FLUX.1 Kontext：面向潜在空间中上下文图像生成与编辑的流匹配

Flow Matching for In-Context Image Generation and Editing in Latent Space

Black Forest Labs*

Stephen Batifol Andreas Blattmann Frederic Boesel Saksham Consul Cyril Diagne Tim Dockhorn Jack English Zion English Patrick Esser Sumith Kulal Kyle Lacey Yam Levi Cheng Li Dominik Lorenz Jonas Müller Dustin Podell Robin Rombach Harry Saini Axel Sauer Luke Smith

摘要

We present evaluation results for *FLUX.1 Kontext*, a generative flow matching model that unifies image generation and editing. The model generates novel output views by incorporating semantic context from text and image inputs. Using a simple sequence concatenation approach, *FLUX.1 Kontext* handles both local editing and generative in-context tasks within a single unified architecture. Compared to current editing models that exhibit degradation in character consistency and stability across multiple turns, we observe that *FLUX.1 Kontext* improved preservation of objects and characters, leading to greater robustness in iterative workflows. The model achieves competitive performance with current state-of-the-art systems while delivering significantly faster generation times, enabling interactive applications and rapid prototyping workflows. To validate these improvements, we introduce *KontextBench*, a comprehensive benchmark with 1026 image-prompt pairs covering five task categories: local editing, global editing, character reference, style reference and text editing. Detailed evaluations show the superior performance of *FLUX.1 Kontext* in terms of both single-turn quality and multi-turn consistency, setting new standards for unified image processing models.

我们介绍 *FLUX.1 Kontext* 的评测结果。该模型是一个生成式流匹配模型，统一了图像生成和编辑。模型通过结合文本和图像输入的语义上下文来生成新颖的输出图像。使用简单的序列拼接方法，*FLUX.1 Kontext* 能在单一统一架构内处理局部编辑和生成式上下文任务。与当前编辑模型在多轮交互中出现的角色一致性和稳定性退化相比，*FLUX.1 Kontext* 显著改进了对象和角色的保存，在迭代工作流中表现出更强的鲁棒性。该模型在性能上与当前最先进系统相竞争，同时生成速度明显加快，使交互式应用和快速原型设计工作流成为可能。为了验证这些改进，我们引入了 *KontextBench*，一个包含 1026 个图像-提示词对的综合基准，涵盖五个任务类别：局部编辑、全局编辑、角色参考、风格参考和文本编辑。详细评测表明，*FLUX.1*

本文为 AI 辅助翻译版本，仅供学习交流；引用请以英文原文为准：arXiv:2506.15742。

*Please cite this works as "Black Forest Labs (2025)"

Kontext 在单轮质量和多轮一致性两个方面都表现出色，为统一图像处理模型设立了新标准。

1 引言 (Introduction)

Images are a foundation of modern communication and form the basis for areas as diverse as social media, e-commerce, scientific visualization, entertainment, and memes. As the volume and speed of visual content increases, so does the demand for intuitive but faithful and accurate image editing. Professional and casual users expect tools that preserve fine detail, maintain semantic coherence, and respond to increasingly natural language commands. The advent of large-scale generative models has changed this landscape, enabling purely text-driven image synthesis and modifications that were previously impractical or impossible [11, 17, 41, 42, 40, 2, 12, 49, 21].

图像是现代通信的基础，广泛应用于社交媒体、电子商务、科学可视化、娱乐和网络文化等领域。随着视觉内容数量和速度的增加，对直观而忠实准确的图像编辑工具的需求也随之增长。专业用户和普通用户都期待能够保留细节、保持语义连贯，并响应自然语言命令的工具。大规模生成式模型的出现改变了这一格局，使纯文本驱动的图像合成和修改成为现实，这些任务在过去是不切实际或不可能的 [11, 17, 41, 42, 40, 2, 12, 49, 21]。

Traditional image processing pipelines work by directly manipulating pixel values or by applying geometric and photometric transformations under explicit user control [14, 55]. In contrast, generative processing uses deep learning models and their learned representations to synthesize content that seamlessly fits into the new scene. Two complementary capabilities are central to this paradigm

传统图像处理管道通过直接操纵像素值或在用户显式控制下应用几何和光度变换来工作 [14, 55]。相比之下，生成式处理使用深度学习模型及其学到的表示来合成无缝融入新场景的内容。这个范式的核心包括两个互补的能力：

- **局部编辑。** Local, limited modifications that keep the surrounding context intact (e.g. changing the color of a car while preserving the background or replacing the background while keeping the subject in the foreground). Generative inpainting systems such as LaMa [54], Latent Diffusion inpainting [42], RePaint [33], Stable Diffusion Inpainting variants[2], and FLUX.1 Fill[3] make such context-aware edits instantaneous; see also Palette [45] and Paint-by-Example [57]. Beyond inpainting, ControlNet [61] enables mask-guided background replacement, while DragGAN [37] offers interactive point-based geometric manipulation.

  保持周围背景不变的局部、有限的修改（例如在保留背景的情况下改变汽车的颜色，或在保留前景主体的情况下替换背景）。生成式图像修复系统如 LaMa [54]、潜在扩散图像修复 [42]、RePaint [33]、Stable Diffusion Inpainting 变体[4] 和 FLUX.1 Fill[5] 使得这类上下文感知的编辑瞬间完成；另见 Palette [45] 和 Paint-by-Example [57]。除了修复外，ControlNet [61] 支持掩码引导的背景替换，而 DragGAN [37] 提供交互式基于点的几何操纵。

[2] https://huggingface.co/runwayml/stable-diffusion-inpainting
[3] https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev
[4] https://huggingface.co/runwayml/stable-diffusion-inpainting
[5] https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev

(a) **Context image** generated with FLUX.1. **上下文图像**由 FLUX.1 生成。

(b) Image context from 图 1a: *"The bird is now sitting in a bar and enjoying a beer."* 图 1a 的图像上下文："这只鸟现在坐在酒吧里喝啤酒。"

(c) Image context from 图 1b: *"There are now two of these birds."* 图 1b 的图像上下文："现在有两只这样的鸟。"

(d) From 图 1c: *"Watch them from behind."* 图 1c 的后续："从后面看它们。"

(e) From 图 1c: *"The two bird characters are now sitting in a movie theater."* 图 1c 的后续："两个鸟类角色现在坐在电影院里。"

(f) From 图 1c: *"The two bird characters are now grocery shopping."* 图 1c 的后续："两个鸟类角色现在正在杂货店购物。"

(g) From 图 1f: *"The two bird characters are now celebrating a successful launch."* 图 1f 的后续："两个鸟类角色现在正在庆祝成功启动。"

图 1: Consistent character synthesis with *FLUX.1 Kontext*. Generated images can be used iteratively as context for new generations, enabling applications such as storyboard generation and iterative narrative creation. 利用 *FLUX.1 Kontext* 进行一致的角色合成。生成的图像可以迭代用作新生成的上下文，支持故事板生成和迭代叙事创作等应用。

- **生成式编辑。** Extraction of a visual concept (e.g. a particular figure or logo), followed by its faithful reproduction in new environments, potentially synthesized under a new viewpoint or rendering in a new visual context. Similarly to *in-context learning* in large language models, where the network learns a task from the examples provided in the prompt without any parameter updates [6], the generator adapts its output to the conditioning context on the fly. This property enables personalization of generative image and video models without the need for finetuning [43] or LoRA training [19, 27, 20]. Early works on such training-free subject-driven image synthesis include *IP-Adapter* [58] or retrieval-augmented diffusion variants [7, 3].

  提取视觉概念（如特定人物或标志），然后在新环境中忠实地再现它，可能在新视点或新视觉背景下合成。与大型语言模型中的上下文学习类似，网络从提示词中提供的示例学习任务，不需任何参数更新 [6]，生成器能动态调整输出以适应条件背景。此特性支持对生成式图像和视频模型的个性化，无需微调 [43] 或 LoRA 训练 [19, 27, 20]。这类无训练的主体驱动图像合成的早期工作包括 *IP-Adapter* [58] 或检索增强扩散变体 [7, 3]。

近期进展。 InstructPix2Pix [5] and subsequent work [4] demonstrated the promise of synthetic instruction-response pairs for fine-tuning a diffusion model for image editing, while learning-free methods for personalized text-to-image synthesis[13, 43, 24] enable image modification with off-the-shelf, high-performance image generation models [42, 40]. Subsequent instruction-driven editors such as Emu Edit [51], OmniGen [56], HiDream-E1[15] and ICEdit [62] – extend these ideas to refined datasets and model architectures. Huang et al. [20] introduce in-context LoRAs for diffusion transformers on specific tasks, where each task needs to train dedicated LoRA weights. Novel proprietary systems embedded in multimodal LLMs (e.g., GPT-Image[36] and Gemini Native Image Gen [23]) further blur the line between dialog and editing. Generative platforms such as Midjourney[35] and RunwayML[44] integrate these advances into end-to-end creative workflows.

