, object removal, background replacement, spatial editing, style transfer, and general instruction-based editing. This prevents any single edit type from dominating the update. Each editing rollout is scored by RationalRewards [83], a reasoning reward model that first produces a multi-dimensional critique and then emits a scalar preference. The critique evaluates four aspects: instruction adherence, preservation of source content outside the edited region, physical and perceptual plausibility, and the quality of rendered text. The four aspect scores are averaged into a normalized reward in $[0, 1]$. The joint post-training runs for 300 optimizer steps and yields the aligned editing checkpoint, ***Mage-Flow-Edit***.

Diffusion-NFT objective. For both tasks, each condition c is used to sample online candidates from the current generator. The candidates are scored by the task-specific rubric and reward evaluator, and the raw rewards are normalized within each prompt group to obtain an optimality probability $r_i^{(s)} \in [0, 1]$, where larger values indicate higher-quality samples. Diffusion-NFT then optimizes a reward-weighted flow-matching objective with implicit positive and negative policies:

$$\mathcal{L}_{\text{NFT}}^{(s)}(\theta) = \mathbb{E}_{c, x_{0,i} \sim \pi_{\text{old}}(\cdot|c), t} \left[\underbrace{r_i^{(s)} \left\| v_{\theta}^{+}(x_{i,t}, t, c) - v_{i,t} \right\|_2^2}_{\text{positive match}} + \underbrace{(1 - r_i^{(s)}) \left\| v_{\theta}^{-}(x_{i,t}, t, c) - v_{i,t} \right\|_2^2}_{\text{negative match}} \right],$$

where $v_{i,t}$ denotes the target velocity of the forward process, and v_{θ}^{+} and v_{θ}^{-} denote the implicit positive and negative policies defined by Diffusion-NFT. The positive branch pulls the model toward high-reward samples, while the negative branch suppresses low-reward samples. For *Mage-Flow*, c is a text prompt and the reward comes from the rubric-selected generation evaluator. For *Mage-Flow-Edit*, c can be either a generation prompt or an editing condition; the former uses the same generation reward evaluators, while the latter uses RationalRewards on the source image, instruction, and edited result.

Table 7 Effect of mixing generation data during editing training. We compare editing models trained with and without generation data. The effect is modest for the full-step editor; for the 4-step Turbo editor, generation data improves ImgEdit, while GEdit changes are mixed.

Model	Generation Data	ImgEdit	GEEdit-EN	GEEdit-CN	TextEdit-Syn	TextEdit-Real
<i>Mage-Flow-Edit</i>	✓	4.34	8.127	8.123	14.14	16.26
<i>Mage-Flow-Edit</i>	-	4.34	7.991	8.062	14.29	15.95
<i>Mage-Flow-Edit-Turbo</i>	✓	4.38	8.271	8.264	12.77	15.41
<i>Mage-Flow-Edit-Turbo</i>	-	4.20	7.984	8.050	12.66	15.20

5.3 Few-step Distillation

To reduce inference cost, we distill the RL-aligned checkpoints into 4-step Turbo models: ***Mage-Flow-Turbo*** for text-to-image generation and ***Mage-Flow-Edit-Turbo*** for instruction-based editing. In both tasks, the student is initialized from the corresponding frozen teacher checkpoint. As illustrated in Fig. 13, our distillation framework combines Decoupled DMD [15] with adversarial perceptual guidance inspired by SenseFlow [16].

The distillation objective builds on Decoupled DMD, which separates two signals that are coupled in standard distribution-matching distillation: a classifier-free-guidance augmentation (CA) term that steers the student toward the guided teacher direction, and a distribution-matching (DM) regularizer defined through a trainable fake-score network. Instead of sharing a single timestep distribution, Decoupled DMD assigns independent noise schedules to the two terms, improving stability in the few-step regime.

However, compressing the sampling trajectory to only four denoising steps makes perceptual quality more fragile. As shown in Table 6, adversarial perceptual guidance provides clear gains for generation and text editing, while its effect on general editing benchmarks is more mixed: it improves ImgEdit slightly, but does not uniformly improve GEdit. We therefore add an adversarial perceptual term based on a feature discriminator operating in frozen DINOv2 and CLIP feature spaces [68, 69]. This term acts as a lightweight regularizer on top of Decoupled DMD, with the generator updated once every five discriminator updates. Overall, the student is optimized by the total gradient:

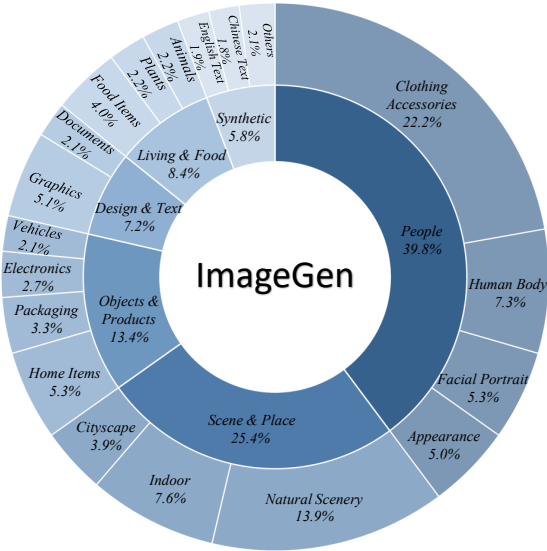
$$\nabla \mathcal{L}_{\text{Total}} = \underbrace{\Delta_{\text{CA}} + \Delta_{\text{DM}}}_{\nabla \mathcal{L}_{\text{D-DMD}}} + \lambda_{\text{GAN}} \nabla \mathcal{L}_{\text{GAN}}, \quad (1)$$

$$\Delta_{\text{CA}} = (w - 1)(T_{\text{cond}}^{\text{real}}(x_{\tau_{\text{ca}}}) - T_{\text{uncond}}^{\text{real}}(x_{\tau_{\text{ca}}})), \quad \Delta_{\text{DM}} = T_{\text{cond}}^{\text{real}}(x_{\tau_{\text{dm}}}) - T_{\text{cond}}^{\text{fake}}(x_{\tau_{\text{dm}}}).$$

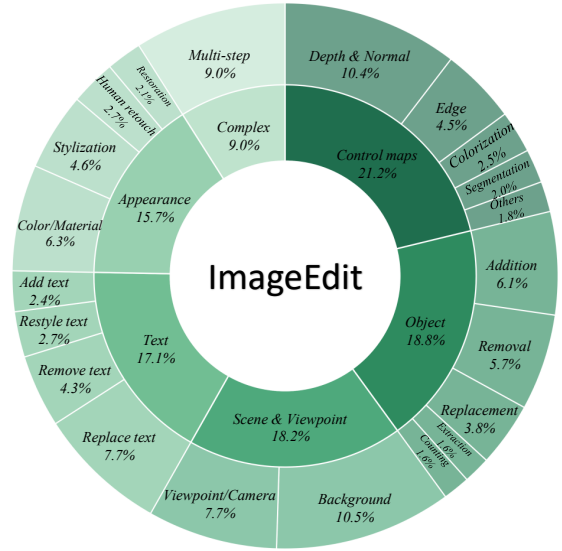
Here, Δ_{CA} is the CFG-augmentation term formed from the teacher’s conditional and unconditional predictions $T_{\text{cond}}^{\text{real}}$ and $T_{\text{uncond}}^{\text{real}}$, while Δ_{DM} is the distribution-matching term against the fake-score prediction $T_{\text{cond}}^{\text{fake}}$. The two terms are evaluated at independent noise levels τ_{ca} and τ_{dm} . We use guidance scale $w = 7.5$ and adversarial weight $\lambda_{\text{GAN}} = 0.13$.

Finding 4: Adversarial perceptual guidance is most beneficial for preserving generation and text-editing quality under a highly compressed four-step trajectory; its gains on general editing benchmarks are benchmark-dependent rather than uniform.

Generation distillation. For *Mage-Flow-Turbo*, we distill the student from the frozen *Mage-Flow* teacher on roughly 200K curated high-quality prompt–image pairs. The set spans six broad content categories: People, Scene & Place, Objects & Products, Living & Food, Design & Text, and Synthetic, whose distribution is shown in Fig. 14(a). With the combined Decoupled-DMD



(a) Generation distillation set.



(b) Editing distillation set.

Figure 14 Composition of the distillation sets. The inner rings show the main categories, and the outer rings show the corresponding sub-categories. (a) The generation distillation set covers diverse content categories such as people, scenes, products, design, food, and synthetic data. (b) The editing distillation set covers various edit-type categories such as object editing, scene and viewpoint changes, text editing, appearance modification, control-map editing, and complex edits.

and adversarial perceptual objective, *Mage-Flow-Turbo* preserves the visual quality and prompt-following behavior of the multi-step teacher while reducing sampling to a four-step rectified-flow trajectory.

Editing distillation. For *Mage-Flow-Edit-Turbo*, we use the same distillation framework with two task-specific adjustments. First, the real branch of the feature discriminator uses target images from editing triples rather than images from a separate real-image corpus, aligning the adversarial signal with the edited-image distribution. Second, the student is trained on a 3:1 mixture of editing and generation data. The editing samples teach the shortened model to follow edit instructions, while the retained generation samples preserve open-ended generation ability and improve robustness on edits that require large visual changes. In total, the editing student is distilled on roughly 250K editing samples spanning six categories: Object, Scene & Viewpoint, Text, Appearance, Control maps, and Complex. The category distribution before mixing with generation samples is shown in Fig. 14(b).

To further verify the role of generation data during editing training, we compare models trained with and without generation data in Table 7. The full-step editor changes only marginally. For the Turbo variant, adding generation data improves ImgEdit from 4.20 to 4.38, whereas the two GEdit splits remain comparable and do not move uniformly. These results suggest that generation data primarily helps broad category-wise editing robustness under four-step compression, rather than uniformly improving every editing metric.

Overall, this unified recipe allows both *Mage-Flow-Turbo* and *Mage-Flow-Edit-Turbo* to inherit the quality and alignment of their multi-step teachers while reducing inference to only 4 steps.

Finding 5: Mixing generation data during editing training has its clearest effect on broad editing robustness for the four-step model, as reflected by ImgEdit, while GEdit remains comparable and does not improve uniformly.

Table 8 Summary of text-to-image generation results across eight benchmarks. Steps is the number of denoising steps at evaluation. Models are grouped into closed-source and open-source. The best open-source score in each column is in **bold**. The second-best results are underlined.

Type	Model	#Params	Steps	GenEval	DPG	TIIF-Short	TIIF-Long	CVTG-2K	OneIG-EN	OneIG-CN	LongText-EN	LongText-CN
Closed-Source	Seedream 3.0 [90]	—	—	0.84	88.27	86.02	84.31	0.592	0.530	0.528	0.896	0.878
	Seedream 4.0 [3]	—	—	0.84	88.63	—	—	0.892	0.573	0.554	0.936	0.946
	GPT-Image-1 [1]	—	—	0.84	85.15	89.15	88.29	0.857	0.533	0.474	0.956	0.619
	Nano-Banana-Pro [2]	—	—	0.83	87.16	—	—	0.779	0.580	0.570	0.981	0.949
	Kolors 2.0 [91]	—	—	—	—	—	—	—	0.434	0.426	0.258	0.329
Open-Source Unified	BAGEL [44]	14B	50	0.82	85.07	71.50	71.70	0.356	0.361	0.370	0.373	0.310
	Janus-Pro-7B [42]	7B	—	0.80	84.19	66.50	65.02	—	—	—	0.019	0.006
	Emu3-Gen [92]	8B	—	0.54	80.60	—	—	—	—	—	—	—
	OmniGen2 [36]	4B	50	0.80	83.57	—	—	—	0.475	—	0.561	0.059
	InternVL-U [93]	4B	20	0.85	85.18	74.90	73.90	0.623	0.500	0.500	0.738	0.860
	UniWorld-V1 [37]	19B	30	0.80	81.38	—	—	—	—	—	—	—
	Ovis-U1 [94]	3.6B	50	<u>0.89</u>	83.72	66.70	68.20	0.093	0.340	0.340	0.030	0.051
	SD3.5-Large [20]	8B	28	0.71	84.96	76.10	68.71	0.655	0.462	0.248	0.473	0.009
Open-Source Specialist	HiDream-I1-Full [95]	17B	50	0.83	85.89	79.33	71.88	0.740	0.477	0.337	0.543	0.024
	FLUX.1-dev [96]	12B	50	0.66	83.84	71.09	71.78	0.496	0.434	0.245	0.607	0.005
	FLUX.1-Krea-dev [97]	12B	50	0.72	86.59	80.36	81.67	0.444	0.443	0.271	0.693	0.002
	FLUX.2-dev [6]	32B	50	0.87	87.57	88.82	<u>88.10</u>	0.893	0.551	0.516	0.963	0.757
	FLUX.2-Klein-Base-4B [6]	4B	50	0.78	83.02	79.94	80.01	0.656	0.485	0.366	0.554	0.071
	FLUX.2-Klein-Base-9B [6]	9B	50	0.83	85.29	81.47	84.52	0.655	0.544	0.400	0.872	0.227
	FLUX.2-Klein-4B [6]	4B	4	0.83	85.53	78.91	79.04	0.628	0.500	0.364	0.649	0.068
	FLUX.2-Klein-9B [6]	9B	4	0.86	86.20	85.22	84.13	0.424	0.538	0.406	0.872	0.226
	Qwen-Image [5]	20B	50	0.87	88.32	<u>86.14</u>	86.83	0.829	0.539	0.548	0.943	0.946
	JoyAI-Image [98]	16B	50	—	88.05	—	—	0.874	0.542	0.521	0.963	0.963
	HunyuanImage-3.0 [7]	80B	50	0.72	86.10	—	—	0.765	—	—	—	—
	LongCat-Image [22]	6B	50	0.87	86.80	80.93	81.30	0.866	0.516	0.518	0.885	<u>0.956</u>
	Z-Image-Base [4]	6B	50	0.84	88.14	80.20	83.04	0.867	<u>0.546</u>	<u>0.535</u>	0.935	0.936
	Z-Image-Turbo [4]	6B	8	0.82	84.86	77.73	80.05	0.859	0.528	0.507	0.917	0.926
	Lens-Base [99]	3.8B	50	0.70	85.51	80.33	83.49	0.635	0.527	0.500	0.830	0.741
	Lens-RL [99]	3.8B	20	0.85	<u>88.19</u>	84.23	84.92	0.843	0.526	0.497	0.901	0.817
	Lens-Turbo [99]	3.8B	4	0.83	87.13	82.20	81.81	0.882	0.520	0.489	0.909	0.860
	<i>Mage-Flow-Base</i>	4B	30	0.79	86.26	82.50	83.19	0.851	0.542	0.509	0.904	0.792
	<i>Mage-Flow</i>	4B	20	0.90	86.49	82.19	84.70	<u>0.887</u>	0.536	0.505	<u>0.944</u>	0.823
	<i>Mage-Flow-Turbo</i>	4B	4	0.88	85.48	83.58	84.16	0.873	0.523	0.491	0.911	0.801

6 Experiments

6.1 Experimental Setup

Text-to-image generation. We evaluate *Mage-Flow* on eight widely used text-to-image benchmarks covering prompt following, fine-grained generation, and text rendering. GenEval [84], DPG-Bench [85], and TIIF-Bench [86] measure prompt following and compositional understanding. OneIG [87] evaluates fine-grained generation on English and Chinese splits. CVTG-2K [88] focuses on multi-region text rendering, and LongText [89] evaluates long-form English and Chinese text rendering. All results are computed at native 1024^2 resolution using the official evaluation protocols. Unless otherwise specified, *Mage-Flow-Base* uses 30 denoising steps, *Mage-Flow* uses 20 steps, and *Mage-Flow-Turbo* uses 4 steps.

