Paragraph
graph

NextStep-1：迈向大规模连续标记的自回归图像生成

NextStep-Team, StepFun

Homepage: https://stepfun.ai/research/en/nextstep1
Github: https://github.com/stepfun-ai/NextStep-1
Huggingface: NextStep-1 Collections

Abstract

Prevailing autoregressive (AR) models for text-to-image generation either rely on heavy, computationally-intensive diffusion models to process continuous image tokens, or employ vector quantization (VQ) to obtain discrete tokens with quantization loss. In this paper, we push the autoregressive paradigm forward with **NextStep-1**, a 14B autoregressive model paired with a 157M flow matching head, training on discrete text tokens and continuous image tokens with next-token prediction objectives. NextStep-1 achieves state-of-the-art performance for autoregressive models in text-to-image generation tasks, exhibiting strong capabilities in high-fidelity image synthesis. Furthermore, our method shows strong performance in image editing, highlighting the power and versatility of our unified approach. To facilitate open research, we will release our code and models to the community.

1. 介绍

在大型语言模型中，自回归模型取得了显著的成功（Brown et al., 2020; OpenAI, 2025a; Radford et al., 2018, 2019），这激发了将其扩展到文本生成图像的任务。通过将多模态输入统一成一个序列，自回归图像生成模型（Chen et al., 2025b; Fan et al., 2024; Sun et al., 2023, 2024b,c; Wang et al., 2024b; Yu et al., 2022）提供了一种可扩展且灵活的方法，用于进行文本生成图像，能够自然而然地适应各种条件信号。

然而，大多数现有的文本到图像生成自回归方法（Chen et al., 2025b; Dong et al., 2024; Sun et al., 2024a,b; Tong et al., 2024; Wang et al., 2024b）要么依赖于复杂的扩散模型，要么采用矢量量化（VQ）（Eslami et al., 2021; Yu et al., 2023; Zheng et al., 2022）将图像标记化为离散视觉标记，这面临包括曝光偏差（Han et al., 2025）和次优图像标记化（Li et al., 2024c）在内的限制。尽管最近使用连续潜表示（Fan et al., 2024; Li et al., 2024c）的努力显示出了希望，但自回归模型与最先进的扩散方法（Esser et al., 2024; Labs, 2024; Podell et al., 2024）之间仍存在显著性能差距，尤其是在图像质量和一致性方面。

在本文中，我们引入了 NextStep-1 ，这是一种基于下一步令牌预测范式的简单但有效的自回归模型，在文本到图像生成任务中实现了最先进的性能。全面的评估确认了其在一系列具有挑战性的基准测试中的竞争力表现。具体来说，NextStep-1 在 WISE（Niu et al., 2025）中取得了 0.54，在 GenAI-Bench（Lin et al., 2024）的高级提示中取得了 0.67，在 DPG-Bench（Hu et al., 2024）中取得了 85.28，在 OneIG-Bench（Chang et al., 2025）的英语提示中取得了 0.417。这些结果展示了其在多种场景中的能力，从短提示和长提示到需要世界知识的任务。除了生成之外，NextStep-1 的多功能性也通过其在基于指令的图像编辑中的出色表现得到了验证，NextStep-1-Edit ，在 GEdit-Bench（Liu et al., 2025）的英语提示中取得了 6.58 的竞赛分数，在 ImgEdit-Bench（Ye et al., 2025）中取得了 3.71。我们在 Fig. 1 中展示了其定性表现。

NextStep-1 是一个由 14 亿个参数组成的自回归模型，它由一个 Transformer 骨干网络、用于离散文本标记的标准语言建模头、用于连续图像标记的轻量流匹配头和一个图像标记器组成。

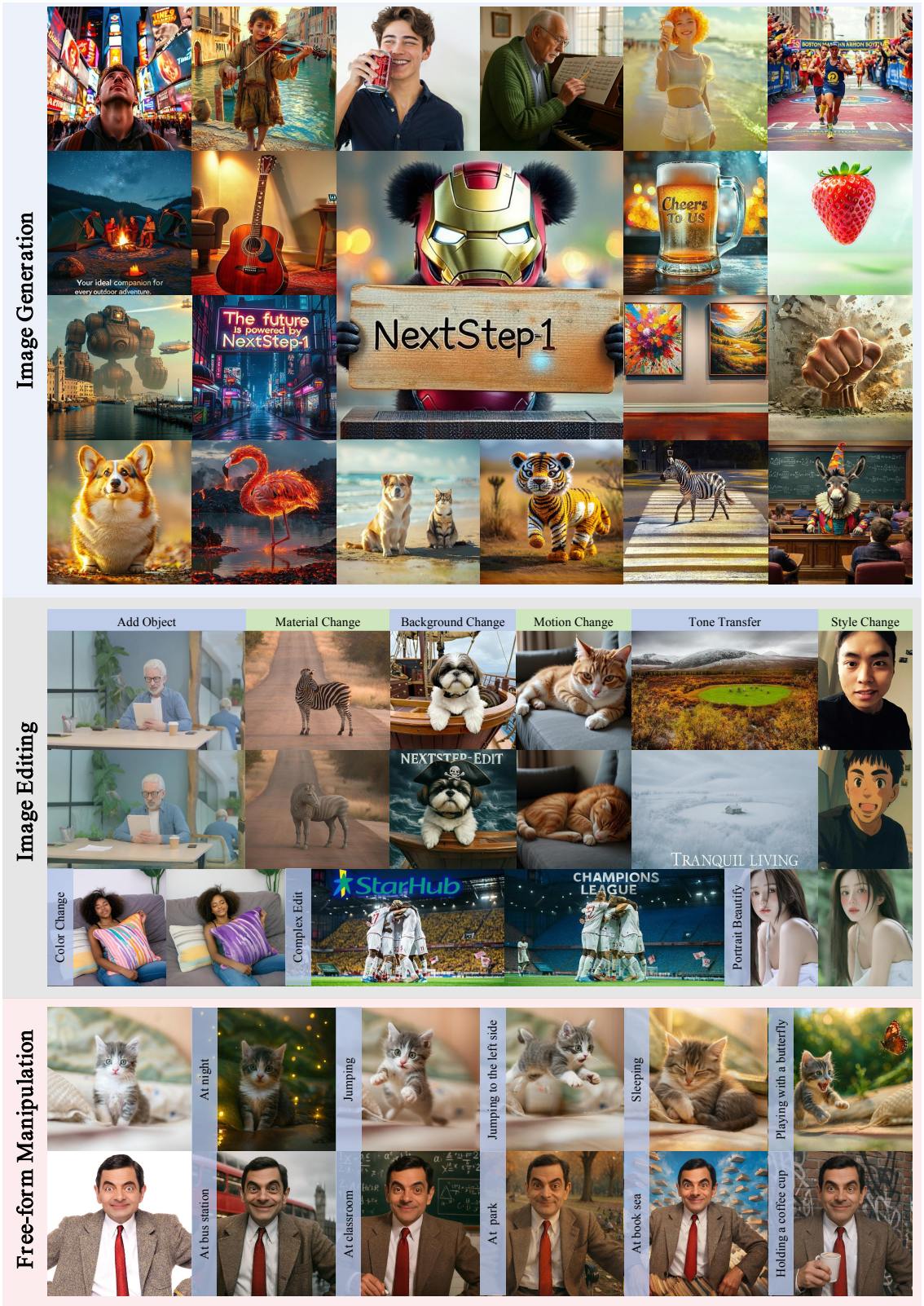

Figure 1 | 高保真图像生成、多样化图像编辑和复杂自由形式操作中的 NextStep-1 概述。

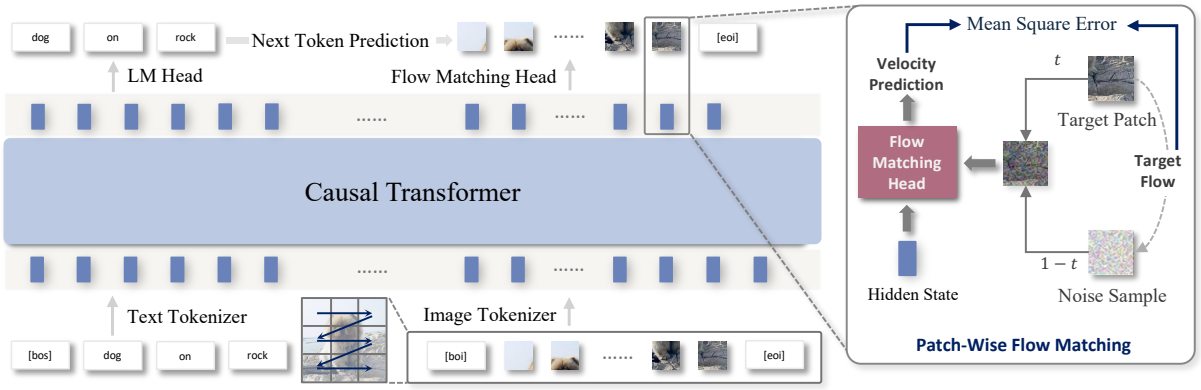

Figure 2 | 是关于 NextStep-1 框架的概述。NextStep-1 采用了一种因果变压器来处理标记化的文本和图像令牌。在训练过程中，流匹配头根据输出的隐藏状态预测从噪声样本到下一个目标图像块的连续流。在推理时，这允许通过迭代地引导噪声来生成图像，以创建下一个图像块。

流匹配头是一个基于 MLP 的 1.57 亿参数模型，利用流匹配目标进行训练，遵循了 (Li et al., 2024c) 的方法。在自回归建模中，高维潜在空间对于实现高图像质量至关重要，但往往会导致训练不稳定和发散。我们的图像标记器通过增强连续图像标记的鲁棒性和促进良好分散、归一化的潜在空间，解决了这一权衡，从而确保即使在更高维度（例如，16 个通道）下也能稳定收敛。实验证实，这种设计对于 16 通道潜在空间的稳定和有效训练是至关重要的。

