

# JanusFlow: 融合自回归与整流流以实现统一的多模态理解与生成

Yiyang Ma<sup>1,2</sup> Xingchao Liu<sup>1,†</sup> Xiaokang Chen<sup>1,†</sup> Wen Liu<sup>1,†</sup> Chengyue Wu<sup>1,3</sup> Zhiyu Wu<sup>1,2</sup>  
Zizheng Pan<sup>1</sup> Zhenda Xie<sup>1</sup> Haowei Zhang<sup>1</sup> Xingkai Yu<sup>1</sup> Liang Zhao<sup>1</sup> Yisong Wang<sup>1,4</sup>  
Jiaying Liu<sup>2</sup> Chong Ruan<sup>1,‡</sup>

<sup>1</sup>DeepSeek-AI <sup>2</sup>Peking University <sup>3</sup>The University of Hong Kong <sup>4</sup>Tsinghua University

† 同等贡献, ‡ 通讯作者

项目主页: <https://github.com/deepseek-ai/Janus>

## Abstract

我们提出了 **JanusFlow**, 这是一个强大的框架, 能够在单一模型中统一图像理解与生成。JanusFlow 引入了一种极简架构, 将自回归语言模型与整流流 (rectified flow, 生成建模中的最先进方法) 相结合。我们的核心发现表明, 整流流可以直接在大型语言模型框架内进行训练, 无需进行复杂的架构修改。为了进一步提升我们统一模型的性能, 我们采用了两项关键策略: (i) 解耦理解与生成编码器, 以及 (ii) 在统一训练过程中对齐它们的表示。大量实验表明, JanusFlow 在各自领域内达到了与专用模型相当或更优的性能, 同时在标准基准测试中显著优于现有的统一方法。这项工作代表了迈向更高效、更通用的视觉-语言模型的重要一步。

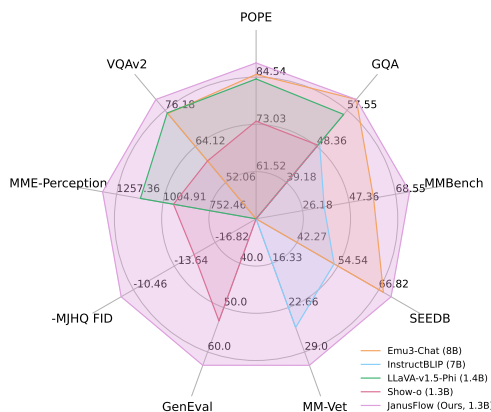
## 1. 引言

大型语言模型 (LLMs) 已展现出学习多样化知识并泛化至新场景的卓越能力 [1, 7, 8, 69, 91]。借助这些能力, 研究人员开发了专门用于图像理解 [2, 15, 47, 49, 56, 58] 和文本到图像生成 [23, 73, 76, 79] 的复杂模型。

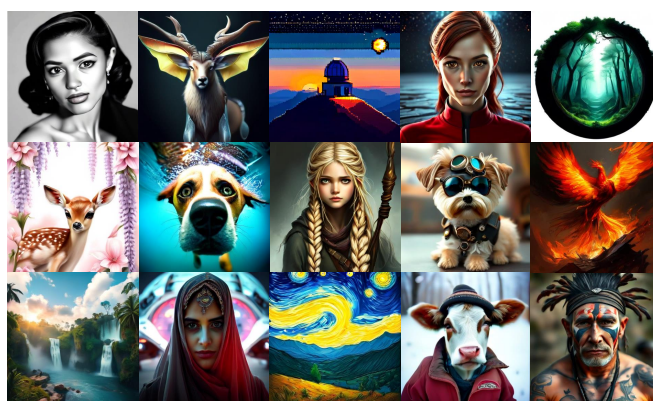
近年来, 该领域的研究重心已转向构建能够同时处理这两项任务的统一系统。一个主要方向是利用预训练的文本到图像模型进行高质量生成, 同时训练 LLM 为这些模型生成条件 [19, 25–27, 87]。然而, 这种方法引入了架构复杂性, 并且通过维护独立的 LLM 和生成组件, 可能会限制模型的能力。替代方案 [88, 97, 99, 100, 108] 则提议训练单个 LLM 来同时完成这两项任务, 通常结合扩散模型 [32, 83] 或向量量化自回归模型 [22, 86]。

我们的方法建立在 rectified flow 模型的最新突破之上 [3, 23, 55, 61, 62], 该模型为生成式建模提供了一个简洁的框架, 同时展现出卓越的实证性能 [23, 36, 45]。基于这些进展, 我们提出了 **JanusFlow**, 这是一个强大的统一多模态模型, 它将 rectified flow 与 LLM 架构无缝集成。

---



(a) 基准测试性能.



(b) 视觉生成结果.

图 1 | 基于 JanusFlow 的多模态理解与图像生成。JanusFlow 在视觉理解基准测试上超越了最先进的统一多模态模型以及多种特定任务的理解模型。它同时也具备生成高质量图像的能力。图像分辨率为  $384 \times 384$ 。

遵循极简设计原则，我们的架构仅需一个轻量级的编码器和解码器，即可使 LLM 适配 rectified flow 操作。为了优化 JanusFlow 的性能，我们实施了两项关键策略：首先，我们为理解和生成任务分别保留独立的视觉编码器，以避免任务干扰，从而提升理解能力。其次，我们在训练过程中对齐生成模块与理解模块之间的中间表示，以增强生成过程中的语义连贯性。

JanusFlow 与现有的统一方法相比，在多模态理解和文本到图像生成方面均展现出最先进的性能，甚至超越了多种专用方法。具体而言，在文本到图像生成基准测试 MJHQ FID-30k [48]、GenEval [28] 和 DPG-Bench [34] 上，JanusFlow 分别取得了 9.51、0.63 和 80.09% 的分数，超越了包括 SDv1.5 [77] 和 SDXL [73] 在内的成熟文本到图像模型。在多模态理解基准测试中，JanusFlow 在 MMBench [63]、SeedBench [46] 和 GQA [35] 上分别取得了 74.9、70.5 和 60.3 的分数，超越了 LLaVA-v1.5 [56] 和 Qwen-VL-Chat [4] 等专用模型。值得注意的是，这些结果是在仅包含 1.3B 参数的紧凑 LLM 架构下取得的。

## 2. 相关工作

**基于流生成模型的视觉生成。** 近年来，扩散模型 [32, 83] 在视觉生成领域取得了显著进展，催生了一系列令人瞩目的模型，如 [67, 73, 76–79]。在此基础上，基于流的生成模型 [3, 55, 61] 作为一种简化的替代框架应运而生。这些方法近期推动了先进视觉生成模型 [23, 36] 的发展，这些模型在实现更快采样速度的同时取得了更优的实证性能。我们的工作表明，整流流（rectified flow） [60–62] 可以有效地集成到大语言模型（LLMs）中，从而构建出在理解和生成任务上均表现出色的统一模型。

**面向理解与生成的统一模型。** 多模态大语言模型（MLLMs）的发展实现了文本与视觉信息的有效融合。基于强大的大语言模型 [7, 91, 92]，近期的多模态大语言模型 [2, 15, 49, 56, 58, 64] 已展现出卓越的多模态理解能力。当前研究日益关注能够同时处理视觉理解与生成任务的架构。一种方法是通过预训练的扩散模型来扩展多模态大语言模型 [19, 25–27, 87, 101]。然而，这些系统本质上是将扩散模型作为外部工具使用，多模态大语言模型仅生成图像生成的条件，自身并不具备直接的生成能力。这种分离架构往往导致其性能不如独立的扩散模型 [25, 87]。另一类工作 [88, 97, 99, 100, 108] 旨在训练单一的大语言模型以同时完成这两项任务。其中许多方法采用向量量化 [22, 86] 将图像转换为离散标记（tokens），从而实现统一的自回归处理 [88, 97]。尽管这些方法易于实现，但其性能本质上受限于图像标记化的质量。