Instruction-based image editing. We evaluate *Mage-Flow-Edit* on three instruction-based image editing benchmarks. ImgEdit-Bench [103] covers nine representative editing skills, including Add, Adjust, Extract, Replace, Remove, Background, Style, Hybrid, and Action, and reports category-level performance on a 0–5 scale. GEdit-Bench [34] evaluates editing quality from three perspectives: Semantic Consistency, Perceptual Quality, and Overall Score, on both English and Chinese subsets, with each metric measured on a 0–10 scale. TextEdit-Bench [104] focuses on text-centric editing and evaluates five dimensions: Instruction Following, Text Accuracy, Visual Consistency, Layout Preservation, and Semantic Expectation. Each dimension is scored on a 0–5 scale, with the overall score computed by summing the five metric means (out of 25). Unless otherwise specified, *Mage-Flow-Edit-Base* and *Mage-Flow-Edit* are evaluated with 30 denoising steps, while *Mage-Flow-Edit-Turbo* uses only 4 denoising steps.

6.2 Quantitative Results

Table 9 Quantitative results on GenEval [84] and DPG-Bench [85].

Type	Model	#Params	GenEval							DPG-Bench					
			Single	Two	Counting	Colors	Position	Attr.Bind	Overall	Global	Entity	Attribute	Relation	Other	Overall
Closed-Source	Seedream 3.0 [90]	–	0.99	0.96	0.91	0.93	0.47	0.80	0.84	94.31	92.65	91.36	92.78	88.24	88.27
	Seedream 4.0 [3]	–	0.99	0.92	0.72	0.91	0.76	0.74	0.84	87.17	92.41	92.29	93.33	95.48	88.63
	GPT-Image-1 [1]	–	0.99	0.92	0.85	0.92	0.75	0.61	0.84	88.89	88.94	89.84	92.63	90.96	85.15
	Nano-Banana-Pro [2]	–	1.00	0.96	0.71	0.84	0.86	0.65	0.83	91.00	92.85	91.56	92.39	89.93	87.16
Open-Source Unified	BAGEL [44]	14B	0.99	0.94	0.81	0.88	0.64	0.63	0.82	88.94	90.37	91.29	90.82	88.67	85.07
	Janus-Pro-7B [42]	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80	86.90	88.90	89.40	89.32	89.48	84.19
	Emu3-Gen [92]	8B	0.98	0.71	0.34	0.81	0.17	0.21	0.54	85.21	86.68	86.84	90.22	83.15	80.60
	Show-o2 [100]	7B	1.00	0.87	0.58	0.92	0.52	0.62	0.76	–	–	–	–	–	–
	OmniGen2 [36]	4B	1.00	0.95	0.64	0.88	0.55	0.76	0.80	88.81	88.83	90.18	89.37	90.27	83.57
	InternVL-U [93]	4B	0.99	0.94	0.74	0.91	0.77	0.74	0.85	90.39	90.78	90.68	90.29	88.77	85.18
	UniWorld-V1 [37]	19B	0.99	0.93	0.79	0.89	0.49	0.70	0.80	83.64	88.39	88.44	89.27	87.22	81.38
	Ovis-U1 [94]	3.6B	0.98	0.98	0.90	0.92	0.79	0.75	0.89	82.37	90.08	88.68	93.35	85.20	83.72
Open-Source Specialist	SD3.5-Large [20]	8B	0.98	0.89	0.73	0.83	0.34	0.47	0.71	84.75	89.93	88.19	93.23	89.70	84.96
	HiDream-1I-Full [95]	17B	1.00	0.98	0.79	0.91	0.60	0.72	0.83	76.44	90.22	89.48	93.74	91.83	85.89
	FLUX.1-dev [96]	12B	0.98	0.81	0.74	0.79	0.22	0.45	0.66	74.35	90.00	88.96	90.87	88.33	83.84
	FLUX.1-Krea-dev [97]	12B	0.99	0.93	0.69	0.82	0.30	0.59	0.72	87.54	92.08	89.54	94.85	87.20	86.59
	FLUX.2-dev [6]	32B	1.00	0.99	0.79	0.93	0.73	0.78	0.87	92.20	91.36	93.28	93.52	89.72	87.57
	FLUX.2-Klein-Base-4B [6]	4B	0.99	0.87	0.81	0.90	0.53	0.59	0.78	91.61	88.72	90.30	91.23	88.48	83.02
	FLUX.2-Klein-Base-9B [6]	9B	1.00	0.90	0.87	0.93	0.65	0.63	0.83	89.76	90.34	90.66	93.31	86.58	85.29
	FLUX.2-Klein-4B [6]	4B	1.00	0.92	0.88	0.86	0.67	0.64	0.83	86.54	90.18	91.80	90.85	91.51	85.53
	FLUX.2-Klein-9B [6]	9B	0.99	0.96	0.89	0.91	0.70	0.69	0.86	88.94	91.95	89.53	92.74	92.18	86.20
	Qwen-Image [5]	20B	0.99	0.92	0.89	0.88	0.76	0.77	0.87	91.32	91.56	92.02	94.31	92.73	88.32
	JoyAI-Image [98]	16B	–	–	–	–	–	–	–	–	–	–	–	–	88.05
	HunyuanImage-3.0 [7]	80B	1.00	0.92	0.48	0.82	0.42	0.63	0.72	92.12	92.53	89.13	92.13	91.92	86.10
	LongCat-Image [22]	6B	0.99	0.98	0.86	0.86	0.75	0.73	0.87	89.10	92.54	92.00	93.28	87.80	86.80
	Z-Image-Base [4]	6B	1.00	0.94	0.78	0.93	0.62	0.77	0.84	93.39	91.22	93.16	92.22	91.52	88.14
	Z-Image-Turbo [4]	6B	1.00	0.95	0.77	0.89	0.65	0.68	0.82	91.29	89.59	90.14	92.16	88.68	84.86
	Lens-Base [99]	3.8B	0.99	0.79	0.58	0.83	0.52	0.46	0.70	88.68	91.24	91.93	91.89	89.65	85.51
	Lens-RL [99]	3.8B	1.00	0.95	0.83	0.89	0.73	0.72	0.85	91.07	92.94	91.93	93.39	92.78	88.19
	Lens-Turbo [99]	3.8B	0.99	0.94	0.84	0.86	0.66	0.68	0.83	92.19	93.44	90.34	95.11	90.89	87.13
	Mage-Flow-Base	4B	0.99	0.88	0.68	0.90	0.63	0.65	0.79	92.66	92.38	89.34	88.33	91.87	86.26
	Mage-Flow	4B	1.00	0.97	0.89	0.89	0.93	0.73	0.90	91.57	92.41	90.04	91.17	91.70	86.49
	Mage-Flow-Turbo	4B	1.00	0.97	0.80	0.88	0.90	0.73	0.88	80.88	89.42	91.87	91.34	91.19	85.48

Table 10 Quantitative Results on TIIF-Bench testmini [86].

Type	Model	Overall		Basic Following								Advanced Following								Designer					
				Avg		Attribute		Relation		Reasoning		Avg		Attribute + Relation		Attribute + Reasoning		Relation + Reasoning		Style		Text		Real World	
		short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long	short	long
Closed-Source	Seedream 3.0 [90]	86.02	84.31	87.07	84.93	90.50	90.00	89.85	85.94	80.86	78.86	79.16	80.60	79.76	81.82	77.23	78.85	75.64	78.64	100.00	93.33	97.17	87.78	83.21	83.58
	GPT-Image-1 [1]	89.15	88.29	90.75	89.66	91.33	87.08	84.57	84.57	96.32	97.32	88.55	88.35	87.07	89.44	87.22	83.96	85.59	83.21	90.00	93.33	89.83	86.83	89.73	93.46
	DALL-E 3 [101]	74.96	70.81	78.72	78.50	79.50	79.83	80.82	78.82	75.82	76.82	73.39	67.27	73.45	67.20	72.01	71.34	63.59	60.72	89.66	86.67	66.83	54.83	72.93	60.99
	MidJourney v7 [102]	68.74	65.69	77.41	76.00	77.58	79.83	82.07	76.82	72.57	69.32	64.66	60.53	67.20	62.70	81.22	71.59	60.72	64.59	83.33	80.00	24.83	20.83	68.83	63.61
Open-Source Unified	BAGEL [44]	71.50	71.70	81.80	80.10	82.50	83.50	83.00	79.90	79.90	76.80	70.20	72.20	74.40	75.00	67.40	70.10	72.00	74.90	86.70	83.30	29.40	33.90	68.30	67.90
	Janus-Pro-7B [42]	66.50	65.02	79.33	78.25	79.33	82.33	78.32	73.32	80.32	79.07	59.71	58.82	66.07	56.20	70.46	70.84	67.22	59.97	60.00	70.00	28.83	33.83	65.84	60.25
	InternVL-U [93]	74.90	73.90	82.30	81.50	86.00	81.50	84.10	82.20	76.70	80.90	73.50	72.70	75.30	76.20	70.40	67.60	75.50	75.80	93.30	83.30	47.50	50.70	65.30	66.80
	Ovis-U1 [94]	66.70	68.20	77.80	79.40	83.50	81.50	80.10	81.40	69.90	75.20	67.40	67.80	71.80	68.30	66.80	73.80	69.00	65.90	83.30	86.70	8.10	12.70	67.20	68.70
Open-Source Specialist	FLUX.1-dev [96]	71.09	71.78	83.12	78.65	87.05	83.17	87.25	80.39	75.01	72.39	65.79	68.54	67.07	73.69	73.84	73.34	69.09	71.59	66.67	66.67	43.83	52.83	70.72	71.47
	FLUX.1-Krea-dev [97]	80.36	81.67	84.26	83.76	86.00	84.00	82.93	86.18	83.47	80.99	73.85	76.15	79.89	77.01	74.75	79.21	76.43	78.98	73.33	80.00	66.52	70.14	93.66	94.78
	FLUX.2-Klein-Base-4B [6]	79.94	80.01	82.99	80.21	89.33	84.83	83.74	85.37	74.38	69.42	73.60	76.28	77.01	78.16	68.32	77.23	67.52	71.97	93.33	80.00	77.38	76.47	94.03	90.67
	FLUX.2-Klein-Base-9B [6]	81.47	84.52	82.49	85.28	84.00	85.33	83.74	87.80	79.34	82.64	78.32	82.02	78.16	77.59	75.74	78.53	73.89	83.44	76.67	80.00	84.16	87.78	89.18	90.67
	FLUX.2-Klein-4B [6]	78.91	79.04	81.22	80.96	79.33	80.00	83.74	83.74	80.99	79.34	73.34	73.56	77.01	79.31	72.28	74.11	67.52	75.16	76.67	73.33	75.11	67.42	91.79	92.16
	FLUX.2-Klein-9B [6]	85.22	84.13	88.32	85.17	90.67	84.00	86.99	90.24	86.78	81.36	80.87	81.12	78.24	77.01	74.75	78.71	76.43	81.53	93.10	90.00	90.05	85.07	93.28	91.42
	Qwen-Image [5]	86.14	86.83	86.18	87.22	90.50	91.50	88.22	90.78	79.81	79.38	79.30	80.88	79.21	78.94	78.85	81.69	75.57	78.59	100.00	100.00	92.76	89.14	90.30	91.42
	Z-Image-Base [4]	80.20	83.04	78.36	82.79	79.50	86.50	80.45	79.94	75.13	81.94	72.89	77.02	72.91	77.56	66.99	73.82	73.89	75.62	90.00	93.33	94.84	93.21	88.06	85.45
	Z-Image-Turbo [4]	77.73	80.05	81.85	81.59	86.50	87.00	82.88	79.99	76.17	77.77	68.32	74.69	72.04	75.24	60.22	73.33	68.90	71.92	83.33	93.33	83.71	84.62	85.82	77.24
	Lens-Base [99]	80.33	83.49	82.90	83.70	86.50	83.00	82.37	85.20	79.81	8.98	74.35	80.22	71.24	80.18	75.87	82.70	74.40	76.96	90.00	93.33	71.04	73.76	91.79	93.28
	Lens-RL [99]	84.23	84.92	88.03	89.21	91.00	89.50	91.18	90.54	81.90	87.58	80.75	83.27	84.53	82.82	79.30	85.01	78.97	83.90	66.67	66.67	90.50	84.62	94.03	93.66
	Lens-Turbo [99]	82.20	81.81	88.48	88.41	92.00	91.00	90.49	88.17	82.94	86.06	78.49	79.62	81.35	83.34	74.45	76.72	81.24	81.49	56.67	56.67	87.78	81.00	92.91	91.79
	Mage-Flow-Base	82.50	83.19	83.36	85.34	87.00	90.50	88.86	85.63	74.21	79.90	77.04	80.81	78.62	80.10	77.48	82.98	72.86	79.74	86.67	83.33	84.62	75.11	92.16	91.42
	Mage-Flow	82.19	84.70	83.44	85.72	88.00	86.00	86.19	86.67	76.12	84.50	77.35	80.86	79.61	79.51	76.72	81.97	74.66	80.28	80.00	80.00	83.26	88.24	95.15	95.15
	Mage-Flow-Turbo	83.58	84.16	85.11	88.45	88.00	92.00	87.52	88.86	79.81	84.50	77.07	78.70	82.32	80.90	74.28	77.39	71.86	76.61	86.67	80.00	89.59	86.88	92.12	90.36

Table 11 Quantitative results of multi-region English text rendering on CVTG-2K [88].

Type	Model	#Params	2-reg.	3-reg.	4-reg.	5-reg.	Avg	NED	CLIP
Closed-Source	Seedream 3.0 [90]	–	0.628	0.596	0.604	0.561	0.592	0.854	0.782
	Seedream 4.0 [3]	–	0.890	0.915	0.899	0.887	0.892	0.951	0.785
	GPT-Image-1 [1]	–	0.878	0.866	0.873	0.822	0.857	0.948	0.798
	Nano-Banana-Pro [2]	–	0.737	0.775	0.786	0.793	0.779	0.875	0.737
Open-Source Unified	BAGEL [44]	14B	0.498	0.391	0.332	0.291	0.356	0.657	0.779
	InternVL-U [93]	4B	0.729	0.660	0.618	0.549	0.623	0.804	0.816
	Ovis-U1 [94]	3.6B	0.133	0.109	0.091	0.065	0.093	0.477	0.725
Open-Source Specialist	SD3.5-Large [20]	8B	0.729	0.682	0.657	0.594	0.655	0.847	0.780
	FLUX.1-dev [96]	12B	0.609	0.553	0.466	0.432	0.496	0.688	0.740
	FLUX.1-Krea-dev [97]	12B	0.527	0.454	0.440	0.407	0.444	0.737	0.773
	FLUX.2-dev [6]	32B	0.926	0.890	<u>0.899</u>	0.873	0.893	0.948	0.810
	FLUX.2-Klein-Base-4B [6]	4B	0.679	0.653	0.667	0.636	0.656	0.825	0.757
	FLUX.2-Klein-Base-9B [6]	9B	0.676	0.653	0.667	0.636	0.655	0.825	0.757
	FLUX.2-Klein-4B [6]	4B	0.637	0.641	0.639	0.601	0.628	0.829	0.773
	FLUX.2-Klein-9B [6]	9B	0.464	0.443	0.416	0.399	0.424	0.733	0.759
	Qwen-Image [5]	20B	0.837	0.836	0.831	0.816	0.829	0.912	0.802
	JoyAI-Image [98]	16B	–	–	–	–	0.874	0.937	0.799
	HunyuanImage-3.0 [7]	80B	0.830	0.763	0.738	0.728	0.765	0.876	0.812
	LongCat-Image [22]	6B	<u>0.913</u>	0.874	0.856	0.831	0.866	0.936	0.786
	Z-Image-Base [4]	6B	0.901	0.872	0.865	0.851	0.867	0.937	0.797
	Z-Image-Turbo [4]	6B	0.887	0.866	0.863	0.835	0.859	0.928	0.805
	Lens-Base [99]	3.8B	0.708	0.656	0.645	0.577	0.635	0.814	0.777
	Lens-RL [99]	3.8B	0.871	0.867	0.847	0.806	0.843	0.935	0.797
	Lens-Turbo [99]	3.8B	0.911	<u>0.899</u>	0.883	<u>0.853</u>	0.882	0.955	0.801
	<i>Mage-Flow-Base</i>	4B	0.908	0.878	0.866	0.789	0.851	0.933	0.815
	<i>Mage-Flow</i>	4B	0.902	0.902	0.902	0.850	<u>0.887</u>	<u>0.950</u>	<u>0.826</u>
	<i>Mage-Flow-Turbo</i>	4B	0.892	0.883	0.878	0.851	0.873	0.945	0.831

Table 12 Quantitative results on OneIG-EN and OneIG-CN [87].