2. 框架

2.1. 使用连续视觉标记的统一多模型生成

NextStep-1 通过简单而直观的架构将成熟的自回归语言建模范式扩展到图像生成，如 Fig. 2 所示。为了将多模态输入统一到单一序列中，图像将通过图像分词器被标记化为连续的图像标记，并与离散的文本标记结合。假设 $x = \{x_0, x_1, ..., x_n\}$ 是多模态标记序列，其中 x_i 是离散的文本标记或连续的视觉标记，则在统一序列下的自回归目标形式化为：

$$p(x) = \prod_{i=1}^{n} p(x_i \mid x_{<i}). \tag{1}$$

。统一的多模态生成任务通过从网络建模的条件分布 $p(x_i \mid x_{<i})$ 中采样下一个标记 x_i 来进行。离散的文本标记是通过语言建模头进行采样的，而连续的图像标记是通过流匹配头进行采样的。

我们的训练目标由两个不同的损失组成：一个用于离散文本标记的标准交叉熵损失，和一个用于连续图像标记的流匹配损失 (Lipman et al., 2023b) 。具体来说，流匹配损失是预测速度向量与目标速度向量之间的均方误差，这些向量将噪声块映射到其对应的干净块。通过优化这两个损失的加权和来端到端地训练模型：

$$\mathcal{L}_{\text{total}} = \lambda_{\text{text}} \mathcal{L}_{\text{text}} + \lambda_{\text{visual}} \mathcal{L}_{\text{visual}} \tag{2}$$

，其中 $\mathcal{L}_{\text{text}}$ 和 $\mathcal{L}_{\text{visual}}$ 分别表示文本和图像标记的损失，这些损失通过超参数 λ_{text} 和 λ_{visual} 进行平衡。

2.2. 模型架构

我们的图像分词器在 flux VAE (Labs, 2024) 的基础上进行了微调，仅使用重建和感知损失。分词器首先将图像编码为 16 通道的潜变量 z ，并应用 $8 \times$ 的空间降采样因子。为了稳定和规范化

潜在空间，我们应用逐通道规范化，将每个通道标准化为零均值和单位方差。此外，为了增强图像分词器的鲁棒性并鼓励更均匀的潜在分布，我们在标准化的潜变量上引入了随机扰动。这种技术是从 σ -VAE (Sun et al., 2024c) 中改编而来的，该技术用于防止方差崩溃。

$$\tilde{z} = \text{Normlization}(z) + \alpha \cdot \varepsilon, \quad \text{where } \alpha \sim \mathcal{U}[0,\ \gamma] \text{ and } \varepsilon \sim \mathcal{N}(0,\ I) \tag{3}$$

其中 ε 是标准高斯噪声，其幅度由从 $[0,\ \gamma\ ]$ 均匀采样的随机因子 α 缩放。γ 是一个控制最大噪声强度的超参数。图像分词器的潜变量被像素洗牌成一个更紧凑的序列。通过 2×2 核的空间到深度转换实现，这将 2×2 空间潜变量展平成通道维度。例如，这会将一个 256×256 图像的潜变量转换成 16×16 的 64 通道标记网格。然后，这个网格被展平成 256 个标记的一维序列，以作为接下来因果变换器的输入。

我们从仅解码器的 Qwen2.5-14B (Yang et al., 2024) 初始化我们的模型，利用其强大的语言理解和推理能力进行文本到图像生成。我们将多模态输入序列组织成如下格式：

{ text } <image_area>h*w <boi> { image } <eoi>...

其中，{ text } 表示离散文本标记，而 { image } 表示连续图像标记。<boi> 和 <eoi> 是标记图像开始和结束的特殊标记。<image_area>h*w 表示关于二维图像标记空间尺寸的元数据。

然后从 LLM 输出的隐藏状态传递到两个轻量级头部，以进行特定模态的损失：

- 语言建模头。我们计算文本隐藏状态的交叉熵损失。
- 逐块流匹配头。根据 (Li et al., 2024c)，我们使用每个逐块图像的隐藏状态作为条件，在时间步 t 对目标块去噪，并通过一个 157M，12 层，1536 隐藏维度的 MLP 来计算逐块流匹配损失 (Lipman et al., 2023a)。

对于位置信息，我们使用标准的一维 RoPE (Su et al., 2024)。尽管有更复杂的二维或多模态 RoPE 替代方案 (Bai et al., 2025; Wang et al., 2024a)，我们发现简单的一维公式在混合文本-图像序列中仍然非常有效，因此为了简单性和效率而保留它。

3. 数据

为了全面装备我们的模型，使其具有广泛和多功能的能力，我们构建了一个由四个主要数据类别组成的多样化训练语料库：纯文本语料库、图文数据、图像到图像的数据、以及交错数据。每个类别都经过精心策划，以扮演不同的角色，培养模型生成能力的不同方面。

3.1. 纯文本语料库

为了保留大型语言模型（LLM）中固有的广泛语言能力，我们在训练过程中纳入了从步骤 3 至 (Wang et al., 2025a) 中抽取的 4000 亿文本标记。

3.2. 图文配对数据

由图像-文本对组成的数据构成了模型从文本生成图像能力的基础。我们开发了一个综合的流程，从多样化的初始来源中策划出一个高质量的大规模数据集。

1. 数据获取：我们从多种来源收集了大规模数据集，包括网络数据、多任务 VQA 数据和富文本文档。
2. 基于质量的筛选：我们随后进行了一项严格的筛选过程，从美学质量、水印存在、清晰度、OCR 检测和文本-图像语义对齐度等方面评估每幅图像。

3. 重新添加标题：在对筛选后的图像去重后，我们使用 Step-1o-turbo [1] 为每张图像生成丰富且详细的中英文标题。

这个多阶段流程最终生成了一个包含 5.5 亿高质量图像-文本对的数据集，为训练一个具有强美感和广泛世界知识的模型提供了基础。

为了实现广泛的实际应用，我们策划了一个高质量的数据集，用于指令指导的图像到图像任务，如视觉感知（Kirillov et al., 2023）、可控图像生成（Zhang et al., 2023b）、图像修复（Labs, 2025）、通用图像编辑（Peng et al., 2024）等。

对于视觉感知和可控图像生成任务，我们通过将 ControlNet（Zhang et al., 2023b）的注释器应用于我们的一部分高质量图像-文本对数据，合成了 100 万样本。对于图像恢复和通用图像编辑，我们收集了 350 万样本，包括来自 GPT-Image-Edit（Wang et al., 2025c）、Step1X-Edit（Liu et al., 2025）和一个专有内部数据集的数据。按照 Step1X-Edit（Liu et al., 2025），所有编辑数据都经过严格的基于 VLM 的过滤流程，该流程评估图像对质量、合理性、一致性和指令对齐性，从而产生约 100 万高质量的指令引导的图像到图像数据用于训练。

3.3. 交错数据

交错数据无缝集成了文本和图像，为不同模态提供了丰富且细腻的序列关联。具体而言，我们知识丰富的交错数据集主要由四个不同类别构成：通用视频交错数据、教程、以角色为中心的场景和多视角数据。

为了赋予我们的模型广泛的世界知识，我们首先构建了一个大规模的，包含 8000 万样本的视频交错数据集。这是通过一个精细的策划流程实现的，该流程受 Step-Video（Ma et al., 2025a）的启发，涵盖了帧提取、去重和字幕生成。此外，遵循 mmtextbook（Zhang et al., 2025）的方法，我们通过利用 ASR 和 OCR 工具收集和处理教程视频。这个组件特别针对文本丰富的真实场景，增强了模型在上下文中的文本理解和生成能力。Fig. 3 中详细介绍的一个关键贡献是我们的以角色为中心的数据集，NextStep-Video-Interleave-5M 。对于这个数据集，我们提取了围绕特定角色的视频帧，并生成了类似（Oliveira and de Matos, 2025）讲故事风格丰富的字幕，从而显著改善了模型的多轮互动能力。最后，为了加强几何推理，我们从两个开源数据集 MV-ImageNet-v2（Han et al., 2024）和 Objaverse-XL（Deitke et al., 2023）中策划了多视图数据，这增强了模型维护多视图一致性的能力。

4. 训练方案

4.1. 训练图像标记器

我们的图像分词器初始化自 Flux.1-dev VAE（Labs, 2024），这是因为该模型在重建性能方面表现出色。我们在 Sec. 3.2 中详述的图像-文本数据集上微调了该模型，以使其适应我们的特定数据分布。为了优化，我们采用了 AdamW 优化器（Loshchilov and Hutter, 2019），并使用 $(\beta_1 = 0.9, \beta_2 = 0.95, \varepsilon = 1 \times 10^{-8})$，以确保收敛的稳定性。模型训练了 50K 步，总批次大小为 512，使用常数学习率 1×10^{-5}，在训练开始前进行 1000 步的线性预热。

4.2. 预训练

我们在 Tab. 1 中详细介绍了预训练的具体超参数和数据比例。具体而言，预训练遵循一个三阶段课程，旨在逐步完善模型的能力。在这些阶段中，除预训练的图像分词器外，所有模型参数都是端到端训练的。