我们的工作专注于开发结合自回归能力与流/扩散模型的统一模型，充分利用它们在视觉生成中已验证的有效性。与类似方法 [100, 107, 108] 相比，JanusFlow 具有三大核心优势：(i) 采用整流流的简单而高效的生成过程；(ii) 通过解耦的视觉编码器解决任务间冲突，从而提升性能；(iii) 借助解耦编码器设计实现的表示对齐正则化，进一步改善了生成质量。

## 3. JanusFlow

在本节中，我们介绍 JanusFlow 的架构以及我们的训练策略。

### 3.1. 背景

**多模态大语言模型。** 给定包含离散 token 序列的数据集  $\mathcal{D}$ ，其中每个序列可表示为  $x = (x_1, \dots, x_\ell)$ ，大语言模型（LLMs）通过自回归方式对序列分布进行建模，

$$\log P_{\theta_{LLM}}(x) = \sum_{i=0}^{\ell-1} \log P_{\theta_{LLM}}(x_{i+1}|x_1, \dots, x_i), \quad (1)$$

其中  $\theta_{LLM}$  表示 LLM 的参数， $\ell$  为序列长度。在大规模数据集上训练后，LLMs 展现出跨多种任务的泛化能力以及遵循多样化指令的能力 [1, 8, 69]。为使这些模型能够处理视觉输入，LLMs 被引入了视觉编码器 [2, 56, 58]。例如，LLaVA [58] 通过投影层将 LLM 与预训练的 CLIP [75] 图像编码器相结合，将提取的图像特征转换到联合嵌入空间中，以便 LLM 能够将其作为词嵌入进

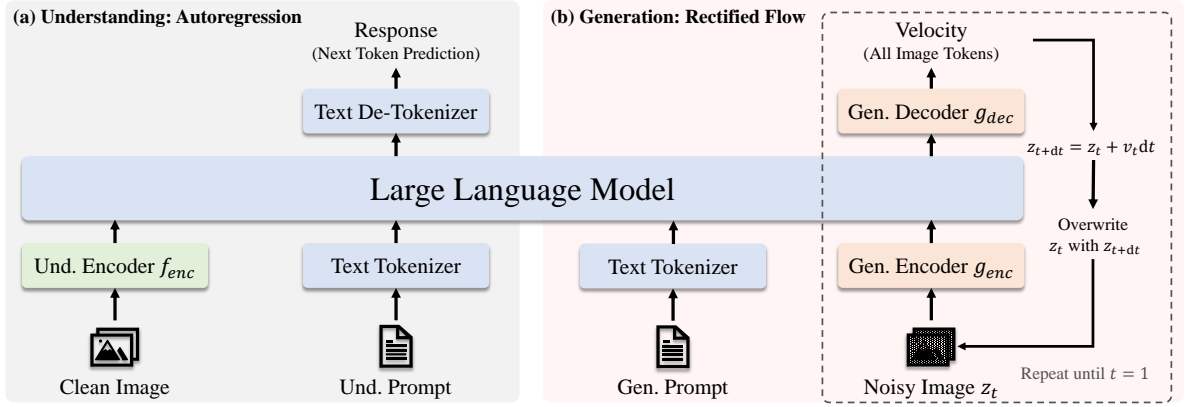


图 2 | 本文提出的 JanusFlow 架构。在视觉理解任务中，LLM 通过自回归下一个 token 预测来生成回复。在图像生成任务中，LLM 结合整流流生成图像。从  $t = 0$  的高斯噪声开始，LLM 通过预测速度向量迭代更新  $z_t$ ，直至达到  $t = 1$ 。为简化起见，我们省略了 VAE 编码器、生成过程中使用的跳跃连接以及  $f_{enc}$  之后的线性层。

行处理。借助大规模多模态数据集和日益强大的 LLMs，该架构推动了先进多模态模型的发展，使其能够应对广泛的视觉-语言任务 [4, 47, 56, 64]。

**整流流。** 对于由连续  $d$  维数据点  $x = (x_1, \dots, x_d)$  组成的数据集  $\mathcal{D}$ ，这些数据点来自未知的数据分布  $\pi_1$ ，整流流 [55, 61] 通过学习定义在时间  $t \in [0, 1]$  上的常微分方程 (ODE) 来对数据分布进行建模：

$$\frac{dz_t}{dt} = v_{\theta_{NN}}(z_t, t), \quad z_0 \sim \pi_0, \quad (2)$$

其中  $\theta_{NN}$  表示速度神经网络的参数， $\pi_0$  是一个简单的分布，通常为标准高斯噪声  $\mathcal{N}(0, I)$ 。该网络通过最小化神经速度与连接  $\pi_0$  和  $\pi_1$  中随机点的直线路径方向之间的欧氏距离来进行训练，

$$\min_{\theta} \mathbb{E}_{t \sim P(t), z_0 \sim \pi_0, x \sim \pi_1} \left[ \left\| v_{\theta_{NN}}(z_t, t) - (x - z_0) \right\|^2 \right], \quad \text{where } z_t = tx + (1-t)z_0. \quad (3)$$

此处， $P(t)$  是时间  $t \in [0, 1]$  上的分布。当网络具有足够的容量且目标函数被完美最小化时，最优速度场  $v_{\theta_{NN}^*}$  将基础分布  $\pi_0$  映射到真实数据分布  $\pi_1$ 。更准确地说，给定  $z_0 \sim \pi_0$ ， $z_1 = \int_0^1 v_{\theta_{NN}^*}(z_t, t) dt$  的分布服从  $\pi_1$ 。尽管其概念简单，但整流流在多种生成建模任务中均展现出优越的性能，包括文本到图像生成 [23]、音频生成 [40] 以及生物结构生成 [38]。

### 3.2. 多模态理解与生成的统一框架

JanusFlow 提出了一个统一的框架，旨在同时解决视觉理解和图像生成任务。接下来，我们将概述 JanusFlow 如何在单一的 LLM 架构中处理这两项任务。

**多模态理解。** 在多模态理解任务中，LLM 处理由交替排列的文本和图像数据组成的输入序列。文本被分词为离散的 token，每个 token 被转换为维度为  $D_{emb}$  的嵌入表示。对于图像，图像编码

器  $f_{enc}$  将每张图像  $x_{im}$  编码为形状为  $H_{im} \times W_{im} \times D_{enc}$  的特征图。该特征图被展平，并通过线性变换层投影为形状为  $H_{im}W_{im} \times D_{emb}$  的嵌入序列。 $H_{im}$  和  $W_{im}$  由图像编码器决定。文本和图像嵌入被拼接起来，形成 LLM 的输入序列，随后 LLM 基于该嵌入序列自回归地预测下一个 token。遵循常见做法 [88, 97, 100]，我们在图像前添加特殊 token  $|B0I|$ ，在图像后添加  $|E0I|$ ，以帮助模型在序列中定位图像嵌入。

**图像生成。** 对于图像生成任务，我们的 LLM 以文本序列  $x^{con}$  为条件，并使用整流流 (rectified flow) 生成相应的图像。为提高计算效率，生成过程在潜在空间中进行，并使用预训练的 SDXL-VAE [73]。