Type	Model	#Params	OneIG-EN						OneIG-CN					
			Alignment	Text	Reasoning	Style	Diversity	Overall	Alignment	Text	Reasoning	Style	Diversity	Overall
Closed-Source	Seedream 3.0 [90]	–	0.818	0.865	0.275	0.413	0.277	0.530	0.793	0.928	0.281	0.397	0.243	0.528
	Seedream 4.0 [3]	–	0.892	0.983	0.347	0.453	0.191	0.573	0.836	0.986	0.304	0.443	0.200	0.554
	GPT-Image-1 [1]	–	0.851	0.857	0.345	0.462	0.151	0.533	0.812	0.650	0.300	0.449	0.159	0.474
	Nano-Banana-Pro [2]	–	0.890	0.940	0.330	0.480	0.250	0.580	0.840	0.980	0.310	0.460	0.240	0.570
	Kolors 2.0 [91]	–	0.820	0.427	0.262	0.360	0.300	0.434	0.738	0.502	0.226	0.331	0.333	0.426
Open-Source Unified	BAGEL [44]	14B	0.769	0.244	0.173	0.367	<u>0.251</u>	0.361	0.672	0.365	0.186	0.357	0.268	0.370
	Show-o2 [100]	7B	0.798	0.002	0.219	0.317	0.186	0.304	–	–	–	–	–	–
	OmniGen2 [36]	4B	0.804	0.680	0.271	0.377	0.242	0.475	–	–	–	–	–	–
	InternVL-U [93]	4B	0.820	0.740	0.270	0.400	0.250	0.500	0.750	0.900	0.230	0.370	0.260	0.500
	Ovis-U1 [94]	3.6B	0.810	0.030	0.220	0.450	0.180	0.340	0.720	0.150	0.210	0.430	0.200	0.340
Open-Source Specialist	SD3.5-Large [20]	8B	0.809	0.629	0.294	0.353	0.225	0.462	0.137	0.200	0.141	0.243	<u>0.519</u>	0.248
	HiDream-1I-Full [95]	17B	0.829	0.707	0.317	0.347	0.186	0.477	0.620	0.205	0.256	0.304	0.300	0.337
	FLUX.1-dev [96]	12B	0.786	0.523	0.253	0.368	0.238	0.434	0.152	0.218	0.116	0.282	0.456	0.245
	FLUX.1-Krea-dev [97]	12B	0.842	0.483	0.282	0.377	0.232	0.443	0.105	0.234	0.114	0.263	0.640	0.271
	FLUX.2-Klein-Base-4B [6]	4B	0.830	0.688	0.262	0.414	0.230	0.485	0.744	0.211	0.240	0.384	0.250	0.366
	FLUX.2-Klein-Base-9B [6]	9B	0.867	0.897	0.299	0.442	0.217	<u>0.544</u>	0.795	0.310	0.250	0.417	0.229	0.400
	FLUX.2-Klein-4B [6]	4B	0.862	0.779	0.272	0.420	0.169	0.500	0.794	0.208	0.247	0.386	0.183	0.364
	FLUX.2-Klein-9B [6]	9B	0.884	0.898	<u>0.306</u>	<u>0.445</u>	0.158	0.538	0.818	0.354	0.280	<u>0.418</u>	0.162	0.406
	Qwen-Image [5]	20B	0.882	0.891	<u>0.306</u>	0.418	0.197	0.539	0.825	0.963	0.267	0.405	0.279	0.548
	JoyAI-Image [98]	16B	–	–	–	–	–	0.542	–	–	–	–	–	0.521
	LongCat-Image [22]	6B	0.854	0.923	0.216	0.365	0.222	0.516	0.812	0.948	0.253	0.355	0.224	0.518
	Z-Image-Base [4]	6B	0.881	<u>0.987</u>	0.280	0.387	0.194	0.546	0.793	0.988	0.266	0.386	0.243	<u>0.535</u>
	Z-Image-Turbo [4]	6B	0.840	0.994	0.298	0.368	0.139	0.528	0.782	<u>0.982</u>	<u>0.276</u>	0.361	0.134	0.507
	Lens-Base [99]	3.8B	0.842	0.908	0.241	0.384	0.262	0.527	0.781	0.904	0.256	0.358	0.200	0.500
	Lens-RL [99]	3.8B	0.877	0.965	0.279	0.348	0.162	0.526	<u>0.824</u>	0.944	0.259	0.316	0.141	0.497
	Lens-Turbo [99]	3.8B	0.868	0.968	0.291	0.314	0.160	0.520	0.802	0.952	0.260	0.284	0.148	0.489
	<i>Mage-Flow-Base</i>	4B	0.863	0.950	0.305	0.431	0.159	0.542	0.800	0.904	0.279	0.396	0.163	0.509
	<i>Mage-Flow</i>	4B	0.869	0.969	<u>0.309</u>	0.411	0.124	0.536	0.812	0.926	0.282	0.385	0.119	0.505
	<i>Mage-Flow-Turbo</i>	4B	0.861	0.949	0.307	0.391	0.105	0.523	0.803	0.909	<u>0.281</u>	0.363	0.099	0.491

Prompt following and compositional understanding. Table 9 shows that *Mage-Flow* achieves the strongest overall GenEval score, with particularly strong results on single-object generation, two-object generation, counting, and positional relations. On DPG-Bench, reported in Table 9, *Mage-Flow* remains competitive with strong open-source specialist models and clearly outperforms most unified baselines of similar or larger scale. Table 10 further shows the same trend under both short- and long-prompt settings: *Mage-Flow* maintains strong instruction-following ability, while *Mage-Flow-Turbo* retains competitive performance despite using only four sampling steps.

Table 13 Summary of image editing results across three benchmarks. Steps denotes the number of denoising steps used during inference. Models are grouped into closed-source and open-source. The best open-source score in each column is in **bold**. The second-best results are underlined.

Type	Model	#Params	Steps	ImgEdit	GEEdit-EN	GEEdit-CN	TextEdit-Syn	TextEdit-Real
Closed-Source	Nano-Banana [2]	–	–	4.29	7.291	7.399	16.54	18.22
	Seedream4.0 [3]	–	–	4.30	7.701	7.692	14.90	18.54
	Seedream4.5 [3]	–	–	4.32	7.820	7.800	–	–
	Nano-Banana-Pro [2]	–	–	4.37	7.738	7.799	–	–
Open-Source Unified	BAGEL [44]	14B	50	3.20	–	–	10.41	12.83
	OmniGen2 [36]	4B	50	3.44	–	–	8.52	10.99
	UniWorld-V1 [37]	19B	30	3.26	–	–	–	–
	Emu3.5 [45]	34B	–	4.41	–	–	9.34	14.05
Open-Source Specialist	Instruct-Pix2Pix [28]	1B	10	1.88	–	–	5.51	5.35
	MagicBrush [29]	1B	100	1.90	–	–	3.86	4.12
	AnyEdit [30]	1B	50	2.45	–	–	–	–
	UltraEdit [31]	2B	50	2.70	–	–	–	–
	OmniGen [32]	3.8B	50	2.96	–	–	–	–
	ICEdit [35]	12B	28	3.05	–	–	–	–
	Step1X-Edit-v1.2 [34]	19B	50	3.95	7.480	7.467	9.26	12.02
	ChronoEdit [39]	14B	50	4.42	–	–	–	–
	FLUX.1-Kontext-dev [33]	12B	28	3.71	6.462	1.857	12.14	14.31
	FLUX.2-dev [6]	32B	50	4.35	7.413	7.278	11.86	14.71
	FLUX.2-Klein-Base-4B [6]	4B	50	3.80	7.081	7.102	11.01	13.79
	FLUX.2-Klein-4B [6]	4B	4	4.01	7.717	7.750	11.84	14.46
	FLUX.2-Klein-Base-9B [6]	9B	50	4.05	7.740	7.745	12.76	15.65
	FLUX.2-Klein-9B [6]	9B	4	4.18	8.040	8.055	12.73	15.75
	Z-Image-Edit [4]	6B	50	4.30	7.570	7.540	–	–
	Qwen-Image-Edit-2509 [5]	20B	50	4.31	7.480	7.467	13.40	15.81
	Qwen-Image-Edit-2511 [5]	20B	50	<u>4.51</u>	7.877	7.819	13.53	16.81
	LongCat-Image-Edit [22]	6B	50	4.45	7.748	7.731	12.46	14.89
	FireRed-Image-Edit-1.0 [8]	20B	50	4.56	7.943	7.887	15.19	17.23
	JoyAI-Image-Edit [98]	16B	50	4.46	8.276	<u>8.125</u>	<u>14.80</u>	<u>17.23</u>
	<i>Mage-Flow-Edit-Base</i>	4B	30	4.28	7.860	7.970	13.63	15.57
	<i>Mage-Flow-Edit</i>	4B	30	4.34	8.127	8.123	14.14	16.26
	<i>Mage-Flow-Edit-Turbo</i>	4B	4	4.38	<u>8.271</u>	8.264	12.77	15.41

Text rendering and bilingual generation. Table 11 reports the CVTG-2K results, where *Mage-Flow* achieves near-best open-source multi-region text-rendering accuracy and substantially outperforms most unified generation models. On LongText, *Mage-Flow* performs strongly on English long-text rendering and remains competitive on Chinese, although the Chinese split still leaves room for further data supplementation. Table 12 show that *Mage-Flow-Base*, *Mage-Flow*, and *Mage-Flow-Turbo* maintain strong alignment and text-rendering ability on both English and Chinese OneIG splits, indicating that the multi-granularity captioning and bilingual data pipeline transfer well to fine-grained generation settings.

6.2.2 Instruction-based image editing

Overall comparison. Table 13 summarizes the main instruction-based editing results across ImgEdit-Bench, GEEdit-Bench, and TextEdit-Bench. With only 4B parameters, *Mage-Flow-Edit* delivers competitive editing quality against substantially larger open-source specialist models while using a much smaller backbone. Our editing models attain the best open-source GEEdit-Bench Chinese score and a near-best English score, and remain competitive on ImgEdit-Bench and TextEdit-Bench against models with several times more parameters. The distilled *Mage-Flow-Edit-Turbo* variant preserves most of this performance with only four denoising steps.

ImgEdit-Bench. Table 14 reports category-wise editing results. The Base, RL-aligned, and Turbo variants reach overall scores of 4.28, 4.34, and 4.38, respectively, outperforming earlier

Table 14 Quantitative results on ImgEdit-Bench [103]. All metrics are evaluated by GPT-4.1 on 0–5 scale. Overall is the average of all scores across tasks.

Type	Model	#Params	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall↑
Closed-Source	Nano-Banana [2]	–	4.62	4.41	3.68	4.34	4.39	4.40	4.18	3.72	4.83	4.29
	Seedream4.0 [3]	–	4.33	4.38	3.89	4.65	4.57	4.35	4.22	3.71	4.61	4.30
	Seedream4.5 [3]	–	4.57	4.65	2.97	4.66	4.46	4.37	4.92	3.71	4.56	4.32
	Nano-Banana-Pro [2]	–	4.44	4.62	3.42	4.60	4.63	4.32	4.97	3.64	4.69	4.37
Open-Source Unified	BAGEL [44]	14B	3.56	3.31	1.70	3.30	2.62	3.24	4.49	2.38	4.17	3.20
	OmniGen2 [36]	4B	3.57	3.06	1.77	3.74	3.20	3.57	4.81	2.52	4.68	3.44
	UniWorld-V1 [37]	19B	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
	Emu3.5 [45]	34B	4.61	4.32	3.96	4.84	4.58	4.35	4.79	3.69	4.57	4.41
Open-Source Specialist	Instruct-Pix2Pix [28]	1B	2.45	1.83	1.44	2.01	1.50	1.44	3.55	1.20	1.46	1.88
	MagicBrush [29]	1B	2.84	1.58	1.51	1.97	1.58	1.75	2.38	1.62	1.22	1.90
	AnyEdit [30]	1B	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
	UltraEdit [31]	2B	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.70
	OmniGen [32]	3.8B	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
	ICEdit [35]	12B	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
	Step1X-Edit-v1.2 [34]	19B	3.91	4.04	2.68	4.48	4.26	3.90	4.82	3.23	4.22	3.95
	ChronoEdit [39]	14B	4.48	4.39	3.49	4.66	<u>4.67</u>	4.57	4.91	3.82	4.83	4.42
	FLUX.1-Kontext-dev [33]	12B	3.99	3.88	2.19	4.27	3.13	3.98	4.51	3.23	4.18	3.71
	FLUX.2-dev [6]	32B	4.50	4.18	3.83	4.65	4.65	4.31	4.88	3.46	4.70	4.35
	FLUX.2-Klein-Base-4B [6]	4B	4.40	4.04	1.96	4.00	3.10	4.23	4.76	3.14	4.74	3.80
	FLUX.2-Klein-4B [6]	4B	4.45	4.37	2.04	4.21	3.93	4.29	4.87	3.49	<u>4.89</u>	4.01
	FLUX.2-Klein-Base-9B [6]	9B	4.40	4.22	2.31	4.50	3.84	4.26	4.89	3.27	<u>4.89</u>	4.05
	FLUX.2-Klein-9B [6]	9B	4.50	4.50	2.32	4.44	4.32	4.42	4.90	3.36	4.96	4.18
	Z-Image-Edit [4]	6B	4.40	4.14	4.30	4.57	4.13	4.14	4.85	3.63	4.50	4.30
	Qwen-Image-Edit-2509 [5]	20B	4.34	4.27	3.42	4.73	4.36	4.37	4.91	3.56	4.80	4.31
	Qwen-Image-Edit-2511 [5]	20B	4.54	<u>4.57</u>	4.13	4.70	4.46	4.36	4.89	4.16	4.81	<u>4.51</u>
	LongCat-Image-Edit [22]	6B	4.44	4.53	3.83	<u>4.80</u>	4.60	4.33	<u>4.92</u>	3.75	4.82	4.45
	FireRed-Image-Edit-1.0 [8]	20B	<u>4.55</u>	4.66	4.34	4.75	4.58	<u>4.45</u>	4.97	<u>4.07</u>	4.71	4.56
	JoyAI-Image-Edit [98]	16B	4.47	4.48	<u>4.31</u>	4.57	4.75	4.33	4.79	3.72	4.69	4.46
	<i>Mage-Flow-Edit-Base</i>	4B	4.28	4.22	4.08	4.45	4.39	4.01	4.82	3.40	4.21	4.28
	<i>Mage-Flow-Edit</i>	4B	4.34	4.27	4.03	4.48	4.50	4.06	4.89	3.49	4.51	4.34
	<i>Mage-Flow-Edit-Turbo</i>	4B	4.38	4.48	3.94	4.48	4.39	4.26	4.91	3.58	4.48	4.38

Table 15 Comparison of Semantic Consistency (G_SC), Perceptual Quality (G_PQ), and Overall Score (G_O) on GEdit-Bench [34]. GPT-4.1 is used as the evaluator, with all scores normalized to a 0–10 scale. G_O is calculated as the geometric mean of G_SC and G_PQ, averaged overall samples.

