多模态自回归（AR）范式最近引起了广泛关注。该范式的特点是一个统一的下一个令牌预测框架，它具有在一个连贯结构中对齐多样性模态和任务的独特优势。此外，它在架构和概念上与大语言模型（LLMs）的相似性促进了最先进的训练和推理技术的无缝应用。虽然这个范式在多年前就被提出了，但随着诸如生成对抗网络（GANs）和扩散模型等替代方法的广泛采用，它逐渐失去了显著性。然而，在 2024 年，Chameleon 重新激活并扩展了这个范式，使其有希望的潜力重新回到研究社区的前沿。基于 Chameleon，Lumina-mGPT 作为第一个开源项目，展示了仅解码器的 AR 模型可以生成具有灵活纵横比的高分辨率图像，并达到与 SDXL 水平扩散模型相当的质量。

在这一领域的后续发展主要沿着两个主要方向推进：模型调整和资源投资的规模化。从建模的角度来看，Show-o（Xie et al., 2025c）和 Transfusion（Zhou et al., 2025）通过扩散方法替代了图像生成中的传统自回归范式，创建了混合的扩散-自回归框架。同时，Janus（Wu et al., 2025）和 Janus-Pro（Chen et al., 2025）通过整合连续视觉编码器（SigLIP（Zhai et al., 2023））来增强图像理解。虽然这些修改带来了各自的好处，但它们也在视觉理解和生成之间引入了不一致性，从而限制了它们在更灵活交互（如统一生成）方面的适应性。相比之下，Emu3（Wang et al., 2024b）保留了纯自回归架构并强调大规模的资源投资。然而，它在图像生成质量方面仍然表现不佳，并且在其内在非常适合于纯自回归架构的统一生成能力方面仍未得到充分探索。

在这篇论文中，我们介绍了 Lumina-mGPT 2.0，这是一种纯粹的 AR 模型，其性能可与最先进的扩散模型媲美，如 SANA（Xie et al., 2025a）和 DALL · E 3（Betker et al., 2023），在各种文本到图像（T2I）基准测试中表现卓越。基于其卓越的图像生成能力，我们进一步利用 AR 架构的内在优势探讨统一生成，证明 GPT-4o 图像生成的部分能力（如图 **??** 所示）可以通过适度的计算资源（例如，4 到 5 周使用 64 个 A100 GPU）实现。值得注意的是，与 Janus-Pro（Chen et al., 2025）和 Lumina-mGPT（Liu et al., 2024a）等以往工作不同，我们的方法不依赖于任何预训练权重，这不仅强调了我们方法的高效性，还使我们在设计模型架构和其它组件时没有预训练模型或其许可限制的约束，而能自由设计。

1. 相关工作

1.1. 扩散模型

扩散模型在文本到图像生成中表现出色。在这个框架中，模型的任务是预测添加到连续潜表示中的高斯噪声。最初，扩散模型主要依赖于 U-Net 架构（Rombach et al., 2022; Podell et al., 2024）。然而，随着 Transformer 架构的成功应用，扩散 Transformer（DiT）（Peebles and Xie, 2022）被引入作为 U-Net 的替代方案，标志着向更高效、可扩展的扩散模型迈出的重要一步（Rombach et al., 2022; Xie et al., 2025a; Zhuo et al., 2024; Chen et al., 2023; Betker et al., 2023; Labs, 2024; Liu et al., 2025）。基于这些基础模型，在优化采样质量和效率（Song and Ermon, 2019; Song et al., 2021a; Lu et al., 2022; Yi et al., 2024）以及开发下游任务方面取得了显著进展，如可控生成（Zhang et al., 2023; Zhao et al., 2023）、主题驱动的生成（Ruiz et al., 2023; Xiao et al., 2024）和图像编辑（Brooks et al., 2023; Couairon et al., 2023）。

自回归(AR)模型，利用 LLMs（Brown et al., 2020; Achiam et al., 2023; Touvron et al., 2023; Liu et al., 2024b; Xin et al., 2025）的强大扩展能力，已通过如 VQGAN（Esser et al., 2021）、CogView（Ding et al., 2021）和 Parti（Yu et al., 2022b）等开创性工作扩展到图像生成。这些方法采用两阶段过程，其中图像标记器首先将连续图像编码为离散标记，然后通过一个变压器解码器通过下一个标记预测来生成图像。在此基础之上，后续的研究如 LlamaGen（Sun et al., 2024）、Emu3（Wang et al., 2024b）、Chameleon（Team, 2024）、Lumina-mGPT（Liu et al., 2024a）和 Loong（Wang et al., 2024c）扩展了 AR 模型，表明 AR 模型的生成性能与扩散模型相当。此外，最近的研究（Li et al., 2024b; Fan et al., 2024; Zhou et al., 2025; Xie et al., 2025c）探索了将 AR 模型与扩散技术相结合的混合方法。

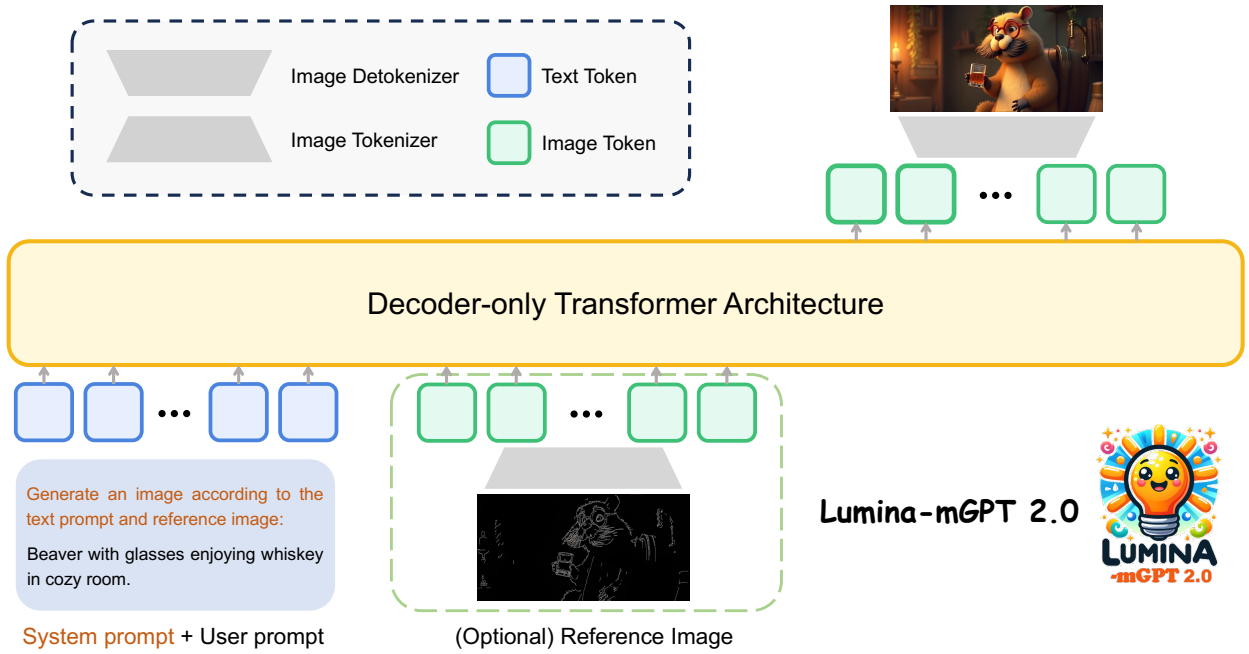

Figure 1: Lumina-mGPT 2.0 的解码器专用 Transformer 架构。该架构利用自回归建模进行图像合成，并支持条件图像输入，通过集成文本提示和可选参考图像，促进广泛的生成任务。