[1] https://platform.stepfun.com/docs/llm/vision

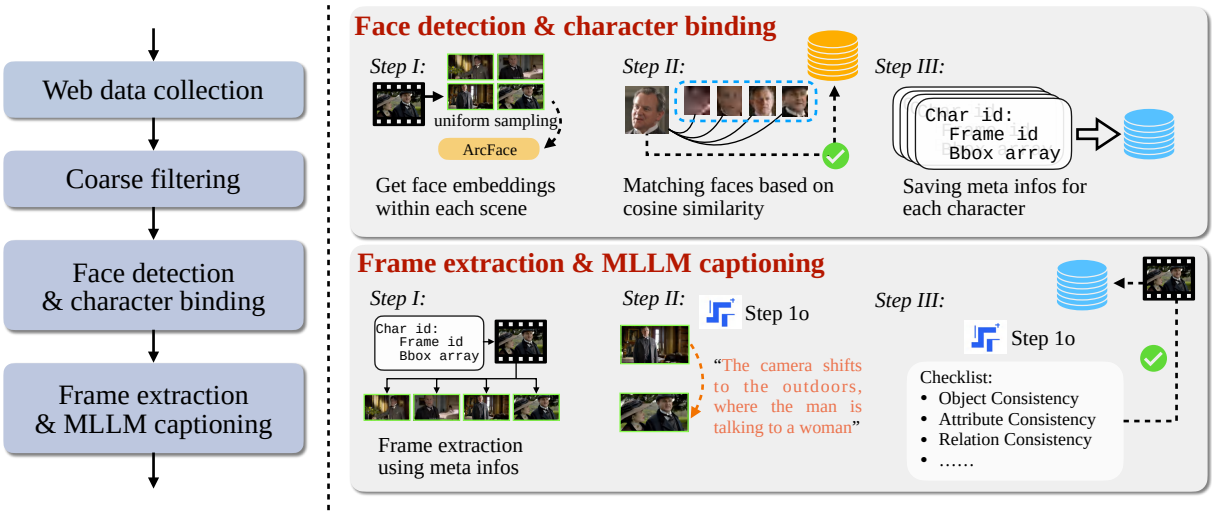

Figure 3 | 以字符为中心的数据处理。

Table 1 | NextStep-1 的 训练方案。

	Pre-Training			Post-Training	
	Stage1	Stage2	Annealing	SFT	DPO
Hyperparameters					
Learning Rate (Min, Max)	1×10^{-4}	1×10^{-5}	$(0, 1\times10^{-5})$	$(0, 1\times10^{-5})$	2×10^{-6}
LR Scheduler	Constant	Constant	Cosine	Cosine	Constant
Weight Decay	0.1	0.1	0.1	0.1	0.1
Loss Weight (CE : MSE)	(0.01 : 1)	(0.01 : 1)	(0.01 : 1)	(0.01 : 1)	-
Training Steps	200K	100K	20K	10K	300
Warm-up Steps	5K	5K	0	500	200
Sequence Length per Rank	16K	16K	16K	8K	-
Image Area (Min, Max)	256 × 256	(256 × 256, 512 × 512)	(256 × 256, 512 × 512)	(256 × 256, 512 × 512)	(256 × 256, 512 × 512)
Image Tokens (Min, Max)	256	(256, 1024)	(256, 1024)	(256, 1024)	(256, 1024)
Training Tokens	1.23T	0.61T	40B	5B	-
Data Ratio					
Text-only Corpus	0.2	0.2	0.2	0	-
Image-Text Pair Data	0.6	0.6	0.6	0.9	-
Image-to-Image Data	0.0	0.0	0.1	0.1	-
Interleaved Data	0.2	0.2	0.1	0	-

在这个初始阶段，模型学习对图像结构和构图的基础理解。为提高计算效率，所有图像均被调整尺寸并随机裁剪到固定的 256 × 256 的分辨率。训练课程由多样的数据混合组成：20 % 仅文本语料，60 % 图像-文本对，以及 20 % 交错数据。该阶段大约消耗了 1.23T 的 tokens。

阶段 2。我们采用动态分辨率策略，将模型训练在一系列更高的分辨率，目标是 256 × 256 和 512 × 512 基础区域。此策略利用不同的纵横比桶以提高计算效率。在这一阶段，我们通过更多富含文本和视频交错的数据丰富数据混合，利用模型增强的能力以这些分辨率处理精细细节。

在预训练的最后阶段，我们进行退火阶段，以提高模型在高度精选的数据集上的能力。这是通过在一个高质量的 2000 万样本子集上训练一个轮次来实现的，这些样本通过对美学评分、图像清晰度、语义相似性、水印等应用更严格的过滤阈值从 Sec. 3.2 中选择出来。这个退火步骤显著提高了模型的最终输出，增强了整体图像结构、构图、纹理和美学吸引力。

4.3. 后训练

在广泛的语料库上进行预训练以建立通用模型之后，后续训练用于使模型输出与人类偏好和下游任务对齐。我们通过两阶段过程实现这种对齐：监督微调（SFT）随后是直接偏好优化（DPO）(Rafailov et al., 2023) 。每个阶段的超参数在 Tab. 1 中。

Table 2 | 对 GenEval（Ghosh et al., 2023）、GenAI-Bench（Lin et al., 2024）和 DPG-Bench（Hu et al., 2024）上的图像-文本对齐的比较。* 结果为重写后的结果。† 结果为自我链式思考（Self-CoT）结果。

Method	GenEval ↑	GenAI-Bench ↑		DPG-Bench ↑
		Basic	Advanced	
Proprietary				
DALL-E 3 (Betker et al., 2023)	0.67	0.90	0.70	83.50
Seedream 3.0 (Gao et al., 2025)	0.84	-	-	88.27
GPT4o (OpenAI, 2025b)	0.84	-	-	85.15
Diffusion				
Stable Diffusion 1.5 (Rombach et al., 2022)	0.43	-	-	-
Stable Diffusion XL (Podell et al., 2024)	0.55	0.83	0.63	74.65
Stable Diffusion 3 Medium (Esser et al., 2024)	0.74	0.88	0.65	84.08
Stable Diffusion 3.5 Large (Esser et al., 2024)	0.71	0.88	0.66	83.38
PixArt-Alpha (Chen et al., 2024)	0.48	-	-	71.11
Flux.1-dev (Labs, 2024)	0.66	0.86	0.65	83.79
Transfusion (Zhou et al., 2025)	0.63	-	-	-
CogView4 (Z.ai, 2025)	0.73	-	-	85.13
Lumina-Image 2.0 (Qin et al., 2025)	0.73	-	-	87.20
HiDream-I1-Full (Cai et al., 2025)	0.83	0.91	0.66	85.89
Mogao (Liao et al., 2025)	0.89	-	0.68	84.33
BAGEL (Deng et al., 2025)	0.82/ 0.88 †	0.89/0.86 †	0.69/ 0.75 †	85.07
Show-o2-7B (Xie et al., 2025b)	0.76	-	-	86.14
OmniGen2 (Wu et al., 2025b)	0.80/0.86*	-	-	83.57
Qwen-Image (Wu et al., 2025a)	0.87	-	-	88.32
AutoRegressive				
SEED-X (Ge et al., 2024)	0.49	0.86	0.70	-
Show-o (Xie et al., 2024)	0.53	0.70	0.60	-
VILA-U (Wu et al., 2024)	-	0.76	0.64	-
Emu3 (Wang et al., 2024b)	0.54/0.65*	0.78	0.60	80.60
Fluid (Fan et al., 2024)	0.69	-	-	-
Infinity (Han et al., 2025)	0.79	-	-	86.60
Janus-Pro-7B (Chen et al., 2025b)	0.80	0.86	0.66	84.19
Token-Shuffle (Ma et al., 2025b)	0.62	0.78	0.67	-

监督微调 (SFT)。SFT 阶段增强了模型的指令跟随能力，并使其输出与人类偏好一致。SFT 数据集总共由 500 万样本组成，分为三个部分：1）一个语义一致性高且视觉吸引力强的有人类选择的图文对语料库，通过加入其他生成模型的图像，以蒸馏方式提升模型处理复杂和富有想象力的提示的能力；2）链式思维（CoT）数据（Deng et al., 2025; Wei et al., 2022），通过在最终图像创建前加入基于语言的推理步骤，改进文本到图像的生成；3）来自 **??** 的高质量指令引导的图像到图像数据，以增强模型的图像编辑能力。

为了使我们的模型符合人类偏好，我们采用了直接策略优化（DPO）（Rafailov et al., 2024），这是一种受扩散-DPO（Wallace et al., 2024）启发的方法。为此，我们从经过挑选的大约 20,000 个不同的提示中构建了两种不同类型的偏好数据集。

1. 标准 DPO 数据集：对于每个提示 c，我们直接使用 SFT 模型生成 16 个候选图像。这些图像然后由 ImageReward（Xu et al., 2023）进行评分，形成偏好对 (y^w, y^l)，其中获胜图像 y^w 从前 4 个候选者中随机抽取，而失败图像 y^l 从剩余的 12 个中随机抽取。
2. Self-CoT DPO 数据集：为了增强模型的推理能力，我们引入了一个明确的推理步骤。对于每个提示 c，我们首先提示模型生成详细的文本化 CoT，然后将其扩展到原始提示中。使用这个增强了 CoT 的提示，我们按照与上述相同的流程来形成一个偏好对 (y^w, y^l)。

Table 3 | 在 OneIG-Bench（Chang et al., 2025）上的英文提示比较。

Method	Alignment	Text	Reasoning	Style	Diversity	Overall ↑
Proprietary						
Imagen3 (Baldridge et al., 2024)	0.843	0.343	0.313	0.359	0.188	0.409
Recraft V3 (team, 2024)	0.810	0.795	0.323	0.378	0.205	0.502
Kolors 2.0 (team, 2025)	0.820	0.427	0.262	0.360	0.300	0.434
Seedream 3.0 (Gao et al., 2025)	0.818	0.865	0.275	0.413	0.277	0.530
Imagen4 (deepmind Imagen4 team, 2025)	0.857	0.805	0.338	0.377	0.199	0.515
GPT-4o (OpenAI, 2025b)	0.851	0.857	0.345	0.462	0.151	0.533
Diffusion						
Stable Diffusion 1.5 (Rombach et al., 2022)	0.565	0.010	0.207	0.383	0.429	0.319
Stable Diffusion XL (Podell et al., 2024)	0.688	0.029	0.237	0.332	0.296	0.316
Stable Diffusion 3.5 Large (Stability-AI, 2024)	0.809	0.629	0.294	0.353	0.225	0.462
Flux.1-dev (Labs, 2024)	0.786	0.523	0.253	0.368	0.238	0.434
CogView4 (Z.ai, 2025)	0.786	0.641	0.246	0.353	0.205	0.446
SANA-1.5 1.6B (PAG) (Xie et al., 2025a)	0.762	0.054	0.209	0.387	0.222	0.327
SANA-1.5 4.8B (PAG) (Xie et al., 2025a)	0.765	0.069	0.217	0.401	0.216	0.334
Lumina-Image 2.0 (Qin et al., 2025)	0.819	0.106	0.270	0.354	0.216	0.353
HiDream-I1-Full (Cai et al., 2025)	0.829	0.707	0.317	0.347	0.186	0.477
BLIP3-o (Chen et al., 2025a)	0.711	0.013	0.223	0.361	0.229	0.307
BAGEL (Deng et al., 2025)	0.769	0.244	0.173	0.367	0.251	0.361
Show-o2-1.5B (Xie et al., 2025b)	0.798	0.002	0.219	0.317	0.186	0.304
Show-o2-7B (Xie et al., 2025b)	0.817	0.002	0.226	0.317	0.177	0.308
OmniGen2 (Wu et al., 2025b)	0.804	0.680	0.271	0.377	0.242	0.475
Qwen-Image (Wu et al., 2025a)	0.882	0.891	0.306	0.418	0.197	0.539
AutoRegressive						
Emu3 (Wang et al., 2024b)	0.737	0.010	0.193	0.361	0.251	0.311
Janus-Pro (Chen et al., 2025b)	0.553	0.001	0.139	0.276	0.365	0.267