Type	Model	#Params	GEdit-Bench-EN			GEdit-Bench-CN		
			G_SC↑	G_PQ↑	G_O↑	G_SC↑	G_PQ↑	G_O↑
Closed-Source	Nano-Banana [2]	–	7.396	8.454	7.291	7.540	8.424	7.399
	Seedream4.0 [3]	–	8.143	8.124	7.701	8.159	8.074	7.692
	Seedream4.5 [3]	–	8.268	8.167	7.820	8.254	8.167	7.800
	Nano-Banana-Pro [2]	–	8.102	8.344	7.738	8.135	8.306	7.799
Open-Source Specialist	Step1X-Edit-v1.2 [34]	19B	7.974	7.714	7.480	7.988	7.679	7.467
	FLUX.1-Kontext-dev [33]	12B	7.045	7.206	6.462	1.606	7.863	1.857
	FLUX.2-dev [6]	32B	7.835	8.064	7.413	7.697	8.046	7.278
	FLUX.2-Klein-Base-4B [6]	4B	7.574	7.640	7.081	7.526	7.759	7.102
	FLUX.2-Klein-4B [6]	4B	8.142	8.041	7.717	8.119	8.101	7.750
	FLUX.2-Klein-Base-9B [6]	9B	8.300	7.870	7.740	8.243	7.960	7.745
	FLUX.2-Klein-9B [6]	9B	8.592	7.993	8.040	8.550	8.106	8.055
	Z-Image-Edit [4]	6B	8.110	7.720	7.570	8.030	7.800	7.540
	Qwen-Image-Edit-2509 [5]	20B	7.974	7.714	7.480	7.988	7.679	7.467
	Qwen-Image-Edit-2511 [5]	20B	8.297	<u>8.202</u>	7.877	8.252	<u>8.134</u>	7.819
	LongCat-Image-Edit [22]	6B	8.128	8.177	7.748	8.141	8.117	7.731
	FireRed-Image-Edit-1.0 [8]	20B	8.363	8.245	7.943	8.287	8.227	7.887
	JoyAI-Image-Edit [98]	16B	8.829	8.120	8.276	8.618	8.110	<u>8.125</u>
	<i>Mage-Flow-Edit-Base</i>	4B	8.685	7.495	7.860	8.772	7.584	7.970
	<i>Mage-Flow-Edit</i>	4B	<u>8.893</u>	7.766	8.127	<u>8.927</u>	7.685	8.123
	<i>Mage-Flow-Edit-Turbo</i>	4B	8.965	7.970	<u>8.271</u>	8.948	8.002	8.264

instruction-tuned models and remaining competitive with recent MMDiT-based editing systems. The Turbo variant is strongest among our models on this benchmark despite using only four denoising steps, indicating that few-step distillation preserves the teacher’s broad editing behavior.

Table 16 Comparison of Instruction Following (IF), Text Accuracy (TA), Visual Consistency (VC), Layout Preservation (LP), and Semantic Expectation (SE) on the synthetic and real-world subsets of TextEdit-Bench [104]. All metrics are scored by GPT-4o on a 0–5 scale, and Overall denotes the sum of the five metric means (out of 25). For the synthetic subset, we use reference-based evaluation with ground-truth edited images, measuring preservation fidelity only on non-edited regions. For the real-world subset, where paired references are unavailable, we adopt non-reference evaluation by comparing outputs with original inputs outside the masked regions to quantify unintended changes.

Type	Model	#Params	Synthetic						Real-World					
			IF↑	TA↑	VC↑	LP↑	SE↑	Overall↑	IF↑	TA↑	VC↑	LP↑	SE↑	Overall↑
Closed-Source	Nano-Banana [2]	–	2.90	3.25	3.40	4.46	2.53	16.54	3.18	3.60	3.94	4.77	2.73	18.22
	Seedream4.0 [3]	–	2.67	3.44	2.97	3.25	2.57	14.90	3.64	3.96	3.90	4.42	2.62	18.54
Open-Source Unified	BAGEL [44]	14B	1.81	2.31	1.83	2.94	1.52	10.41	2.21	2.50	2.75	4.22	1.15	12.83
	OmniGen2 [36]	4B	1.12	1.74	1.72	3.08	0.86	8.52	1.67	2.22	2.45	3.63	1.02	10.99
	Emu3.5 [45]	34B	1.40	2.41	1.98	1.90	1.65	9.34	2.82	3.00	3.10	3.73	1.40	14.05
Open-Source Specialist	Instruct-Pix2Pix [28]	1B	0.90	1.14	1.14	1.61	0.72	5.51	0.56	1.22	1.09	1.69	0.79	5.35
	MagicBrush [29]	1B	0.73	0.73	0.60	1.18	0.62	3.86	0.59	0.80	0.64	1.42	0.67	4.12
	Step1X-Edit-v1.2 [34]	19B	1.40	1.44	1.88	3.40	1.14	9.26	1.96	2.11	2.81	4.06	1.08	12.02
	LongCat-Image-Edit [22]	6B	2.18	2.27	2.51	3.87	1.64	12.46	2.92	2.96	3.32	4.43	1.26	14.89
	FLUX.1-Kontext-dev [33]	12B	1.80	1.92	2.44	4.43	1.55	12.14	2.64	2.73	3.18	4.64	1.12	14.31
	FLUX.2-dev [6]	32B	1.81	1.99	2.44	4.14	1.49	11.86	2.68	2.81	3.39	4.42	1.41	14.71
	FLUX.2-Klein-Base-4B [6]	4B	1.62	1.82	2.20	4.12	1.25	11.01	2.48	2.62	3.06	4.54	1.10	13.79
	FLUX.2-Klein-4B [6]	4B	1.84	1.90	2.39	4.21	1.51	11.84	2.58	2.76	3.30	4.60	1.22	14.46
	FLUX.2-Klein-Base-9B [6]	9B	2.09	2.21	2.58	4.25	1.64	12.76	3.03	3.04	3.51	4.68	1.40	15.65
	FLUX.2-Klein-9B [6]	9B	2.13	2.25	2.63	4.12	1.61	12.73	3.02	3.14	3.58	4.62	1.39	15.75
	Qwen-Image-Edit-2509 [5]	20B	2.35	2.41	2.76	4.20	1.68	13.40	3.13	3.14	3.59	4.65	1.30	15.81
	Qwen-Image-Edit-2511 [5]	20B	2.49	2.54	2.74	4.01	1.75	13.53	3.46	<u>3.48</u>	3.67	4.61	<u>1.58</u>	<u>16.81</u>
	FireRed-Image-Edit-1.0 [8]	20B	2.88	2.85	3.17	4.46	<u>1.85</u>	15.19	3.54	3.53	3.95	<u>4.74</u>	1.47	17.23
	JoyAI-Image-Edit [98]	16B	<u>2.68</u>	<u>2.76</u>	<u>2.92</u>	4.49	1.95	<u>14.80</u>	<u>3.52</u>	3.42	<u>3.79</u>	4.83	1.67	17.23
	<i>Mage-Flow-Edit-Base</i>	4B	2.53	2.55	2.65	4.20	1.70	13.63	3.18	3.14	3.41	4.61	1.23	15.57
	<i>Mage-Flow-Edit</i>	4B	2.55	2.59	2.75	<u>4.48</u>	1.77	14.14	3.24	3.24	3.55	<u>4.74</u>	1.48	16.26
	<i>Mage-Flow-Edit-Turbo</i>	4B	2.26	2.31	2.52	3.98	1.69	12.77	3.17	3.11	3.42	4.56	1.15	15.41

GEEdit-Bench. As shown in Table 15, all three variants achieve strong overall scores on both language splits. *Mage-Flow-Edit* reaches 8.127 on English and 8.123 on Chinese, while *Mage-Flow-Edit-Turbo* further reaches 8.271 and 8.264 with only four denoising steps. Both variants outperform Qwen-Image-Edit-2511 [5], FireRed-Image-Edit-1.0 [8], and LongCat-Image-Edit [22] under the same benchmark protocol.

TextEdit-Bench. TextEdit-Bench isolates text-editing ability, which is only partially reflected by general editing benchmarks such as ImgEdit-Bench and GEEdit-Bench. Table 16 consolidates the available results into one table. *Mage-Flow-Edit* is the strongest of our variants, reaching 14.142 on Synthetic and 16.255 on Real-World, while precise layout preservation and complex text replacement remain important directions for improvement.

6.3 Qualitative Results

6.3.1 Text-to-image visualizations

Fig. 15 to 19 show representative *Mage-Flow* generations at native resolutions and aspect ratios. The first three galleries cover broad photorealistic and imaginative generation, including cuisine, landmarks, general concepts, and portraits. The last two galleries focus on English and Chinese text rendering, where *Mage-Flow* produces legible text on packaging, posters, magazine covers, and signage rather than text-like textures.

6.3.2 Instruction-based editing visualizations

Fig. 12 provides a high-level overview of *Mage-Flow-Edit* across semantic content editing, appearance transformation, restoration, and structure-aware output tasks. Fig. 2 and 3 then present

two compact source-centric overviews in which each source image is paired with multiple edited outputs. All showcased edit types are supported bidirectionally. As shown in Fig. 20 to 24, the six larger multi-source galleries provide broader coverage of content and object editing, scene and camera transformations, appearance, human-centered and creative editing, composite multi-task edits, low-level control, and paired degradation–restoration tasks. Throughout these galleries, dark labels denote source images and blue labels indicate the edit type applied to each output.

7 Conclusion

We presented the *Mage-Flow* generative stack, a compact and efficient 4B-scale framework for text-to-image generation and instruction-based image editing. Instead of relying on continued backbone scaling, *Mage-Flow* is built through the joint design of a lightweight latent tokenizer, a native-resolution diffusion Transformer, and system-level training infrastructure. *Mage-VAE* substantially reduces the cost of latent encoding and decoding while preserving reconstruction fidelity. NR-MMDiT supports flexible training and inference across resolutions and aspect ratios through native-resolution packing. Fused CUDA kernels further remove memory-bound overhead from the repeated blocks of the tokenizer, text encoder, and diffusion backbone. Together, these designs make large-scale native-resolution training practical under a fixed compute budget.

On this foundation, we build a complete generation-and-editing model family, including Base, RL-aligned, and Turbo variants for both *Mage-Flow* and *Mage-Flow-Edit*. Diffusion-NFT improves alignment with prompt-following, text-rendering, aesthetic, and editing preferences, while Decoupled-DMD distillation with adversarial perceptual guidance produces 4-step Turbo models for low-latency inference. Across standard generation and editing benchmarks, the resulting models achieve competitive performance against substantially larger open-source systems, while maintaining low memory usage and fast inference in both generation and editing settings.

These results show that strong visual generation does not necessarily require tens-of-billions-parameter backbones. With an efficient tokenizer, native-resolution modeling, and carefully optimized training infrastructure, a 4B-scale stack can serve as a practical and research-friendly foundation for image generation, controllable editing, post-training alignment, and vertical-domain applications. Future work will explore more robust multi-image editing, stronger multilingual long-text rendering, and tighter integration with agentic visual creation workflows.

Contributor List

Contributors: Xinjie Zhang^{*†}, Peng Zhang^{*}, Shicheng Zheng^{*}, Jinghao Guo^{*}, Zhaoyang Jia^{*}, Yifei Shen^{*}, Xun Guo, Yuxuan Luo, Jiahao Li, Wenxuan Xie, Fanyi Pu, Xiaoyi Zhang, Kaichen Zhang, Zongyu Guo, Tianci Bi, Dongnan Gui, Zhening Liu, Zimo Wen, Zihan Zheng, Senqiao Yang, Xiao Li, Jinglu Wang, Bin Li, Yan Lu

^{*}Equal Contribution. [†]Project Lead (xinjiezhang@microsoft.com).



Figure 15 *Mage-Flow* gallery: cuisine & world landmarks. Native-resolution *Mage-Flow-Base* samples of plated dishes from a range of cuisines alongside iconic architecture, justified into a mosaic and shown uncropped at their generated aspect ratios. Consistent lighting, material detail, and depth of field hold across close-up food and wide architectural scenes.

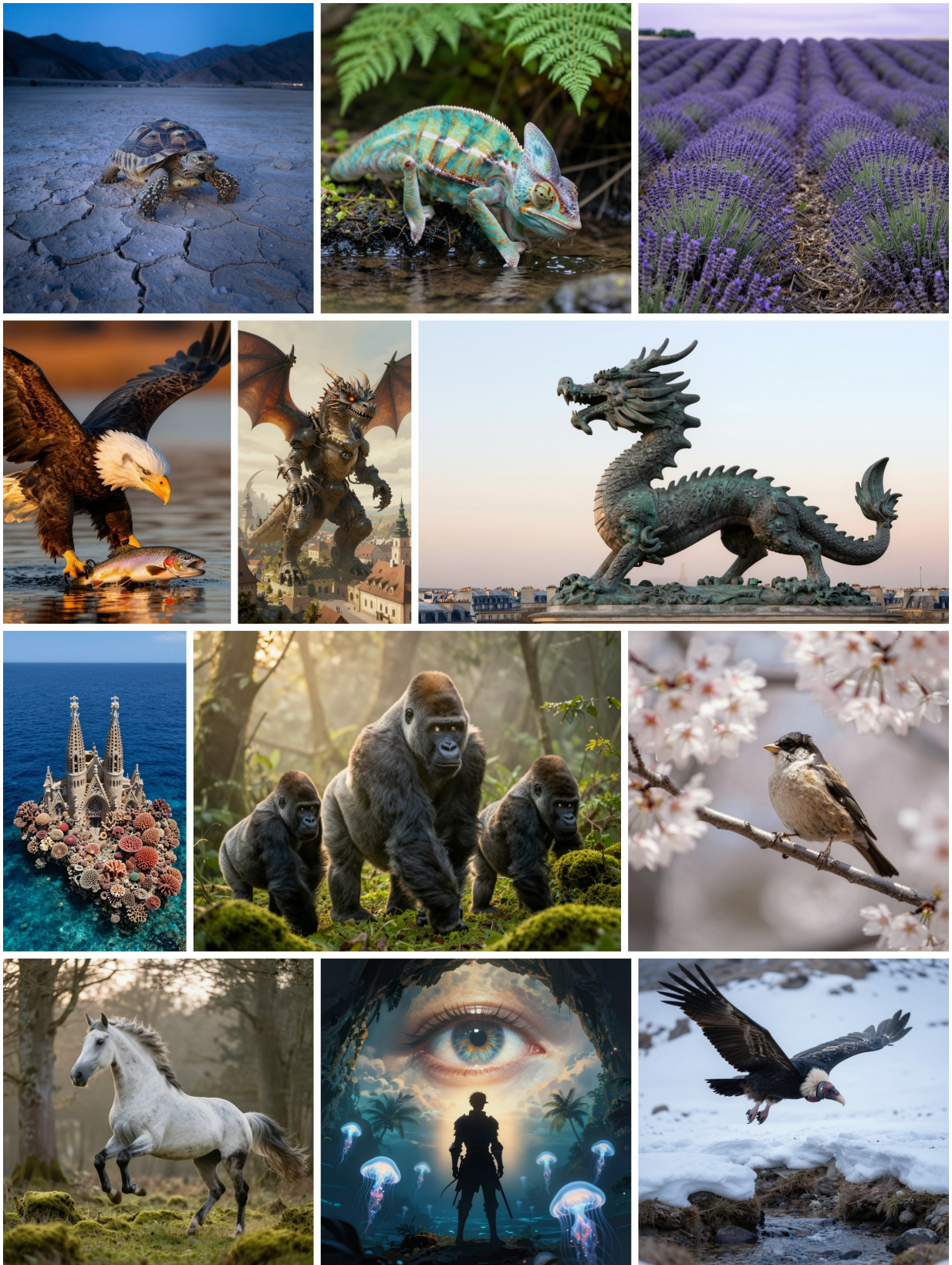


Figure 16 *Mage-Flow* gallery: general & imaginative concepts. A mix of wildlife, natural scenery, and fantastical subjects (mechanical and sculptural creatures, surreal composites), demonstrating broad subject coverage and coherent structure under widely varying styles.



Figure 17 *Mage-Flow* gallery: portraits & people. Human subjects spanning ages, genders, ethnicities, and cultural dress, in both studio and environmental settings, with plausible anatomy, natural skin and hair detail, and controlled lighting.