2. 重新审视 Lumina-mGPT

Lumina-mGPT（Liu et al., 2024a）代表了一系列开源的多模态自回归模型，这是首个具备高质量、高分辨率以及灵活长宽比图像生成能力的模型。在 Lumina-mGPT 的核心有两个设计：灵活递进监督微调（FP-SFP），一种通过逐步提高图像分辨率的训练方案；以及明确的图像表示（Uni-Rep），这是一种专门的图像表示机制，用于处理 1D 扁平化图像标记固有的二维形状不确定性，并构建了 Lumina-mGPT 在可变长宽比下理解和生成图像能力的基础。然而，虽然这项工作取得了重要进展，但也存在某些不足之处：

从加载预训练模型的约束。Lumina-mGPT 从预训练的 Chameleon（Team, 2024）检查点初始化。正如其他加载预训练权重的工作（Chen et al., 2025）所面临的，这一选择带来了关键的限制：模型架构及其附带的组件（如图像/文本分词器）必须严格符合预训练模型，这不仅阻碍了定制化的需求（例如创建不同大小的模型），而且也限制了追求更佳选择的空间。此外，预训练模型还带来了许可证限制，限制了其在更广泛商业和现实世界中的应用。

模型中的多任务冲突。除了基础的 T2I 生成任务，Lumina-mGPT 需要经过单独的微调，以扩展其在不同下游生成任务中的能力。它将这些任务视为具有不同检查点的条件扩展，而不是在一个框架内将它们整合到统一的训练范式中。这种分离阻碍了多任务目标与主要图像生成任务的有效对齐，限制了模型整体的一致性和效率。

有限推理时间优化。多模态自回归模型依赖于数千步的下一个令牌预测，导致显著的计算开销和延长的推理时间。值得注意的是，存在许多推理优化技术（Teng et al., 2025; Guo et al., 2025），可以显著减少采样时间或潜在地提高生成质量。然而，在 Lumina-mGPT 中，这一方向的探索是缺失的。

性能逊于 SOTA 模型。尽管图像生成能力已经在其领域取得了一个里程碑，但仍然落后于最新的扩散模型，例如 Lumina-Image 2.0（Qin et al., 2025）、Sana（Xie et al., 2025a）和 DALL · E 3（Betker et al., 2023），这为进一步改进留有空间。

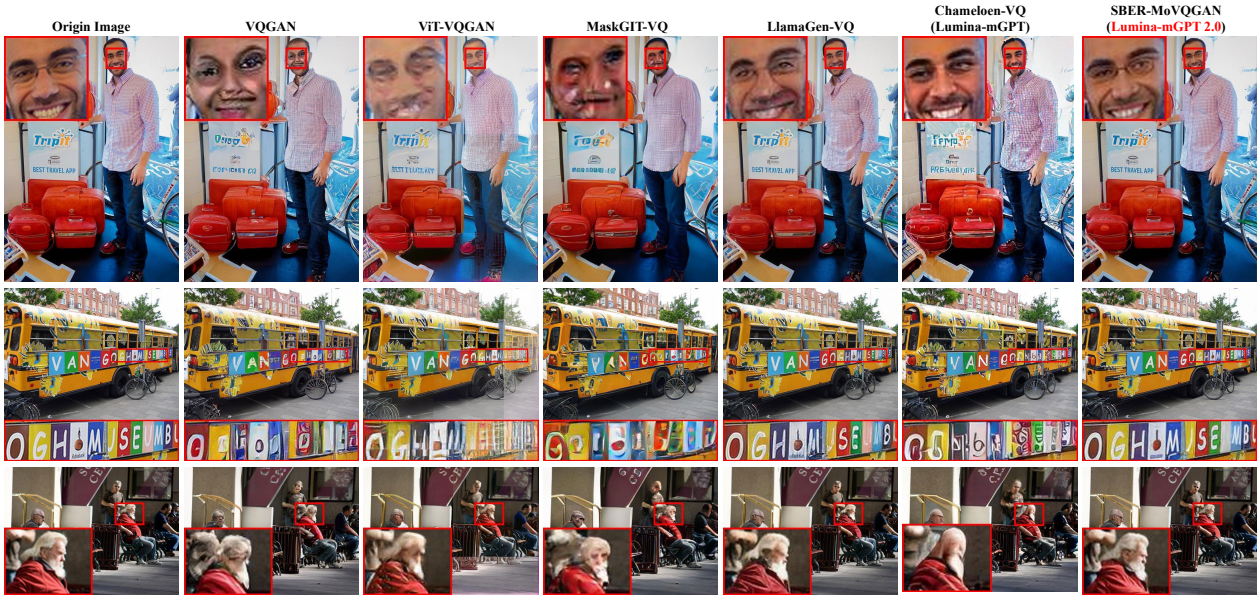

Figure 2: 不同图像标记器的图像重建结果。我们展示了被 红色框 突出显示的选定区域的具体细节，以展示各种图像标记器的性能。

Tokenizer	Year	Ratio	PSNR ↑	SSIM ↑	LPIPS ↓
VQGAN (Esser et al., 2021)	2021	16 × 16	18.70	0.48	0.17
ViT-VQGAN (Yu et al., 2022a)	2022	8 × 8	18.88	0.60	0.16
MaskGIT-VQ (Chang et al., 2022)	2022	16 × 16	18.55	0.47	0.19
SBER-MoVQGAN (Razzhigaev et al., 2023)	2023	8 × 8	22.77	0.63	0.08
LlamaGen-VQ (Sun et al., 2024)	2024	8 × 8	21.91	0.61	0.09
Chameleon-VQ (Team, 2024)	2024	16 × 16	18.63	0.47	0.18

Table 1: 流行图像分词器的重建性能。我们主要在 MS-COCO 数据集上评估各种图像分词器的重建指标（例如，PSNR、SSIM 和 LPIPS）(Lin et al., 2014) 。

3. Lumina-mGPT 2.0

在本节中，我们介绍了我们的 Lumina-mGPT 2.0 的方法论，该方法论具有三个特点：1）独立架构，2）统一各种生成任务，以及 3）优化的推理策略。

3.1. 独立架构

仅解码器 Transformer 从头训练。Lumina-mGPT 2.0 保持了与其前身 Lumina-mGPT (Liu et al., 2024a) 相似的结构设计，继续使用仅解码器的 Transformer 架构，如图 1 所示。与依赖于微调预训练的 Chameleon 7B 和 34B 模型的 Lumina-mGPT 相比，Lumina-mGPT 2.0 是一个完全独立开发的模型。具体而言，Lumina-mGPT 2.0 采用了从零开始训练的范式，参数随机初始化，从而带来了几个优势：1）偏差减少：从头开始训练最大程度地减少了通常从预训练模型继承的偏差，因此提高了图像生成性能。2）灵活架构：这种方法允许在模型设计中进行灵活适应。例如，我们为 T2I 社区提供了一个具有 2B 参数的更轻量级版本的 Lumina-mGPT 2.0。此外，它提供了根据需要集成改进的图像和文本分词器的灵活性。3）许可证独立性：由于不依赖于 Chameleon 模型，Lumina-mGPT 2.0 避免了任何潜在的许可限制。