InstructPix2Pix [5] 和后续工作 [4] 证明了合成指令-响应对对于微调扩散模型进行图像编辑的潜力，而学习自由的个性化文本到图像合成方法 [13, 43, 24] 支持使用现成的高性能图像生成模型进行图像修改 [42, 40]。随后的指令驱动编辑器如 Emu Edit [51]、OmniGen [56]、HiDream-E1 [15] 和 ICEdit [62] 将这些思想扩展到更精良的数据集和模型架构。Huang et al. [20] 引入了用于扩散变换器特定任务的上下文 LoRA，其中每个任务需要训练专门的 LoRA 权重。嵌入在多模态大型语言模型中的新型专有系统（如 GPT-Image [36] 和 Gemini Native Image Gen [23]）进一步模糊了对话和编辑之间的界线。生成式平台如 Midjourney [35] 和 RunwayML [44] 将这些进展整合到端到端的创意工作流中。

近期方法的不足。 In terms of results, current approaches struggle with three major shortcomings: (i) instruction-based methods trained on synthetic pairs inherit the shortcomings of their generation pipelines, limiting the variety and realism of achievable edits; (ii) maintaining the accurate appearance of characters and objects across multiple edits remains an open problem, hindering story-telling and brand-sensitive applications; (iii) in addition to lower quality compared to denoising-based approaches, autoregressive editing models integrated into large multimodal systems often come with long runtimes that are incompatible with interactive use.

就结果而言，当前方法存在三个主要不足：(i) 在合成对上训练的指令基方法继承了其生成管道的不足，限制了可实现编辑的多样性和真实性；(ii) 在多次编辑中保持角色和对象的准确外观仍是一个开放问题，妨碍了叙事和品牌敏感应用；(iii) 自回归编辑模型与大型多

(a) Input image 输入图像

(b) *"remove the thing from her face"* "去掉她脸上的东西"

(c) *"she is now taking a selfie in the streets of Freiburg, it's a lovely day out.''* "她现在在弗赖堡街道上自拍，天气很好。''

(d) *"it's now snowing, everything is covered in snow.''* "现在下雪了，一切都被雪覆盖了。''

图 2: **Iterative, instruction-driven editing.** Starting from a reference photo (a), our model successively applies three natural-language edits—first removing an occlusion (b), then relocating the subject to Freiburg (c), and finally transforming the scene into snowy weather (d). Character, pose, clothing, and overall photographic style remain consistent throughout the sequence. **迭代指令驱动编辑**。从参考照片开始 (a)，我们的模型连续应用三次自然语言编辑——首先去掉遮挡物 (b)，然后将对象重新定位到弗赖堡 (c)，最后将场景变换为雪天 (d)。在整个序列中，角色、姿态、服装和整体摄影风格保持一致。

模态系统集成时，除了与基于去噪的方法相比质量较低外，还常常伴随与交互使用不兼容的长运行时间。

我们的方案。We introduce *FLUX.1 Kontext*, a flow-based generative image processing model that matches or exceeds the quality of state-of-the-art black-box systems while overcoming the above limitations. *FLUX.1 Kontext* is a simple flow matching model trained using only a velocity prediction target on a concatenated sequence of context and instruction tokens.

我们介绍 *FLUX.1 Kontext*，一个基于流的生成式图像处理模型，在性能上与当前最先进的黑箱系统相当或超越，同时克服了上述限制。*FLUX.1 Kontext* 是一个简单的流匹配模型，仅在上下文和指令词元的拼接序列上使用速度预测目标进行训练。

In particular, *FLUX.1 Kontext* offers:

特别地，*FLUX.1 Kontext* 具有以下特点：

- **角色一致性：** Character consistency: *FLUX.1 Kontext* excels at character preservation, including multiple, iterative edit turns.

  FLUX.1 Kontext 在角色保存方面表现出色，包括多轮迭代编辑。

- **交互式速度：** Interactive speed: *FLUX.1 Kontext* is *fast*. Both text-to-image and image-to-image application reach speeds for synthesising an image at 1024×1024 of 3–5 seconds.

  FLUX.1 Kontext 速度快。文本到图像和图像到图像应用在 1024×1024 分辨率下都能在 3-5 秒内合成一张图像。

- **迭代应用：** Iterative application: Fast inference and robust consistency allow users to refine an image through multiple successive edits with minimal visual drift.

  快速推理和鲁棒一致性使用户能通过多次连续编辑来完善图像，同时最小化视觉漂移。

2 FLUX.1

FLUX.1 is a rectified flow transformer [12, 30, 32] trained in the latent space of an image autoencoder [42]. We follow Rombach et al. [42] and train a convolutional autoencoder with an adversarial objective from scratch. By scaling up the training compute and using 16 latent channels, we improve the reconstruction capabilities compared to related models; see 第2节. Furthermore, FLUX.1 is built from a mix of double stream and single stream [38] blocks. Double stream blocks employ separate weights for image and text tokens, and *mixing* is done by applying the attention operation over the concatenation of tokens. After passing the sequences through the double stream blocks, we concatenate them and apply 38 single stream blocks to the image and text tokens. Finally, we discard the text tokens and decode the image tokens.

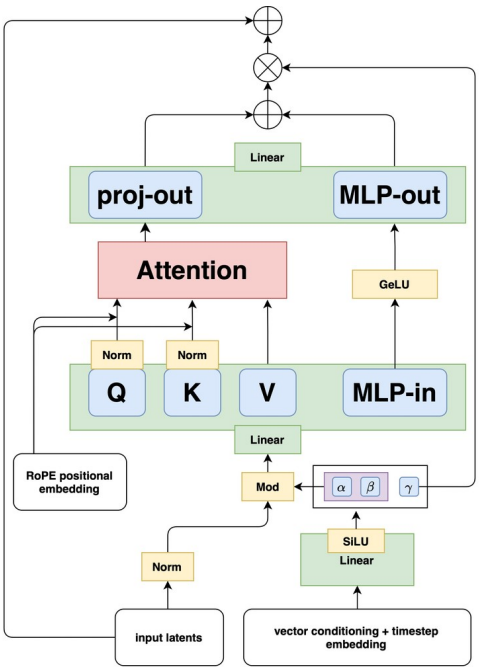


图 3: A fused DiT block equipped with rotary positional embeddings 配备旋转位置编码的融合 DiT 模块

FLUX.1 是一个修正流变换器 [12, 30, 32]，在图像自编码器的潜在空间中训练 [42]。我们遵循 Rombach et al. [42] 从头开始训练一个卷积自编码器，使用对抗目标。通过扩大训练计算并使用 16 个潜在通道，我们改进了与相关模型相比的重建能力；见 第2节。此外，FLUX.1 由双流和单流块的混合组成 [38]。双流块对图像和文本词元使用单独的权重，混合通过对词元拼接应用注意力操作完成。在通过双流块后，我们拼接它们并对图像和文本词元应用 38 个单流块。最后，我们丢弃文本词元并解码图像词元。

To improve GPU utilization of single stream blocks, we leverage *fused* feed-forward blocks in-

模型	PDist ↓	SSIM ↑	PSNR ↑
Flux-VAE	**0.332** ± 0.003	**0.896** ± 0.004	**31.1** ± 0.08
SD3-VAE [12]	0.452 ± 0.004	0.858 ± 0.005	29.6 ± 0.07
SD3-TAE[6]	0.746 ± 0.004	0.774 ± 0.014	27.9 ± 0.06
SDXL-VAE [40]	0.890 ± 0.005	0.748 ± 0.006	25.9 ± 0.07
SD-VAE[7]	0.949 ± 0.005	0.720 ± 0.004	25.0 ± 0.07

表 1: Reconstruction quality comparison across different VAE architectures. All metrics computed on 4096 ImageNet images. Values are mean ± standard error (rounded). See also Appendix B. 不同 VAE 架构的重建质量比较。所有指标在 4096 个 ImageNet 图像上计算。数值为平均值 ± 标准误（四舍五入）。另见 Appendix B。

spired by Dehghani et al. [8], which i) reduce the number of modulation parameters in a feedforward block by a factor of 2 and ii) fuse the attention input- and output linear layers with that of the MLP, leading to larger matrix-vector multiplications and thus more efficient training and inference. We utilize factorized three–dimensional Rotary Positional Embeddings (3D RoPE) [53]. Every latent token is indexed by its space-time coordinates (t, h, w) (with $t \equiv 0$ for single image inputs). See 图 3 for a visualization.

为了改进单流块的 GPU 利用率，我们利用受 Dehghani et al. [8] 启发的融合前馈块，这些块 i) 将前馈块中的调制参数数量减少 2 倍，ii) 将注意力输入-输出线性层与 MLP 融合，导致更大的矩阵-向量乘法，从而更高效的训练和推理。我们使用分解的三维旋转位置嵌入（3D RoPE） [53]。每个潜在词元由其时空坐标 (t, h, w) 索引（单个图像输入时 $t \equiv 0$）。见 图 3 的可视化。

3 FLUX.1 Kontext

Our goal is to learn a model that can generate images conditioned jointly on a text prompt and a reference images. More formally, we aim to approximate the conditional distribution

$$p(x \mid y, c) \tag{1}$$

where x is the target image, y is a context image (or $\varnothing$), and c is a natural-language instruction. Unlike classic text-to-image generation, this objective entails learning *relations between images themselves*, mediated by c, so that the same network can >i) perform image-driven edits when $y \neq \varnothing$, and (ii) create new content from scratch when $y = \varnothing$.