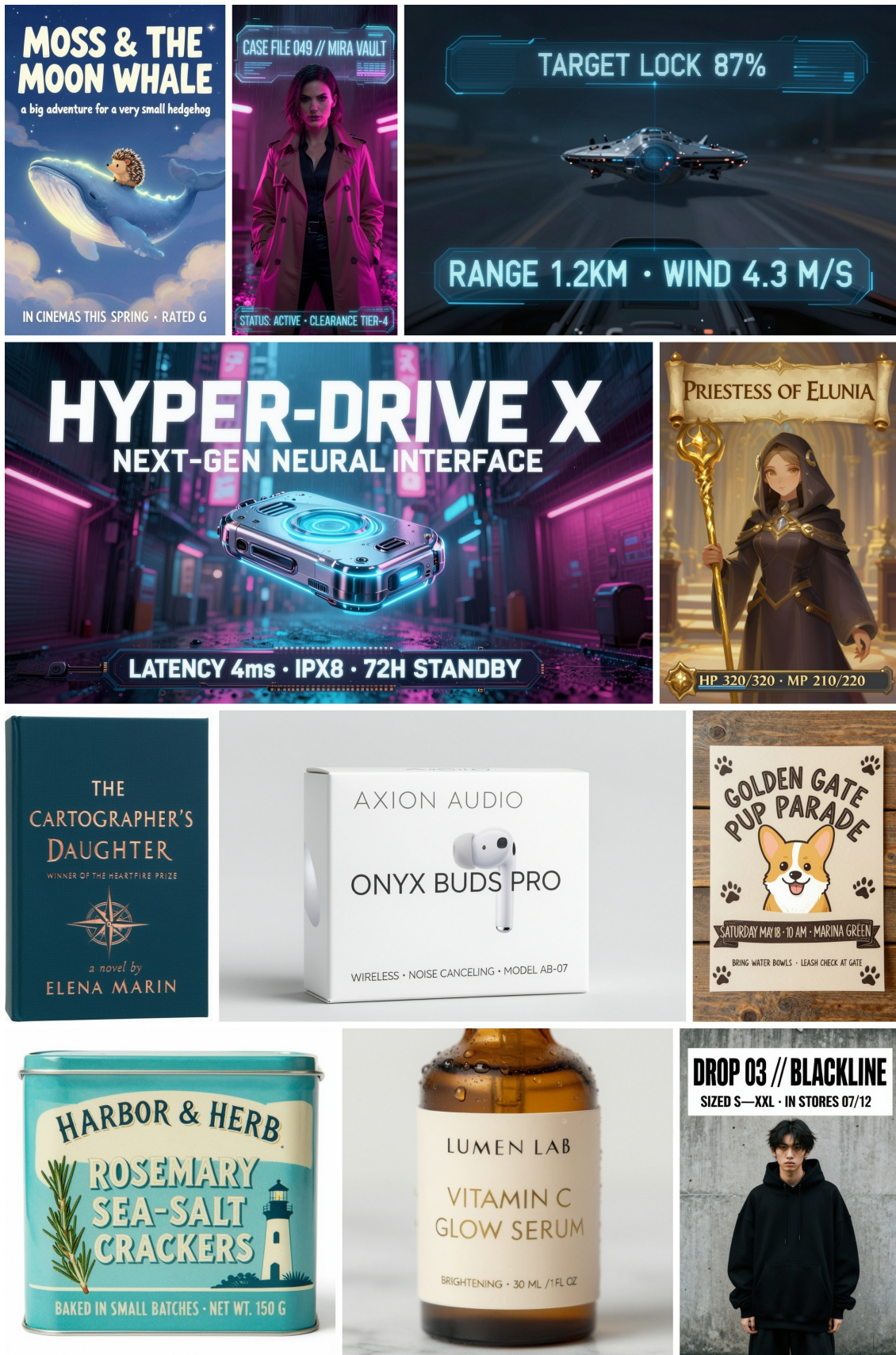


Figure 18 *Mage-Flow* gallery: English text rendering. Product packaging, book and magazine covers, posters, and signage. Multi-line English copy—brand names, taglines, and fine print—is reproduced *verbatim* from the prompt and stays legible down to the smallest lines at full resolution.



Figure 19 Mage-Flow gallery: Chinese & multilingual text rendering. Signage, packaging, and covers carrying Chinese—and several bilingual Chinese/English—layouts. The model reproduces the prompted characters, including denser glyphs, without the stroke-level artifacts typical of diffusion models.

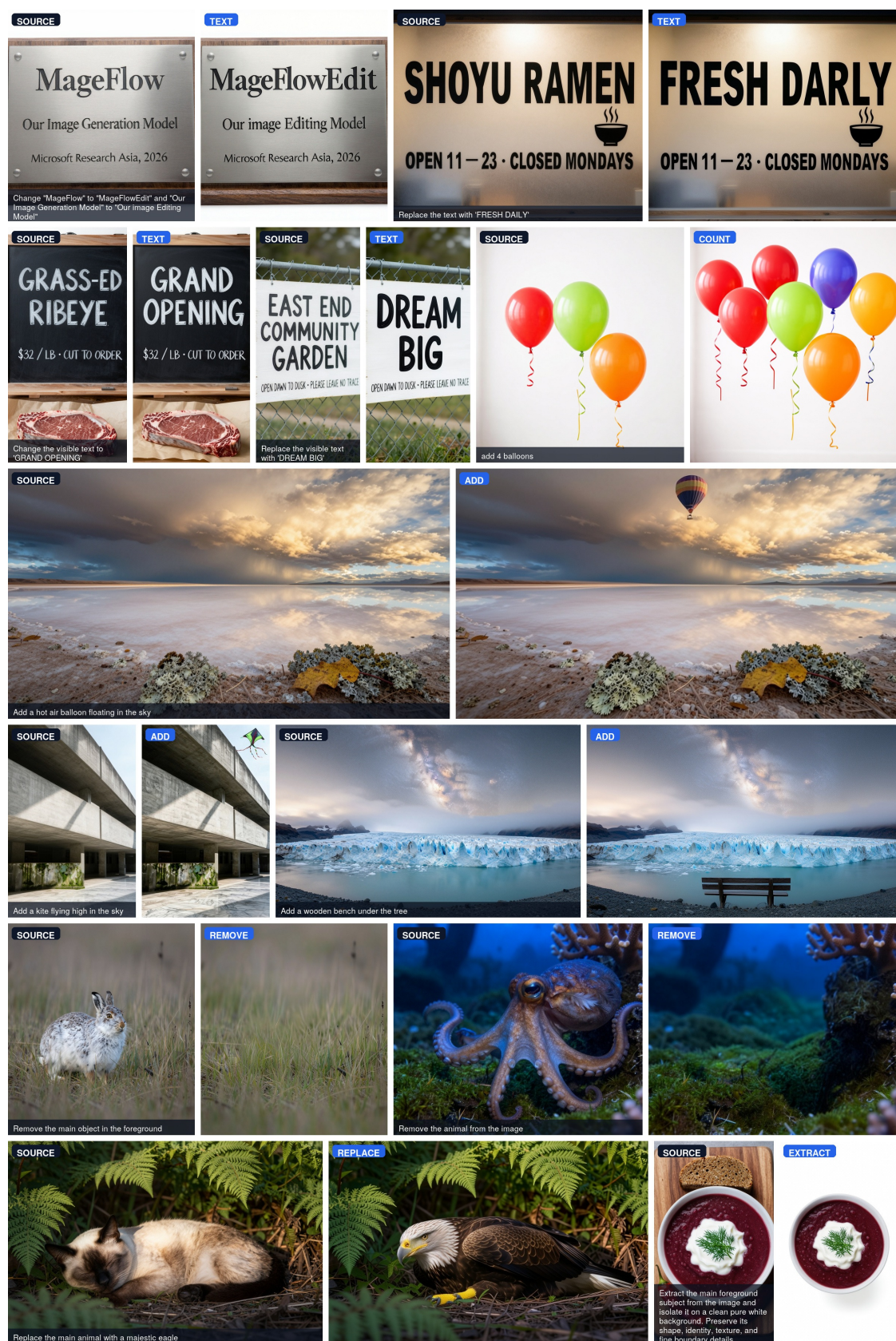


Figure 20 Qualitative gallery of *Mage-Flow-Edit*: localized content and object editing.

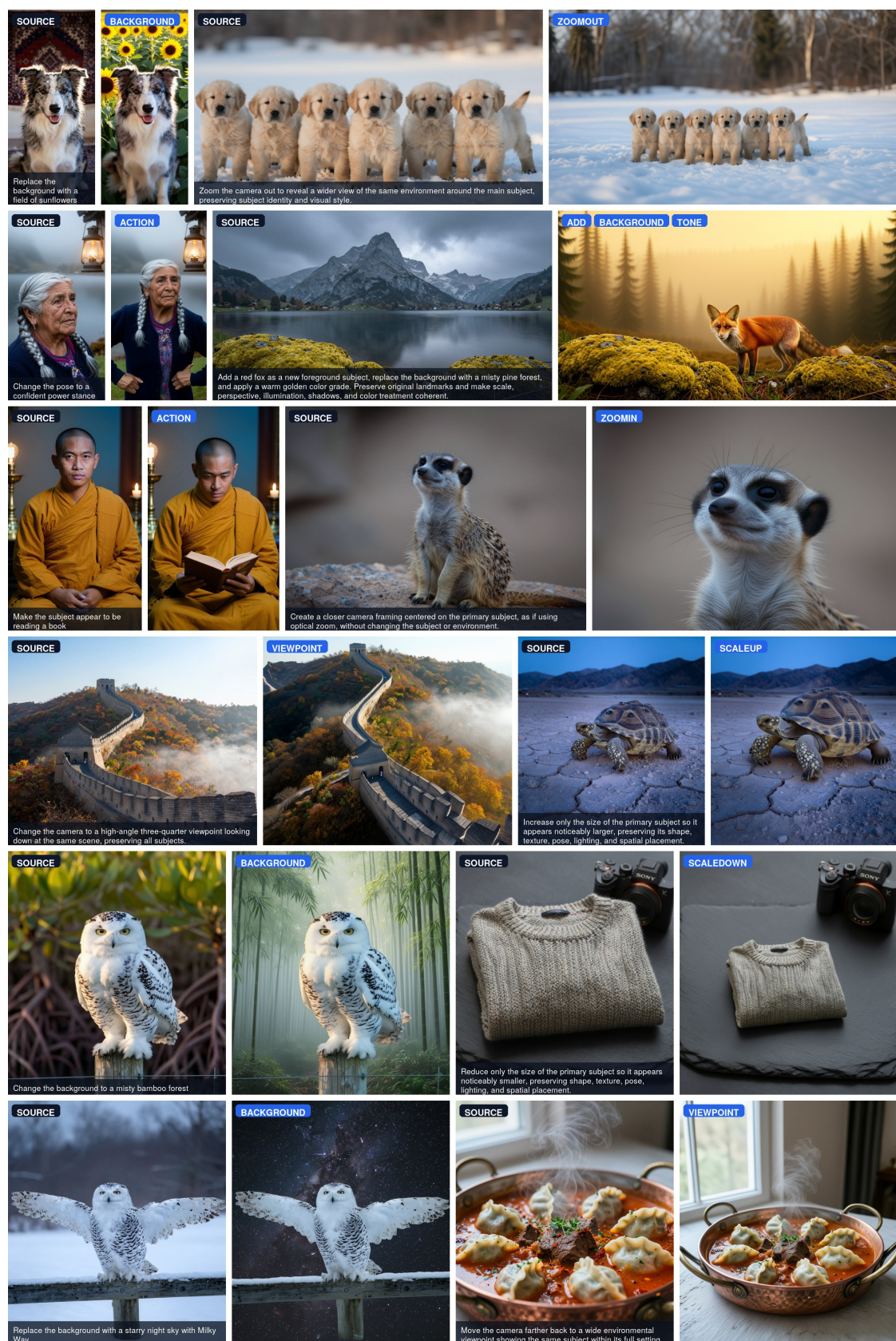


Figure 21 Qualitative gallery of *Mage-Flow-Edit*: scene, subject, and camera transformations.

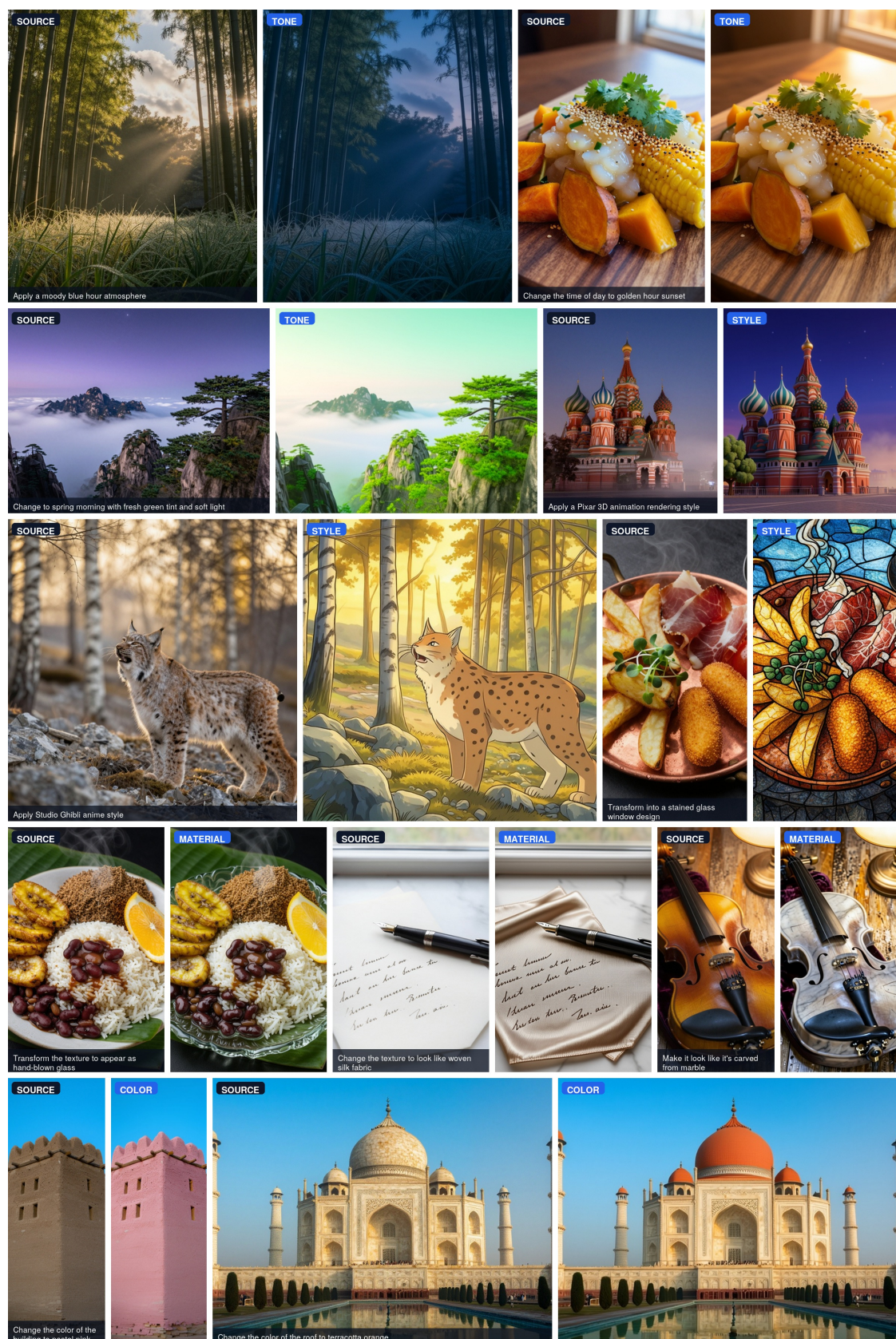


Figure 22 Qualitative gallery of *Mage-Flow-Edit*: appearance and artistic rendering.