复活 SBER-MoVQGAN 图像标记器。图像标记器的重建质量在决定生成质量的上限方面起着至关重

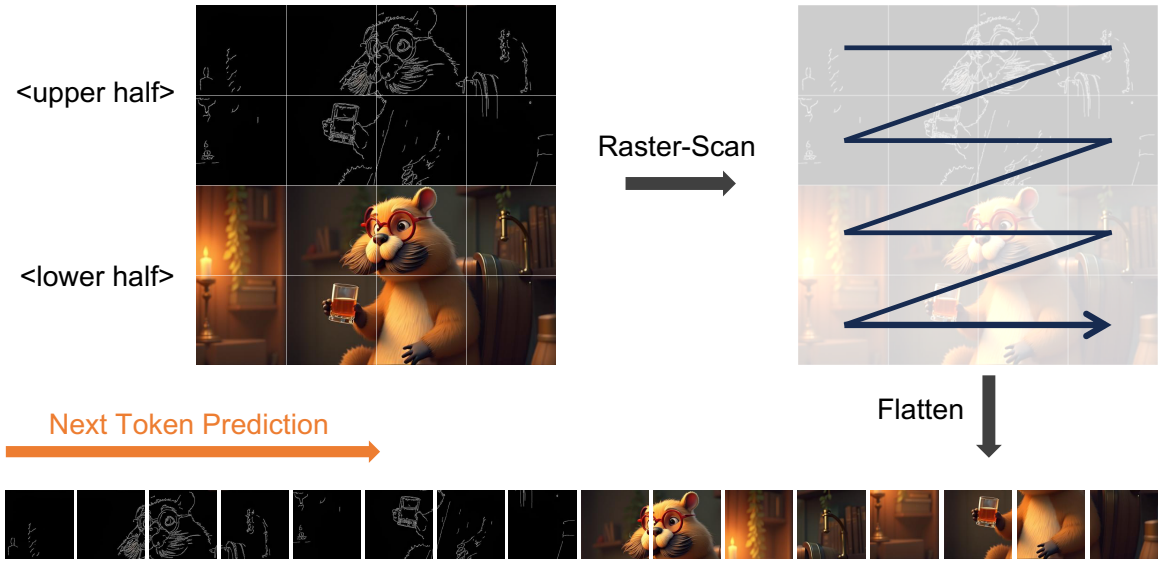

Figure 3: 通过自回归光栅扫描方案统一不同的生成任务。该模型首先生成图像的上半部分（或在推理过程中给定的参考图像），其作为生成下半部分的上下文引导。

Model	Params	Vocabulary Size	Hidden Size	Intermediate Size	Heads	KV Heads	Layers
Lumina-mGPT 2.0	2B	171,385	2,048	8,192	32	32	32
Lumina-mGPT 2.0	7B	171,385	4,096	11,008	32	32	32

Table 2: Lumina-mGPT 2.0 的架构配置。该模型从 2B 扩展到 7B 参数，主要的扩展策略是增加模型的维度。

要的作用。由于 Lumina-mGPT 2.0 是一个独立的模型，它允许灵活使用标记器。为了实现高质量的生成，我们对 AR 框架中使用的流行图像标记器进行了全面的重建质量分析，包括 VQGAN（Esser et al., 2021），ViT-VQGAN（Yu et al., 2022a），MaskGIT-VQ（Chang et al., 2022），LlamaGen-VQ（Sun et al., 2024），SBER-MoVQGAN（Razzhigaev et al., 2023）和在 Lumina-mGPT 中使用的 Chameleon-VQ（Team, 2024）。表 1 和图 2 中展示的 MS-COCO 数据集（Lin et al., 2014）上的对比结果表明，SBER-MoVQGAN 当前是图像重建的 SOTA 模型。因此，Lumina-mGPT 2.0 采用 SBER-MoVQGAN 以确保优越的生成性能。然而，一个挑战是 8 × 8 的下采样比率，这导致了更长的图像标记序列，并增加了推理时间和成本。

不使用预训练的文本编码器。在 Lumina-mGPT 2.0 中，我们使用基于标记的格式同时表示文本和图像数据，如图 1 所示。这种方法使得 Lumina-mGPT 2.0 区别于一些常规的 AR 方法（Sun et al., 2024; Yu et al., 2022b），后者通常使用预训练的文本编码器来获取编码的文本特征，然后通过 MLP 将这些特征投射到模型中。相反，Lumina-mGPT 2.0 利用 QwenTokenizer（Wang et al., 2024a）直接将文本编码为离散标记。这种方法简化了流程，转变为纯粹的下一个标记预测范式，从而消除了加载预训练文本编码器的必要性。

模型扩展。为了展示我们独立自回归图像建模的可扩展性，我们提供了 Lumina-mGPT 2.0 系列中的两个模型尺寸：2B 和 7B。每个模型的超参数详见表 2 。扩展主要涉及增加模型的维度。在实验过程中，我们观察到随着模型规模的增加，训练损失的收敛速度加快，生成图像的质量在连贯性、细粒度细节以及对精细化提示的忠实度方面显著改善，详见第 4.2 节。这些进展强调了我们模型的稳健扩展特性。

Text-to-Image Generation	Generate an image of { width } × { height } according to the following text prompt: { User Text Prompt }
Subject-Driven Generation	Generate a dual-panel image of { width } × { height } where the < 上半部分 > displays: { Object Description } , while the < 下半部分 > shows the image according to the object and following prompt: { Subject Driven Prompt } .
Image Editing	Generate a dual-panel image of { width } × { height } where the < 上半部分 > displays an image: { Image Description } , while the < 下半部分 > shows an image according to the upper part: { Editing Instruction }
Controllable Generation	Generate a dual-panel image of { width } × { height } where the < 上半部分 > shows a < 控制任务 > image, while the < 下半部分 > displays the image according to the upper part and following prompt: { User Text Prompt }
Dense Prediction Tasks	Generate a dual-panel image of { width } × { height } where the < 的上半部分 > displays the image according to the following description: { Image Description } , while the < 下半部分 > shows a < 控制任务 > image according to the upper part.

Table 3: Lumina-mGPT 2.0 的系统提示。< 下半部分 > 和 < 上半部分 > 是特殊字符，指示图像的位置。< 控制任务 > 表示下游任务，包括边缘检测、深度估计、姿态估计等。

3.2. 统一多样化生成任务

Lumina-mGPT 2.0 的自回归架构促进了在联合序列生成框架内统一多种视觉任务。具体来说，我们利用了自回归方法的一个关键优势：它自然的以光栅扫描形式的图像令牌生成顺序。这确保了图像的上部分首先生成，为后续生成的下部区域提供上下文指导，正如图 3 所示。

基于这一特性，我们将各种文本-图像到图像的任务整合到我们的框架中，包括以主体为导向的生成、图像编辑、可控生成和密集预测。此外，这种特殊范式还可以生成图像对，如图 6 所示。给定多个图像，我们简单地将它们垂直拼接成一个图像网格以进行联合建模，从而确保上部图像区域在生成过程中充当上下文。对于可控生成，条件图像放置在顶部，而生成的输出放置在其下方。同样，对于诸如深度估计之类的密集预测任务，原始图像被置于顶部，相应的标签图放置在底部。为了进一步区分这些任务类型，我们利用了 < 系统的提示 > ，如表 3 所示。

这种统一范式促进了原生的多任务训练（所有任务都被视为文本到图像生成），使模型能够同时学习各种视觉任务而无需额外的架构更改。根据在自回归模型中的标准 T2I 训练，损失仅针对图像标记计算，文本标记保持不变。在推理时，这种公式提供了一种灵活且动态的提示构建机制。用户可以明确指定 < 系统提示 > 以控制任务类型，决定是否提供参考图像作为指导条件，如图 1 所示。这种能力无缝弥合了可控生成与无条件生成之间的差距，拓宽了 Lumina-mGPT 2.0 作为视觉通才的原生多任务能力。