生成过程首先从潜在空间中采样形状为  $H_{latent} \times W_{latent} \times D_{latent}$  的高斯噪声  $z_0$ ，随后通过生成编码器  $g_{enc}$  将其处理为形状为  $H_{gen}W_{gen} \times D_{emb}$  的嵌入序列。该序列与表示当前时间步  $t$  (初始时  $t = 0$ ) 的时间嵌入相拼接，得到长度为  $H_{gen}W_{gen} + 1$  的序列。与采用各种注意力掩码策略的先前方法 [100, 108] 不同，我们发现因果注意力 (causal attention) 已足够，因为我们的初步实验表明，其他掩码方案并未带来性能提升。LLM 对应于  $z_0$  的输出通过生成解码器  $g_{dec}$  转换回潜在空间，生成形状为  $H_{latent} \times W_{latent} \times D_{latent}$  的速度向量。状态通过标准的欧拉求解器进行更新，

$$z_{t+dt} = z_t + v(z_t, t)dt, \quad (4)$$

其中  $dt$  为用户定义的步长。我们将输入中的  $z_0$  替换为  $z_{dt}$  并迭代该过程，直到得到  $z_1$ ，随后由 VAE 解码器将其解码为最终图像。为提升生成质量，我们在计算速度时采用无分类器引导 (classifier-free guidance, CFG):

$$v(z_t, t) = wv(z_t, t | x^{con}) + (1 - w)v(z_t, t | \emptyset), \quad (5)$$

其中  $v(z_t, t | \emptyset)$  表示无文本条件推断的速度， $w \geq 1$  控制 CFG 的强度。经验表明，增大  $w$  可获得更高的语义对齐度 [23, 62, 73, 77]。与多模态理解类似，我们在序列开头添加特殊 token  $|B0I|$  以指示图像生成的开始。

**两项任务的编码器解耦。** 先前在联合 LLM 训练框架内统一自回归生成与扩散模型的方法 [100, 108]，对理解和生成任务均采用相同的编码器 ( $f_{enc}$  和  $g_{enc}$ )。例如，Zhou et al. [108] 使用共享的 U-Net 或线性编码器在相同的 VAE 潜在空间中执行这两项任务，而 Xie et al. [100] 则利用 MAGVIT-v2 [102] 将图像块编码为离散 token 以服务于两项任务。

然而，近期关于统一自回归模型的研究表明，这种共享编码器设计并非最优 [97]，尤其是在通过向量量化 token 的自回归来生成图像的模型中。借鉴这些见解，JanusFlow 采用了解耦的编码器设计。具体而言，我们使用预训练的 SigLIP-Large-Patch/16 [106] 模型作为  $f_{enc}$ ，以提取用于多模态理解的语义连续特征；同时，出于有效性考虑，我们使用从零初始化的独立 ConvNeXt 模块 [96] 作为生成任务的  $g_{enc}$  和  $g_{dec}$ 。遵循现有实践 [5, 14, 93]，我们在  $g_{enc}$  和  $g_{dec}$  之间引入了长跳跃连接 (long skip connection)。我们在第 4.5 节中的对照实验表明，这种解耦编码器设计

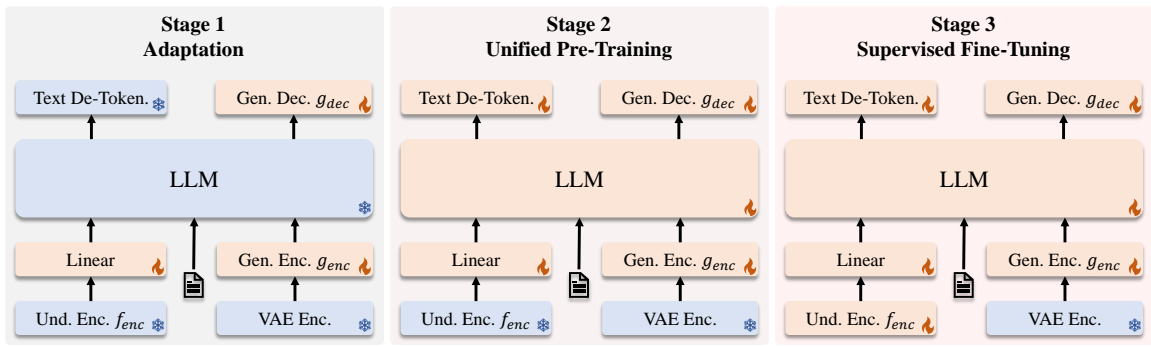


图 3 | JanusFlow 的三个训练阶段。可训练模块以火焰标记，冻结模块以雪花标记。

显著提升了我们统一模型的性能。JanusFlow 的完整架构如图 2所示。

### 3.3. 训练方案

如图 3 所示，我们在三个连续的阶段中训练模型，具体细节如下。

**阶段 1：随机初始化组件的适配。** 在第一阶段，我们仅专注于训练随机初始化的组件：线性层、生成编码器和生成解码器。该阶段旨在使这些新模块能够与预训练的 LLM 和 SigLIP 编码器有效协同工作，实质上充当了新引入组件的初始化阶段。

**阶段 2：统一预训练。** 在适配阶段之后，我们训练除视觉编码器之外的整个模型，这与先前的方法保持一致 [58, 64]。训练过程包含三种数据类型：多模态理解、图像生成和纯文本数据。我们最初分配较高比例的多模态理解数据以建立模型的理解能力。随后，我们增加图像生成数据的比例，以满足基于扩散的模型的收敛需求 [18, 72]。

**阶段 3：监督微调 (SFT)。** 在最后阶段，我们使用指令微调数据对预训练模型进行微调，该数据包含对话、特定任务对话以及高质量的文本条件图像生成示例。在此阶段，我们还解冻了 SigLIP 编码器的参数 [64, 90, 97]。该微调过程使模型能够有效响应用户在多模态理解和图像生成任务上的指令。

### 3.4. 训练目标

训练 JanusFlow 涉及两种类型的数据：多模态理解数据和图像生成数据。这两种数据均包含两个部分：“条件 (condition)”和“响应 (response)”。“条件”指任务的提示 (*e.g.*, 生成任务中的文本提示和理解任务中的图像)，而“响应”指这两个任务对应的输出。数据可格式化为  $x = (x^{con}, x^{res})$ ，其中上标 *con* 表示“条件”，*res* 表示“响应”。我们将整个序列  $x$  的长度记为  $\ell$ ， $x^{con}$  的长度记为  $\ell_{con}$ ， $x^{res}$  的长度记为  $\ell_{res}$ 。我们使用  $\theta$  表示 JanusFlow 中所有可训练参数的集合，包括 LLM、 $f_{enc}$ 、 $g_{enc}$ 、 $g_{dec}$  以及线性变换层。

**自回归目标。** 对于多模态理解任务， $x^{res}$  仅包含文本 token。JanusFlow 采用最大似然原理进行训练，

$$\mathcal{L}_{AR}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}_{und}} \left[ \sum_{i=\ell_{con}}^{\ell-1} \log P_{\theta}(x_{i+1} | x_1, \dots, x_i) \right], \quad (6)$$

其中期望是在多模态理解数据集  $\mathcal{D}_{und}$  中的所有  $(x^{con}, x^{res})$  对上计算的，且仅对  $x^{res}$  中的 token 计算损失。

**整流流目标。** 对于图像生成任务， $x^{con}$  由文本 token 组成， $x^{res}$  为对应的图像。JanusFlow 采用整流流目标进行训练，

$$\mathcal{L}_{RF}(\theta) = \mathbb{E}_{x \sim \mathcal{D}_{gen}, t \sim P(t), z_0 \sim \mathcal{N}(0, I)} \left[ \|v_{\theta}(z_t, t | x^{con}) - (x^{res} - z_0)\|^2 \right], \quad (7)$$