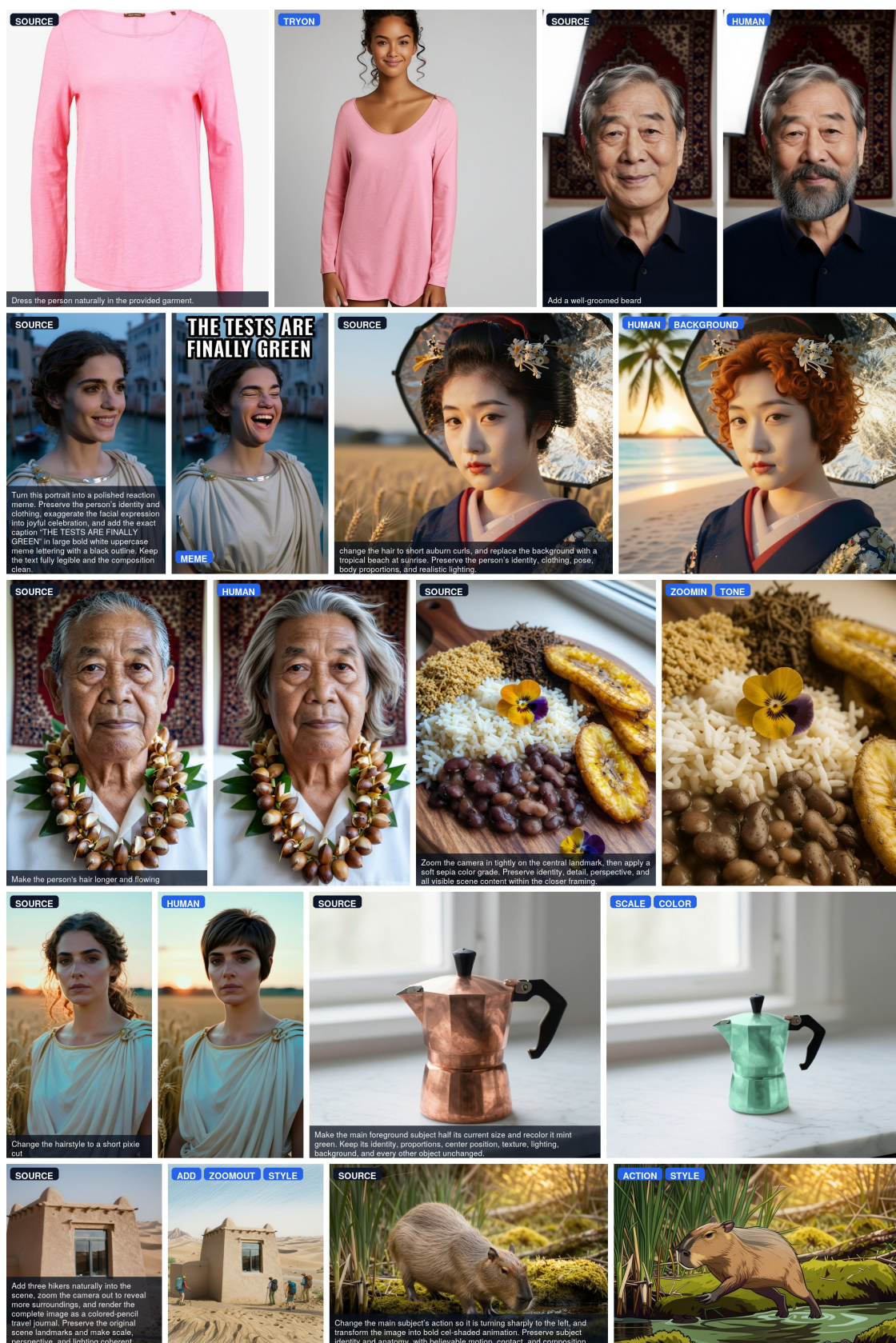


Figure 23 Qualitative gallery of *Mage-Flow-Edit*: human-centered and creative editing.

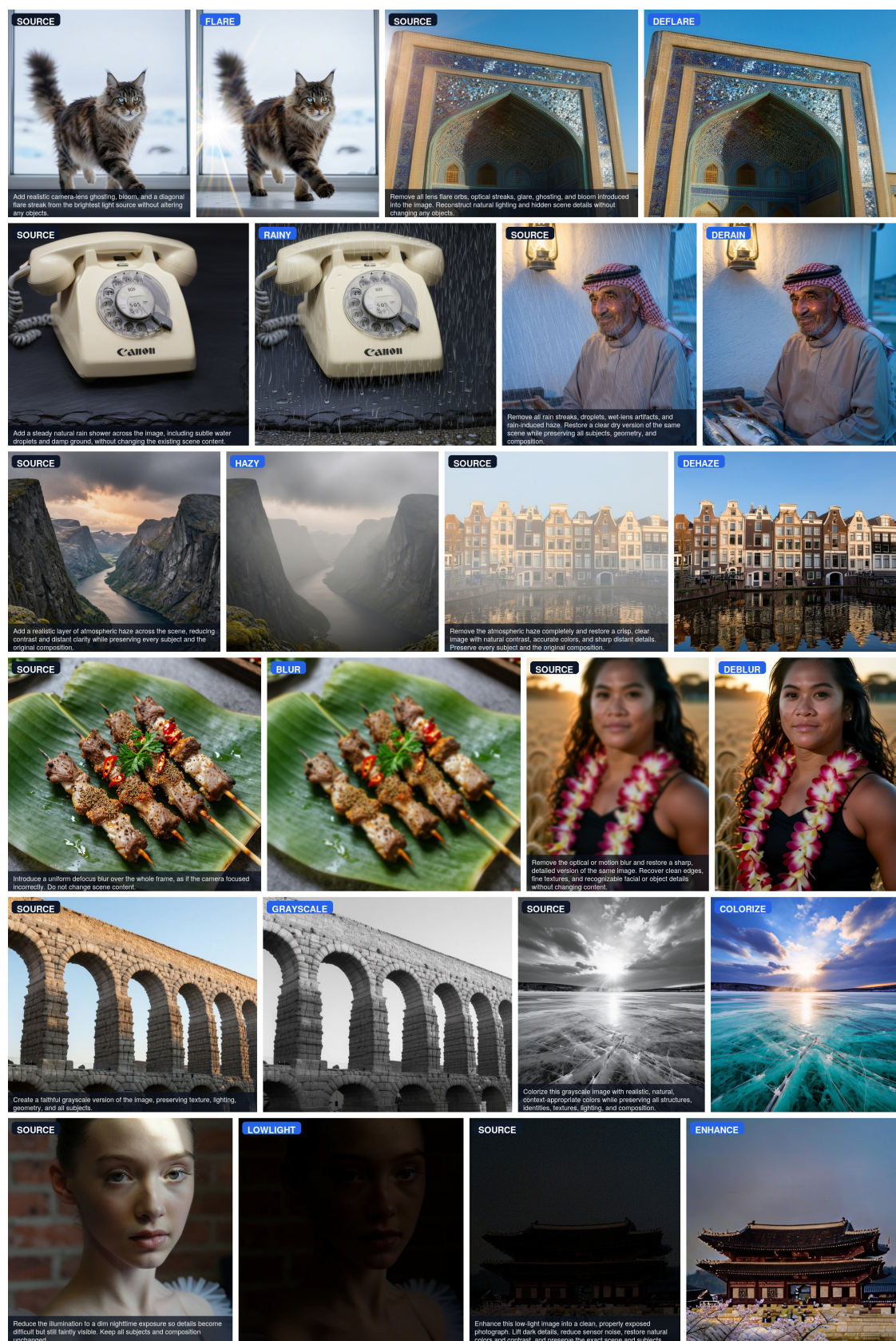


Figure 24 Qualitative gallery of *Mage-Flow-Edit*: bidirectional degradation and restoration.

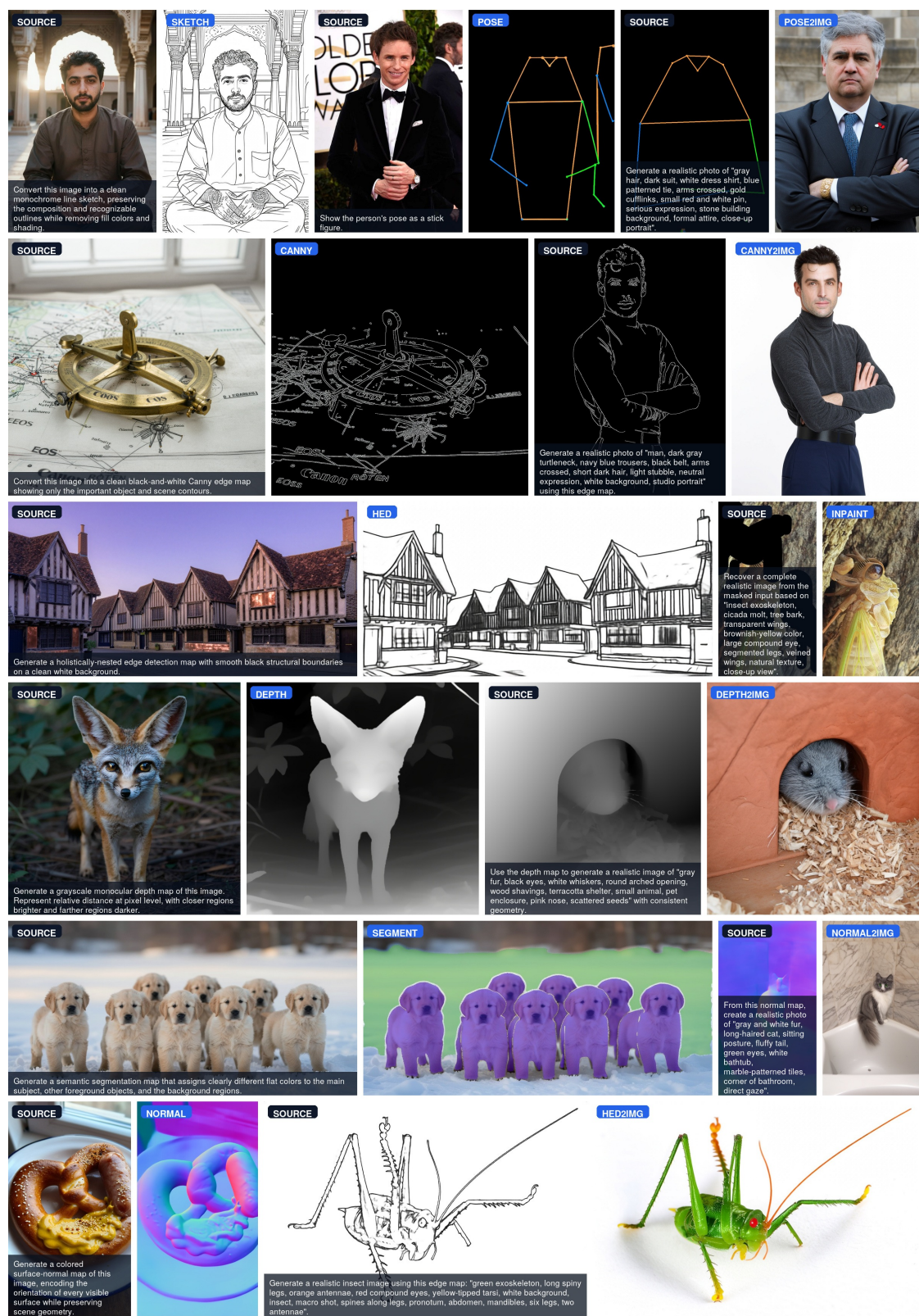


Figure 25 Qualitative gallery of *Mage-Flow-Edit*: low-level vision and conditional reconstruction.

A Detailed *Mage*-VAE Training Process

As described in Section 3.1.3, the training of *Mage*-VAE proceeds in three stages. Let x denote an input image and

$$z^a = \mathcal{P}(\mathcal{E}_{\text{FLUX}}(x)) \quad (2)$$

denote its anchor latent, where $\mathcal{E}_{\text{FLUX}}$ is the frozen FLUX.2-VAE encoder and $\mathcal{P}$ denotes latent patchification.

Stage I: Multi-step flow-matching pre-training. We first pre-train both the encoder and decoder as multi-step diffusion models using flow matching in the $\mathcal{X}$ -prediction parameterization. For a target y_0 , we sample $\epsilon \sim \mathcal{N}(0, I)$ and $t \sim \mathcal{U}(0, 1)$, and construct the linear interpolation

$$y_t = (1 - t)y_0 + t\epsilon. \quad (3)$$

Instead of directly predicting the velocity field, the network predicts the clean endpoint y_0 :

$$\mathcal{L}_{\text{FM}}^{\mathcal{X}}(\theta) = \mathbb{E}_{y_0, \epsilon, t} \left[\|\mathcal{X}_{\theta}(y_t, t \mid c) - y_0\|_2^2 \right]. \quad (4)$$

The corresponding velocity field used for ODE sampling can be recovered as

$$v_{\theta}(y_t, t \mid c) = \frac{y_t - \mathcal{X}_{\theta}(y_t, t \mid c)}{t}. \quad (5)$$

For the decoder, the prediction target and condition are $(y_0, c) = (x, z^a)$, so that the model learns to reconstruct image pixels from anchor latents. For the encoder, they are $(y_0, c) = (z^a, x)$, such that the encoder learns to generate patchified FLUX.2-VAE anchor latents from image pixels. The Stage-I objective is therefore

$$\mathcal{L}_{\text{I}} = \mathcal{L}_{\text{FM}, \text{dec}}^{\mathcal{X}} + \mathcal{L}_{\text{FM}, \text{enc}}^{\mathcal{X}}. \quad (6)$$

Stage II: One-step decoder distillation. We next distill the multi-step decoder into a one-step decoder $\mathcal{D}_{\psi}$. Given an anchor latent z^a , the reconstruction is obtained with a single network evaluation:

$$\hat{x} = \mathcal{D}_{\psi}(z^a). \quad (7)$$

We combine pixel-level and perceptual reconstruction objectives,

$$\mathcal{L}_{\text{rec}} = \|x - \hat{x}\|_1 + \mathcal{L}_{\text{LPIPS}}(x, \hat{x}), \quad (8)$$

with a DINOv2-projected GAN loss [76] and a DMD loss [64]. The projected discriminator operates on features extracted by a frozen DINOv2 network, encouraging the reconstruction to match the real-image distribution in a perceptually meaningful feature space. Meanwhile, DMD transfers the distribution-level prior of the compression-oriented diffusion teacher [75] to the one-step decoder. The complete Stage-II decoder objective is

$$\mathcal{L}_{\text{II}} = \underbrace{\|x - \hat{x}\|_1 + \mathcal{L}_{\text{LPIPS}}(x, \hat{x})}_{\mathcal{L}_{\text{rec}}} + 0.01 \mathcal{L}_{\text{GAN}}^{\text{DINO}} + 0.1 \mathcal{L}_{\text{DMD}}. \quad (9)$$

The diffusion teacher and DINOv2 feature extractor remain frozen during distillation, while the projected discriminator and the fake-score model used by DMD are optimized with their respective objectives.

Stage III: Joint one-step VAE optimization. Finally, we convert the encoder into a one-step model and optimize it together with the distilled decoder. Let

$$q_\phi(z | x) \quad (10)$$

denote the latent distribution produced by the one-step encoder and let $q_a(z | x)$ denote the corresponding frozen FLUX.2-VAE anchor-latent distribution. We regularize the learned latent space using

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_x [D_{\text{KL}}(q_\phi(z | x) \| q_a(z | x))]. \quad (11)$$

This preserves the structure of the pretrained anchor latent space and prevents the encoder distribution from drifting during perceptual fine-tuning.

Stage III.1: Encoder warm-up with a fixed decoder. We first freeze the one-step decoder and optimize only the encoder. For

$$z \sim q_\phi(z | x), \quad \hat{x} = \mathcal{D}_\psi(z), \quad (12)$$

the encoder is trained using the complete Stage-II reconstruction and perceptual objective together with the anchor-latent KL regularizer:

$$\mathcal{L}_{\text{III.1}} = \mathcal{L}_{\text{II}} + 0.1 \mathcal{L}_{\text{KL}}, \quad \psi \text{ fixed}. \quad (13)$$

Although the decoder parameters are frozen, gradients from all components of $\mathcal{L}_{\text{II}}$ are back-propagated through the decoder to train the encoder. This warm-up stage aligns the one-step encoder with the already distilled decoder without perturbing the decoder’s reconstruction ability.

Stage III.2: End-to-end fine-tuning. After encoder warm-up, we unfreeze the decoder and jointly optimize the complete one-step encoder–decoder pipeline:

$$\mathcal{L}_{\text{III.2}} = \mathcal{L}_{\text{II}} + 0.1 \mathcal{L}_{\text{KL}}, \quad \phi, \psi \text{ jointly optimized}. \quad (14)$$

End-to-end fine-tuning allows the encoder and decoder to co-adapt while the KL term maintains compatibility with the anchor-latent space. The resulting *Mage-VAE* provides one-step encoding and decoding, high-fidelity reconstruction, and a compact, regularized latent space suitable for subsequent *Mage-Flow* training.

B Recaptioning System Prompt

As described in Section 4.1, we recaption each retained image at multiple granularities with Qwen3-VL-32B-Instruct [13]. The four caption layers introduced there, namely the entity, phrase, composition, and photographic descriptions, are produced by the single system prompt below in one pass and returned as a JSON object. The prompt fixes the objectivity and consistency principles, a step-by-step captioning procedure, the JSON output schema for the four layers, and a style guide, so that the four granularities stay mutually consistent.

Multi-granularity recaptioning system prompt

You are a world-class multi-granularity image captioning expert.