3.3. 优化的推理策略

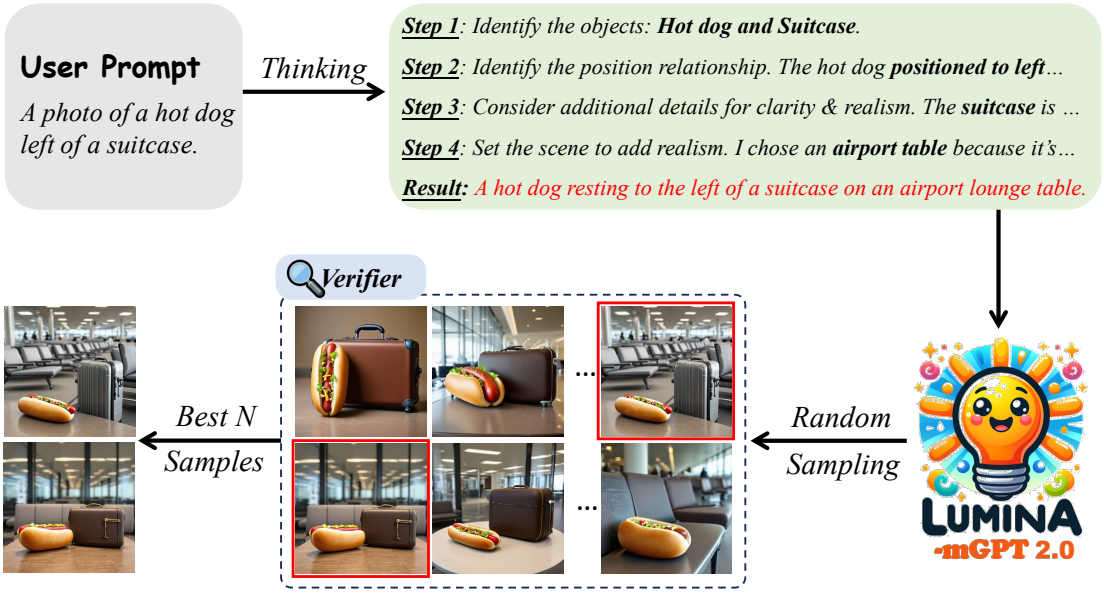

Figure 4: 高质量采样流程。我们首先考虑用户的提示，并对其进行详细说明以增强清晰度和连贯性。随后，我们使用最佳的 N 策略从生成的候选中选择最优图像。

3.3.1. 高质量采样

生成前的思考。图像生成类似于艺术创作，需要在执行之前进行深刻的思考、概念化和迭代的完善。然而，当前的图像生成模型通常缺乏这种创造前的推理，而是将文本提示视为直接指令，而不是一个渐进的思维过程。实际上，用户的提示往往是模糊的、含糊不清的，或者在生成连贯且有意义的图像所需的关键细节上存在不足。受到了在大规模语言模型（LLMs）中的链式思维（CoT）推理显著进展的启发（Wei et al., 2022; Wang et al., 2023），我们引入了一种“生成前思考”的范式。

具体来说，与其直接将用户提示输入到 Lumina-mGPT 2.0 中，我们首先通过一个大型语言模型（GPT-4o）进行处理。该大型语言模型（LLM）被提示系统地分析和理解用户的潜在意图，通过逐步推理，最终生成一个经过优化的提示，具有增强的连贯性、描述丰富性和清晰性。例如，当遇到一个无意义的提示时，模型会推断出一个合理的解释；当面临歧义时，它会进行澄清和阐述；当提示过于简单时，它会通过加入基本元素丰富描述，细节如图 5 。在这个过程中，LLM 作为一个推理引擎，逐步优化提示，就像艺术家迭代发展其愿景一样。通过整合这一反思性推理过程，我们的方法确保最终的提示不仅结构更为严谨而且更具表现力，同时也更忠实于用户的原始意图。

推理时间缩放。最近的研究开始调查扩散模型中的推理时间缩放行为（Ma et al., 2025; Xie et al., 2025b; Zhuo et al., 2025）以及混合 AR 和扩散模型（Guo et al., 2025）。在这些进展的基础上，我们首次探索在我们的 Lumina-mGPT 2.0 中纯粹的 AR 推理时间缩放。具体来说，给定一个文本提示，Lumina-mGPT 2.0 首先以随机方式生成一组多样的候选图像。随后，我们采用最佳 N 选择策略，其中验证器评估生成的图像并选择最优的一个。鉴于图像的固有复杂性和嵌入在文本条件中的丰富语义信息，更全面地评估生成质量显得尤为重要。为此，我们整合了多个验证器，包括 VQAScore [1]（Li et al., 2024a），LAION-AestheticScore [2]（Schuhmann et al., 2022），PickScore [3]（Kirstain et al., 2023）。

[1]VQAScore 是一种评分方法，旨在评估生成模型的提示跟随能力。
[2]LAION-AestheticScore 是一种用于评估美学质量的评分方法。
[3]PickScore 是一种旨在评估人类偏好的评分机制。

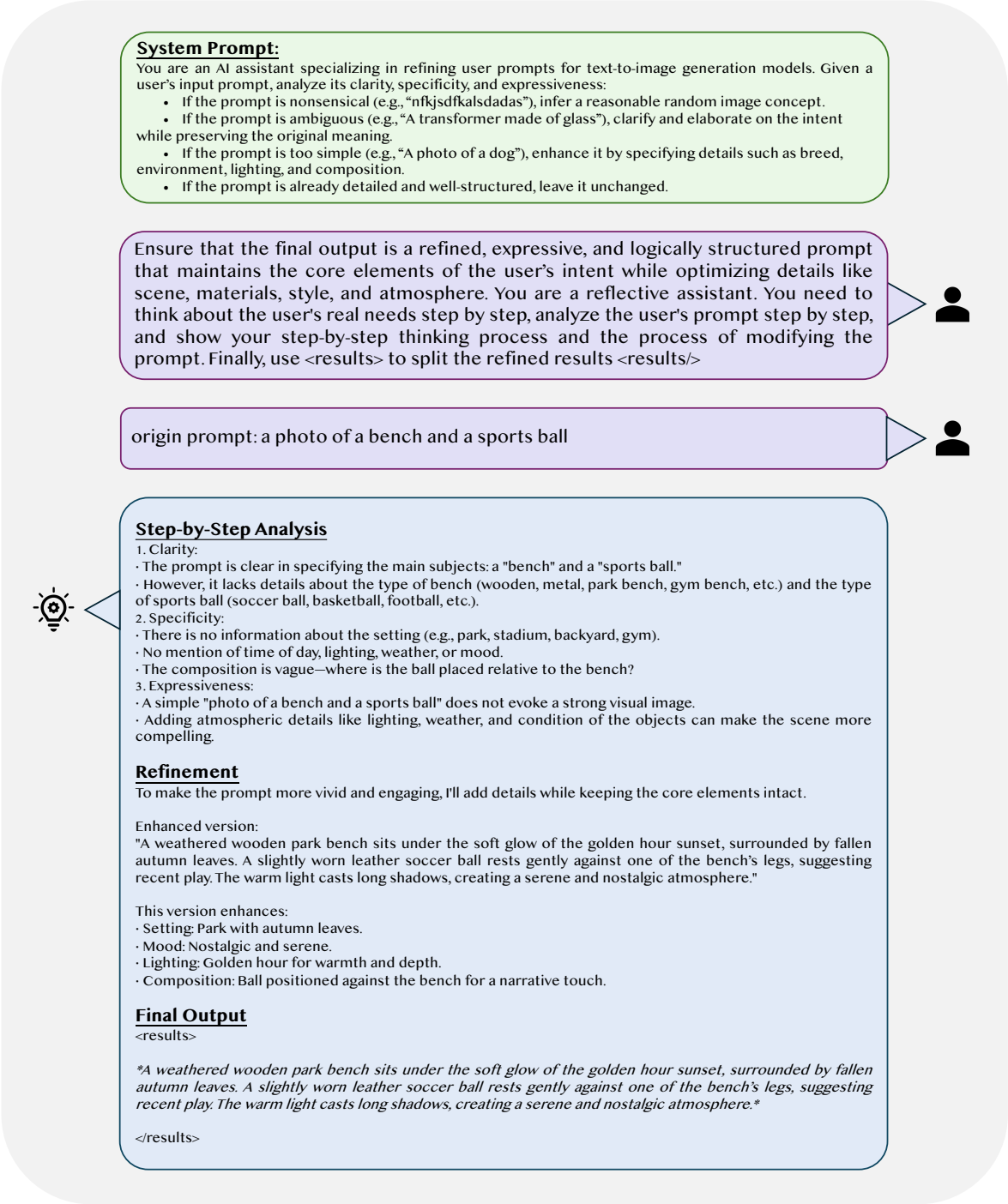

Figure 5: 一个调用 GPT-4o 来完成思维过程的例子。这种方法不同于传统的提示重写过程，因为它以一个周到的、逐步的思维过程为中心。

3.3.2. 加速度采样

模型量化。为了优化 GPU 内存使用和加速推理，我们使用 TorchAo （torchao maintainers and contributors, 2024) 对 Lumina-mGPT 2.0 的前向解码模块进行训练后量化。该方法将模型权重量化为具有 128 元分组的 4 位整数，同时将激活张量保持在 bfloat16 精度以减轻潜在的质量退化。利用 PyTorch 2.0 （Ansel et al., 2024) 中的本地编译工具包，我们通过 torch.compile 的减小开销模式引入量化操作，促进内核自动调优和静态图优化。值得注意的是，这种优化是在无需对模型架构进行