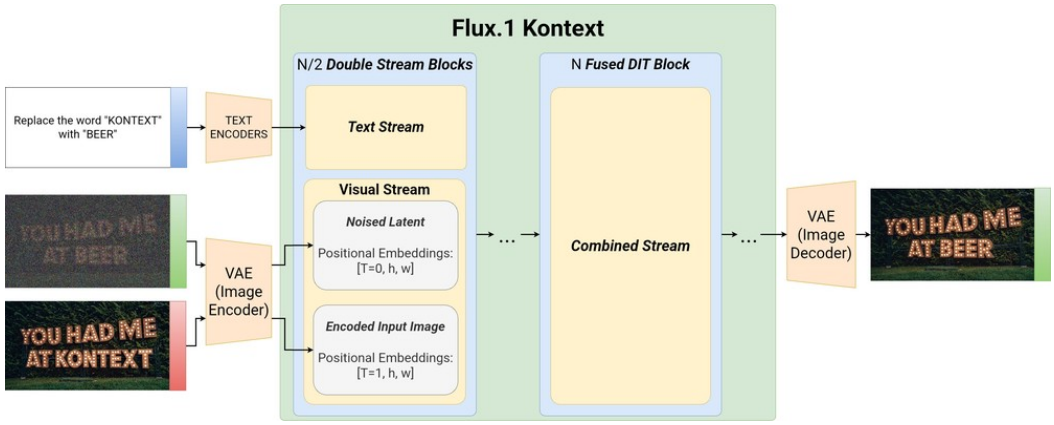


图 4: High-level overview of *FLUX.1 Kontext*, with input and context image on the left. Details in 第3节. *FLUX.1 Kontext* 的高层概览，输入图像和上下文图像在左侧。详见 第3节。

我们的目标是学习一个模型，该模型可以根据文本提示和参考图像的联合条件生成图像。更正式地，我们旨在近似条件分布

$$p(x \mid y, c) \tag{2}$$

其中 x 是目标图像，y 是上下文图像（或 $\varnothing$），c 是自然语言指令。与经典的文本到图像生成不同，这个目标需要学习图像之间的关系，由 c 中介，使得同一网络可以 (i) 在 $y \neq \varnothing$ 时执行图像驱动编辑，(ii) 在 $y = \varnothing$ 时从零开始创建新内容。

To that end, let $x \in \mathcal{X}$ be an output (target) image, $y \in \mathcal{X} \cup \{\varnothing\}$ an optional *context* image, and $c \in \mathcal{C}$ a text prompt. We model the conditional distribution $p_\theta(x \mid y, c)$ such that the same network handles *in-context and local edits* when $y \neq \varnothing$ and free *text-to-image generation* when $y = \varnothing$. Training starts from a FLUX.1 text-to-image checkpoint, and we collect and curate millions of relational pairs $(x \,|\, y, c)$ for optimization. In practice, we do not model images in pixel space but instead encode them into a token sequence as discussed in the following paragraph.

为此，设 $x \in \mathcal{X}$ 为输出（目标）图像，$y \in \mathcal{X} \cup \{\varnothing\}$ 为可选的上下文图像，$c \in \mathcal{C}$ 为文本提示。我们建模条件分布 $p_\theta(x \mid y, c)$，使得同一网络在 $y \neq \varnothing$ 时处理上下文和局部编辑，在 $y = \varnothing$ 时处理自由的文本到图像生成。训练从 FLUX.1 文本到图像检查点开始，我们为优化收集和策展了数百万个关系对 $(x \,|\, y, c)$。在实践中，我们不在像素空间建模图像，而是将其编码为词元序列，如下一段讨论的。

词元序列构造。 Images are encoded into latent tokens by the frozen FLUX auto-encoder. These context image tokens y are then appended to the image tokens x and fed into the visual stream of the model. This simple *sequence concatenation* (i) supports different input/output resolutions and aspect ratios, and (ii) readily extends to multiple images $y_1, y_2, \ldots, y_N$. Channel-wise concatenation of x and y was also tested but in initial experiments we found this design choice to perform worse.

图像由冻结的 FLUX 自编码器编码为潜在词元。这些上下文图像词元 y 然后附加到图像词元 x 并送入模型的视觉流。这种简单的序列拼接 (i) 支持不同的输入/输出分辨率和纵横比，(ii) 易于扩展到多个图像 $y_1, y_2, \ldots, y_N$。我们也测试了 x 和 y 的通道方向拼接，但在初始实验中我们发现这种设计选择性能较差。

We encode positional information via 3D RoPE embeddings, where the embeddings for the context y receive a constant offset for all context tokens. We treat the offset as a *virtual time step* that cleanly separates the context and target blocks while leaving their internal spatial structure intact. Concretely, if a token position is denoted by the triplet $\mathbf{u} = (t, h, w)$, then we set $\mathbf{u}_x = (0, h, w)$ for the target tokens and for context tokens we set

我们通过 3D RoPE 嵌入编码位置信息，其中上下文 y 的嵌入对所有上下文词元接收常数偏移。我们将偏移视为虚拟时间步，能干净地分离上下文和目标块，同时保持其内部空间结构完整。具体来说，如果词元位置由三元组 $\mathbf{u} = (t, h, w)$ 表示，那么我们为目标词元设置 $\mathbf{u}_x = (0, h, w)$，为上下文词元设置

$$\mathbf{u}_{y_i} = (i, h, w), \qquad i = 1, \ldots, N, \tag{3}$$

修正流目标。 We train with a rectified flow–matching loss

$$\mathcal{L}_\theta = \mathbb{E}_{t\sim p(t),\, x, y, c}\big[\, \| v_\theta(z_t, t, y, c) - (\varepsilon - x) \|_2^2 \big], \tag{4}$$

where z_t is the linearly interpolated latent between x and noise $\varepsilon \sim \mathcal{N}(0, 1)$; $z_t = (1-t)x + t\varepsilon$. We use a logit normal shift schedule (see Appendix A.2) for $p(t; \mu, \sigma = 1.0)$, where we change the mode μ depending on the resolution of the data during training. When sampling pure text–image pairs ($y = \varnothing$) we omit all tokens y, preserving the text-to-image generation capability of the model.

我们使用修正流匹配损失进行训练

$$\mathcal{L}_\theta = \mathbb{E}_{t\sim p(t),\, x, y, c}\big[\, \|v_\theta(z_t, t, y, c) - (\varepsilon - x)\|_2^2\big], \tag{5}$$

其中 z_t 是 x 和噪声 $\varepsilon \sim \mathcal{N}(0, 1)$ 之间的线性插值潜在向量；$z_t = (1-t)x + t\varepsilon$。我们为 $p(t; \mu, \sigma = 1.0)$ 使用对数正态移位计划（见 Appendix A.2），其中我们根据训练期间数据的分辨率改变模式 μ。当采样纯文本-图像对（$y = \varnothing$）时，我们省略所有词元 y，保留模型的文本到图像生成能力。

对抗性扩散蒸馏 Sampling of a flow matching model obtained by optimizing 式 (5) typically involves solving an ordinary or stochastic differential equation [30, 1], using 50–250 guided [16] network evaluations. While samples obtained through such a procedure are of good quality for a well-trained model v_Θ, this comes with a few potential drawbacks: First, such multi-step sampling is slow, rendering model-serving at scale expensive and hindering low-latency, interactive applications. Moreover, guidance may occasionally introduce visual artifacts such as over-saturated samples. We tackle both challenges using latent *adversarial diffusion distillation* (LADD) [47, 48, 49], reducing the number of sampling steps while increasing the quality of the samples through adversarial training.

通过优化 式 (5) 得到的流匹配模型的采样通常涉及求解常微分或随机微分方程 [30, 1]，使用 50-250 次引导 [16] 网络评估。虽然通过这样的程序为良好训练的模型 v_Θ 获得的样本质量好，但这带来了几个潜在的缺点：首先，这样的多步采样速度慢，使大规模模型服务成本昂贵，并妨碍低延迟交互应用。此外，引导偶尔会引入视觉伪影，如过饱和样本。我们使用潜在对抗性扩散蒸馏（LADD）[47, 48, 49] 解决这两个挑战，减少采样步数，同时通过对抗训练增加样本质量。

实现细节。 Starting from a pure text-to-image checkpoint, we jointly fine-tune the model on image-to-image and text-to-image tasks following 式 (5). While our formulation naturally covers multiple input images, we focus on single context images for conditioning at this time. *FLUX.1 Kontext* [pro] is trained with the flow objective followed by LADD [48]. We obtain *FLUX.1 Kontext* [dev] through guidance-distillation into a 12B diffusion transformer following the techniques outlined in Meng et al. [34]. To optimize *FLUX.1 Kontext* [dev] performance on edit tasks, we focus exclusively on image-to-image training, i.e. do not train on

the pure text-to-image task for *FLUX.1 Kontext* [dev]. We incorporate safety training measures including classifier-based filtering and adversarial training to prevent the generation of non-consensual intimate imagery (NCII) and child sexual abuse material (CSAM).

从纯文本到图像检查点开始，我们在遵循式 (5) 的图像到图像和文本到图像任务上联合微调模型。虽然我们的公式自然覆盖多个输入图像，但我们目前仅专注于单个上下文图像的条件化。*FLUX.1 Kontext* [pro] 使用流目标训练，其后是 LADD [48]。我们通过指导蒸馏获得 *FLUX.1 Kontext* [dev]，蒸馏进一个 12B 扩散变换器，遵循 Meng et al. [34] 概述的技术。为了优化 *FLUX.1 Kontext* [dev] 在编辑任务上的性能，我们仅专注于图像到图像训练，即不在 *FLUX.1 Kontext* [dev] 上进行纯文本到图像任务训练。我们结合了安全训练措施，包括基于分类器的过滤和对抗训练，以防止生成非自愿私密影像（NCII）和儿童性虐待材料（CSAM）。

We use FSDP2 [29] with mixed precision: all-gather operations are performed in `bfloat16` while gradient reduce-scatter uses `float32` for improved numerical stability. We use selective activation checkpointing [26] to reduce maximum VRAM usage. To improve throughput, we use *Flash Attention 3* [50] and regional compilation of individual Transformer blocks.