其中  $z_t = tx^{res} + (1-t)z_0$ 。遵循 Stable Diffusion 3 [23]，我们将时间分布  $P(t)$  设置为 logit-正态分布。为了支持 CFG 推理，我们在训练过程中随机丢弃 10% 的文本提示。

**表示对齐正则化。** 近期工作 [103] 表明，对齐扩散 Transformer 与语义视觉编码器之间的中间表示能够提升扩散模型的泛化能力。我们解耦的视觉编码器设计使得能够高效地将此对齐实现为正则化项。具体而言，对于生成任务，我们将理解编码器  $f_{enc}$  的特征与 LLM 的中间特征进行对齐，

$$\mathcal{L}_{REPA}(\theta, \varphi) = -\mathbb{E}_{x \sim \mathcal{D}_{gen}} \left[ \text{sim}(\text{stop\_grad}(f_{enc}(x^{res})), h_{\varphi}(q_{\theta}(z_t))) \right], \quad (8)$$

其中  $q_{\theta}(z_t)$  表示给定输入  $z_t$  时的 LLM 中间表示， $h_{\varphi}$  是一个小型可训练 MLP，用于将  $q_{\theta}(z_t)$  投影到维度  $D_{enc}$ 。函数  $\text{sim}(\cdot, \cdot)$  计算嵌入向量之间逐元素余弦相似度的均值。在计算损失之前，我们将  $h_{\varphi}(q_{\theta}(z_t))$  重塑为  $H_{gen} \times W_{gen} \times D_{enc}$ 。为简化实现，我们有意调整了  $g_{enc}$  和  $g_{dec}$  的配置，以确保  $H_{gen} = H_{im}$  且  $W_{gen} = W_{im}$ 。 $\mathcal{L}_{REPA}$  的梯度不会通过理解编码器进行反向传播。该对齐损失有助于 LLM 的内部特征空间（在给定噪声输入  $z_t$  时）与理解编码器的语义特征空间对齐，从而在推理阶段从新的随机噪声和文本条件生成图像时提升生成质量。

**总结。** 所有三个目标均应用于各个训练阶段。多模态理解任务使用  $\mathcal{L}_{AR}$ ，而图像生成任务采用组合损失  $\mathcal{L}_{RF} + \mathcal{L}_{REPA}$ 。详细的实验设置见第 4.1 节。

## 4. 实验

我们进行了广泛的实验，以评估 JanusFlow 在多模态理解和生成任务中的能力。首先，我们介绍实验设置与实现细节。随后，我们展示了在多模态理解和图像生成标准基准上的结果。最后，我们进行了消融实验以验证我们的关键设计选择。

表 1 | 所提 JanusFlow 的超参数。数据比例表示多模态理解数据、图像生成数据和纯文本数据的占比。在阶段 2 的初始 10,000 步中，我们采用 30 : 50 : 20 的数据比例以提升理解能力。

	阶段 1	阶段 2	阶段 3
学习率	$1.0 \times 10^{-4}$	$1 \times 10^{-4}$	$2.0 \times 10^{-5}$
学习率调度器	固定	固定	固定
权重衰减	0.0	0.0	0.0
梯度裁剪	1.0	1.0	1.0
优化器	AdamW ( $\beta_1 = 0.9, \beta_2 = 0.95$ )		
预热步数	2,000	2,000	1,000
训练步数	10,000	390,000	26,000
批次大小	512	512	256
数据比例	50 : 50 : 0	14 : 80 : 6	21 : 70 : 9

#### 4.1. 实验设置与实现细节

我们的框架基于 DeepSeek-LLM (1.3B) [7, 64] 的一个增强版本<sup>1</sup>。该大语言模型 (LLM) 包含 24 个 Transformer 块，支持 4,096 的序列长度。在我们的模型中，理解和生成任务均使用分辨率为 384 的图像。

对于多模态理解，我们采用 SigLIP-Large-Patch/16 [106] 作为  $f_{enc}$ 。对于图像生成，我们利用预训练的 SDXL-VAE [73] 作为其潜在空间。生成编码器  $g_{enc}$  包含一个  $2 \times 2$  的图像分块层，后接两个 ConvNeXt [96] 块和一个线性层。生成解码器  $g_{dec}$  由两个 ConvNeXt 块、一个用于上采样特征图的像素重排层 (pixel-shuffle layer) 以及一个线性层组成。我们的 SigLIP 编码器包含  $\sim 300M$  参数。 $g_{enc}$  和  $g_{dec}$  是轻量级模块，总共包含  $\sim 70M$  参数。表 1 详细列出了每个训练阶段的超参数。在对齐正则化中，我们使用第 6 个块之后的 LLM 特征作为  $q_{\theta}(z_t)$ ，并使用一个三层 MLP 作为  $h_{\varphi}$ 。我们采用衰减率为 0.99 的指数移动平均 (EMA) 以确保训练稳定性。

在数据预处理方面，我们对理解和生成数据采用不同的处理方式。对于理解任务，我们通过将长边调整至目标尺寸并将图像填充为正方形来保留所有图像信息。对于生成任务，我们将短边调整至目标尺寸，并应用随机正方形裁剪以避免填充伪影。在训练期间，为了提高训练效率，我们将多个序列打包成一个长度为 4,096 的单一序列。我们的实现基于使用 PyTorch [74] 的 HAI-LLM 平台 [31]。训练在 NVIDIA A100 GPU 上进行，每个模型需要约  $\sim 1,600$  个 A100 GPU 天。

#### 4.2. 训练数据设置

我们遵循 Janus [97] 构建训练数据。各训练阶段的数据配置如下。

**第一阶段和第二阶段的数据。** 我们框架的前两个阶段使用三类数据：多模态理解数据、图像生成数据和纯文本数据。

<sup>1</sup>与 Janus [97] 中使用的版本相比，该版本在扩展的文本语料库上进行了训练，已被证明在多项选择题基准（如 MMBench [63] 和 SEED Bench [46]）上具有更好的性能。我们的初步实验表明，它对视觉生成质量的影响微乎其微。

表 2 | GenEval 基准上的性能。“Gen.”表示“生成”，“Unified”表示统一理解与生成模型。使用外部预训练生成模型的模型标有<sup>†</sup>。

类型	方法	参数量	单对象	双对象	计数	颜色	位置	颜色属性	总体 <sup>↑</sup>
仅生成	LlamaGen [86]	0.8B	0.71	0.34	0.21	0.58	0.07	0.04	0.32
	LDM [77]	1.4B	0.92	0.29	0.23	0.70	0.02	0.05	0.37
	SDv1.5 [77]	0.9B	0.97	0.38	0.35	0.76	0.04	0.06	0.43
	PixArt- $\alpha$ [9]	0.6B	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1 [77]	0.9B	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2 [76]	6.5B	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	Emu3-Gen [95]	8B	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	SDXL [73]	2.6B	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	IF-XL [17]	4.3B	0.97	0.74	0.66	0.81	0.13	0.35	0.61
DALL-E 3 [6]	-	0.96	0.87	0.47	0.83	0.43	0.45	0.67	
统一模型	Chameleon [88]	34B	-	-	-	-	-	-	0.39
	LWM [59]	7B	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	SEED-X <sup>†</sup> [27]	17B	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	Show-o [100]	1.3B	0.95	0.52	0.49	0.82	0.11	0.28	0.53
	Janus [97]	1.3B	0.97	0.68	0.30	0.84	0.46	0.42	0.61
	Transfusion [108]	7.3B	-	-	-	-	-	-	0.63
	<b>JanusFlow (Ours)</b>	1.3B	0.97	0.59	0.45	0.83	0.53	0.42	0.63