Methods	# Params	GenEval ↑				DPG ↑			
		Two Obj.	Counting	Color Attri.	Overall	Entity	Relation	Attribute	Overall
Diffusion Models									
SDv1.5 (Rombach et al., 2022)	0.9B	-	-	-	0.40	74.23	73.49	75.39	63.18
Lumina-Next (Zhuo et al., 2024)	1.7B	0.49	0.38	0.15	0.46	83.78	89.78	82.67	75.66
SDv2.1 (Rombach et al., 2022)	0.9B	0.51	0.44	0.50	0.47	-	-	-	68.09
PixArt- α (Chen et al., 2023)	0.6B	0.50	0.44	0.07	0.48	79.32	82.57	78.60	71.11
SDXL (Podell et al., 2024)	2.6B	0.74	0.39	0.23	0.55	82.43	86.76	80.91	74.65
SD3-medium (Esser et al., 2024)	2B	0.74	0.63	0.36	0.62	91.01	80.70	88.83	84.08
Sana-1.6B (Xie et al., 2025a)	1.6B	0.77	0.62	0.47	0.66	91.50	91.90	88.90	84.80
DALL-E3 (Betker et al., 2023)	-	0.87	0.47	0.45	0.67	89.61	90.58	88.39	83.50
OmniGen (Xiao et al., 2025)	3.8B	0.86	0.64	0.55	0.70	-	-	-	-
Lumina-Image 2.0 (Qin et al., 2025)	2.6B	0.87	0.67	0.62	0.73	91.97	94.85	90.20	87.20
AutoRegressive Models									
LlamaGen (Sun et al., 2024)	0.8B	0.34	0.21	0.04	0.32	-	-	-	65.16
Chameleon (Team, 2024)	7B	-	-	-	0.39	-	-	-	-
Show-o (Xie et al., 2025c)	1.3B	0.52	0.49	0.28	0.53	-	-	-	67.48
Emu3 (Wang et al., 2024b)	8.0B	0.71	0.34	0.21	0.54	86.68	90.22	86.84	80.60
Infinity (Han et al., 2025)	2B	0.85	-	0.57	0.73	-	90.76	-	83.46
Janus-Pro-1B (Chen et al., 2025)	1.5B	0.82	0.51	0.56	0.73	88.63	88.98	88.17	82.63
Janus-Pro-7B (Chen et al., 2025)	7B	0.89	0.59	0.66	0.80	88.90	89.32	89.40	84.19
Lumina-mGPT (Liu et al., 2024a)	7B	0.77	0.27	0.32	0.56	86.60	91.29	84.61	79.70
Lumina-mGPT 2.0 †	2B	0.83	0.50	0.54	0.68	87.37	90.03	84.79	82.05
Lumina-mGPT 2.0 †	7B	0.92	0.57	0.72	0.80	88.94	91.70	88.08	84.30

Table 4: 在 GenEval (Ghosh et al., 2024) 和 DPG (Hu et al., 2024) 基准上，不同 T2I 模型的性能对比。† 表示 GenEval 是我们采样优化策略的结果。我们分别对扩散和 AR 的 最佳 和 第二 结果进行了突出显示。

任何修改的情况下实现的。

推测 Jacobi 解码。我们通过采用推测 Jacobi 解码（SJD） (Teng et al., 2025) 来改进采样策略，该方法将确定性 Jacobi [4] (Santilli et al., 2023; Song et al., 2021b) 迭代与随机采样相结合。SJD 引入了一种概率收敛准则，该准则根据草稿和目标标记分布之间的似然比来接受标记，从而在保留样本多样性的同时实现并行解码。

在实践中，我们的目标是通过共同利用模型量化和 SJD 来加速采样。然而，一个关键挑战来自 SJD 对动态 KV 缓存的固有需求。在传统的自回归解码中，KV 缓存通过在生成标记时附加新的键和值张量动态增长。SJD 通过引入迭代细化扩展了这一点，其中标记可能会根据收敛标准被接受或拒绝，这需要一个灵活的缓存来处理可变的序列长度和标记回退。这种动态行为与编译操作施加的限制相冲突，例如在 torch.compile 中的那些，它们要求一个静态的、预分配的 KV 缓存以生成优化的内核并最小化重新编译的开销。

为了解决这个问题，我们提出了一种用于 SJD 的静态 KV 缓存和静态因果注意力掩码，从而实现与静态编译框架的兼容性。KV 缓存预先分配固定大小的缓冲区，并使用基于指针的方法来管理有效的序列长度，避免动态调整大小。类似地，提前计算了固定大小的注意力掩码，以支持解码阶段，并在推理过程中使用指针调整掩码，确保高效和并行的令牌预测。该设计在满足静态内存要求的同时保持 SJD 的效率。