Your task is to analyze the given image and produce a structured, detailed, and objective description following these principles:

Core Principles

1. Absolute Objectivity: Describe only what is visually present. Avoid subjective words like "beautiful" or "sad." Express aesthetic qualities through concrete descriptions of color, light, shadow, and composition.
2. Physical and Logical Consistency: Ensure all elements adhere to real-world physics and spatial logic (e.g., shadows align with light sources, objects rest naturally).
3. Structured Description: Start from the overall scene, then move to background, mid-ground, and foreground. Use directional terms like "left side," "center," "background" to clarify spatial layout.
4. Present Tense: Describe as if observing the image now (e.g., "A man stands," "Light falls on...").
5. Rich and Specific Language: Use precise adjectives for quantity, size, shape, color, and texture. Avoid vague terms.

Steps

1. Identify the most important part of the image: Specify the subject(s), their color, quantity, and position.
2. Describe the background: Indicate environment type (e.g., mountains, sea, urban, indoor cafe), lighting, and atmosphere.
3. Detail composition and style: Include artistic medium, perspective, and any stylistic elements (e.g., oil painting, anime, photorealistic).
4. Extract subject information as phrases: Provide a concise set of keywords describing the subject.
5. Summarize the image information: Combine all details into a coherent summary.

Output Format (JSON with Four Layers and Guidelines)

```
{
  "Entity": "Describe the most important subject(s) in the image with clear identification. This should be a short phrase naming the main object or scene (e.g., 'Hollywood Sign in Los Angeles', 'A vibrant mosaic artwork').",
  "Phrase": "Provide a set of concise keywords or short phrases that capture the subject's attributes, colors, shapes, and related objects. These should be comma-separated and reflect what is visually present (e.g., 'iconic landmark, Hollywood Sign, cloudy sky, modern house, natural surroundings'). Very similar phrase only appear once.",
  "Composition": "Write one single sentence that objectively summarizes the main subject and its context, mentioning only what is visually present in the image, including the setting or environment, without adding lighting, perspective, texture, or any subjective interpretation.",
  "Photographic": "Provide a detailed artistic and technical description of the image in several coherent sentences, explicitly mentioning the image category, inferred style (reinforced), any recognized text (OCR), main entity, named entities, lighting and color conditions, camera composition, and notable photography techniques, ensuring the description is objective and visually grounded."
}
```

Style Guide

- Infer the style from the image content (e.g., photorealistic, cinematic, anime, oil painting, minimalist, cyberpunk).
- If the style is unclear, default to photorealistic photography.
- Apply the inferred style consistently in the Composition and Photographic layers.
- Mention the style explicitly in the Photographic description and reinforce it through lighting, perspective, and atmosphere details.

C Reward Models

This appendix details the four reward evaluators used in the Diffusion-NFT post-training stage (Section 5.2). Three of them score text-to-image generation samples: text rendering, aesthetic quality, and semantic understanding; these are used both for *Mage-Flow* post-training and for the generation stream of *Mage-Flow-Edit*. The fourth, RationalRewards, scores instruction-based editing samples in the editing stream of *Mage-Flow-Edit*. Each RL prompt carries a capability tag that routes it to exactly one evaluator, which turns a generated or edited image into a scalar reward in $[0, 1]$; rewards from different evaluators are never mixed within a prompt. Below we describe each evaluator together with a concrete example of how it scores.

Text rendering (PaddleOCR-VL-1.5). Every text-rendering prompt is annotated with a set of target strings $\mathcal{T} = \{t_1, \dots, t_n\}$ that must appear, legibly and correctly spelled, in the image. We render the sample, run PaddleOCR-VL-1.5 [82] on it to obtain a recognized string, and split that string on line and punctuation separators into segments, normalizing each segment by removing spaces and lower-casing it to form a set $\mathcal{S}$. Each target $t \in \mathcal{T}$ is normalized in the same way, with spaces removed and lower-cased, and we write $\tilde{t}$ for the resulting normalized target. A target that occurs verbatim as a substring of some segment scores 1; otherwise we take its smallest character-level edit distance to any segment, capped at the target length, and convert it to a similarity. Writing $d(t) = \min_{s \in \mathcal{S}} \min(\text{Lev}(s, \tilde{t}), |\tilde{t}|)$ for that best capped distance, where $\text{Lev}(\cdot, \cdot)$ is the character-level Levenshtein (edit) distance and $|\tilde{t}|$ the target length, the per-target score is

$$p(t) = \begin{cases} 1, & \text{if } \tilde{t} \subseteq s \text{ for some } s \in \mathcal{S}, \\ \max(1 - d(t)/|\tilde{t}|, 0), & \text{otherwise,} \end{cases} \quad (15)$$

and the OCR reward is the mean over the prompt’s targets, $r_{\text{ocr}} = \frac{1}{n} \sum_{i=1}^n p(t_i) \in [0, 1]$. Capping the distance at $|\tilde{t}|$ bounds the penalty of a missing or badly garbled string to one full target length, so an absent target scores 0 and a perfect rendering scores 1. For example, for the target “First Place Winner” (normalized to `firstplacewinner`, length $|\tilde{t}| = 16$) an exact reading scores 1, a single dropped character (“First Place Winer”, $d = 1$) scores $1 - 1/16 \approx 0.94$, and absent or unrecognizable text scores 0. The recognizer takes no system prompt: it is queried with a single user message containing the rendered image together with the literal instruction `OCR:`, decoded greedily at temperature 0.

Aesthetic quality (Qwen3.5-27B). Aesthetic-quality prompts are scored by a Qwen3.5-27B [81] judge that grades the generated image against a bank of detailed quality criteria. Each criterion is posed as an independent binary (Yes=1/No=0) question, and the reward is the mean of the answers, so it takes evenly spaced values in $[0, 1]$; a criterion that does not apply, for example a hand-anatomy check on an image with no hands, returns 1 by convention. Grading quality as many focused yes/no items, rather than eliciting a single opaque score, yields a stable and less reward-hackable signal. The judge is driven by the system prompt and per-criterion user template below.

Aesthetic-quality judge: system prompt and user template

System.

You are evaluating whether an AI-generated image satisfies the following visual quality criterion.

You will receive a user prompt that was used to generate the image and a visual quality evaluation criterion.

Your job is to return 1 if the image fully satisfies the criterion, or 0 if it clearly fails. Do not explain or elaborate.

Only output: 1 or 0.

User (one call per criterion).

The prompt of the generated image that needs to be evaluated is: {user_prompt}

Evaluation Criterion ({key}): {criterion}

Scoring Instructions:

- Return 1 if the image fully satisfies the criterion.
- Return 0 if the image clearly fails.

Output: 1 or 0 only

The quality criteria plugged into {criterion} span general photographic quality and detailed human-anatomy checks:

Aesthetic-quality judge: quality criteria

Naturalness:

Image looks like a realistic photograph with correct perspective, natural shadows, and physically plausible lighting.

Detail and Clarity:

Textures, edges, and fine details are sharp and coherent; no smearing, noise, or blur in the main subjects.

No Artifacts:

Image is free from visible distortions, watermarks, digital-rendering glitches, or impossible geometry.

Face and Eye Anatomy:

If a person's face is visible, both eyes are anatomically correct: two matched pupils with iris, symmetric placement, no warped, duplicate, or missing eyes, no asymmetric distortion. If no face is visible, return 1.

Hand and Finger Correctness:

If hands are visible, each hand has exactly five distinct fingers with no fusion, no extra or missing digits, and plausible knuckle structure and finger lengths. If no hand is visible, return 1.

Body Proportions and Limbs:

If a body is shown, limbs (arms, legs) and torso are anatomically plausible: correct count (2 arms + 2 legs), plausible joint articulation, no extra or missing limbs, reasonable head-to-body ratio. If no body is shown, return 1.

Skin Texture Realism:

If human skin is visible, texture shows natural variation (pores, subtle tone differences); not plastic, smeared, over-smoothed, or blurred to a wax-like appearance. If no skin is visible, return 1.

Pose Coherence:

If a person is posed, the pose is physically plausible: limbs do not intersect impossibly, weight distribution makes sense, no floating limbs or contorted joints. If no person is shown, return 1.

Semantic understanding (Qwen3.5-27B). Semantic-understanding prompts are scored by the same Qwen3.5-27B judge under the same binary system prompt and per-criterion user template as the aesthetic reward above; the only difference is that each criterion is a yes/no alignment check derived from the prompt rather than a quality criterion. The judge answers each check independently, and the reward is the fraction answered “yes”. Decomposing alignment into fine-grained faithfulness checks gives a dense, interpretable signal for multi-object scenes. For each prompt, the checks cover:

Semantic-understanding judge: evaluation criteria

- **Objects:** is every object the prompt names present, with none of the required objects missing?
- **Attributes:** does each object have the specified color, material, shape, and state?
- **Counts:** is the exact number of each object present, neither too few nor too many?
- **Spatial relations:** are the objects positioned and arranged relative to one another as the prompt specifies?
- **Actions and interactions:** are the described actions, poses, and interactions between subjects correctly depicted?
- **Scene and setting:** do the background, environment, and overall composition match the prompt?
- **Faithfulness:** is the image free of hallucinated objects, attributes, or changes the prompt did not request?

Instruction-based editing. Editing rollouts are scored by RationalRewards [83], a reasoning reward model that is shown the source image, the instruction, and the edited result, and rates

Table 17 Training data mixture for *Mage-Flow*. The configuration balances downstream domain performance with generic generative capabilities.

Source Dataset	Images	Packs	Weight
SciFormaData-700K (1024px)	646,132	161,536	50%
SciFormaData-700K (768px)	663,353	82,944	20%
Natural scenes	357,009	74,269	15%
Text rendering	74,866	15,360	10%
Poster designs	38,290	10,084	5%
Total	1,779,650	344,193	100%

four aspects on a 1–4 scale: (i) text faithfulness (adherence to the edit instruction); (ii) image faithfulness (preservation of source content outside the edited region); (iii) physical and visual quality (plausibility and freedom from artifacts); and (iv) text rendering (legibility of any edited text, marked not-applicable when the edit involves no text). Writing $a_j \in [1, 4]$ for the applicable aspect scores, the reward linearly maps their mean from $[1, 4]$ onto $[0, 1]$:

$$r_{\text{edit}} = \text{clip}\left(\frac{\bar{a} - 1}{3}, 0, 1\right), \quad \bar{a} = \frac{1}{|\mathcal{A}|} \sum_{j \in \mathcal{A}} a_j, \quad (16)$$

where $\mathcal{A}$ is the set of applicable aspects, so an all-4 edit scores 1 and an all-1 edit scores 0.

D Application: Scientific Diagram Generation

The *Mage-Flow* model features exceptional fine-tuning efficiency, enabling rapid adaptation to highly demanding, structure-critical visual domains. To demonstrate this transferability, we evaluate *Mage-Flow-Base* on scientific diagram generation, a representative task introduced by [105]. Producing these figures serves as a rigorous stress test for generative models, as academic schematics require synchronous correctness across dense text rendering, directional arrow routing, and global layout topology.

Joint Training Strategy. Unlike the original two-stage fine-tuning in SciForma [105], we implement a single-stage data mixture. As shown in Table 17, we construct a balanced data mixture to preserve general generative fidelity while building architectural competence. This joint recipe incorporates the **SciForma-700K** dataset alongside natural scenes, targeted text-rendering subsets, and poster designs.

Optimization Setup. We fine-tune the 4B *Mage-Flow-Base* checkpoint for 130k steps on an 8× NVIDIA B200 GPU machine. The optimization is conducted via Supervised Fine-Tuning (SFT) within the native *Mage-VAE* latent space at a 1024-pixel native-aspect-ratio regime. We apply sequence packing with a fixed length of 20,480 tokens, utilize the AdamW optimizer with a constant learning rate of 1e-5 and a 500-step warmup. An Exponential Moving Average (EMA) with a decay rate of 0.99 is maintained throughout every 100 steps. We denote the resulting checkpoint as *Mage-Flow-SciForma*.

Quantitative Results. We evaluate the fine-tuning efficiency on the **SciFormaBench-2K** evaluation suite [105], which assesses diagram quality across three structural dimensions: component, arrow, and text accuracy.

As outlined in Table 18, our 4B *Mage-Flow-Base* already demonstrates a stronger structural prior than the zero-shot FLUX.2-klein-Base baselines. This advantage is further amplified via fine-tuning; *Mage-Flow-SciForma* achieves an Overall score of **61.61**, outperforming its zero-shot

Table 18 Quantitative results on the SciFormaBench-2K benchmark ($\uparrow$). Metrics are reported on a 0–100 scale. The baseline models are evaluated in a zero-shot manner. *Mage-Flow-SciForma* represents our fine-tuned Model on the data mixture from Table 17.

Model	# Params	Component	Arrow	Text	Overall
FLUX.2-klein-Base (Zero-shot)	4B	50.04	18.45	10.95	27.05
FLUX.2-klein-Base (Zero-shot)	9B	51.50	25.20	23.60	33.87
<i>Mage-Flow-Base</i> (Zero-shot)	4B	56.79	35.82	28.17	40.80
<i>Mage-Flow-SciForma</i> (Ours)	4B	70.49	60.52	52.60	61.61

counterpart (40.80) by 20.81 points, with component and arrow accuracies increased by 13.70 and 24.70 points, respectively. These results validate that the 4B *Mage-Flow-Base* can be efficiently fine-tuned for specialized downstream applications. While this single-stage strategy establishes a viable baseline, the performance can be further enhanced by expanding to multi-stage fine-tuning or incorporating the proposed M-DPO algorithm [105].

Qualitative Analysis. Figure 26 presents the generation results of *Mage-Flow-SciForma* across various scientific diagram archetypes. The results show precise global layouts and granular annotations. Specifically, the model generates neural network architectures with skip-connections, renders unrolled transformer blocks, and routes processing pipelines with directional vectors. Notably, it renders legible mathematical indices and step-by-step text descriptors inside nodes and dense workflows.

Figure 27 compares *Mage-Flow-Base*, *Mage-Flow-SciForma*, and SciForma-9B-Base to evaluate the impact of fine-tuning. While *Mage-Flow-Base* outlines basic layout components, it exhibits limitations in text legibility and arrow semantics. After joint fine-tuning, *Mage-Flow-SciForma* corrects these layout issues, rendering precise text and correct directional arrows. Furthermore, our 4B model produces schematic renderings comparable to SciForma-9B-Base, indicating that lightweight models can close the performance gap against larger domain-specific counterparts.

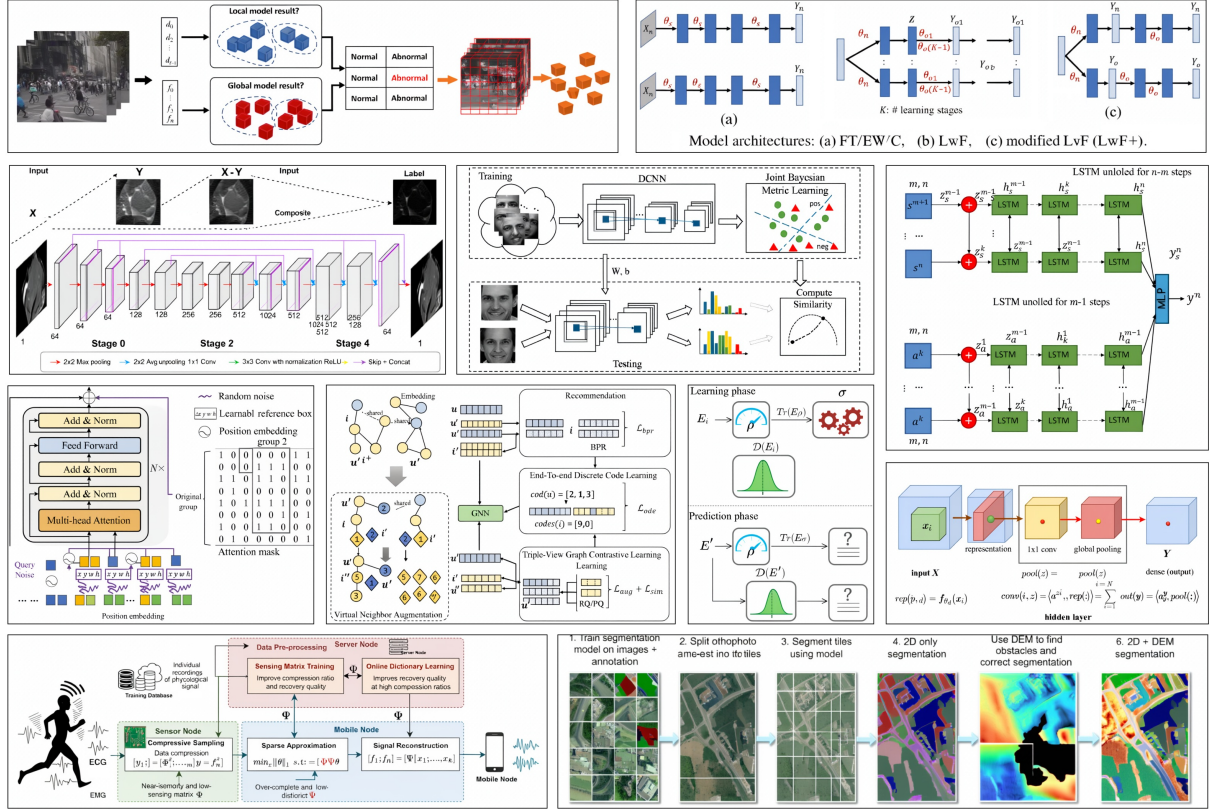


Figure 26 Scientific diagram generation results of *Mage-Flow-SciForma*. Examples include neural network layouts, unrolled temporal modules, graph contrastive frameworks, and text-annotated processing pipelines with legible text and correct directional arrows.