- 多模态理解数据。**该类型数据包含几个子类别：(a) 图像描述数据。我们整合了来自 [20, 41, 50, 51, 53, 82] 的描述数据集，并使用开源多模态理解模型为来自 [16, 43] 的图像生成额外描述。数据集名称见补充材料。数据遵循模板格式，*e.g.*, “<image> 生成这张图片的描述。<caption>'”。(b) 图表数据。我们直接采用 DeepSeek-VL [64] 训练数据中的图表数据。(c) 任务数据。我们利用 ShareGPT4V [11] 数据以促进预训练期间的基本问答能力，其结构为“<image>< 问题 >< 答案 >'”。(d) 图文交错数据。该子类别的数据来源于 [42, 84]。
- 图像生成数据。**我们的图像生成数据集结合了来自 [16, 21, 41, 43, 68, 71, 82, 85] 的高质量图像和 200 万条内部数据。我们使用多模态理解模型为其添加机器生成的描述以进行增强。我们根据宽高比和美学分数对 [16, 82] 中的图像进行过滤，保留了原始数据集的约 20%。25% 的数据包含单句描述。这类数据有助于模型处理短提示词。所有数据点均格式化为“< 提示词 >< 图像 >'”。
- 纯文本数据。**我们直接使用 DeepSeek-LLM [7] 的文本语料库。

**第三阶段的数据。**监督微调 (SFT) 阶段同样使用三类数据：

- 多模态指令数据。**我们利用来自 [29, 33, 35, 47, 65, 80] 的指令微调数据集。
- 图像生成数据。**我们将来自 [16, 82, 85] 的高质量图文对重新格式化为指令格式：“用户:< 用户提示 >\n\n助手:< 图像 >'”。
- 纯文本数据。**我们直接整合了来自 [47] 的纯文本数据。

表 3 | DPG-Bench 上的性能。除我们的方法外，本表中的方法均为专用生成模型。

方法	全局	实体	属性	关系	其他	总体 ↑
SDv1.5 [77]	74.63	74.23	75.39	73.49	67.81	63.18
PixArt- $\alpha$ [9]	74.97	79.32	78.60	82.57	76.96	71.11
Lumina-Next [110]	82.82	88.65	86.44	80.53	81.82	74.63
SDXL [73]	83.27	82.43	80.91	86.76	80.41	74.65
Playground v2.5 [48]	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-DiT [54]	84.59	80.59	88.01	74.36	86.41	78.87
PixArt- $\Sigma$ [10]	86.89	82.89	88.94	86.59	87.68	80.54
Emu3-Gen [95]	85.21	86.68	86.84	90.22	83.15	80.60
<b>JanusFlow (Ours)</b>	87.03	87.31	87.39	89.79	88.10	80.09

### 4.3. 评估设置

**图像生成。** 我们使用视觉质量和语义准确性指标来评估生成的图像。在视觉质量评估方面，我们采用 Fréchet Inception Distance [30] (FID) 指标，并计算 30,000 张生成图像与 MJHQ 数据集 [48] 中对应参考图像之间的 FID。FID 的计算遵循 GigaGAN [39] 的实现。为了评估语义准确性，我们使用了两个专用框架：GenEval [28] 和 DPG-Bench [34]。这些框架旨在评估生成图像是否准确包含了输入提示中指定的对象和关系，从而对生成能力进行全面评估。

**多模态理解。** 我们在一系列多样化的视觉语言基准测试上评估 JanusFlow 的多模态理解能力，以检验其通用理解性能，包括 POPE [52]、MME [24]、MMBench [63]、SEEDBench [46]、VQA<sub>v2</sub> [29]、GQA [35]、MM-Vet [104]、MMMUS [105]、ChartQA [70] 和 TextVQA [81]

### 4.4. 定量结果

**图像生成性能。** 我们报告了在 GenEval、DPG-Bench 和 MJHQ FID-30k 上的性能。在表 2 中，我们给出了 GenEval 上的对比结果，包括所有子任务的得分和总体得分。JanusFlow 取得了 0.63 的总体得分，超越了之前的统一框架以及包括 SDXL [73] 和 DALL-E 2 [76] 在内的多个专用生成模型。在表 3 中，我们展示了 DPG-Bench 上的结果及相应对比。需要注意的是，表 3 中的所有方法除我们的模型外均为专用生成模型。GenEval 和 DPG-Bench 上的结果证明了我们模型的指令遵循能力。我们在表 4 中给出了 MJHQ FID-30k 的对比结果。用于计算 FID 的采样图像是在 CFG 系数  $w = 2$  和采样步数 30 的条件下生成的。我们对 CFG 系数和采样步数进行了遍历搜索，并将结果提供在附录中。我们的方法在所有 1.3B LLM 模型中取得了最佳性能。结果证明，整流流 (rectified flow) 能够提升生成图像的质量，优于 Janus [97] 等自回归模型。

**多模态理解性能。** 我们在表 5 中展示了我们的方法与其他方法 (包括专用理解模型和统一理解

表 4 | MJHQ FID-30k 结果。与我们的模型规模相近的模型以蓝色背景标出。JanusFlow 在 1.3B 模型中取得了最佳的 FID。

方法	参数量	FID↓
LWM [59]	7B	17.77
VILA-U 256 [99]	7B	12.81
VILA-U 384 [99]	7B	7.69
Show-o [100]	1.3B	15.18
Janus [97]	1.3B	10.10
<b>JanusFlow (Ours)</b>	1.3B	9.51

表 5 | 多模态理解基准测试上的方法对比。“Und.” 表示“理解”，“Unified” 表示统一理解与生成模型。采用外部预训练生成模型的模型以 <sup>†</sup> 标记。LLM 参数量与我们相近的模型在虚线下方以蓝色背景标出。

类型	模型	LLM 参数量	POPE	MME-P	MMB <sub>dev</sub>	SEED	VQAv2 <sub>test</sub>	GQA	MMMU	MM-Vet	ChartQA	TextVQA
仅理解	MobileVLM [12]	2.7B	84.9	1288.9	59.6	-	-	59.0	-	-	-	47.5
	MobileVLM-V2 [13]	2.7B	84.7	1440.5	63.2	-	-	61.1	-	-	-	57.5
	LLaVA-Phi [109]	2.7B	85.0	1335.1	59.8	-	71.4	-	-	28.9	-	48.6
	LLaVA [58]	7B	76.3	809.6	38.7	33.5	-	-	-	25.5	-	-
	LLaVA-v1.5 [56]	7B	85.9	1510.7	64.3	58.6	78.5	62.0	35.4	31.1	-	58.2
	InstructBLIP [15]	7B	-	-	36.0	53.4	-	49.2	-	26.2	-	50.1
	Qwen-VL-Chat [4]	7B	-	1487.5	60.6	58.2	78.2	57.5	-	-	66.3	61.5
	LLaVA-NeXT [57]	7B	-	1519.3	-	-	-	-	35.1	-	54.8	-
	Qwen2-VL [94]	7B	-	-	-	-	-	-	54.1	62.0	83.0	84.3
	IDEFICS-9B [44]	8B	-	-	48.2	-	50.9	38.4	-	-	-	25.9
	Emu3-Chat [95]	8B	85.2	-	58.5	68.2	75.1	60.3	31.6	-	68.6	64.7
	InstructBLIP [15]	13B	78.9	1212.8	-	-	-	49.5	-	25.6	-	50.7
	LLaVA-v1.5-Phi-1.5 [100]	1.3B	84.1	1128.0	-	-	75.3	56.5	30.7	-	-	-
	MobileVLM [12]	1.4B	84.5	1196.2	53.2	-	-	56.1	-	-	-	41.5
	MobileVLM-V2 [13]	1.4B	84.3	1302.8	57.7	-	-	59.3	-	-	-	52.1
	统一模型	Gemini-Nano-1 [89]	1.8B	-	-	-	-	62.7	-	26.3	-	53.6
LWM [59]		7B	75.2	-	-	-	55.8	44.8	-	9.6	-	-
VILA-U [99]		7B	85.8	1401.8	-	59.0	79.4	60.8	-	33.5	-	60.8
Chameleon [88]		7B	-	-	-	-	-	-	22.4	8.3	-	-
DreamLLM <sup>†</sup> [19]		7B	-	-	-	-	72.9	-	-	36.6	-	41.8
LaVIT <sup>†</sup> [37]		7B	-	-	-	-	66.0	46.8	-	-	-	-
Emu <sup>†</sup> [87]		13B	-	-	-	-	52.0	-	-	-	-	-
NEXT-GPT <sup>†</sup> [98]		13B	-	-	-	-	66.7	-	-	-	-	-
Show-o [100]		1.3B	73.8	948.4	-	-	59.3	48.7	25.1	-	-	-
Janus [97]		1.3B	87.0	1338.0	69.4	63.7	77.3	59.1	30.5	34.3	-	-
JanusFlow (Ours)		1.3B	88.0	1333.1	74.9	70.5	79.8	60.3	29.3	30.9	64.6	55.5