4. 实验

[4] Jacobi 解码通过反复预测多个标记直至收敛，加速自回归推理，而无需微调。

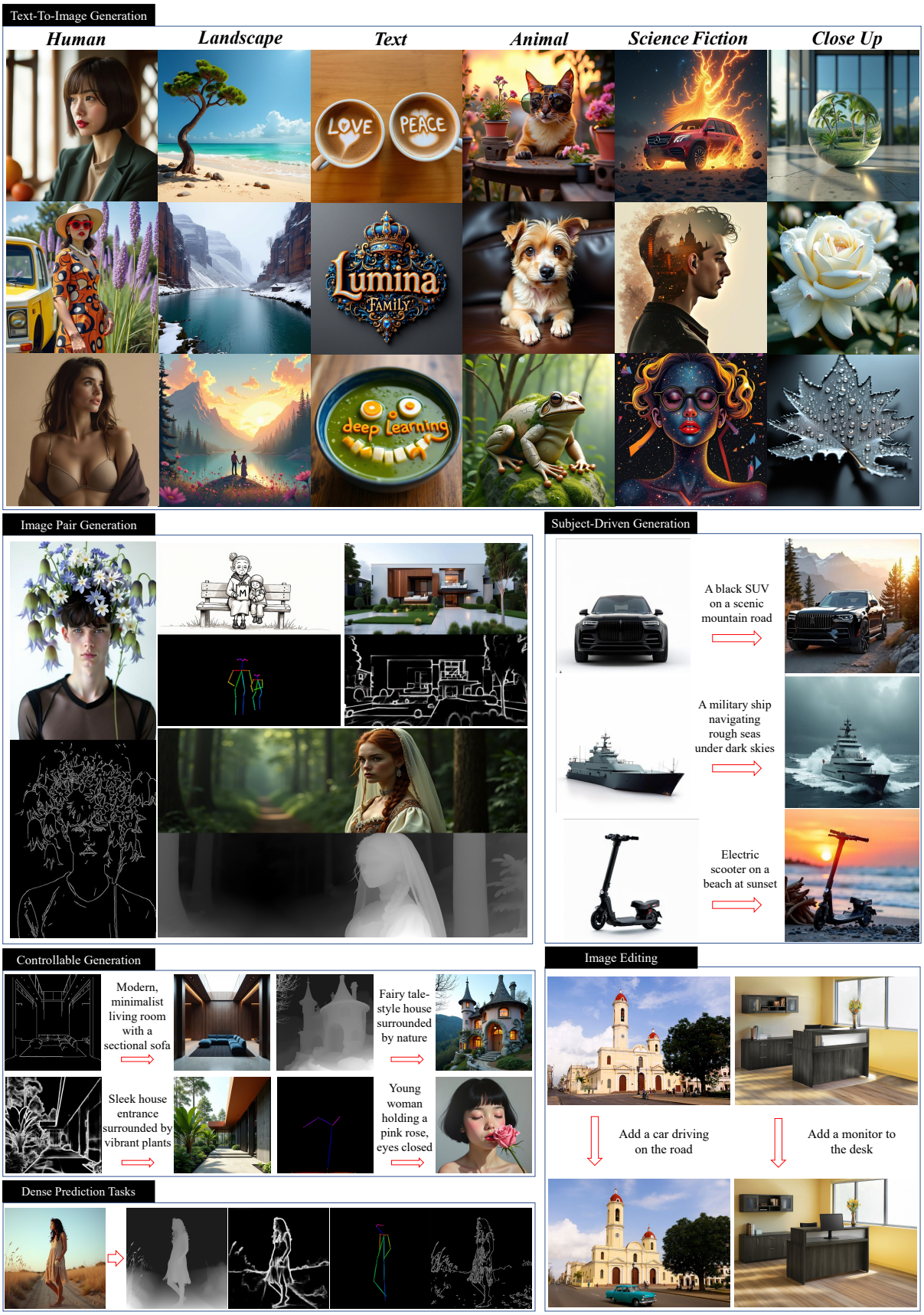

Figure 6: Lumina-mGPT 2.0 的文本到图像生成和多任务生成结果。

4.1. 实现细节

训练数据。我们的 T2I 训练数据集是从 Lumina-Image 2.0 (Qin et al., 2025) 中提取的一个子集，包含真实和合成数据。该数据集经过了用 OmniCaptioner (Lu et al., 2025) 的精心过滤和重新标注。对于多任务训练，使用了不同的数据集：主题驱动的生成任务使用了 Subject200K (Tan et al., 2025)，编辑任务使用了 OminiEdit (Wei et al., 2025)，并且可控生成和密集预测任务是通过从 T2I 数据集中随机抽取 200,000 条目开发的。

训练细节。最近的研究 (Qin et al., 2025) 已经表明，采用分层数据显著提高了性能，特别是在高保真图像合成中。因此，我们的训练流程采用三阶段的分层策略：首先在 256px 下以 2e-4 的学习率进行模型的预训练，然后在 512px 和 768px 下以 2e-5 的学习率进行微调。通过梯度累积，全局批次大小在 512 到 1,024 之间动态调整。Lumina-mGPT 2.0 每个阶段的训练数据结构如下：256px 阶段使用 5000 万数据，512px 阶段使用 1900 万数据，768px 阶段使用 800 万数据。随着数据集规模的缩小，数据质量逐步提高。所有训练均在 64 块 A100 GPU 上进行分布式训练。

评估细节。我们采用多种 T2I 评估基准来评估 Lumina-mGPT 2.0，包括 GenEval (Ghosh et al., 2024) 和 DPG (Hu et al., 2024)，其测试集分别包含 553 和 1065 个提示。这些指标主要关注评估文本-图像对齐。此外，我们使用 GenEval 进行全面的消融研究，以验证我们的推理策略的有效性。对于多任务评估，我们采用由 VisualCloze (Li et al., 2025a) 提出的基准，这使我们能够分别使用 1000 个样本来评估可控生成和主题驱动生成的能力。

文本到图像生成。我们在表 4 中将 Lumina-mGPT 2.0 与先进的 T2I 生成方法在两个基准上进行比较。我们的模型实现了与 AR 模型和基于扩散的模型（包括 Emu3 (Wang et al., 2024b)、Janus Pro (Chen et al., 2025)，甚至 Lumina-Image 2.0 (Qin et al., 2025)）相当或更高的性能。值得注意的是，Lumina-mGPT 2.0 取得了 0.80 的 GenEval 得分，位居当前顶级生成模型之中。它尤其在与“两个物体”和“颜色属性”相关的 GenEval 测试中表现出色。此外，Lumina-mGPT 2.0 获得了 84.30 的 DPG 得分，超越了之前为 AR 模型设定的上限。

原生多任务。我们主要评估 Lumina-mGPT 2.0 在可控生成方面的能力，如表 **??** 所示，以及主体驱动生成方面的能力，如表 5 所示。结果显示，Lumina-mGPT 2.0 作为一个通用的通才模型表现出色。在可控生成领域，它实现了顶级性能，在 Canny 和 Depth 条件下表现出高度的结构依从性，同时保持卓越的图像质量和文本一致性。在主体驱动任务中，Lumina-mGPT 2.0 在维护主体身份方面超过了所有竞争对手，并在图像一致性和文本对齐方面取得了令人印象深刻的结果。

Method	DINOv2 ↑	CLIP-I ↑	CLIP-T ↑
OminiControl (Tan et al., 2025)	73.17	87.70	33.53
OneDiffusion (Le et al., 2025)	73.88	86.91	34.85
OmniGen (Xiao et al., 2025)	67.73	83.43	34.53
Lumina-mGPT (Liu et al., 2024a)	60.94	70.63	30.16
Lumina-mGPT 2.0	76.60	87.37	33.90

Table 5: 针对主体驱动图像生成的定量比较。我们报告了文本对齐和风格一致性的 clip 分数。专家在 灰色 中以阴影显示。在其余方法中，我们强调了 最佳 和 第二 的结果。

4.1.1. 定性表现

文本到图像生成。我们在图 6 中展示了 T2I 生成结果，展示了该模型在合成高保真视觉内容方面的熟练程度，涵盖广泛的类别。Lumina-mGPT 2.0 能够有效生成逼真的人物、令人叹为观止的景观，以及具有卓越细节的复杂文本设计。此外，它在生动呈现栩栩如生的动物、富有想象力的科幻场景以及高度详细的特写镜头方面表现出色。这些结果突出了该模型准确解释提示的能力，捕捉丰富的纹理、动态的光效和引人注目的构图。

此外，我们对比分析了 Lumina-mGPT 2.0 与 Janus Pro 及其前身 Lumina-mGPT 的 T2I 结果，如图 7 所示。Lumina-mGPT 2.0 在真实性、细节和连贯性方面显示出显著改进，不仅超过了其前身，还超越了 Janus Pro。生成的图像具有更清晰的纹理、更精确的光照和增强的构图，使其在视觉上

Figure 7: 对比 Lumina-mGPT 2.0、Lumina-mGPT 和 Janus Pro 之间的文本到图像的视觉效果

Figure 8: 可控/主体驱动生成在 Lumina-mGPT 2.0、Lumina-mGPT、OneDiffusion 和 OmniGen 之间的视觉比较。控制输入包括 Canny（第一行）和 Depth（第二行）。

更具吸引力，并与自然美学更为一致。有趣的是，这些发现与从表 4 得出的结论不同，在该表中，Lumina-mGPT 2.0 和 Janus Pro 在 GenEval 基准测试中表现相当。此基准测试主要依赖于 VLM 模型来评估文本与图像之间的对齐程度，而未明确考虑图像生成的质量和美学。

原生多任务。除了文本生成图像（T2I）之外，Lumina-mGPT 2.0 还显示了显著的多任务能力，如图 6 所示。结果表明，Lumina-mGPT 2.0 天生支持广泛的图像到图像生成任务，包括主体驱动生成、图像编辑和可控生成（例如，从 canny 到图像，从深度到图像，从姿态到图像，从 hed 到图像），这一切都不需要额外的模块 (Li et al., 2025b; Yao et al., 2024) 或额外的微调阶段 (Chung et al., 2025)。此外，Lumina-mGPT 2.0 可以高效生成特定任务的图像对，以增强其他模型的图像到图像任务训练的数据集，同时为各种密集预测任务提供强大的支持。