我们使用 FSDP2 [29] 与混合精度：all-gather 操作在 `bfloat16` 中执行，而梯度 reduce-scatter 使用 `float32` 以获得改进的数值稳定性。我们使用选择性激活检查点 [26] 来减少最大 VRAM 使用。为了改进吞吐量，我们使用 *Flash Attention 3* [50] 和单个 Transformer 块的区域编译。

4 评估与应用 (Evaluations & Applications)

In this section, we evaluate *FLUX.1 Kontext*'s performance and demonstrate its capabilities. We first introduce KontextBench, a novel benchmark featuring real-world image editing challenges crowd-sourced from users. We then present our primary evaluation: a systematic comparison of *FLUX.1 Kontext* against state-of-the-art text-to-image and image-to-image synthesis methods, where we demonstrate competitive performance

across diverse editing tasks. Finally, we explore *FLUX.1 Kontext*'s practical applications, including iterative editing workflows, style transfer, visual cue editing, and text editing.

在本节中, 我们评估 *FLUX.1 Kontext* 的性能并展示其能力。我们首先介绍 KontextBench, 一个包含从用户众源的真实图像编辑挑战的新型基准。然后我们呈现我们的主要评估: *FLUX.1 Kontext* 与当前最先进的文本到图像和图像到图像合成方法的系统比较, 我们展示了在多样化编辑任务中的竞争性能。最后, 我们探索 *FLUX.1 Kontext* 的实际应用, 包括迭代编辑工作流、风格转移、视觉线索编辑和文本编辑。

4.1 KontextBench – 上下文任务的众源真实世界基准

Existing benchmarks for editing models are often limited when it comes to capturing real-world usage. InstructPix2Pix [5] relies on synthetic Stable Diffusion samples and GPT-generated instructions, creating inherent bias. MagicBrush [60], while using authentic MS-COCO images, is constrained by DALLE-2's [41] capabilities during data collection. Other benchmarks like Emu-Edit [51] use lower-resolution images with unrealistic distributions and focus solely on editing tasks, while DreamBench [39] lacks broad coverage and GEdit-bench [31] does not represent the full scope of modern multimodal models. IntelligentBench [9] remains unavailable with only 300 examples of uncertain task coverage.

现有的编辑模型基准在捕获真实世界使用方面通常受到限制。InstructPix2Pix [5] 依赖于合成 Stable Diffusion 样本和 GPT 生成的指令, 造成内在偏差。MagicBrush [60] 虽然使用真实的 MS-COCO 图像, 但受到数据收集期间 DALLE-2 [41] 能力的限制。其他基准如 Emu-Edit [51] 使用低分辨率、分布不真实的图像并仅专注于编辑任务, 而 DreamBench [39] 覆盖范围不广, GEdit-bench [31] 不代表现代多模态模型的全部范围。IntelligentBench [9] 仍然不可用, 仅有 300 个示例且任务覆盖范围不确定。

To address these gaps, we compile *KontextBench* from crowd-sourced real-world use cases. The benchmark comprises 1026 unique image-prompt pairs derived from 108 base images including personal photos, CC-licensed art, public domain images, and AI-generated content. It spans five core tasks: local instruction editing (416 exam-

ples), global instruction editing (262), text editing (92), style reference (63), and character reference (193). We found that the scale of the benchmark provides a good balance between reliable human evaluation and comprehensive coverage of real-world applications. We will publish this benchmark including the samples of *FLUX.1 Kontext* and all reported baselines.

为了解决这些差距，我们从众源真实世界使用案例编译了 *KontextBench*。该基准由 1026 个独特的图像-提示词对组成，源自 108 个基础图像，包括个人照片、CC-许可艺术、公有领域图像和人工智能生成的内容。它跨越五个核心任务：局部指令编辑（416 个示例）、全局指令编辑（262 个）、文本编辑（92 个）、风格参考（63 个）和角色参考（193 个）。我们发现该基准的规模在可靠的人工评估和全面覆盖真实世界应用之间提供了良好的平衡。我们将发布此基准，包括 *FLUX.1 Kontext* 的样本和所有报告的基线。

4.2 当前最先进方法的比较 (State-of-the-Art Comparison)

FLUX.1 Kontext is designed to perform both text-to-image (T2I) and image-to-image (I2I) synthesis. We evaluate our approach against the strongest proprietary and open-weight models in both domains. We evaluate *FLUX.1 Kontext [pro]* and *[dev]*. As stated above, for *[dev]* we exclusively focus on image-to-image tasks. Additionally, we introduce *FLUX.1 Kontext [max]*, which uses more compute to improve generative performance.

FLUX.1 Kontext 设计用于执行文本到图像（T2I）和图像到图像（I2I）合成。我们针对两个领域中最强的专有和开放权重模型评估我们的方法。我们评估 *FLUX.1 Kontext [pro]* 和 *[dev]*。如上所述，对于 *[dev]*，我们仅专注于图像到图像任务。此外，我们引入 *FLUX.1 Kontext [max]*，它使用更多计算来改进生成性能。

图像到图像结果。 For image editing evaluation, we assess performance across multiple editing tasks: image quality, local editing, *character reference* (CREF), *style reference* (SREF), text editing, and computational efficiency. CREF enables consistent generation of specific characters or objects across novel settings, whereas SREF allows style transfer from reference images while maintaining semantic control. We compare different APIs and

find that our models offer the fastest latency (cf. 图 7b), outperforming related models by up to an order of magnitude in speed difference. In our human evaluation (图 8), we find that *FLUX.1 Kontext [max]* and *[pro]* are the best solution in the categories local and text editing, and for general CREF. We also calculate quantitative scores for CREF, to asses changes in facial characteristics between input and output images we use AuraFace[8] to extract facial embeddings before and after and edit and compare both, see 图 8f. In alignment with our human evaluations, *FLUX.1 Kontext* outperforms all other models. For global editing and SREF, *FLUX.1 Kontext* is second only to gpt-image-1, and Gen-4 References, respectively.

Overall, *FLUX.1 Kontext* offers state-of-the-art character consistency, and editing capabilities, while outperforming competing models such as GPT-Image-1 by up to an order of magnitude in speed.

对于图像编辑评估，我们在多个编辑任务中评估性能：图像质量、局部编辑、角色参考（CREF）、风格参考（SREF）、文本编辑和计算效率。CREF 支持在新环境中一致地生成特定角色或对象，而 SREF 允许从参考图像进行风格转移，同时保持语义控制。我们比较了不同的 API，发现我们的模型提供最快的延迟（参见 图 7b），在速度差异上超过相关模型一个数量级。在我们的人工评估（图 8）中，我们发现 *FLUX.1 Kontext [max]* 和 *[pro]* 是局部和文本编辑以及一般 CREF 类别中的最佳解决方案。我们还计算了 CREF 的定量分数，为了评估输入和输出图像之间的面部特征变化，我们使用 AuraFace[9] 来提取编辑前后的面部嵌入并比较两者，见 图 8f。与我们的人工评估一致，*FLUX.1 Kontext* 超过所有其他模型。对于全局编辑和 SREF，*FLUX.1 Kontext* 分别仅次于 gpt-image-1 和 Gen-4 References。

总的来说，*FLUX.1 Kontext* 提供最先进的角色一致性和编辑能力，同时在速度上超过如 GPT-Image-1 等竞争模型一个数量级。

文本到图像结果。 Current T2I benchmarks predominantly focus on general preference, typically asking questions like *"which image do you prefer?"*. We observe that this broad evaluation criterion often favors a characteristic "AI aesthetic" meaning over-saturated colors, excessive focus on central subjects, pronounced bokeh effects, and convergence toward homogeneous styles. We term this phenomenon *bakeyness*. To address this limitation, we decompose T2I evaluation into five distinct dimensions: prompt following, aesthetic (*"which image do you find more aesthetically pleasing"*), realism (*"which image looks more real"*), typography accuracy, and inference speed. We evaluate on 1 000 diverse test prompts compiled from academic benchmarks (DrawBench [46], PartiPrompts [59]) and real user queries. We refer to this benchmark

[8] https://huggingface.co/fal/AuraFace-v1
[9] https://huggingface.co/fal/AuraFace-v1

as Internal-T2I-Bench in the following. In addition, we complement this benchmark with additional evaluations on GenAI bench [28].