5. 模型性能

我们在几个具有代表性的基准上全面评估了 NextStep-1 的文本到图像（T2I）生成性能，每个基准针对图像生成的不同方面，包括视觉-文本对齐和世界知识。

如表 2 所示，我们在三个关键基准上评估了 NextStep-1 的提示跟随能力。在 GenEval（Ghosh et al., 2023）上，NextStep-1 得分为 0.63（使用 Self-CoT 时为 0.73），表现出强大的计数、定位和空间对齐能力。在 GenAI-Bench（Li et al., 2024a）上，它在基本提示上得分 0.88，在高级提示上得分 0.67（使用 Self-CoT 时分别为 0.9 和 0.74），进一步验证了其强大的组合能力。这些结果表明 NextStep-1 是一个出色的自回归图像生成模型，其性能可与一些扩散模型（如 Stable Diffusion 3.5 Large（Stability-AI, 2024）和 BAGEL（Deng et al., 2025））相媲美。最后，在 DPG-Bench（Hu et al., 2024）基准上评估长上下文、多物体场景时，NextStep-1 得分为 85.28，证实了其在复杂提示下的可靠组合保真度。

为了进行细粒度分析，我们使用英文提示在 OneIG-Bench（Chang et al., 2025）上评估了我们的模型。此基准测试评估了多个领域的性能，例如对齐、文本渲染、推理和风格控制。如 Tab. 3 所示，NextStep-1 实现了 0.417 的总体得分。这个结果显著超过了其自回归同行，例如 Emu3（Wang et al., 2024b）（ 0.311 ）和 Janus-Pro（Chen et al., 2025b）（ 0.267 ）。

为了评估 NextStep-1 将世界知识整合到图像生成中的能力，我们使用 WISE 基准（Niu et al., 2025），该基准强调事实基础和语义理解。如表 4 所示，NextStep-1 在自回归模型中取得了最佳性能，整体得分为 0.54（自我共事促发器为 0.67），也超过了大多数扩散模型。值得注意的是，在提示重写协议下，其得分增加到 0.79（自我共事促发器为 0.83）。总体而言，这些结果展示了 NextStep-1 强大的知识感知语义对齐和跨领域推理能力。

Table 4 | 世界知识推理在 WISE (Niu et al., 2025) 上的比较。† 结果与 Self-CoT。

Model	Cultural	Time	Space	Biology	Physics	Chemistry	Overall ↑	Overall (Rewrite) ↑
Proprietary								
GPT-4o (OpenAI, 2025b)	0.81	0.71	0.89	0.83	0.79	0.74	0.80	-
Diffusion								
Stable Diffusion 1.5 (Rombach et al., 2022)	0.34	0.35	0.32	0.28	0.29	0.21	0.32	0.50
Stable Diffusion XL (Podell et al., 2024)	0.43	0.48	0.47	0.44	0.45	0.27	0.43	0.65
Stable Diffusion 3.5 Large (Stability-AI, 2024)	0.44	0.50	0.58	0.44	0.52	0.31	0.46	0.72
PixArt-Alpha (Chen et al., 2024)	0.45	0.50	0.48	0.49	0.56	0.34	0.47	0.63
Playground v2.5 (Li et al., 2024b)	0.49	0.58	0.55	0.43	0.48	0.33	0.49	0.71
Flux.1-dev (Labs, 2024)	0.48	0.58	0.62	0.42	0.51	0.35	0.50	0.73
MetaQuery-XL (Pan et al., 2025)	0.56	0.55	0.62	0.49	0.63	0.41	0.55	-
BAGEL (Deng et al., 2025)	0.44/ 0.76 †	0.55/ 0.69 †	0.68/0.75 †	0.44/ 0.65 †	0.60/ 0.75 †	0.39/ 0.58 †	0.52/ 0.70 †	0.71/0.77 †
Qwen-Image (Wu et al., 2025a)	0.67	0.67	0.80	0.62	0.79	0.41	0.67	-
AutoRegressive								
Show-o-512 (Xie et al., 2024)	0.28	0.40	0.48	0.30	0.46	0.30	0.35	0.64
VILA-U (Wu et al., 2024)	0.26	0.33	0.37	0.35	0.39	0.23	0.31	-
Emu3 (Wang et al., 2024b)	0.34	0.45	0.48	0.41	0.45	0.27	0.39	0.63
Janus-Pro-7B (Chen et al., 2025b)	0.30	0.37	0.49	0.36	0.42	0.26	0.35	0.71

Table 5 | 在 GEdit-Bench（完整集） (Liu et al., 2025) 和 ImgEdit-Bench (Ye et al., 2025) 上的图像编辑性能比较。G_SC、G_PQ 和 G_O 指的是由 GPT-4.1 (OpenAI, 2025a) 评价的指标。性能评估基于 1:1 的纵横比 NextStep-1-Edit 。

Model	GEdit-Bench-EN (Full Set) ↑			GEdit-Bench-CN (Full Set) ↑			ImgEdit-Bench ↑
	G_SC	G_PQ	G_O	G_SC	G_PQ	G_O	
Proprietary							
Gemini 2.0 (Gemini2, 2025)	6.87	7.44	6.51	5.26	7.60	5.14	-
Doubao (Shi et al., 2024)	7.22	7.89	6.98	7.17	7.79	6.84	-
GPT-4o (OpenAI, 2025b)	7.74	8.13	7.49	7.52	8.02	7.30	4.20
Flux.1-Kontext-pro (Labs et al., 2025)	7.02	7.60	6.56	1.11	7.36	1.23	-
Open-source							
Instruct-Pix2Pix (Brooks et al., 2023)	3.30	6.19	3.22	-	-	-	1.88
MagicBrush (Zhang et al., 2023a)	4.52	6.37	4.19	-	-	-	1.83
AnyEdit (Yu et al., 2024a)	3.05	5.88	2.85	-	-	-	2.45
OmniGen (Xiao et al., 2024)	5.88	5.87	5.01	-	-	-	2.96
OmniGen2 (Wu et al., 2025b)	7.16	6.77	6.41	-	-	-	3.44
Step1X-Edit v1.0 (Liu et al., 2025)	7.13	7.00	6.44	7.30	7.14	6.66	3.06
Step1X-Edit v1.1 (Liu et al., 2025)	7.66	7.35	6.97	7.65	7.40	6.98	-
BAGEL (Deng et al., 2025)	7.36	6.83	6.52	7.34	6.85	6.50	3.42
Flux.1-Kontext-dev (Labs et al., 2025)	-	-	6.26	-	-	-	3.71
GPT-Image-Edit (Wang et al., 2025c)	-	-	7.24	-	-	-	3.80

5.1. 图像编辑的性能

编辑基准的定量结果。我们通过在 **??** 上对 100 万高质量的仅编辑数据进行微调开发了 NextStep-1-Edit ,其展示了与先进的基于扩散的模型竞争的性能。如 Tab. 5 所示,NextStep-1-Edit 在 GEdit-Bench-EN (Liu et al., 2025) 上获得了 6.58 的分数，在 ImgEdit-Bench (Ye et al., 2025) 上获得了 3.71 的分数，表明其具有很强的实际编辑能力。

6. 讨论

6.1. 图像生成的决定因素是什么：AR Transformer 还是 FM Head？

我们框架的一个关键架构区别在于其使用流匹配目标对连续图像标记进行直接、自回归建模。当前流行的图像生成自回归模型 (Chen et al., 2025a; Dong et al., 2024; Sun et al., 2023, 2024b; Zhou et al., 2025) 通常依赖于用于整个图像的繁重扩散模型：自回归模型首先生成一个语义嵌

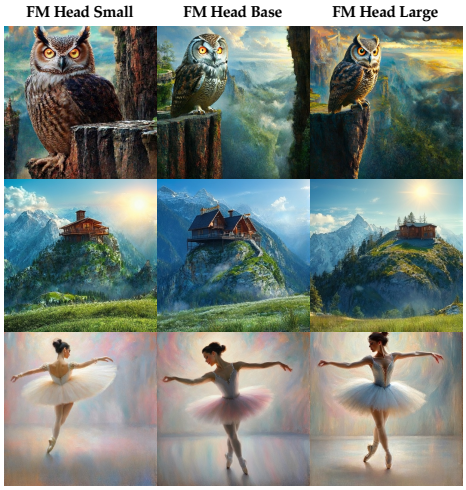

Figure 4 | 在不同流匹配头下生成的图像。

Table 6 | Configurations for different flow-matching heads.

	Layers	Hidden Size	# Parameters
FM Head Small	6	1024	40M
FM Head Base	12	1536	157M
FM Head Large	24	2048	528M

Table 7 | Quantitative results for different flow-matching head configurations. All variants are finetuned from the baseline with a newly initialized head.