此外，我们进行多任务生成视觉比较，包括 Lumina-mGPT、OneDiffusion 和 OmniGen，如图 8 所示。Lumina-mGPT 2.0 在可控生成和主题驱动生成任务中展示了出色的性能。

4.2. 消融研究

模型增长的效果。在 Lumina-mGPT 2.0 中，我们将模型从 20 亿参数扩展到 70 亿参数。为了评估这种扩展的影响，我们从三个角度进行分析：(1) 基准性能：如表格 4 所示，70 亿参数的模型在多个基准测试中，包括 GenEval 和 DPG，都表现优于 20 亿参数的变体。(2) 训练损失：如图 9 所示，较大的 70 亿模型相比较小的 20 亿模型显示出训练损失收敛明显更快。(3) 视觉质量：如图 7 所示，70 亿参数模型生成的结果表现出更高的稳定性、增强的视觉保真度和更丰富的细节。

思考后生成的效果。在图像生成之前，我们调用 GPT-4o API 对输入提示进行深入分析，充分理解其含义，并生成增强提示。图 5 展示了逐步思考过程和生成的增强提示实例。为了评估该方法的有效性，我们在 GenEval 基准上进行了消融实验，如表 6 所示。结果表明，经过思考过程的提示平均提升了 4 %，尤其在位置和颜色归因能力上都有显著增强，各提高了 8 %。这些发现表明，该方法能更有效地支持图像生成过程，更加贴近用户的意图。

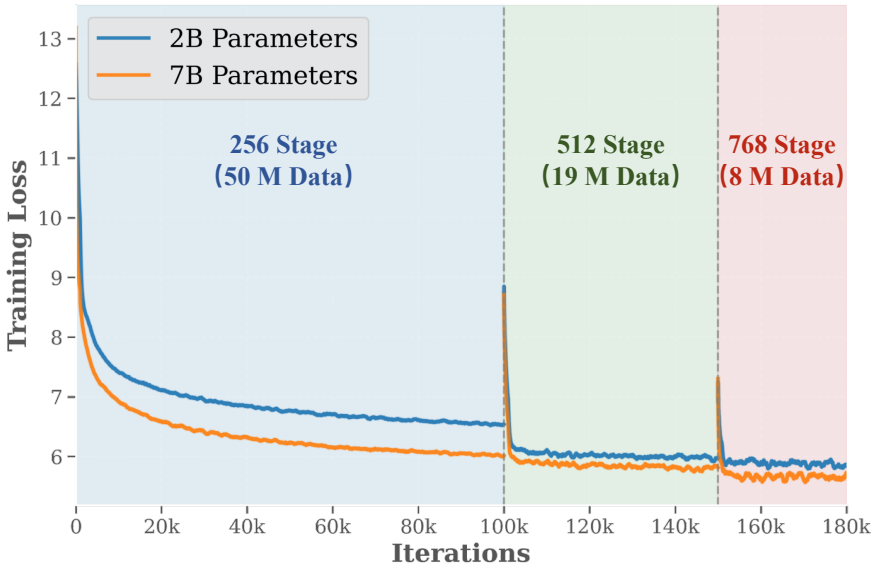

Figure 9: 三阶段训练损失曲线用于 Lumina-mGPT 2.0。训练损失曲线比较了两个模型规模的性能，分别为 2B 和 7B 参数。7B 模型在整个训练过程中始终达到比 2B 模型更低的训练损失，表明其收敛性更好。

Method	# Params.	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall
LlamaGen (Sun et al., 2024)	0.8B	0.71	0.34	0.21	0.58	0.07	0.04	0.32
Emu-3 (Wang et al., 2024b)	8B	0.98	0.71	0.34	0.81	0.17	0.21	0.54
Janus-Pro-7B (Chen et al., 2025)	7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80
Lumina-mGPT (Liu et al., 2024a)	7B	0.98	0.77	0.27	0.82	0.17	0.32	0.56
Lumina-mGPT-2.0	7B	0.99	0.87	0.44	0.85	0.44	0.54	0.69
+ Thinking Before Generation	7B	1.00	0.87	0.49	0.85	0.52	0.62	0.73
+ Inference-Time Scaling	7B	1.00	0.92	0.57	0.88	0.70	0.72	0.80

Table 6: 高质量采样策略的消融研究。实验在 GenEval（Ghosh et al., 2024）上进行。

推理时缩放的效果。我们将推理时缩放整合到 Lumina-mGPT-2.0 中，并在 GenEval 基准上评估其性能，与其他大规模图像生成模型进行比较，具体如表 6 所示。通过从一组生成的 16 张图像中选取样本，推理缩放模型在总体准确率上比简单的单张图像生成提高了 11 %。这些提升在“两个物体”、“计数”、“位置”和“颜色归因”子任务中尤为显著。这些发现强调，即使模型容量受限，牺牲推理效率也可以显著提升生成质量和准确性。

加速采样策略的效果。在图 10 中，我们通过整合模型量化和推测雅可比解码（SJD）策略评价了 Lumina-mGPT 2.0 的采样效率。实验结果表明，模型量化在保持视觉保真度的同时，将采样时间减少了 48 %，并将 GPU 内存消耗减少了 47 %。在此基础上，SJD 通过其并行解码机制进一步提高了效率，实现了 72 % 的采样时间减少。这些采样策略有效解决了 Lumina-mGPT 2.0 的采样速度慢的问题，这是自回归生成模型中的一个常见挑战，从而使其在实际应用中更为用户友好。

5. 结论

本文介绍了 Lumina-mGPT 2.0，这是一种独立的、仅解码的自回归图像生成模型。Lumina-mGPT 2.0 完全从零开始训练，没有引用任何预训练模型的权重。在文本到图像的生成任务中，它在标准基准测试中达到了可与最先进模型相媲美的性能，同时在合成图像中表现出卓越的视觉质量。此外，Lumina-mGPT 2.0 原生支持多种下游任务，提高了其灵活性和对更广泛研究社区的可用性。

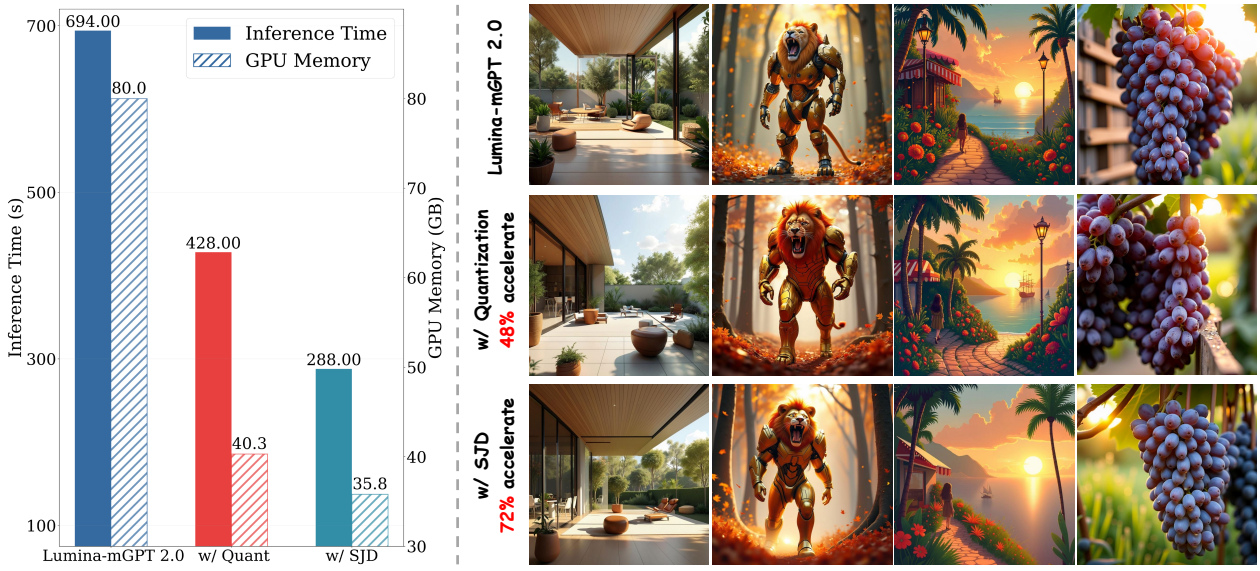

Figure 10: 关于高效采样策略的消融研究。结果表明，SJD 和模型量化都显著减少了推理时间和 GPU 内存，同时仍然生成高质量的图像。

限制。尽管已努力优化推理过程，Lumina-mGPT 2.0 仍旧遇到需要数分钟的采样时间，这个挑战在所有基于 AR 的生成模型中都是普遍存在的，可能导致不理想的用户体验。目前，Lumina-mGPT 2.0 依赖于外部的大型语言模型进行思考过程。未来的改进目标是使 Lumina-mGPT 2.0 能够执行自主思考。此外，Lumina-mGPT 2.0 目前的重点是多模态生成，计划中的更新将扩展其能力，以包括多模态理解。

我们感谢滕曜和何也飞在加速采样方面的帮助。我们也感谢郭子瑜关于“在生成前先思考”的宝贵建议。

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 , 2023.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, et al. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems , 2024.