当前的 T2I 基准主要关注一般偏好，通常提出“你更喜欢哪张图像？”这样的问题。我们观察到这种广泛的评估标准经常倾向于特征化的”AI 美学”，意指过饱和颜色、过度关注中心主体、明显的焦外效果以及向同质化风格的收敛。我们称这种现象为 *bakeyness*（”AI 味”）。为了解决这一限制，我们将 T2I 评估分解为五个不同的维度：提示词遵循、美学（“你认为哪张图像在美学上更令人愉悦”）、真实感（“哪张图像看起来更真实”）、排版准确性和推理速度。我们在从学术基准（DrawBench [46]、PartiPrompts [59]）和真实用户查询编译的 1000 个多样化测试提示词上进行评估。我们在下文中将此基准称为 Internal-T2I-Bench。此外，我们用 GenAI bench [28] 上的额外评估来补充此基准。

In T2I, FLUX.1 Kontext demonstrates balanced performance across evaluation categories (see 图 9). Although competing models excel in certain domains, this often comes at the expense of other categories. For instance, Recraft delivers strong aesthetic quality but limited prompt adherence, whereas GPT-Image-1 shows the inverse overall performance pattern. *FLUX.1 Kontext* consistently improves performance across categories over its predecessor FLUX1.1 [pro]. We also observe progressive gains from FLUX.1 Kontext [pro] to FLUX.1 Kontext [max]. We highlight samples in 图14.

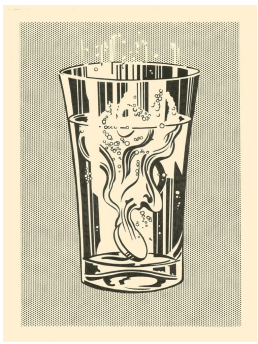
在 T2I 中，FLUX.1 Kontext 在评估类别中展示均衡的性能（见 图 9）。虽然竞争模型在某些领域表现出色，但这常常以其他类别的性能下降为代价。例如，Recraft 提供强大的美学质量但提示词遵循有限，而 GPT-Image-1 显示相反的整体性能模式。*FLUX.1 Kontext* 相比其前身 FLUX1.1 [pro] 在所有类别中一致改进性能。我们还观察到从 FLUX.1 Kontext [pro] 到 FLUX.1 Kontext [max] 的渐进增益。我们在图14 中突出了样本。

4.3 迭代工作流 (Iterative Workflows)

Maintaining character and object consistency across multiple edits is crucial for brand-sensitive and storytelling applications. Current state-of-the-art approaches suffer from noticeable visual drift: characters lose identity and objects lose defining features with each edit. In 图12, we demonstrate

character identity drift across edit sequences produced by *FLUX.1 Kontext*, Gen-4, and GPT-Image-high. We additionally compute the cosine similarity of AuraFace [10, 22] embeddings between the input and images generated via successive edits, highlighting the slower drift of *FLUX.1 Kontext* relative to competing methods. Consistency is essential: marketing needs stable brand characters, media production demands asset continuity, and e-commerce must preserve product details. Applications enabled by *FLUX.1 Kontext*'s reliable consistency are shown in 图10 and 图11.

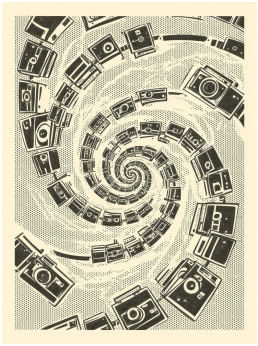
在多次编辑中保持角色和对象的一致性对于品牌敏感和叙事应用至关重要。当前最先进的方法存在明显的视觉漂移问题：角色失去身份，对象随着每次编辑失去定义特征。在 图12 中，我们展示了由 *FLUX.1 Kontext* 、Gen-4 和 GPT-Image-high 生成的编辑序列中的角色身份漂移。我们还计算了输入和通过连续编辑生成的图像之间 AuraFace [10, 22] 嵌入的余弦相似性，突出了 *FLUX.1 Kontext* 相对于竞争方法的更慢漂移。一致性至关重要：营销需要稳定的品牌角色，媒体制作需要资产连续性，电子商务必须保留产品细节。*FLUX.1 Kontext* 的可靠一致性实现的应用如 图10 和 图11 所示。



(a) Input image 输入图像

(b) *"Using this style, a kid on a bicycles rolls through desert ruins, spotlights scanning ancient scrolls projected as holographic sandstorms."* "用这种风格，一个骑自行车的小孩穿过沙漠废墟，聚光灯扫过投影成全息沙尘暴的古老卷轴。"

(c) *"Using this style, a grand piano made of shifting mirrors performs itself for an audience of empty velvet chairs in zero-gravity."* "用这种风格，一架由移动镜子组成的大钢琴为空洞的天鹅绒椅子组成的观众表演，在零重力下。"



(d) *"Using this style, a spiral of vintage cameras captures its own collapse, each flash freeze-framing a different timeline."* "用这种风格，一螺旋的复古相机捕捉了自己的崩溃，每次闪光都冻结了一条不同的时间线。"

(e) Input image 输入图像

(f) *"Using this style, a half-folded metropolis hangs from steel strings over an ink-wash ocean while cranes of light sketch new streets in mid-air."* "用这种风格，一个半折的城市悬挂在钢弦上，悬在墨水洗海洋上方，而光的起重机在半空中勾勒出新的街道。"

(g) *"Using this style, a dapper octopus conducts a jazz duo of owls on a shimmering moonlit bandstand."* "用这种风格，一只得体的章鱼在闪闪发光的月光舞台上指挥猫头鹰的爵士二重奏。"

(h) *"Using this style, a bunny, a dog and a cat are having a tea party seated around a small white table."* "用这种风格，一只兔子、一条狗和一只猫坐在一个小白桌周围举办茶话会。"

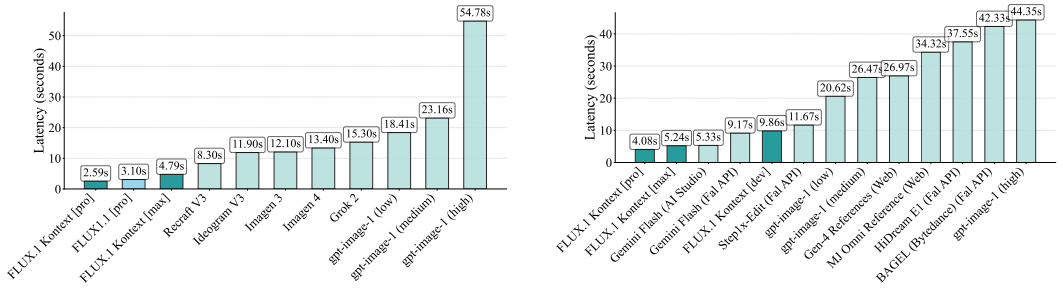
图 5: **Style Reference.** Given an input image, the model extracts its artistic style and applies it to generate diverse new scenes while preserving the original stylistic characteristics. **风格参考。**给定一个输入图像，模型提取其艺术风格并将其应用于生成多种新场景，同时保留原始的风格特征。

(a) Input image 输入图像

(b) *"extract only the skirt over a white background, product photography style"* "仅提取裙子，白色背景，产品摄影风格"

(c) *"show me an extreme closeup of the fabric"* "给我布料的极端特写"

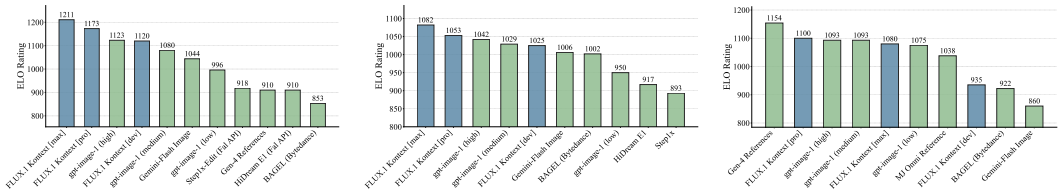
图 6: **Product Photography.** (a) Input image showing the full outfit. (b) Extracted skirt on a white background in a product-photography style. (c) Extreme close-up of the skirt's fabric, highlighting texture and pattern details. **产品摄影**。(a) 显示完整服装的输入图像。(b) 提取的裙子，白色背景，产品摄影风格。(c) 裙子布料的极端特写，突出纹理和图案细节。



(a) Text-to-image inference latency 文本转图像推理延迟

(b) Image-to-image inference latency 图像转图像推理延迟

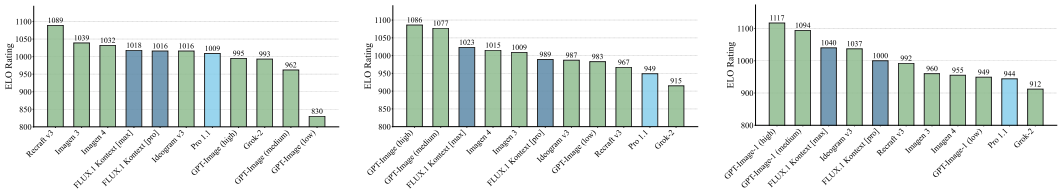
图 7: Median inference latency [seconds] for 1024×1024 generation across models (lower is better). FLUX.1 Kontext achieves competitive speeds for both text-to-image and image-to-image tasks. 所有模型 1024×1024 生成的中位推理延迟 [秒]（越低越好）。FLUX.1 Kontext 在文本转图像和图像转图像任务中都实现了竞争力的速度。



(a) Text Editing 文本编辑　(b) Local Editing 局部编辑　(c) Style Reference 风格参考

(d) Global Editing 全局编辑　(e) Character Reference 角色参考　(f) AuraFace Similarity AuraFace 相似度