	GenEval	GenAI-Bench	DPG-Bench
Baseline	0.59	0.77	85.15
w/ FM Head Small	0.55	0.76	83.46
w/ FM Head Base	0.55	0.75	84.68
w/ FM Head Large	0.56	0.77	85.50

入，然后用于条件化一个扩散模型，该模型在单次去噪过程中生成一整个图像。相比之下，我们的模型是逐个生成图像块，以显著更轻量的流匹配模型来建模每个块的分布。我们认为，这使得我们的框架建立在纯粹的自回归范式下，通过下一个标记预测（NTP）建模，而不是仅由变压器协调的扩散模型。

我们的实验一个关键发现是模型对其流匹配头大小的惊人不敏感。我们对此进行了消融实验，涉及三个不同大小的头（小型、基础和大型）。在每次实验中，我们重新初始化并只训练该头部 10,000 步。虽然模型大小存在显著差异，但这三个头部产生的结果却惊人地相似（Tab. 7 ，Fig. 4 ）。这种对头部大小的不敏感强烈表明，变换器主干在执行条件分布 $p(x_i \mid x_{<i})$ 的核心生成建模。流匹配头类似于语言模型中的 LM 头，主要作用是作为一个轻量级采样器，将变换器的上下文预测转化为一个连续的符号。因此，基本的生成逻辑驻留在变换器的自回归 NTP 过程之中。

6.2. 分词器是图像生成的关键

缓解强分类器无指导下的不稳定性。在基于 VAE 的自回归模型中，一个已知的失败模式是视觉瑕疵的出现，例如灰色斑块，尤其是在强分类器自由引导比例下（Fan et al., 2024）。虽然先前的工作假设这种不稳定性源于一维位置嵌入的间断性，但我们的分析表明，根本原因在于高引导比例下令牌级别分布变化的放大。

在推理时，引入 CFG 以增强条件保真度。引导预测 $\tilde{v}$ 通过插值计算得出：

$$\tilde{v}(x|y) = (1-w) \cdot v_\theta(x|\varnothing) + w \cdot v_\theta(x|y) \tag{4}$$

，其中 $v_\theta(x|\varnothing)$ 和 $v_\theta(x|y)$ 分别是无条件和有条件预测，w 是引导比例。在扩散模型中，由于潜在变量通常是标准化的，因此使用高引导比例进行推理是稳定的，这确保有条件和无条件预测维持一致的比例。然而，在逐字级的自回归模型中，对整个潜在张量进行全局标准化，并不能确保每个字的统计一致性。因此，由于条件和无条件预测之间的小差异被较大的引导比例放大，从而导致生成字序列的统计数据出现显著偏移。

我们在 Fig. 5 中通过实验证明了这一现象。在一个适中的指导尺度 1.5 下，每个 token 的平均值和方差在整个生成过程中保持稳定。相比之下，在 3.0 的高指导尺度下，这两个统计量在后期 token 上显著偏离，这是直接与视觉伪影的出现相对应的分布变化。我们的分词器设计结合了通道归一化（见 Equation (3) ），直接解决了这一问题，通过强制每个 token 的统计稳定性。

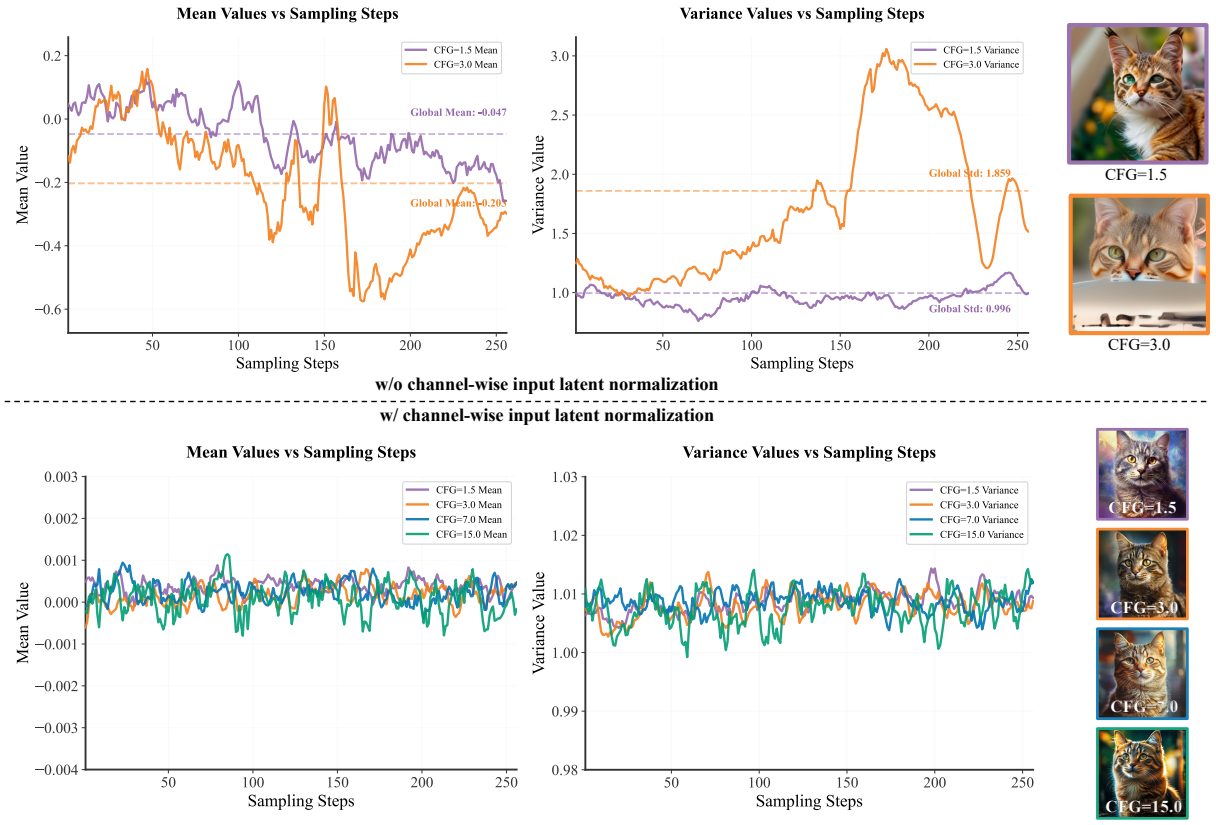

Figure 5 | 每个标记均值和方差在不同采样步骤下的演变，比较两个 CFG 设置。在 CFG=1.5 时，均值和方差分别接近 0 和 1，表明稳定性。在 CFG=3.0 时，它们显著漂移，导致图像质量下降。通过归一化，输出潜变量的分布在所有 CFG 设置下保持稳定。

这个简单但至关重要的设计选择缓解了不稳定性，使得在不降低图像质量的情况下使用强引导成为可能。

一个正则化潜在空间对于生成是至关重要的 我们工作的一个关键发现是生成损失与自回归模型的最终合成质量之间的逆相关性。具体来说，在标记器训练期间应用较高的噪声强度（在 Equation (3) 上的 γ ）会增加生成损失，但却出人意料地改善了生成图像的质量。例如，NextStep-1 使用在 $\gamma = 0.5$ 上训练的标记器，其产生了最高的生成损失，但却生成了最高保真度的图像。相反，为低生成损失训练的标记器导致自回归模型产生的输出类似于纯噪声。

我们将这种现象归因于噪声正则化培养了一个良好条件的潜在空间。这个过程增强了两个关键特性：标记解码器对潜在扰动（Fig. 6 ）的鲁棒性和更分散的潜在分布（Fig. 7 ），以往的工作也发现这对生成有益（Sun et al., 2024c; Yang et al., 2025; Yao et al., 2025）。虽然尚不清楚鲁棒性或分散性是否起关键作用，但这些结果强调了基于噪声的正则化的实际好处，并指出了未来分析的有希望方向。

重建质量是生成质量的上限。图像标记器的重建保真度从根本上决定了最终生成图像的质量上限，特别是对精细细节和纹理的影响。这个原则已经在最近的许多研究中得到了验证（Dai et al., 2023; Esser et al., 2024; Labs, 2024），导致在扩散范式中出现了一种趋势，即在具有卓越重建性能的 VAE 的基础上构建生成模型（例如，PSNR $>$ 30）。相比之下，基于 VQ 的自回归模型历来难以突破这一门槛，如 Tab. 8 所示。虽然重建和生成质量之间的权衡常常被讨论（Yao et al., 2025），但我们的工作成功地将自回归模型应用于高保真度的连续 VAE，弥合了这一差距。

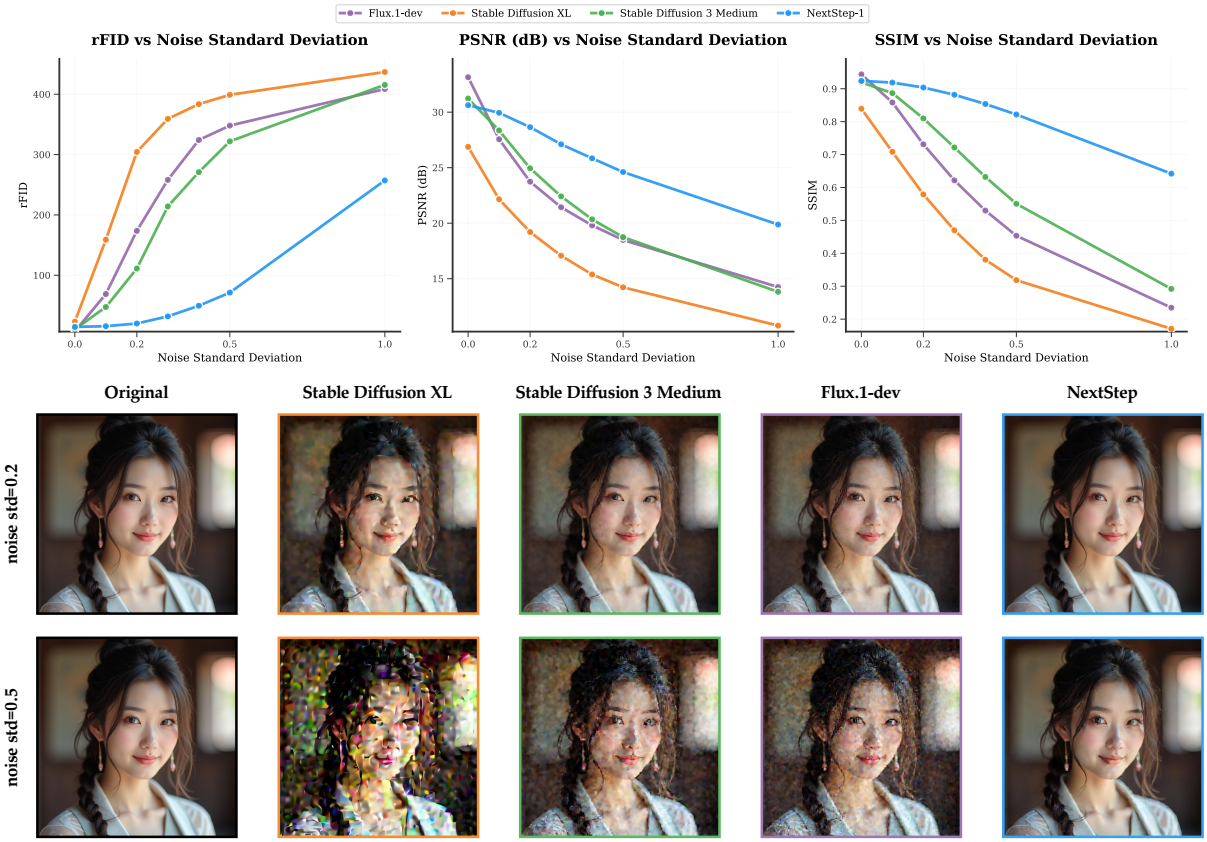

Figure 6 | 噪声扰动对图像分词器性能的影响。顶部面板显示定量指标（rFID ↓、PSNR ↑ 和 SSIM ↑）与噪声强度的关系。底部面板展示在噪声标准差为 0.2 和 0.5 时的质量重建示例。