表 6 | 消融实验。带有 <sup>†</sup> 的模块权重在训练过程中被冻结。“Exp.” 表示“实验”。本表中的“FID”指 MJHQ FID-10k, CFG 因子  $w = 7.5$ , 步数为 30。“CLIP” 表示基于 CLIP-ViT-Large-Patch/14 主干网络的 CLIP 相似度。实验 F 是训练 JanusFlow 的最终配置。

实验 ID	REPA	模型设置			训练迭代次数	评估基准				
		理解模块	生成模块	类型		POPE <sup>†</sup>	VQAv2 <sub>val</sub> <sup>†</sup>	GQA <sup>†</sup>	FID <sup>†</sup>	CLIP <sup>†</sup>
A	×	SigLIP	VAE <sup>†</sup> +ConvNeXt	统一	50,000	82.40	69.62	54.43	19.84	24.94
B	✓	共享	VAE <sup>†</sup> +ConvNeXt	统一	50,000	78.13	53.94	44.04	18.05	26.38
C	✓	VAE+ConvNeXt	VAE <sup>†</sup> +ConvNeXt	统一	50,000	75.30	55.41	44.44	17.53	26.32
D	✓	SigLIP	-	仅理解	13,000	85.03	69.10	54.23	-	-
E	✓	-	VAE <sup>†</sup> +ConvNeXt	仅生成	37,000	-	-	-	16.69	26.89
<b>F</b>	✓	SigLIP	VAE <sup>†</sup> +ConvNeXt	统一	50,000	84.73	69.20	54.83	17.61	26.40

与生成模型) 的对比结果。我们的模型在参数量相近的所有模型中取得了最佳性能，甚至超越了多个规模更大的专用理解方法。我们的结果表明，该方法成功融合了自回归 LLM 与整流流，在理解与生成两方面均取得了令人满意的性能。

#### 4.5. 消融实验

我们进行了全面的消融实验，以验证我们关键设计选择的有效性。出于计算效率的考虑，所有消融实验均在  $256 \times 256$  分辨率的图像上进行<sup>2</sup>。除仅理解和仅生成的变体外，所有模型均在我们的统一预训练数据集上训练 50,000 次迭代；这些变体根据其在预训练阶段各自的数据比例，训练了相应减少的迭代次数。这些消融实验的定量结果如表 6 所示。

**表示对齐的影响。** 实验 A 与 F 的对比证明了在训练过程中引入表示对齐正则化 [103] 的显著优

<sup>2</sup>基于  $256 \times 256$  的消融实验中的理解编码器同样为 SigLIP-Large-Patch/16，该模型在  $256 \times 256$  图像上进行了预训练。



A corgi's head depicted as an explosion of a nebula, with vibrant cosmic colors like deep purples, blues, and pinks swirling around. The corgi's fur blends seamlessly into the nebula, with stars and galaxies forming the texture of its fur. Bright bursts of light emanate from its eyes, and faint constellations can be seen in the background, giving the image a surreal, otherworldly feel.



Beautiful surreal symbolism the mesmerizing vision of a Cleopatra Queen of Egypt, mesmerizing brown eyes, black hair and ethereal features, radiating celestial aura, super high definition, true lifelike color, perfect exposure, razor sharp focus, golden ratio, soft reflections, bokeh effect, fine art photography, cinematic compositing, authentic, professional.



A lone figure in dark robes ascends worn stone steps toward a glowing light in an ancient temple entrance. Ornate arches, lush greenery, and intricate carvings adorn the scene, evoking a mystical, high-fantasy atmosphere reminiscent of works by artists like Randy Vargas, with cinematic lighting and epic storytelling.

图 4 | JanusFlow 的图像生成结果。我们的模型能够生成与文本提示语义一致的高质量图像。

势。具体而言，采用表示对齐训练的模型在 MJHQ 数据集上表现出明显更低的 FID 分数和更高的 CLIP 分数，表明图像质量和语义对齐均得到了同步提升。重要的是，由于我们引入了大语言模型 (LLM) 以及在  $g_{enc}$  和  $g_{dec}$  之间增加了额外的跳跃连接，我们的架构与 [103] 中考察的先前研究 [66, 72] 有所不同。表示对齐在我们修改后的架构中的有效性，表明其具有广泛的适用性和跨不同网络结构的泛化能力。

**解耦视觉编码器的影响。**在多模态理解中使用强大的预训练视觉编码器的有效性。实验 B、C 和 F 的对比展示了为理解和生成任务使用独立视觉编码器的优势。在实验 B 中，我们遵循与 Transfusion [108] 类似的设计，在 SDXL-VAE 潜在空间中为理解和生成编码器实现共享的 ConvNeXt 模块。实验 C 采用架构和初始化参数相同的独立编码器，但分别进行训练。这些配置之间的性能差异验证了解耦视觉编码器对于提升我们统一模型能力的必要性。此外，实验 C 和 F 的优异结果凸显了利用预训练语义视觉编码器进行多模态理解任务的优势。

**与仅理解/仅生成模型的公平比较。**为了建立有意义的基准，我们在相同的条件下评估了任务专用模型——使用相同的预训练数据集、基础设施和超参数。实验 D 和 E 代表了这些专用模型，其训练数据量与表 6 中的统一模型相匹配。实验 F 与这些任务专用基线之间微小的性能差距表明，我们的统一框架成功地将理解和生成能力整合在一起，且未对任一任务的性能造成显著妥协。

#### 4.6. 定性结果

我们展示了本方法在图像生成与理解任务上的定性评估结果。图 1(b) 与图 4 展示了 JanusFlow 的图像生成能力。这些结果不仅体现了生成图像的高视觉质量，也验证了本框架忠实遵循多样化指令的能力。在多模态理解方面，图 5 展示了若干示例对话，体现了本模型在不同场景下的理解能力。这些交互过程充分展现了模型在自然语言对话中对视觉内容进行理解与推理的能力。