图 8: **Image-to-image evaluation on KontextBench**. We show evaluation results across six in-context image generation tasks. *FLUX.1 Kontext* [pro] consistently ranks among the top performers across all tasks, achieving the highest scores in Text Editing and Character Preservation. **在 KontextBench 上的图像转图像评估**。我们展示了六个上下文图像生成任务的评估结果。*FLUX.1 Kontext* [pro] 在所有任务中始终排名靠前，在文本编辑和角色保留方面达到最高分数。



(a) Aesthetics (Internal-T2I-Bench) 美学 (Internal-T2I-Bench)　(b) Prompt Following (Internal-T2I-Bench) 提示词遵循 (Internal-T2I-Bench)　(c) Typography (Internal-T2I-Bench) 排版 (Internal-T2I-Bench)

(d) Realism (Internal-T2I-Bench) 真实感 (Internal-T2I-Bench)　(e) Aesthetics (GenAI) 美学 (GenAI)　(f) Prompt Following (GenAI) 提示词遵循 (GenAI)

图 9: **Text-to-image evaluation on Internal-t2i-bench.** We report evaluation results across multiple quality dimensions. *FLUX.1 Kontext* models demonstrate competitive performance across aesthetics, prompt following, typography, and realism benchmarks. **在 Internal-T2I-Bench 上的文本转图像评估**。我们报告了多个质量维度的评估结果。*FLUX.1 Kontext* 模型在美学、提示词遵循、排版和真实感基准测试中表现出有竞争力的性能。

(a) Input image 输入图像

(b) *"make me a matching flower vase, product photography set against a white wall, sitting on a wooden desk, put some nice flowers in it"* "给我一个匹配的花瓶，产品摄影背景是白墙，放在木制办公桌上，放一些漂亮的花在里面"

(c) *"change the vase base color to black"* "把花瓶底座颜色改成黑色"

图 10: **Iterative, product-style editing.** Starting from the reference bowl (a), our model first generates a matching flower vase in a tabletop studio setting with fresh flowers (b), and subsequently changes the vase' s base color to black while preserving the floral pattern, lighting, and composition (c). **迭代产品风格编辑**。从参考碗开始 (a)，我们的模型首先在桌面工作室设置中生成匹配的花瓶和新鲜花卉 (b)，随后将花瓶的底座颜色改为黑色，同时保留花卉图案、照明和构图 (c)。

(a) Input image 输入图像　(b) *"tilt her head towards the camera"* "把她的头向镜头倾斜"　(c) *"make her laugh"* "让她笑"

图 11: **Sequential, facial-expression editing.** Beginning with the profile reference (a), our model first reorients the subject toward the camera (b) and then changes her expression to a spontaneous laugh (c), while preserving background, clothing and lighting. **顺序面部表情编辑。**从侧面参考开始 (a)，我们的模型首先重新定位被摄体朝向镜头 (b)，然后将她的表情改为自发的笑容 (c)，同时保留背景、服装和照明。

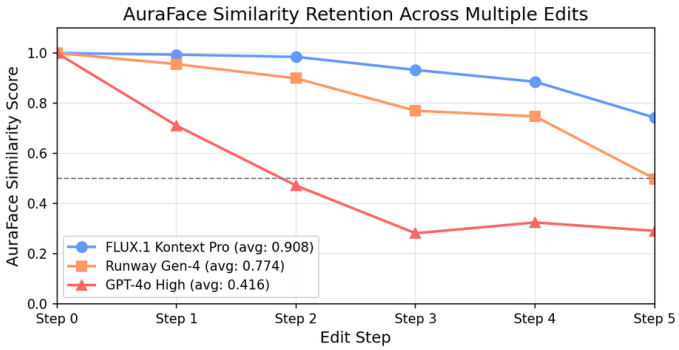


图 12: Iterative editing based on the same starting image with the same prompts and different models (top: *FLUX.1 Kontext*, middle: gpt-image-1, bottom: Runway Gen4). Below are face similarity scores between the edited images and the edited images at different steps. For the last edit ("Add sunglasses"), a relative drop is expected due to partial exclusion of the face. 以同一张起始图像、相同提示词、不同模型进行的迭代编辑（上：*FLUX.1 Kontext*，中：gpt-image-1，下：Runway Gen4）。下方是各步编辑结果之间的人脸相似度分数。最后一步编辑（“加上墨镜”）因人脸被部分遮挡，相似度出现相对下降属于预期。

(a) Input image 输入图像 (b) *"Add hats in the boxes"* "在盒子里加帽子" (c) *Input image* 输入图像 (d) *"Replace 'MONTREAL' with 'FREIBURG'"* "用 *'FREIBURG'* 替换 *'MONTREAL'*" (e) *Input image* 输入图像 (f) *"Replace 'SYNC & BLOOM' with 'FLUX & JOY'"* "用 *'FLUX & JOY'* 替换 *'SYNC & BLOOM'*"

图 13: *FLUX.1 Kontext* is able to leverage visual cues like bounding boxes and edit text while keeping the style. *FLUX.1 Kontext* 能够利用边界框等视觉提示编辑文本，同时保持风格。

图 14: Text-to-image samples by *FLUX.1 Kontext* with low *bakeyness*, diverse styles, and accurate typography. *FLUX.1 Kontext* 生成的文本转图像样本，具有低 bakeyness、多样化的风格和准确的排版。

4.4 专门应用 (Specialized Applications)

FLUX.1 Kontext supports several applications beyond standard generation. Style reference (SREF), first popularized by Midjourney [35] and commonly implemented via IP-Adapters [58], enables style transfer from reference images while maintaining semantic control (see 第4.2节). Additionally, the model supports intuitive editing through visual cues, responding to geometric markers like red ellipses to guide targeted modifications. It also provides sophisticated text editing capabilities, including logo refinement, spelling corrections, and style adaptations while preserving surrounding context. We demonstrate style reference in 图5 and visual cue-based editing in 图13.

FLUX.1 Kontext 支持超越标准生成的多个应用。风格参考（SREF），首次由 Midjourney [35] 推广，通常通过 IP-Adapters [58] 实现，支持从参考图像进行风格转移，同时保持语义控制（见 第4.2节）。此外，该模型通过视觉线索支持直观编辑，响应如红色椭圆等几何标记来引导目标修改。它还提供复杂的文本编辑能力，包括标志优化、拼写更正和风格调适，同时保留周围上下文。我们在 图5 中展示风格参考，在 图13 中展示基于视觉线索的编辑。

5 讨论 (Discussion)

We introduced *FLUX.1 Kontext*, a flow matching model that combines in-context image generation and editing in a single framework. Through simple sequence concatenation and training recipes, *FLUX.1 Kontext* achieves state-of-the-art performance while addressing key limitations such as character drift during multi-turn edits, slow inference, and low output quality. Our contributions include a unified architecture that handles multiple processing tasks, superior character consistency across iterations, interactive speed, and KontextBench: A real-world benchmark with 1 026 image-prompt pairs. Our extensive evaluations reveal that *FLUX.1 Kontext* is comparable to proprietary systems while enabling fast, multi-turn creative workflows.

我们介绍了 *FLUX.1 Kontext*，一个在单一框架中结合上下文图像生成和编辑的流匹配模型。通过简单的序列拼接和训练方法，*FLUX.1 Kontext* 在实现最先进性能的同时解决了关键限制，如多轮编辑期间的角色漂移、缓慢推理和低输出质量。我们的贡献包括处理多个处理任务的统一架构、跨迭代的优越角色一致性、交互式速度，以及 KontextBench：一

个包含 1026 个图像-提示词对的真实世界基准。我们的广泛评估表明，*FLUX.1 Kontext* 与专有系统相当，同时支持快速、多轮创意工作流。

限制。 *FLUX.1 Kontext* exhibits a few limitations in its current implementation. Excessive multi-turn editing can introduce visual artifacts that degrade image quality, see 图 15. The model occasionally fails to follow instructions accurately, ignoring specific prompt requirements. In addition, the distillation process can introduce visual artifacts that impact the fidelity of the output.

FLUX.1 Kontext 在其当前实现中存在一些限制。过度的多轮编辑可能引入视觉伪影，降低图像质量，见 图 15。该模型有时无法准确遵循指令，忽视特定的提示词要求。此外，蒸馏过程可能引入视觉伪影，影响输出的保真度。

未来工作 **Future work** should focus on extending to multiple image inputs, further scaling, and reducing inference latency to unlock real-time applications. A natural extension of our approach is to include edits in the video domain. Most importantly, reducing degradation during multi-turn editing would enable infinitely fluid content creation. The release of *FLUX.1 Kontext* and KontextBench provides a solid foundation and a comprehensive evaluation framework to drive unified image generation and editing.