Table 8 | 在 ImageNet-1K 256 × 256 (Deng et al., 2009) 上重建性能的比较。

Tokenizer	Latent Shape	PSNR ↑	SSIM ↑
Discrete Tokenizer			
SBER-MoVQGAN (270M) (Zheng et al., 2022)	32x32	27.04	0.74
LlamaGen (Sun et al., 2024a)	32x32	24.44	0.77
VAR (Tian et al., 2024)	680	22.12	0.62
TiTok-S-128 (Yu et al., 2024b)	128	17.52	0.44
Sefltok (Wang et al., 2025b)	1024	26.30	0.81
Continuous Tokenizer			
Stable Diffusion 1.5 (Rombach et al., 2022)	32x32x4	25.18	0.73
Stable Diffusion XL (Podell et al., 2024)	32x32x4	26.22	0.77
Stable Diffusion 3 Medium (Esser et al., 2024)	32x32x16	30.00	0.88
Flux.1-dev (Labs, 2024)	32x32x16	31.64	0.91

6.3. 限制和挑战

虽然 NextStep-1 成功地展示了自回归模型可以在高维连续潜在空间中运行，并实现了与扩散模型相当的生成质量，但这种方法也带来了独特的稳定性挑战。我们观察到，当从具有低维潜在空间（例如，空间下采样因子为 8 且潜在通道数为 4）的 VAE 转换到具有高维空间（例如，空间下采样因子为 8 且潜在通道数为 16）时，会出现几种不同的生成伪影。虽然前一种配置产生稳定的输出，但后一种配置有时会展现出故障模式，如 Fig. 8 所示。

虽然根本原因仍然是一个未解的问题，但我们识别了几个可能的促成因素：(1) 在生成后期

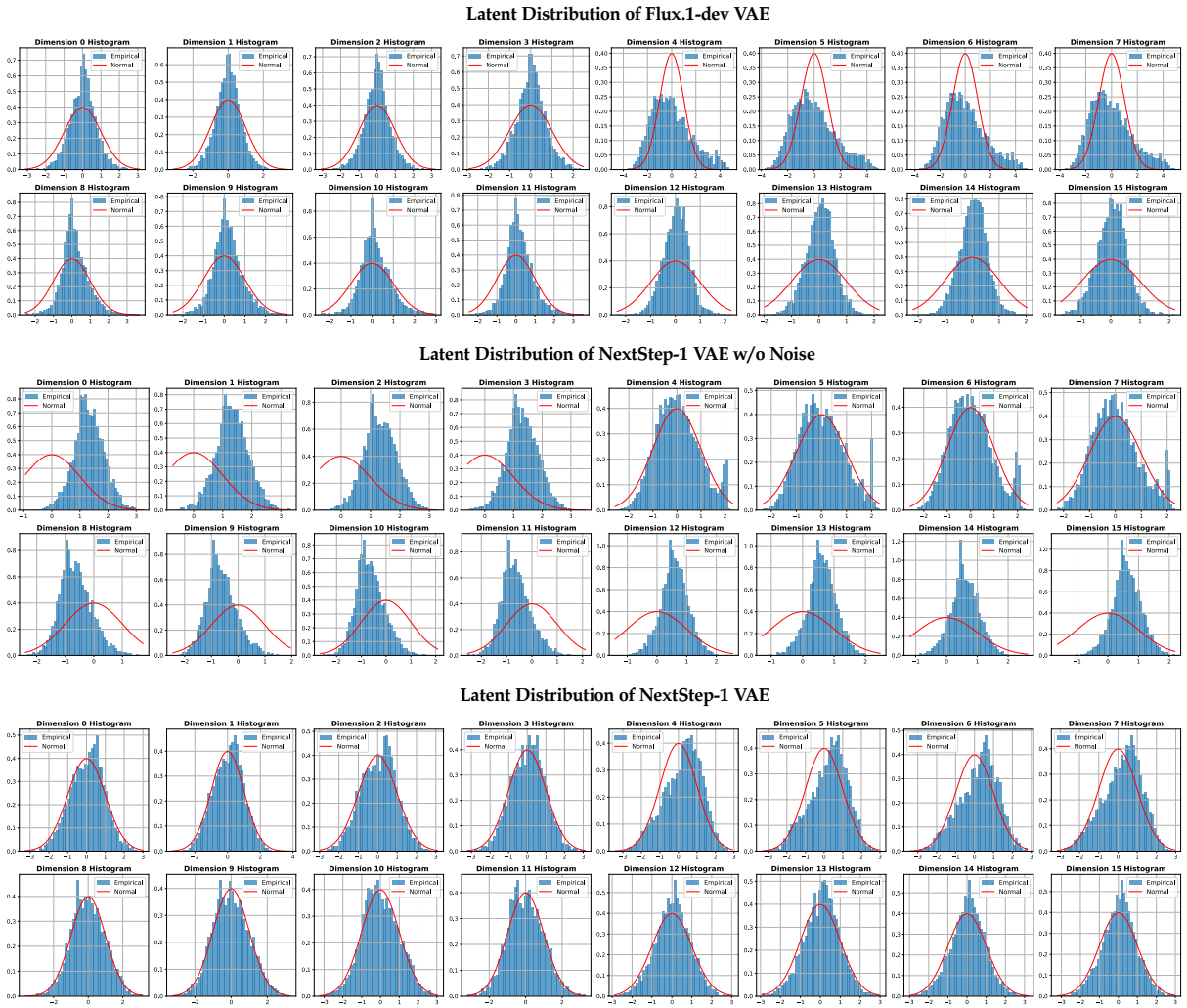

Figure 7 | 16 个通道中三种 VAE 变体的潜在分布：Flux.1-dev、NextStep-1 不带噪声和 NextStep-1。蓝色条形图显示经验直方图；红色线条表示标准正态分布。NextStep-1 VAE 与正态分布最佳对齐，反映了分散的潜在分布。

出现的局部噪声或块状伪影可能源于数值不稳定性；(2) 图像上的全局噪声可能反映了训练不足，意味着额外的训练可以缓解该问题；(3) 微妙的网格状伪影可能揭示了一维位置编码在捕捉二维空间关系上的局限性。

有关一款 H100 GPU（983 TFLOPS，3.36 TB/s 带宽）在批量大小为 1 时每个 token 延迟的理论分析，详细见 Tab. 9 ，其分解了各个组件的贡献。结果显示，主要瓶颈在于大型语言模型的串行解码，而流匹配头的多步采样也构成了每个 token 生成成本的很大一部分。这些观察结果提出了两个加速推理的有潜力的方向。首先，可以通过减少其参数数量、应用蒸馏技术以实现少步骤生成（Meng et al., 2023），或使用更先进的少步骤采样器（Lu et al., 2022, 2025）来提高流匹配头的效率。其次，通过借鉴大型语言模型领域的最新进展，例如投机解码（Leviathan et al., 2023）或多 token 预测（Gloeckle et al., 2024），可以加速自回归主干在图像 token 生成领域的应用。

我们的框架在扩展至高分辨率图像生成时面临两个主要挑战，尤其是与扩散模型相比，而扩散模型在这一领域受益于成熟的技术 。首先，自回归生成的严格顺序性要求在较高分辨率下进行大量训练步骤才能收敛。相比之下，扩散模型在每次迭代中并行地细化整个图像，使得更直接地利用二维空间归纳偏差。其次，最近为高分辨率扩散模型开发的技术，例如时间步移位，很