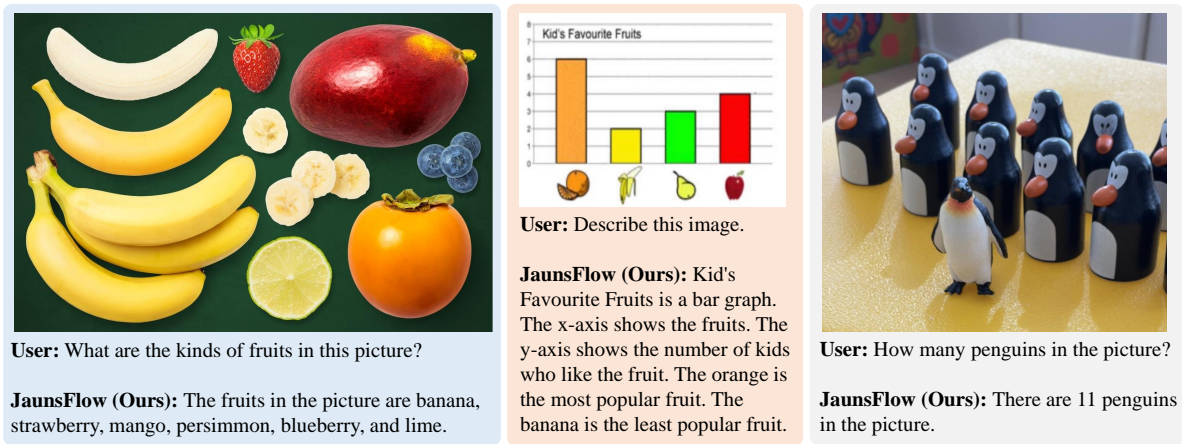


图 5 | 基于 JanusFlow 的视觉理解。我们的模型能够有效处理各种视觉理解任务，例如问答、情节解读和物体计数。

附录中提供了更多定性示例，以进一步展示 JanusFlow 的通用性与有效性。

## 5. 结论

我们提出了 JanusFlow，这是一个统一的框架，成功地将自回归模型与整流流模型相融合，用于多模态理解与生成任务。大量实验结果表明，该统一框架的性能与特定任务模型相当。成功整合这些底层架构截然不同的模型，不仅有效应对了当前多模态学习面临的挑战，也为未来训练统一模型的研究开辟了新的方向。

## 参考文献

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 technical report. [arXiv preprint arXiv:2303.08774](#), 2023.
- [2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. In [Proc. Annu. Conf. Neural Inf. Process. Systems](#), 2022.
- [3] M. Albergo and E. Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In [Proc. Int’l Conf. Learning Representations](#), 2023.
- [4] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-VL: A frontier large vision-language model with versatile abilities. [arXiv preprint arXiv:2308.12966](#), 2023.
- [5] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, and J. Zhu. All are worth words: A ViT backbone for diffusion models. In [Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition](#), 2023.
- [6] 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. [Computer Science](#), 2023.
- [7] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu, et al. DeepSeek LLM: Scaling open-source language models with longtermism. [arXiv preprint arXiv:2401.02954](#), 2024.
- [8] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. [arXiv preprint arXiv:2303.12712](#), 2023.
- [9] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, et al. PixArt-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. [arXiv preprint arXiv:2310.00426](#), 2023.
- [10] 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. [arXiv preprint arXiv:2403.04692](#), 2024.
- [11] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin. ShareGPT4V: Improving large multi-modal models with better captions. [arXiv preprint arXiv:2311.12793](#), 2023.

- [12] X. Chu, L. Qiao, X. Lin, S. Xu, Y. Yang, Y. Hu, F. Wei, X. Zhang, B. Zhang, X. Wei, et al. MobileVLM: A fast, reproducible and strong vision language assistant for mobile devices. arXiv preprint arXiv:2312.16886, 2023.
- [13] X. Chu, L. Qiao, X. Zhang, S. Xu, F. Wei, Y. Yang, X. Sun, Y. Hu, X. Lin, B. Zhang, et al. MobileVLM V2: Faster and stronger baseline for vision language model. arXiv preprint arXiv:2402.03766, 2024.
- [14] K. Crowson, S. A. Baumann, A. Birch, T. M. Abraham, D. Z. Kaplan, and E. Shippole. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In Proc. Int’l Conf. Machine Learning, 2024.
- [15] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2023.
- [16] dclure. LAION-Aesthetics-UMAP, 2022. URL <https://huggingface.co/datasets/dclure/laion-aesthetics-12m-umap>.
- [17] DeepFloyd. DeepFloyd IF, 2023. URL <https://huggingface.co/DeepFloyd/IF-I-XL-v1.0>.
- [18] P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2021.
- [19] 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 Proc. Int’l Conf. Learning Representations, 2024.
- [20] echo840. Detailed caption, 2023. URL [https://huggingface.co/datasets/echo840/Detailed\\_Caption](https://huggingface.co/datasets/echo840/Detailed_Caption).
- [21] B. Egan, A. Redden, XWAVE, and SilentAntagonist. DALLE-3 1 million+ high quality captions, 2024. URL <https://huggingface.co/datasets/ProGamerGov/synthetic-dataset-1m-dalle3-high-quality-captions>.
- [22] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2021.
- [23] 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 Proc. Int’l Conf. Machine Learning, 2024.

- [24] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, Y. Wu, and R. Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394, 2024.
- [25] Y. Ge, Y. Ge, Z. Zeng, X. Wang, and Y. Shan. Planting a SEED of vision in large language model. arXiv preprint arXiv:2307.08041, 2023.
- [26] Y. Ge, S. Zhao, Z. Zeng, Y. Ge, C. Li, X. Wang, and Y. Shan. Making LLaMA SEE and draw with SEED tokenizer. arXiv preprint arXiv:2310.01218, 2023.
- [27] 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.
- [28] D. Ghosh, H. Hajishirzi, and L. Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2024.
- [29] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in VQA matter: Elevating the role of image understanding in visual question answering. In Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2017.
- [30] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. Proc. Annu. Conf. Neural Inf. Process. Systems, 2017.
- [31] High-flyer. HAI-LLM: Efficient and lightweight training tool for large models, 2023. URL <https://www.high-flyer.cn/en/blog/hai-llm>.
- [32] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2020.
- [33] Y.-C. Hsiao, F. Zubach, G. Baechler, V. Carbune, J. Lin, M. Wang, S. Sunkara, Y. Zhu, and J. Chen. ScreenQA: Large-scale question-answer pairs over mobile app screenshots. arXiv preprint arXiv:2209.08199, 2022.
- [34] 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.
- [35] D. A. Hudson and C. D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2019.

- [36] Y. Jin, Z. Sun, N. Li, K. Xu, H. Jiang, N. Zhuang, Q. Huang, Y. Song, Y. Mu, and Z. Lin. Pyramidal flow matching for efficient video generative modeling. arXiv preprint arXiv:2410.05954, 2024.
- [37] Y. Jin, K. Xu, L. Chen, C. Liao, J. Tan, Q. Huang, C. Bin, C. Song, D. ZHANG, W. Ou, et al. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. In Proc. Int’l Conf. Learning Representations, 2024.
- [38] B. Jing, B. Berger, and T. Jaakkola. AlphaFold meets flow matching for generating protein ensembles. In Proc. Int’l Conf. Machine Learning, 2024.
- [39] M. Kang, J.-Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park. Scaling up GANs for text-to-image synthesis. In Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2023.
- [40] S. Kim, K. Shih, J. F. Santos, E. Bakhturina, M. Desta, R. Valle, S. Yoon, B. Catanzaro, et al. P-Flow: a fast and data-efficient zero-shot tts through speech prompting. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2024.
- [41] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In Proc. IEEE Int. Conf. Comput. Vision, 2023.
- [42] M. Koupaee and W. Y. Wang. WikiHow: A large scale text summarization dataset. arXiv preprint arXiv:1810.09305, 2018.
- [43] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallocci, A. Kolesnikov, et al. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. Int’l Journal of Computer Vision, 2020.
- [44] H. Laurençon, D. van Strien, S. Bekman, L. Tronchon, L. Saulnier, T. Wang, S. Karamcheti, A. Singh, G. Pistilli, Y. Jernite, et al. Introducing IDEFICS: An open reproduction of state-of-the-art visual language model, 2023, 2023. URL <https://huggingface.co/blog/idefics>.
- [45] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, et al. VoiceBox: Text-guided multilingual universal speech generation at scale. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2024.
- [46] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan. SEED-Bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125, 2023.