应专注于扩展到多个图像输入、进一步扩展以及降低推理延迟以解锁实时应用。我们方法的自然扩展是在视频领域包含编辑。最重要的是，减少多轮编辑期间的退化将支持无限流畅的内容创建。*FLUX.1 Kontext* 和 KontextBench 的发布为推动统一的图像生成和编辑提供了坚实的基础和全面的评估框架。

(a) Input image 输入图像

(b) *"The woman is now wearing a green dress, the painting in the back now shows a beach scene, the text on the TV says "Kontext" now"* "女性现在穿着绿色连衣裙，背面的画现在显示海滩场景，电视上的文字现在说"*Kontext*""

(c) *"...green dress...beach scene...tv says "Kontext"...table cloth is also now red and the lighting is a bit warmer"* "... 绿色连衣裙... 海滩场景... 电视说"*Kontext*"... 台布现在也是红色的，照明有点更温暖"

(d) Input image 输入图像

(e) *"move the coffee to the left"* "把咖啡移到左边"

(f) Input image 输入图像

(g) After six iterative edits. 六次迭代编辑后。

图 15: Editing failure cases. *Top row:* An example for identity degradation: While the center image shows a good edit, preserving character identity the right one (using a slightly modified prompt) comes with significant identity loss. *Middle row:* Scene modification instead of object movement: the model adds milk foam rather than repositioning the mug. *Bottom row:* After six iterative edits, samples can exhibit visible artifacts. 编辑失败案例。顶行：身份退化的示例：中心图像显示了一次好的编辑，保留了角色身份，但右侧的编辑（使用略微修改的提示词）会伴随显著的身份丧失。中间行：场景修改而不是对象移动：模型添加了奶泡而不是重新定位杯子。底行：经过六次迭代编辑后，样本可能出现可见的伪影。

A 使用流匹配的图像生成

A.1 修正流匹配入门

For training our models, we construct forward noising processes in the latent space of an image autoencoder as

$$z_t = a_t x_0 + b_t \varepsilon, \tag{6}$$

with $x_0 \sim p_{data}$, $\varepsilon \sim \mathcal{N}(0, 1)$, and the coefficients a_t and b_t define the log signal-to-noise ratio (log-SNR) [25]

$$\lambda_t = \log \frac{a_t^2}{b_t^2} \tag{7}$$

Further, we use the conditional flow matching loss [30]

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t\sim p(t),\varepsilon\sim\mathcal{N}(0,1)} ||v_\Theta(z_t, t) - \frac{a_t'}{a_t} z_t + \frac{b_t}{2}\lambda_t' \varepsilon||_2^2 \tag{8}$$

For rectified flow models [32], $a_t = 1 - t$ and $b_t = t$, and thus

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t\sim p(t),\varepsilon\sim\mathcal{N}(0,1),x_0\sim p_{data}} ||v_\Theta(z_t, t) + x_0 - \varepsilon||_2^2 \tag{9}$$

and we sample t from a *Logit-Normal Distribution* [12]: $p(t) = \frac{\exp\left(-0.5\cdot(\text{logit}(t)-\mu)^2/\sigma^2\right)}{\sigma\sqrt{2\pi}\cdot(1-t)\cdot t}$, where $\text{logit}(t) = \log \frac{t}{1-t}$. From the definition of the Logit-Normal Distribution, it follows that a random variable $Y = \text{logit}(t) \sim \mathcal{N}(\mu, \sigma)$.

对于训练我们的模型，我们在图像自编码器的潜在空间中构造前向噪声过程

$$z_t = a_t x_0 + b_t \varepsilon, \tag{10}$$

其中 $x_0 \sim p_{data}$，$\varepsilon \sim \mathcal{N}(0, 1)$，系数 a_t 和 b_t 定义对数信噪比（log-SNR） [25]

$$\lambda_t = \log \frac{a_t^2}{b_t^2} \tag{11}$$

此外，我们使用条件流匹配损失 [30]

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t\sim p(t),\varepsilon\sim\mathcal{N}(0,1)} ||v_\Theta(z_t, t) - \frac{a_t'}{a_t} z_t + \frac{b_t}{2}\lambda_t' \varepsilon||_2^2 \tag{12}$$

对于修正流模型 [32]，$a_t = 1 - t$ 和 $b_t = t$，因此

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t\sim p(t),\varepsilon\sim\mathcal{N}(0,1),x_0\sim p_{data}} ||v_\Theta(z_t, t) + x_0 - \varepsilon||_2^2 \tag{13}$$

我们从对数正态分布 [12] 采样 t: $p(t) = \frac{\exp\left(-0.5\cdot(\text{logit}(t)-\mu)^2/\sigma^2\right)}{\sigma\sqrt{2\pi}\cdot(1-t)\cdot t}$，其中 $\text{logit}(t) = \log \frac{t}{1-t}$。根据对数正态分布的定义，随机变量 $Y = \text{logit}(t) \sim \mathcal{N}(\mu, \sigma)$。

A.2 通过对数正态分布表达时间步计划的移位

Previous work on high-resolution image synthesis introduced an additional shift of the timestep sampling (and, equivalently, the log-SNR schedule) via a parameter α [12, 18]. Esser et al. [12] empirically demonstrated that $\alpha = 3.0$ worked best when increasing the image resolution from 256^2 to 1024^2. In the following, we show that this shifting can be expressed via the Logit-Normal Distribution.

之前关于高分辨率图像合成的工作通过参数 α 引入了时间步采样的额外移位（等价地，log-SNR 计划） [12, 18]。Esser et al. [12] 实证证明当将图像分辨率从 256^2 增加到 1024^2 时，$\alpha = 3.0$ 最有效。以下，我们展示这种移位可以通过对数正态分布表达。

Consider the log-SNR of a rectified flow forward process with $\mu = 0$ and $\sigma = 1$:

$$\lambda_t^{0,1} = 2 \log \frac{1-t}{t} = -2\text{logit}(t), \tag{14}$$

where $\text{logit}(t) \sim \mathcal{N}(0, 1)$. Expressing the log-SNR for arbitrary μ and σ gives

$$\lambda_t^{\mu,\sigma} = -2(\sigma \cdot \text{logit}(t) + \mu) = \sigma \cdot \lambda_t^{0,1} - 2\mu\,. \tag{15}$$

The α-shifted log-SNR [12, 18] is obtained as

$$\lambda_t^{\alpha} = \lambda_t^{0,1} - 2 \log \alpha\,. \tag{16}$$

Comparing 式 (18) and 式 (19), we identify $\mu = \log \alpha$ for $\sigma = 1.0$, i.e. a shift of $\alpha = 3.0$ would correspond to a logit-normal distribution with $\mu = \log 3.0 = 1.0986$ and $\sigma = 1.0$.

考虑修正流前向过程的 log-SNR，其中 $\mu = 0$ 和 $\sigma = 1$：

$$\lambda_t^{0,1} = 2 \log \frac{1-t}{t} = -2\text{logit}(t), \tag{17}$$

其中 $\text{logit}(t) \sim \mathcal{N}(0, 1)$。为任意 μ 和 σ 表达 log-SNR 给出

$$\lambda_t^{\mu,\sigma} = -2(\sigma \cdot \text{logit}(t) + \mu) = \sigma \cdot \lambda_t^{0,1} - 2\mu\,. \tag{18}$$

α 移位的 log-SNR [12, 18] 获得为

$$\lambda_t^{\alpha} = \lambda_t^{0,1} - 2 \log \alpha\,. \tag{19}$$

We can further express the shifted log-SNR as a function of shifted timesteps t'

$$\lambda_{t'} = 2 \log \frac{1-t'}{t'} = \sigma \lambda_t^{0,1} - 2\mu = 2\sigma \log \frac{1-t}{t} - 2\mu \tag{20}$$

and solve for t':

$$t' = \frac{e^{\mu}}{e^{\mu} + (1/t - 1)^{\sigma}} \tag{21}$$

For $\sigma = 1.0$ and $\mu = \log \alpha$ this recovers the redistribution function for the timesteps proposed in [12] $t' = \frac{\alpha t}{1+(\alpha-1)t}$, as expected. This generalized shifting formula 18 can be useful both for training and via 23 for inference.

比较 式 (18) 和 式 (19)，我们识别当 $\sigma = 1.0$ 时 $\mu = \log \alpha$，即 $\alpha = 3.0$ 的移位将对应于 $\mu = \log 3.0 = 1.0986$ 和 $\sigma = 1.0$ 的对数正态分布。

我们可以进一步将移位的 log-SNR 表达为移位时间步 t' 的函数

$$\lambda_{t'} = 2 \log \frac{1-t'}{t'} = \sigma \lambda_t^{0,1} - 2\mu = 2\sigma \log \frac{1-t}{t} - 2\mu \tag{22}$$

并求解 t'：

$$t' = \frac{e^{\mu}}{e^{\mu} + (1/t - 1)^{\sigma}} \tag{23}$$

对于 $\sigma = 1.0$ 和 $\mu = \log \alpha$，这恢复了在 [12] 中为时间步提出的重新分布函数 $t' = \frac{\alpha t}{1+(\alpha-1)t}$，如预期。这个广义移位公式 18 对于训练和通过 23 对于推理都可能有用。

B VAE 评估细节

We compare our VAE with related models using three reconstruction metrics, namely, SSIM, PSNR, and the *Perceptual Distance* (PDist) in VGG [52] feature space. All metrics are computed over 4 096 random ImageNet evaluation images at resolution 256×256. 第2节 shows the mean and the standard deviation over the 4 096 inputs.