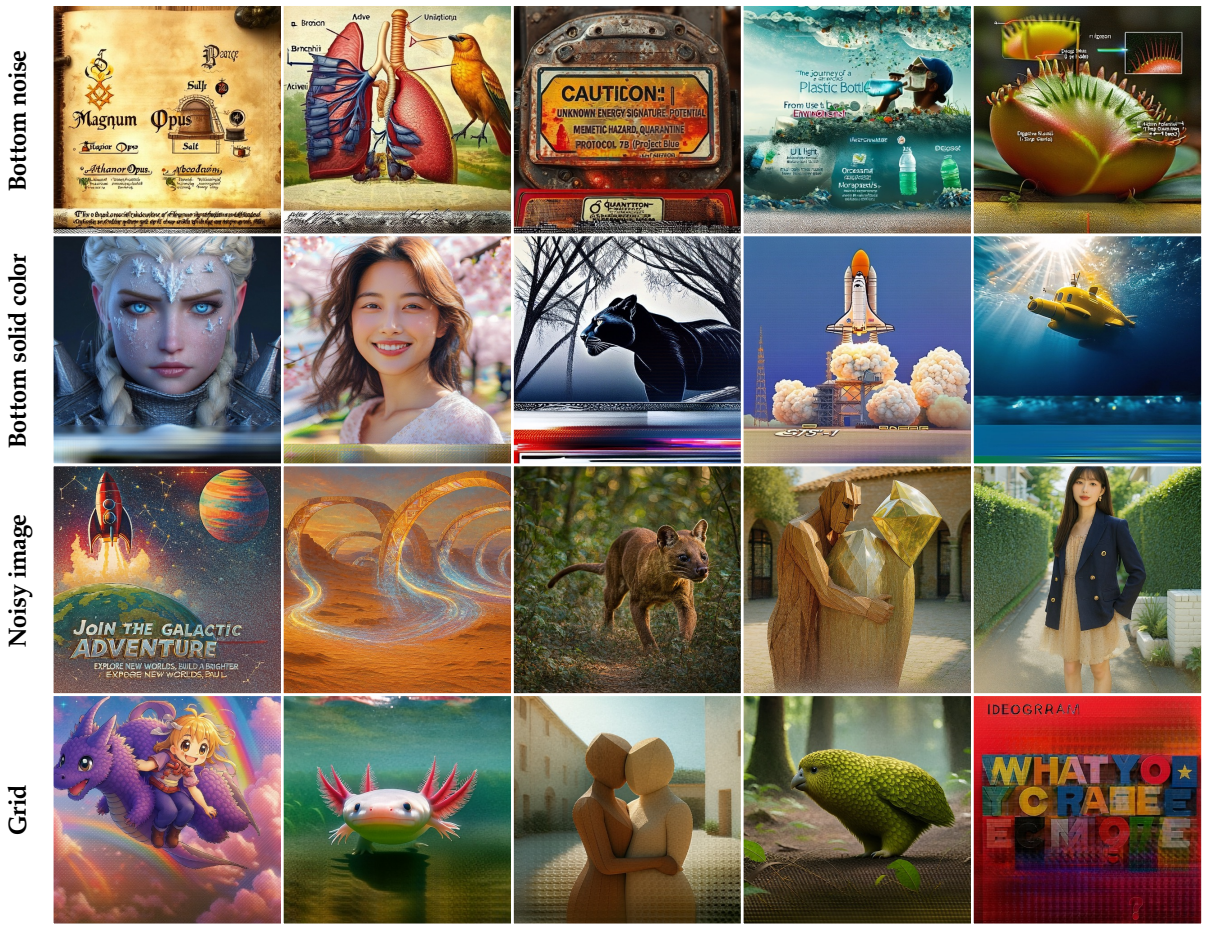

Figure 8 | 高维连续标记的失败案例。

难适应我们的设置。这种限制是由于 Flow Matching Head 主要作为轻量级采样器，而变压器骨干负责核心生成建模；因此，对采样过程的修改对最终输出仅有边际影响。专门为逐块自回归模型设计高分辨率生成策略仍然是未来研究的重要方向。

在我们的自回归框架中，

SFT 中的挑战。进行微调相比于扩散模型提出了独特的挑战。我们观察到，在小型高质量数据集上进行微调会表现出不稳定的动态。与扩散模型相反，后者通常可以适应目标分布并在仅有几千个样本的情况下保持稳定和通用的图像生成，我们的

SFT 中的挑战 微调过程只有在百万样本规模的数据集上训练时才会有显著的改进。对于较小的数据集，模型保持在一个不稳定的平衡状态；要么仅有微小的改进且影响可忽略，要么突然过拟合到目标分布。因此，找到一个既能与目标分布对齐又能保持通用生成能力的中间检查点仍然是一个重大挑战。

我们将研发 NextStep-1 的研究人员称为研究者，而贡献者指的是在数据、系统、平台、早期版本工作或兼职贡献等领域提供支持的人。⋆ 表示核心执行者，† 表示项目负责人。作者按名字的字母顺序排列。

研究人员： 韩春睿 ⋆，李国鹏 ⋆，吴旌玮 ⋆，孙权 ⋆†，蔡妍 ⋆，彭媛 ⋆，葛正 ⋆†，周德育，唐浩淼，周宏宇，刘垦坤

Table 9 | 在 983 TFLOP/s 计算能力和 3.36 TB/s 内存带宽下的 推理延迟分解。

Sequence Length	Last-token Latency (ms)			Accumulated Latency (s)	
	LLM Decoder	LM Head	FM Head	Total	w/o FM Head
256	7.20	0.40	3.40	2.82	1.95
1024	7.23	0.40	3.40	11.31	7.83
4096	7.39	0.40	3.40	45.77	31.86

贡献者： 黄爱林, 王斌, 缪长新, 孙德山, 于恩, 尹福坤, 于刚, 聂浩, 吕浩然, 胡汉鹏, 王佳, 周建, 孙建建, 谭开军, 安康, 林康恒, 赵亮, 陈梅, 邢鹏, 王睿, 刘诗宇, 夏书韬, 尤天昊, 季伟, 曾献方, 韩鑫, 张雪琳, 魏雅娜, 许延明, 江一民, 王英明, 周瑜, 韩玉成, 孟子阳

赞助方： 焦斌星, 姜大新, 张翔宇, 朱义博

我们衷心感谢李天宏和田永龙的深入讨论。

References

S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

J. Baldridge, J. Bauer, M. Bhutani, N. Brichtova, A. Bunner, L. Castrejon, K. Chan, Y. Chen, S. Dieleman, Y. Du, et al. Imagen 3. arXiv preprint arXiv:2408.07009, 2024.

J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, et al. Improving image generation with better captions. OpenAI blog, 2023.

T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. Advances in neural information processing systems (NeurIPS), 2020.

Q. Cai, J. Chen, Y. Chen, Y. Li, F. Long, Y. Pan, Z. Qiu, Y. Zhang, F. Gao, P. Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. arXiv preprint arXiv:2505.22705, 2025.

J. Chang, Y. Fang, P. Xing, S. Wu, W. Cheng, R. Wang, X. Zeng, G. Yu, and H.-B. Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. arXiv preprint arXiv:2506.07977, 2025.

J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, and Z. Li. Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In European Conference on Computer Vision (ECCV), 2024.

J. Chen, Z. Xu, X. Pan, Y. Hu, C. Qin, T. Goldstein, L. Huang, T. Zhou, S. Xie, S. Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568, 2025a.

X. Chen, C. Wu, Z. Wu, Y. Ma, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, C. Ruan, and P. Luo. Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811, 2025b.

X. Dai, J. Hou, C.-Y. Ma, S. Tsai, J. Wang, R. Wang, P. Zhang, S. Vandenhende, X. Wang, A. Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807, 2023.

G. deepmind Imagen4 team. Imagen4, 2025. URL `https://storage.googleapis.com/deepmind-media/Model-Cards/Imagen-4-Model-Card.pdf`.

M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. Advances in Neural Information Processing Systems (NeurIPS), 2023.

C. Deng, D. Zhu, K. Li, C. Gou, F. Li, Z. Wang, S. Zhong, W. Yu, X. Nie, Z. Song, et al. Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683, 2025.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

R. Dong, C. Han, Y. Peng, Z. Qi, Z. Ge, J. Yang, L. Zhao, J. Sun, H. Zhou, H. Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. In International Conference on Learning Representations (ICLR), 2024.

S. M. A. Eslami, S. Liu, A. v. d. Oord, O. Vinyals, M. J. Wainwright, and I. Sutskever. Taming transformers for high-resolution image synthesis. In International Conference on Machine Learning (ICML), 2021.

P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In International Conference on Machine Learning (ICML), 2024.

L. Fan, T. Li, S. Qin, Y. Li, C. Sun, M. Rubinstein, D. Sun, K. He, and Y. Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. arXiv preprint arXiv:2410.13863, 2024.

Y. Gao, L. Gong, Q. Guo, X. Hou, Z. Lai, F. Li, L. Li, X. Lian, C. Liao, L. Liu, et al. Seedream 3.0 technical report. arXiv preprint arXiv:2504.11346, 2025.

Y. Ge, S. Zhao, J. Zhu, Y. Ge, K. Yi, L. Song, C. Li, X. Ding, and Y. Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arxiv:2404.14396, 2024.

G. Gemini2. Experiment with gemini 2.0 flash native image generation, 2025. URL `https://developers.googleblog.com/en/experiment-with-gemini-20-flash-native-image-generation`.

D. Ghosh, H. Hajishirzi, and L. Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In Advances in neural information processing systems (NeurIPS), 2023.

F. Gloeckle, B. Y. Idrissi, B. Rozière, D. Lopez-Paz, and G. Synnaeve. Better & faster large language models via multi-token prediction. arXiv preprint arXiv:2404.19737, 2024.

J. Han, J. Liu, Y. Jiang, B. Yan, Y. Zhang, Z. Yuan, B. Peng, and X. Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025.

X. Han, Y. Wu, L. Shi, H. Liu, H. Liao, L. Qiu, W. Yuan, X. Gu, Z. Dong, and S. Cui. Mvimgnet2. 0: A larger-scale dataset of multi-view images. arXiv preprint arXiv:2412.01430, 2024.

X. Hu, R. Wang, Y. Fang, B. Fu, P. Cheng, and G. Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135, 2024.

A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. B. Girshick. Segment Anything. In IEEE International Conference on Computer Vision (ICCV), 2023.

B. F. Labs. Flux, 2024. URL `https://github.com/black-forest-labs/flux`.

B. F. Labs. Flux.1-fill-dev, 2025. URL https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev.

B. F. Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne, T. Dockhorn, J. English, Z. English, P. Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742, 2025.

Y. Leviathan, M. Kalman, and Y. Matias. Fast inference from transformers via speculative decoding. In International Conference on Machine Learning, pages 19274–19286. PMLR, 2023.

B. Li, Z. Lin, D. Pathak, J. Li, Y. Fei, K. Wu, X. Xia, P. Zhang, G. Neubig, and D. Ramanan. Evaluating and improving compositional text-to-visual generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024a.

D. Li, A. Kamko, E. Akhgari, A. Sabet, L. Xu, and S. Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245, 2024b.

T. Li, Y. Tian, H. Li, M. Deng, and K. He. Autoregressive image generation without vector quantization. In Advances in neural information processing systems (NeurIPS), 2024c.

C. Liao, L. Liu, X. Wang, Z. Luo, X. Zhang, W. Zhao, J. Wu, L. Li, Z. Tian, and W. Huang. Mogao: An omni foundation model for interleaved multi-modal generation. arXiv preprint arXiv:2505.05472, 2025.

Z. Lin, D. Pathak, B. Li, J. Li, X. Xia, G. Neubig, P. Zhang, and D. Ramanan. Evaluating text-to-visual generation with image-to-text generation. arXiv preprint arXiv:2404.01291, 2024.

Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In International Conference on Machine Learning (ICLR), 2023a.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2023b.

S. Liu, Y. Han, P. Xing, F. Yin, R. Wang, W. Cheng, J. Liao, Y. Wang, H. Fu, C. Han, et al. Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761, 2025.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2019.

C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in neural information processing systems, 35:5775–5787, 2022.

C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. Machine Intelligence Research, pages 1–22, 2025.