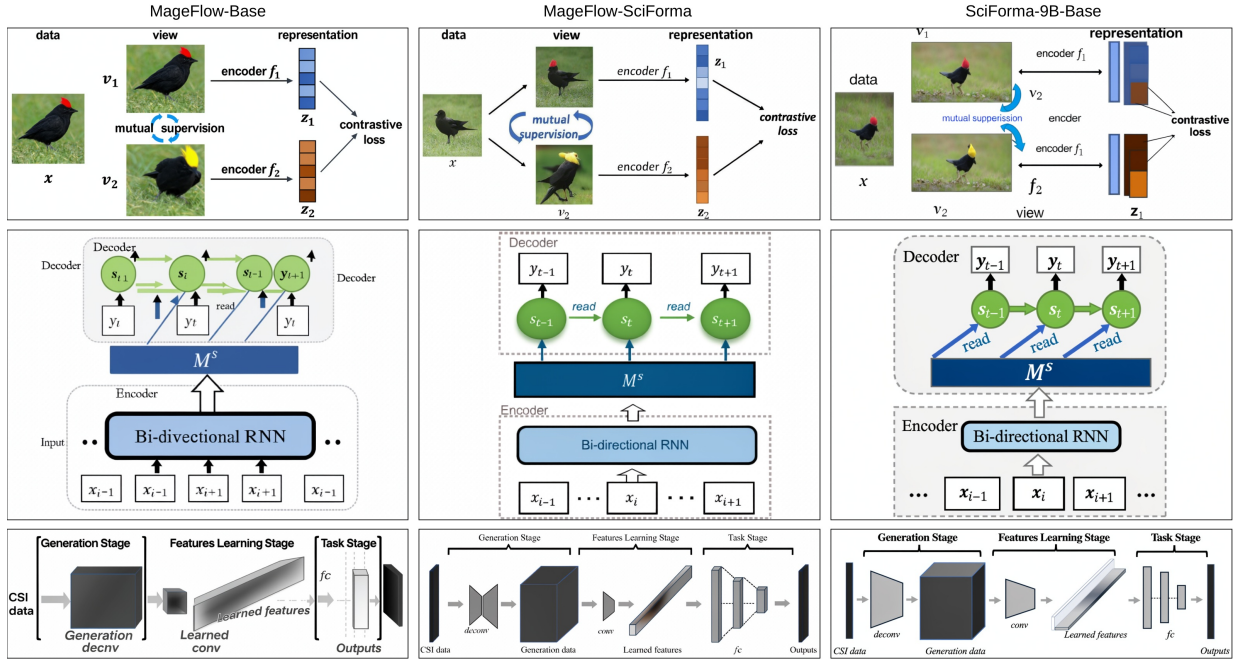


Figure 27 Qualitative comparison on scientific diagram generation. Comparison among zero-shot *MageFlow-Base*, fine-tuned *MageFlow-SciForma*, and *SciForma-9B-Base*. Fine-tuning resolves layout and text errors, allowing our 4B model to match the 9B vertical baseline.

References

- [1] OpenAI. GPT-Image-1. <https://openai.com/index/introducing-4o-image-generation/>, 2025.
- [2] Google. Nano banana pro. <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Image-Model-Card.pdf>, 2025.
- [3] Team Seedream. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.
- [4] Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Feng Yu, Chi Zhang, Shifeng Zhang, and Shilin Zhou. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.
- [5] Chenfei Wu, Jin Bai, Shuai Bai, Zhikai Chen, Kai Dang, Kaiyuan Gao, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.
- [6] Black Forest Labs. FLUX.2: Frontier Visual Intelligence. <https://bfl.ai/blog/flux-2>, 2025.
- [7] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xincheng Deng, et al. HunyuanImage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.
- [8] Super Intelligence Team, Xiaohongshu. FireRed-Image-Edit-1.0 technical report. *arXiv preprint arXiv:2602.13344*, 2026.
- [9] Zhaoyang Jia, Naifu Xue, Zihan Zheng, Jiahao Li, Bin Li, Xiaoyi Zhang, Zongyu Guo, Yuan Zhang, Houqiang Li, and Yan Lu. CoD-Lite: Real-time diffusion-based generative image compression. *arXiv preprint arXiv:2604.12525*, 2026.
- [10] Zidong Wang, Lei Bai, Xiangyu Yue, Wanli Ouyang, and Yiyuan Zhang. Native-resolution image synthesis. *arXiv preprint arXiv:2506.03131*, 2025.
- [11] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.
- [12] Ted Zadouri, Markus Hoehnerbach, Jay Shah, Timmy Liu, Vijay Thakkar, and Tri Dao. Flashattention-4: Algorithm and kernel pipelining co-design for asymmetric hardware scaling. *arXiv preprint arXiv:2603.05451*, 2026.
- [13] Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.
- [14] Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. DiffusionNFT: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025.
- [15] Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Hongsheng Li, and Steven Hoi. Decoupled DMD: CFG augmentation as the spear, distribution matching as the shield. *arXiv preprint arXiv:2511.22677*, 2025.
- [16] Xingtong Ge, Xin Zhang, Tongda Xu, Yi Zhang, Xinjie Zhang, Yan Wang, and Jun Zhang. SenseFlow: Scaling distribution matching for flow-based text-to-image distillation. *arXiv preprint arXiv:2506.00523*, 2025.

- [17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. arXiv:2112.10752.
- [18] 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. arXiv:2307.01952.
- [19] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. arXiv:2212.09748.
- [20] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. arXiv:2403.03206.
- [21] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2025.
- [22] Meituan LongCat Team. LongCat-Image technical report. *arXiv preprint arXiv:2512.07584*, 2025.
- [23] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. arXiv:2108.01073.
- [24] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023. arXiv:2208.01626.
- [25] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. arXiv:2211.09794.
- [26] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. arXiv:2302.05543.
- [27] Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.
- [28] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *CVPR*, 2023. arXiv:2211.09800.
- [29] Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. MagicBrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS*, 2023. arXiv:2306.10012.
- [30] Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. AnyEdit: Mastering unified high-quality image editing for any idea. *arXiv preprint arXiv:2411.15738*, 2024.
- [31] Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. UltraEdit: Instruction-based fine-grained image editing at scale. In *NeurIPS*, 2024. arXiv:2407.05282.
- [32] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. OmniGen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024.

- [33] Black Forest Labs. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.
- [34] Shiyu Liu et al. Step1X-Edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.
- [35] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025.
- [36] Chenyuan Wu, Pengfei Zhou, Shitao Xiao, Jianpeng Zhang, Yong Tang, Zihan Huang, Yiyan Wang, et al. OmniGen2: Towards instruction-aligned multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.
- [37] Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shaodong Wang, Yongyang Ge, Yatian Ning, and Li Yuan. UniWorld-V1: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.
- [38] Bin Xia et al. DreamOmni2: Multimodal instruction-based editing and generation. *arXiv preprint arXiv:2510.06679*, 2025.
- [39] Jiarui Wu et al. ChronoEdit: Towards temporal reasoning for image editing and world simulation. *arXiv preprint arXiv:2510.04290*, 2025.
- [40] Lin Song, Wenbo Li, Guoqing Ma, Wei Tang, Bo Wang, Yuan Zhang, Yijun Yang, Yicheng Xiao, Jianhui Liu, Yanbing Zhang, Guohui Zhang, Wenhui Zhang, Hang Xu, Nan Jiang, Xin Han, Haoze Sun, Maoquan Zhang, Haoyang Huang, and Nan Duan. JoyAI-Image: Awaking spatial intelligence in unified multimodal understanding and generation. *arXiv preprint arXiv:2605.04128*, 2026.
- [41] Xinjie Zhang, Jintao Guo, Shanshan Zhao, Minghao Fu, Lunhao Duan, Jiakui Hu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025.
- [42] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.
- [43] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.
- [44] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.
- [45] Yuying Cui et al. Emu3.5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025.
- [46] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017. arXiv:1711.00937.

- [47] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. arXiv:2012.09841.
- [48] Noam Elata, Tomer Michaeli, and Michael Elad. PSC: Posterior sampling-based compression. In *15th International Conference on Sampling Theory and Applications*, 2025.
- [49] Tongda Xu, Ziran Zhu, Dailan He, Yanghao Li, Lina Guo, Yuanyuan Wang, Zhe Wang, Hongwei Qin, Yan Wang, Jingjing Liu, and Ya-Qin Zhang. Idempotence and perceptual image compression. In *The Twelfth International Conference on Learning Representations*, 2024.
- [50] Anle Ke, Xu Zhang, Tong Chen, Ming Lu, Chao Zhou, Jiawen Gu, and Zhan Ma. Ultra lowrate image compression with semantic residual coding and compression-aware diffusion. *arXiv preprint arXiv:2505.08281*, 2025.
- [51] Jinpei Guo, Yifei Ji, Zheng Chen, Kai Liu, Min Liu, Wang Rao, Wenbo Li, Yong Guo, and Yulun Zhang. Oscar: One-step diffusion codec across multiple bit-rates. *arXiv preprint arXiv:2505.16091*, 2025.
- [52] Naifu Xue, Zhaoyang Jia, Jiahao Li, Bin Li, Yuan Zhang, and Yan Lu. One-step diffusion-based image compression with semantic distillation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
- [53] Tianyu Zhang, Xin Luo, Li Li, and Dong Liu. Stablecodec: Taming one-step diffusion for extreme image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17379–17389, October 2025.
- [54] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *CVPR*, 2024. arXiv:2311.12908.
- [55] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihang Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. *arXiv preprint arXiv:2311.13231*, 2023.
- [56] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *ICLR*, 2024. arXiv:2305.13301.
- [57] 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. arXiv:2304.05977.
- [58] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.
- [59] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. arXiv:2010.02502.
- [60] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. arXiv:2202.00512.
- [61] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023. arXiv:2303.01469.

- [62] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.
- [63] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. InstaFlow: One step is enough for high-quality diffusion-based text-to-image generation. *arXiv preprint arXiv:2309.06380*, 2024.
- [64] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024. arXiv:2311.18828.
- [65] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024. arXiv:2405.14867.
- [66] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.
- [67] Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-SD: Trajectory segmented consistency model for efficient image synthesis. *arXiv preprint arXiv:2404.13686*, 2024.
- [68] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. arXiv:2103.00020.
- [69] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2304.07193.
- [70] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.
- [71] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [72] Yang Ai, Qihang Fan, Xuefeng Hu, Zhenheng Yang, Ran He, and Huaibo Huang. Dico: Revitalizing convnets for scalable and efficient diffusion modeling. *Advances in neural information processing systems*, 38:123582–123618, 2026.
- [73] Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. Deco: Frequency-decoupled pixel diffusion for end-to-end image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 43600–43610, 2026.
- [74] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

- [75] Zhaoyang Jia, Zihan Zheng, Naifu Xue, Jiahao Li, Bin Li, Zongyu Guo, Xiaoyi Zhang, Houqiang Li, and Yan Lu. Cod: A diffusion foundation model for image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 38420–38429, 2026.
- [76] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 34:17480–17492, 2021.
- [77] Challenge on Learned Image Compression. Workshop and challenge on learned image compression (clic 2020). CVPR 2020 Workshop, 2020. URL <https://archive.compression.cc/>.
- [78] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.
- [79] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. *Proc. CVPR*, 2022.
- [80] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.
- [81] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL <https://qwen.ai/blog?id=qwen3.5>.
- [82] Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, et al. Paddleocr-vl-1.5: Towards a multi-task 0.9 b vlm for robust in-the-wild document parsing. *arXiv preprint arXiv:2601.21957*, 2026.
- [83] Haozhe Wang, Cong Wei, Weiming Ren, Jiaming Liu, Fangzhen Lin, and Wenhui Chen. RationalRewards: Reasoning rewards scale visual generation both training and test time. *arXiv preprint arXiv:2604.11626*, 2026.
- [84] Dhruva Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [85] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.
- [86] Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.
- [87] Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *Advances in Neural Information Processing Systems*, 38, 2026.
- [88] Nikai Du, Zhennan Chen, Zhizhou Chen, Shan Gao, Xi Chen, Zhengkai Jiang, Jian Yang, and Ying Tai. Textcrafter: Accurately rendering multiple texts in complex visual scenes. *arXiv e-prints*, pages arXiv–2503, 2025.

- [89] Zigang Geng, Yibing Wang, Yeyao Ma, Chen Li, Yongming Rao, Shuyang Gu, Zhao Zhong, Qinglin Lu, Han Hu, Xiaosong Zhang, et al. X-omni: Reinforcement learning makes discrete autoregressive image generative models great again. *arXiv preprint arXiv:2507.22058*, 2025.
- [90] Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025.
- [91] Kolores Team. Kolores: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024.
- [92] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.
- [93] Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, et al. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing. *arXiv preprint arXiv:2603.09877*, 2026.
- [94] Guo-Hua Wang, Shanshan Zhao, Xinjie Zhang, Liangfu Cao, Pengxin Zhan, Lunhao Duan, Shiyin Lu, Minghao Fu, Xiaohao Chen, Jianshan Zhao, et al. Ovis-u1 technical report. *arXiv preprint arXiv:2506.23044*, 2025.
- [95] Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025.
- [96] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- [97] Krea AI and Black Forest Labs. FLUX.1 Krea [dev]. <https://www.krea.ai/blog/flux-krea-open-source-release>, 2025.
- [98] Lin Song, Wenbo Li, Guoqing Ma, Wei Tang, Bo Wang, Yuan Zhang, Yijun Yang, Yicheng Xiao, Jianhui Liu, Yanbing Zhang, et al. Joyai-image: Awaking spatial intelligence in unified multimodal understanding and generation. *arXiv preprint arXiv:2605.04128*, 2026.
- [99] Microsoft Lens Team. Lens: Rethinking training efficiency for foundational text-to-image models. *arXiv preprint arXiv:2605.21573*, 2026.
- [100] Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.
- [101] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*, <https://cdn.openai.com/papers/dall-e-3.pdf>, 2023.
- [102] Midjourney. Midjourney V7. <https://www.midjourney.com>, 2025.
- [103] Yang Ye, Xianyi He, Zongjian Li, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, Li Yuan, et al. Imgedit: A unified image editing dataset and benchmark. *Advances in Neural Information Processing Systems*, 38, 2026.
- [104] Rui Gui, Yang Wan, Haochen Han, Dongxing Mao, Fangming Liu, Min Li, and Alex Jinpeng Wang. Texteditbench: Evaluating reasoning-aware text editing beyond rendering. *arXiv preprint arXiv:2512.16270*, 2025.

- [105] Yuxuan Luo, Peng Zhang, Xinjie Zhang, Xun Guo, Zhouhui Lian, and Lu Yan. SciForma: Structure-faithful generation of scientific diagrams. *arXiv preprint arXiv:2607.18091*, 2026.