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. , 2023.

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

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 (NeurIPS) , 2020.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2022.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. Proceedings of the International Conference on Learning Representations (ICLR) , 2023.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 , 2025.

Jiwoo Chung, Sangeek Hyun, Hyunjun Kim, Eunseo Koh, MinKyu Lee, and Jae-Pil Heo. Fine-tuning visual autoregressive models for subject-driven generation. Proceedings of the IEEE International Conference on Computer Vision (ICCV) , 2025.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In Proceedings of the International Conference on Learning Representations (ICLR) , 2023.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems (NeurIPS) , 2021.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In Proceedings of the International Conference on Machine Learning (ICML) , 2024.

Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. arXiv preprint arXiv:2410.13863 , 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems (NeurIPS) , 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Communications of the ACM , 2020.

Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. arXiv preprint arXiv:2501.13926 , 2025.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2025.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135 , 2024.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2019.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in Neural Information Processing Systems (NeurIPS) , 2023.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Duong H. Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and Jiasen Lu. One diffusion to generate them all. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2025.

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. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops) , 2024a.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. Advances in Neural Information Processing Systems (NeurIPS) , 2024b.

Zhong-Yu Li, Ruoyi Du, Juncheng Yan, Le Zhuo, Zhen Li, Peng Gao, Zhanyu Ma, and Ming-Ming Cheng. Visualcloze: A universal image generation framework via visual in-context learning. Proceedings of the IEEE International Conference on Computer Vision (ICCV) , 2025a.

Zongming Li, Tianheng Cheng, Shoufa Chen, Peize Sun, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Controlar: Controllable image generation with autoregressive models. Proceedings of the International Conference on Learning Representations (ICLR) , 2025b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV) , 2014.

Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pre-training. arXiv preprint arXiv:2408.02657 , 2024a.

Dongyang Liu, Shicheng Li, Yutong Liu, Zhen Li, Kai Wang, Xinyue Li, Qi Qin, Yufei Liu, Yi Xin, Zhongyu Li, et al. Lumina-video: Efficient and flexible video generation with multi-scale next-dit. arXiv preprint arXiv:2502.06782 , 2025.

Wenze Liu, Le Zhuo, Yi Xin, Sheng Xia, Peng Gao, and Xiangyu Yue. Customize your visual autoregressive recipe with set autoregressive modeling. arXiv preprint arXiv:2410.10511 , 2024b.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in Neural Information Processing Systems (NeurIPS) , 2022.

Yiting Lu, Jiakang Yuan, Zhen Li, Shitian Zhao, Qi Qin, Xinyue Li, Le Zhuo, Licheng Wen, Dongyang Liu, Yuewen Cao, et al. Omnicaptioner: One captioner to rule them all. arXiv preprint arXiv:2504.07089 , 2025.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. arXiv preprint arXiv:2501.09732 , 2025.

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 , 2023.

William S Peebles and Saining Xie. Scalable diffusion models with transformers. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) , 2022.

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. Proceedings of the International Conference on Learning Representations (ICLR) , 2024.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Jiakang Yuan, Xinyue Li, Dongyang Liu, et al. Lumina-image 2.0: A unified and efficient image generative framework. Proceedings of the IEEE International Conference on Computer Vision (ICCV) , 2025.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Proceedings of the International Conference on Machine Learning (ICML) , 2021.

Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) , 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2022.

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. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2023.

Andrea Santilli, Silvio Severino, Emilian Postolache, Valentino Maiorca, Michele Mancusi, Riccardo Marin, and Emanuele Rodolà. Accelerating transformer inference for translation via parallel decoding. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) , 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems (NeurIPS) , 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. Proceedings of the International Conference on Learning Representations (ICLR) , 2021a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. Advances in Neural Information Processing Systems (NeurIPS) , 2019.

Yang Song, Chenlin Meng, Renjie Liao, and Stefano Ermon. Accelerating feedforward computation via parallel nonlinear equation solving. In Proceedings of the International Conference on Machine Learning (ICML) , 2021b.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 , 2024.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. Proceedings of the IEEE International Conference on Computer Vision (ICCV) , 2025.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 , 2024.

Yao Teng, Han Shi, Xian Liu, Xuefei Ning, Guohao Dai, Yu Wang, Zhenguo Li, and Xihui Liu. Accelerating auto-regressive text-to-image generation with training-free speculative jacobi decoding. Proceedings of the International Conference on Learning Representations (ICLR) , 2025.

torchao maintainers and contributors. torchao: Pytorch native quantization and sparsity for training and inference. In https://github.com/pytorch/torchao , 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 , 2023.

Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In Proceedings of the International Conference on Machine Learning (ICML) , 2016.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 , 2024a.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 , 2024b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. Proceedings of the International Conference on Learning Representations (ICLR) , 2023.

Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. Loong: Generating minute-level long videos with autoregressive language models. arXiv preprint arXiv:2410.02757 , 2024c.

Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In Proceedings of the International Conference on Learning Representations (ICLR) , 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems (NeurIPS) , 2022.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2025.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. International Journal of Computer Vision (IJCV) , 2024.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. Proceedings of the International Conference on Learning Representations (ICLR) , 2025a.

Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. Proceedings of the International Conference on Machine Learning (ICML) , 2025b.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. Proceedings of the International Conference on Learning Representations (ICLR) , 2025c.

Yi Xin, Le Zhuo, Qi Qin, Siqi Luo, Yuewen Cao, Bin Fu, Yangfan He, Hongsheng Li, Guangtao Zhai, Xiaohong Liu, et al. Resurrect mask autoregressive modeling for efficient and scalable image generation. arXiv preprint arXiv:2507.13032 , 2025.

Ziyu Yao, Jialin Li, Yifeng Zhou, Yong Liu, Xi Jiang, Chengjie Wang, Feng Zheng, Yuexian Zou, and Lei Li. Car: Controllable autoregressive modeling for visual generation. arXiv preprint arXiv:2410.04671 , 2024.

Mingyang Yi, Aoxue Li, Yi Xin, and Zhenguo Li. Towards understanding the working mechanism of text-to-image diffusion model. Advances in Neural Information Processing Systems (NeurIPS) , 2024.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. In Proceedings of the International Conference on Learning Representations (ICLR) , 2022a.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. Transactions on Machine Learning Research (TMLR) , 2022b.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) , 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) , 2023.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems (NeurIPS) , 2023.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. Proceedings of the International Conference on Learning Representations (ICLR) , 2025.

Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Lirui Zhao, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. Advances in Neural Information Processing Systems (NeurIPS) , 2024.

Le Zhuo, Liangbing Zhao, Sayak Paul, Yue Liao, Renrui Zhang, Yi Xin, Peng Gao, Mohamed Elhoseiny, and Hongsheng Li. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. Proceedings of the IEEE International Conference on Computer Vision (ICCV) , 2025.