我们使用三个重建指标将我们的 VAE 与相关模型进行比较，即 SSIM、PSNR 和 VGG [52] 特征空间中的感知距离（PDist）。所有指标在 256 × 256 分辨率下的 4096 个随机 ImageNet 评估图像上计算。第2节 显示了 4096 个输入上的平均值和标准差。

参考文献

[1] Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants, 2022. 10

[2] 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*, 2(3), 2023. 2

[3] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022. 4

[4] Frederic Boesel and Robin Rombach. Improving image editing models with generative data refinement. In *Tiny Papers @ ICLR*, 2024. URL https://api.semanticscholar.org/CorpusID:271461432. 4

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 4, 12

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 4

[7] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 4

[8] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023. 7

[9] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 12

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 16

[11] Patrick Esser, Robin Rombach, Andreas Blattmann, and Bjorn Ommer. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in neural information processing systems*, 34:3518–3532, 2021. 2

[12] 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, 2024. URL https://arxiv.org/abs/2403.03206. 2, 6, 7, 27, 28

[13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 4

[14] Rafael C Gonzalez. *Digital image processing*. Pearson education india, 2009. 2

[15] HiDream-ai. Hidream-e1: Instruction-based image editing model, 2025. URL https://github.com/HiDream-ai/HiDream-E1. 4

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 10

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 2

[18] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. Simple diffusion: End-to-end diffusion for high resolution images, 2023. 27, 28

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4

[20] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024. 4

[21] Imagen-Team-Google, :, Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluis Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, Hongliang Fei, Nando de Freitas, Yilin Gao, Evgeny Gladchenko, Sergio Gómez Colmenarejo, Mandy Guo, Alex Haig, Will Hawkins, Hexiang Hu, Huilian Huang, Tobenna Peter Igwe, Christos Kaplanis, Siavash Khodadadeh, Yelin Kim, Ksenia Konyushkova, Karol Langner, Eric Lau, Rory Lawton, Shixin Luo, Soňa Mokrá, Henna Nandwani, Yasumasa Onoe, Aäron van den Oord, Zarana Parekh, Jordi Pont-Tuset, Hang Qi, Rui Qian, Deepak Ramachandran, Poorva Rane, Abdullah Rashwan, Ali Razavi, Robert Riachi, Hansa Srinivasan, Srivatsan Srinivasan, Robin Strudel, Benigno Uria, Oliver Wang, Su Wang, Austin Waters, Chris Wolff,

Auriel Wright, Zhisheng Xiao, Hao Xiong, Keyang Xu, Marc van Zee, Junlin Zhang, Katie Zhang, Wenlei Zhou, Konrad Zolna, Ola Aboubakar, Canfer Akbulut, Oscar Akerlund, Isabela Albuquerque, Nina Anderson, Marco Andreetto, Lora Aroyo, Ben Bariach, David Barker, Sherry Ben, Dana Berman, Courtney Biles, Irina Blok, Pankil Botadra, Jenny Brennan, Karla Brown, John Buckley, Rudy Bunel, Elie Bursztein, Christina Butterfield, Ben Caine, Viral Carpenter, Norman Casagrande, Ming-Wei Chang, Solomon Chang, Shamik Chaudhuri, Tony Chen, John Choi, Dmitry Churbanau, Nathan Clement, Matan Cohen, Forrester Cole, Mikhail Dektiarev, Vincent Du, Praneet Dutta, Tom Eccles, Ndidi Elue, Ashley Feden, Shlomi Fruchter, Frankie Garcia, Roopal Garg, Weina Ge, Ahmed Ghazy, Bryant Gipson, Andrew Goodman, Dawid Górny, Sven Gowal, Khyatti Gupta, Yoni Halpern, Yena Han, Susan Hao, Jamie Hayes, Jonathan Heek, Amir Hertz, Ed Hirst, Emiel Hoogeboom, Tingbo Hou, Heidi Howard, Mohamed Ibrahim, Dirichi Ike-Njoku, Joana Iljazi, Vlad Ionescu, William Isaac, Reena Jana, Gemma Jennings, Donovon Jenson, Xuhui Jia, Kerry Jones, Xiaoen Ju, Ivana Kajic, Christos Kaplanis, Burcu Karagol Ayan, Jacob Kelly, Suraj Kothawade, Christina Kouridi, Ira Ktena, Jolanda Kumakaw, Dana Kurniawan, Dmitry Lagun, Lily Lavitas, Jason Lee, Tao Li, Marco Liang, Maggie Li-Calis, Yuchi Liu, Javier Lopez Alberca, Matthieu Kim Lorrain, Peggy Lu, Kristian Lum, Yukun Ma, Chase Malik, John Mellor, Thomas Mensink, Inbar Mosseri, Tom Murray, Aida Nematzadeh, Paul Nicholas, Signe Nørly, João Gabriel Oliveira, Guillermo Ortiz-Jimenez, Michela Paganini, Tom Le Paine, Roni Paiss, Alicia Parrish, Anne Peckham, Vikas Peswani, Igor Petrovski, Tobias Pfaff, Alex Pirozhenko, Ryan Poplin, Utsav Prabhu, Yuan Qi, Matthew Rahtz, Cyrus Rashtchian, Charvi Rastogi, Amit Raul, Ali Razavi, Sylvestre-Alvise Rebuffi, Susanna Ricco, Felix Riedel, Dirk Robinson, Pankaj Rohatgi, Bill Rosgen, Sarah Rumbley, Moonkyung Ryu, Anthony Salgado, Tim Salimans, Sahil Singla, Florian Schroff, Candice Schumann, Tanmay Shah, Eleni Shaw, Gregory Shaw, Brendan Shillingford, Kaushik Shivakumar, Dennis Shtatnov, Zach Singer, Evgeny Sluzhaev, Valerii Sokolov, Thibault Sottiaux, Florian Stimberg, Brad Stone, David Stutz, Yu-Chuan Su, Eric Tabellion, Shuai Tang, David Tao, Kurt Thomas, Gregory Thornton, Andeep Toor, Cristian Udrescu, Aayush Upadhyay, Cristina Vasconcelos, Alex Vasiloff, Andrey Voynov, Amanda Walker, Luyu Wang, Miaosen Wang, Simon Wang, Stanley Wang, Qifei Wang, Yuxiao Wang, Ágoston Weisz, Olivia Wiles, Chenxia Wu, Xingyu Federico Xu, Andrew Xue, Jianbo Yang, Luo Yu, Mete Yurtoglu, Ali Zand, Han Zhang, Jiageng Zhang, Catherine Zhao, Adilet Zhaxybay, Miao Zhou, Shengqi Zhu, Zhenkai Zhu, Dawn Bloxwich, Mahyar Bordbar, Luis C. Cobo, Eli Collins, Shengyang Dai, Tulsee Doshi, Anca Dragan, Douglas Eck, Demis Hassabis, Sissie Hsiao, Tom Hume, Koray Kavukcuoglu, Helen King, Jack Krawczyk, Yeqing Li, Kathy Meier-Hellstern, Andras Orban, Yury Pinsky, Amar Subramanya, Oriol Vinyals, Ting Yu, and Yori Zwols. Imagen 3, 2024. URL https://arxiv.org/abs/2408.07009. 2

[22] isidentical. Introducing auraface: Open-source face recognition and identity preservation models. https://huggingface.co/blog/isidentical/auraface, 2024. Accessed: 2025-05-26. 16

[23] Kat Kampf and Nicole Brichtova. Experiment with gemini 2.0 flash native image generation, 2025. URL https://developers.googleblog.com/en/experiment-with-gemini-20-flash-native-image-generation/. 4

[24] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models, 2023. URL https://arxiv.org/abs/2210.09276. 4

[25] Diederik P Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 27

[26] Vijay Anand Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. Reducing activation recomputation in large transformer models. *Proceedings of Machine Learning and Systems*, 5: 341–353, 2023. 11

[27] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 4

[28] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024. 15

[29] Wanchao Liang, Tianyu Liu, Less Wright, Will Constable, Andrew Gu, Chien-Chin Huang, Iris Zhang, Wei Feng, Howard Huang, Junjie Wang, et al. Torchtitan: One-stop pytorch native solution for production ready llm pre-training. *arXiv preprint arXiv:2410.06511*, 2024. 11

[30] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t. 6, 10, 27

[31] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025. 12

[32] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. 6, 27

[33] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 2

[34] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023. 10, 11

[35] Midjourney. Midjourney, 2025. URL https://www.midjourney.com/home. 4, 24

[36] OpenAI. Introducing 4o image generation, 2025. URL https://openai.com/index/introducing-4o-image-generation/. 4

[37] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 2

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023. doi: 10.1109/iccv51070.2023.00387. URL http://dx.doi.org/10.1109/ICCV51070.2023.00387. 6

[39] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation, 2025. URL https://arxiv.org/abs/2406.16855. 12

[40] 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, 2023. 2, 4, 7

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2, 12

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022. doi: 10.1109/cvpr52688.2022.01042. URL http://dx.doi.org/10.1109/CVPR52688.2022.01042. 2, 4, 6

[43] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023. URL https://arxiv.org/abs/2208.12242. 4

[44] Inc. Runway AI. Runway | tools for human imagination, 2025. URL https://runwayml.com/. 4

[45] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 14, 15

[47] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 2021. 10

[48] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023. 10, 11

[49] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation, 2024. URL https://arxiv.org/abs/2403.12015. 2, 10

[50] Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685, 2024. 11

[51] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. *arXiv preprint arXiv:2311.10089*, 2023. 4, 12

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 29

[53] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063, 2024. 7

[54] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 2

[55] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022. 2

[56] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024. 4

[57] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18381–18391, 2023. 2

[58] Hu Ye, Jun Zhang, Sibo 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. 4, 24

[59] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. URL https://arxiv.org/abs/2206.10789. 14, 15

[60] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing, 2024. URL https://arxiv.org/abs/2306.10012. 12

[61] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2

[62] 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. 4