G. Ma, H. Huang, K. Yan, L. Chen, N. Duan, S. Yin, C. Wan, R. Ming, X. Song, X. Chen, Y. Zhou, D. Sun, D. Zhou, J. Zhou, K. Tan, K. An, M. Chen, W. Ji, Q. Wu, W. Sun, X. Han, Y. Wei, Z. Ge, A. Li, B. Wang, B. Huang, B. Wang, B. Li, C. Miao, C. Xu, C. Wu, C. Yu, D. Shi, D. Hu, E. Liu, G. Yu, G. Yang, G. Huang, G. Yan, H. Feng, H. Nie, H. Jia, H. Hu, H. Chen, H. Yan, H. Wang, H. Guo, H. Xiong, H. Xiong, J. Gong, J. Wu, J. Wu, J. Wu,

J. Yang, J. Liu, J. Li, J. Zhang, J. Guo, J. Lin, K. Li, L. Liu, L. Xia, L. Zhao, L. Tan, L. Huang, L. Shi, M. Li, M. Li, M. Cheng, N. Wang, Q. Chen, Q. He, Q. Liang, Q. Sun, R. Sun, R. Wang, S. Pang, S. Yang, S. Liu, S. Liu, S. Gao, T. Cao, T. Wang, W. Ming, W. He, X. Zhao, X. Zhang, X. Zeng, X. Liu, X. Yang, Y. Dai, Y. Yu, Y. Li, Y. Deng, Y. Wang, Y. Wang, Y. Lu, Y. Chen, Y. Luo, Y. Luo, Y. Yin, Y. Feng, Y. Yang, Z. Tang, Z. Zhang, Z. Yang, B. Jiao, J. Chen, J. Li, S. Zhou, X. Zhang, X. Zhang, Y. Zhu, H.-Y. Shum, and D. Jiang. Step-video-t2v technical report: The practice, challenges, and future of video foundation model, 2025a. URL `https://arxiv.org/abs/2502.10248`.

X. Ma, P. Sun, H. Ma, H. Tang, C.-Y. Ma, J. Wang, K. Li, X. Dai, Y. Shi, X. Ju, et al. Token-shuffle: Towards high-resolution image generation with autoregressive models. arXiv preprint arXiv:2504.17789, 2025b.

C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, and T. Salimans. On distillation of guided diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 14297–14306, 2023.

Y. Niu, M. Ning, M. Zheng, B. Lin, P. Jin, J. Liao, K. Ning, B. Zhu, and L. Yuan. Wise: A world knowledge-informed semantic evaluation for text-to-image generation. arXiv preprint arXiv:2503.07265, 2025.

D. A. Oliveira and D. M. de Matos. Storyreasoning dataset: Using chain-of-thought for scene understanding and grounded story generation. arXiv preprint arXiv:2505.10292, 2025.

OpenAI. Introducing gpt-4.1 in the api. OpenAI Blog, 2025a. URL `https://openai.com/index/gpt-4-1`.

OpenAI. Introducing 4o image generation, 2025b. URL `https://openai.com/index/introducing-4o-image-generation`.

X. Pan, S. N. Shukla, A. Singh, Z. Zhao, S. K. Mishra, J. Wang, Z. Xu, J. Chen, K. Li, F. Juefei-Xu, J. Hou, and S. Xie. Transfer between modalities with metaqueries. arXiv preprint arXiv:2504.06256, 2025.

Y. Peng, Y. Cui, H. Tang, Z. Qi, R. Dong, J. Bai, C. Han, Z. Ge, X. Zhang, and S.-T. Xia. Dreambench++: A human-aligned benchmark for personalized image generation. arXiv preprint arXiv:2406.16855, 2024.

D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In International Conference on Learning Representations (ICLR), 2024.

Q. Qin, L. Zhuo, Y. Xin, R. Du, Z. Li, B. Fu, Y. Lu, J. Yuan, X. Li, D. Liu, et al. Lumina-image 2.0: A unified and efficient image generative framework. arXiv preprint arXiv:2503.21758, 2025.

A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. Improving language understanding by generative pre-training. San Francisco, CA, USA, 2018.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 2019.

R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems (NeurIPS), 2023.

R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. arXiv preprint arXiv:2305.18290, 2024.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

Y. Shi, P. Wang, and W. Huang. Seededit: Align image re-generation to image editing. arXiv preprint arXiv:2411.06686, 2024.

Stability-AI. stable-diffusion-3.5-large, 2024. URL `https://github.com/Stability-AI/sd3.5`.

J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 2024.

P. Sun, Y. Jiang, S. Chen, S. Zhang, B. Peng, P. Luo, and Z. Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525, 2024a.

Q. Sun, Q. Yu, Y. Cui, F. Zhang, X. Zhang, Y. Wang, H. Gao, J. Liu, T. Huang, and X. Wang. Emu: Generative pretraining in multimodality. In International Conference on Learning Representations (ICLR), 2023.

Q. Sun, Y. Cui, X. Zhang, F. Zhang, Q. Yu, Y. Wang, Y. Rao, J. Liu, T. Huang, and X. Wang. Generative multimodal models are in-context learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024b.

Y. Sun, H. Bao, W. Wang, Z. Peng, L. Dong, S. Huang, J. Wang, and F. Wei. Multimodal latent language modeling with next-token diffusion. arXiv preprint arXiv:2412.08635, 2024c.

K. K. team. Kolors2.0, 2025. URL `https://app.klingai.com/cn`.

R. team. Recraft v3, 2024. URL `https://www.recraft.ai/blog/recraft-introduces-a-revolutionary-ai-model-that-thinks-in-design-language`.

K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems (NeurIPS), 2024.

S. Tong, D. Fan, J. Zhu, Y. Xiong, X. Chen, K. Sinha, M. Rabbat, Y. LeCun, S. Xie, and Z. Liu. Metamorph: Multimodal understanding and generation via instruction tuning. arXiv preprint arXiv:2412.14164, 2024.

B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik. Diffusion model alignment using direct preference optimization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

B. Wang, B. Wang, C. Wan, G. Huang, H. Hu, H. Jia, H. Nie, M. Li, N. Chen, S. Chen, et al. Step-3 is large yet affordable: Model-system co-design for cost-effective decoding. arXiv preprint arXiv:2507.19427, 2025a.

B. Wang, Z. Yue, F. Zhang, S. Chen, L. Bi, J. Zhang, X. Song, K. Y. Chan, J. Pan, W. Wu, et al. Selftok: Discrete visual tokens of autoregression, by diffusion, and for reasoning. arXiv preprint arXiv:2505.07538, 2025b.

P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024a.

X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu, et al. Emu3: Next-token prediction is all you need. arXiv preprint arxiv:2409.18869, 2024b.

Y. Wang, S. Yang, B. Zhao, L. Zhang, Q. Liu, Y. Zhou, and C. Xie. Gpt-image-edit-1.5 m: A million-scale, gpt-generated image dataset. arXiv preprint arXiv:2507.21033, 2025c.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems (NeurIPS), 2022.

C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, K. Yan, S.-m. Yin, S. Bai, X. Xu, Y. Chen, et al. Qwen-image technical report. arXiv preprint arXiv:2508.02324, 2025a.

C. Wu, P. Zheng, R. Yan, S. Xiao, X. Luo, Y. Wang, W. Li, X. Jiang, Y. Liu, J. Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871, 2025b.

Y. Wu, Z. Zhang, J. Chen, H. Tang, D. Li, Y. Fang, L. Zhu, E. Xie, H. Yin, L. Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. arXiv preprint arXiv:2409.04429, 2024.

S. Xiao, Y. Wang, J. Zhou, H. Yuan, X. Xing, R. Yan, C. Li, S. Wang, T. Huang, and Z. Liu. Omnigen: Unified image generation. arXiv preprint arXiv:2409.11340, 2024.

E. Xie, J. Chen, Y. Zhao, J. Yu, L. Zhu, C. Wu, Y. Lin, Z. Zhang, M. Li, J. Chen, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. arXiv preprint arXiv:2501.18427, 2025a.

J. Xie, W. Mao, Z. Bai, D. J. Zhang, W. Wang, K. Q. Lin, Y. Gu, Z. Chen, Z. Yang, and M. Z. Shou. Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arxiv:2408.12528, 2024.

J. Xie, Z. Yang, and M. Z. Shou. Show-o2: Improved native unified multimodal models. arXiv preprint arXiv:2506.15564, 2025b.

J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems (NeurIPS), 2023.

A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.

J. Yang, T. Li, L. Fan, Y. Tian, and Y. Wang. Latent denoising makes good visual tokenizers. arXiv preprint arXiv:2507.15856, 2025.

J. Yao, B. Yang, and X. Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025.

Y. Ye, X. He, Z. Li, B. Lin, S. Yuan, Z. Yan, B. Hou, and L. Yuan. Imgedit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275, 2025.

J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, B. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, and Y. Wu. Scaling autoregressive models for content-rich text-to-image generation. In TMLR, 2022.

L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, V. Birodkar, A. Gupta, X. Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. arXiv preprint arXiv:2310.05737, 2023.

Q. Yu, W. Chow, Z. Yue, K. Pan, Y. Wu, X. Wan, J. Li, S. Tang, H. Zhang, and Y. Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. arXiv preprint arXiv:2411.15738, 2024a.

Q. Yu, M. Weber, X. Deng, X. Shen, D. Cremers, and L.-C. Chen. An image is worth 32 tokens for reconstruction and generation. Advances in Neural Information Processing Systems (NeurIPS), 2024b.

T. Z.ai. Cogview4, 2025. URL `https://github.com/THUDM/CogView4`.

K. Zhang, L. Mo, W. Chen, H. Sun, and Y. Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In Advances in neural information processing systems (NeurIPS), 2023a.

L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF international conference on computer vision (ICCV), 2023b.

W. Zhang, H. Zhang, X. Li, J. Sun, Y. Shen, W. Lu, D. Zhao, Y. Zhuang, and L. Bing. 2.5 years in class: A multimodal textbook for vision-language pretraining. arXiv preprint arXiv:2501.00958, 2025.

C. Zheng, T.-L. Vuong, J. Cai, and D. Phung. Movq: Modulating quantized vectors for high-fidelity image generation. Advances in Neural Information Processing Systems (NeurIPS), 2022.

C. Zhou, L. Yu, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, J. Kahn, X. Ma, L. Zettlemoyer, and O. Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. International Conference on Learning Representations (ICLR), 2025.