- [47] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, Y. Li, Z. Liu, and C. Li. LLaVA-OneVision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024.
- [48] 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, 2024.
- [49] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proc. Int’l Conf. Machine Learning, 2023.
- [50] L. Li, Y. Wang, R. Xu, P. Wang, X. Feng, L. Kong, and Q. Liu. Multimodal arXiv: A dataset for improving scientific comprehension of large vision-language models. In Annual Meeting of the Association for Computational Linguistics, 2024.
- [51] X. Li, F. Zhang, H. Diao, Y. Wang, X. Wang, and L.-Y. Duan. DenseFusion-1M: Merging vision experts for comprehensive multimodal perception. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2024.
- [52] Y. Li, Y. Du, K. Zhou, J. Wang, X. Zhao, and J.-R. Wen. Evaluating object hallucination in large vision-language models. In Proc. Conf. on Empirical Methods in Natural Language Process., 2023.
- [53] Z. Li, X. Yang, K. Choi, W. Zhu, R. Hsieh, H. Kim, J. H. Lim, S. Ji, B. Lee, X. Yan, et al. MMSci: A multimodal multi-discipline dataset for phd-level scientific comprehension. In AI for Accelerated Materials Design, 2024.
- [54] Z. Li, J. Zhang, Q. Lin, J. Xiong, Y. Long, X. Deng, Y. Zhang, X. Liu, M. Huang, Z. Xiao, et al. Hunyuan-DiT: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748, 2024.
- [55] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In Proc. Int’l Conf. Learning Representations, 2023.
- [56] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2024.
- [57] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- [58] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2024.

- [59] H. Liu, W. Yan, M. Zaharia, and P. Abbeel. World model on million-length video and language with ringattention. [arXiv preprint arXiv:2402.08268](#), 2024.
- [60] Q. Liu. Rectified flow: A marginal preserving approach to optimal transport. [arXiv preprint arXiv:2209.14577](#), 2022.
- [61] X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In [Proc. Int’l Conf. Learning Representations](#), 2023.
- [62] X. Liu, X. Zhang, J. Ma, J. Peng, et al. InstaFlow: One step is enough for high-quality diffusion-based text-to-image generation. In [Proc. Int’l Conf. Learning Representations](#), 2024.
- [63] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. MMBench: Is your multi-modal model an all-around player? In [Proc. European Conf. Computer Vision](#), 2024.
- [64] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, et al. DeepSeek-VL: towards real-world vision-language understanding. [arXiv preprint arXiv:2403.05525](#), 2024.
- [65] P. Lu, L. Qiu, J. Chen, T. Xia, Y. Zhao, W. Zhang, Z. Yu, X. Liang, and S.-C. Zhu. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. In [Proc. Annu. Conf. Neural Inf. Process. Systems](#), 2021.
- [66] N. Ma, M. Goldstein, M. S. Albergo, N. M. Boffi, E. Vanden-Eijnden, and S. Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. [arXiv preprint arXiv:2401.08740](#), 2024.
- [67] Y. Ma, H. Yang, W. Wang, J. Fu, and J. Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. [arXiv preprint arXiv:2303.09319](#), 2023.
- [68] madebyollin. Megalith-10M, 2024. URL <https://huggingface.co/datasets/madebyollin/megalith-10m>.
- [69] B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, et al. Language models are few-shot learners. [arXiv preprint arXiv:2005.14165](#), 2020.
- [70] A. Masry, X. L. Do, J. Q. Tan, S. Joty, and E. Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In [Annual Meeting of the Association for Computational Linguistics](#), 2022.
- [71] mehdidc. YFCC-15M, 2024. URL <https://huggingface.co/datasets/mehdidc/yfcc15m>.

- [72] W. Peebles and S. Xie. Scalable diffusion models with transformers. In Proc. IEEE Int. Conf. Comput. Vision, 2023.
- [73] 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 Proc. Int’l Conf. Learning Representations, 2024.
- [74] PyTorch-Contributors. PyTorch, 2024. URL <https://pytorch.org>.
- [75] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In Proc. Int’l Conf. Machine Learning, 2021.
- [76] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125, 2022.
- [77] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2022.
- [78] L. Ruan, Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, and B. Guo. MM-Diffusion: Learning multi-modal diffusion models for joint audio and video generation. In Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2022.
- [79] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2022.
- [80] S. Shah, A. Mishra, N. Yadati, and P. P. Talukdar. KVQA: Knowledge-aware visual question answering. In Proc. AAAI Conf. on Artificial Intelligence, 2019.
- [81] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards VQA models that can read. In Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2019.
- [82] V. Singla, K. Yue, S. Paul, R. Shirkavand, M. Jayawardhana, A. Ganjdanesh, H. Huang, A. Bhatele, G. Somepalli, and T. Goldstein. From pixels to prose: A large dataset of dense image captions. arXiv preprint arXiv:2406.10328, 2024.
- [83] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In Proc. Int’l Conf. Learning Representations, 2021.

- [84] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In Proc. ACM SIGIR Conf. Research and Develop. in Info. Retrieval, 2021.
- [85] K. Sun, J. Pan, Y. Ge, H. Li, H. Duan, X. Wu, R. Zhang, A. Zhou, Z. Qin, Y. Wang, et al. JourneyDB: A benchmark for generative image understanding. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2024.
- [86] 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, 2024.
- [87] Q. Sun, Q. Yu, Y. Cui, F. Zhang, X. Zhang, Y. Wang, H. Gao, J. Liu, T. Huang, and X. Wang. Generative pretraining in multimodality. In Proc. Int’l Conf. Learning Representations, 2024.
- [88] C. Team. Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818, 2024.
- [89] G. Team. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.
- [90] S. Tong, E. Brown, P. Wu, S. Woo, M. Middepogu, S. C. Akula, J. Yang, S. Yang, A. Iyer, X. Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. arXiv preprint arXiv:2406.16860, 2024.
- [91] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.
- [92] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. LLaMA 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.
- [93] C. N. Vasconcelos, A. Rashwan, A. Waters, T. Walker, K. Xu, J. Yan, R. Qian, Y. Li, S. LUO, Y. Onoe, et al. Greedy growing enables high-resolution pixel-based diffusion models. Transactions on Machine Learning Research, 2024.
- [94] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024.
- [95] 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, 2024.

- [96] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie. ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders. In Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2023.
- [97] C. Wu, X. Chen, Z. Wu, Y. Ma, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, C. Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. arXiv preprint arXiv:2410.13848, 2024.
- [98] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua. NExT-GPT: Any-to-any multimodal LLM. In Proc. Int’l Conf. Machine Learning, 2024.
- [99] 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.
- [100] 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.
- [101] H. Ye, D.-A. Huang, Y. Lu, Z. Yu, W. Ping, A. Tao, J. Kautz, S. Han, D. Xu, P. Molchanov, et al. X-VILA: Cross-modality alignment for large language model. arXiv preprint arXiv:2405.19335, 2024.
- [102] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, A. Gupta, X. Gu, A. G. Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. In Proc. Int’l Conf. Learning Representations, 2024.
- [103] S. Yu, S. Kwak, H. Jang, J. Jeong, J. Huang, J. Shin, and S. Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. arXiv preprint arXiv:2410.06940, 2024.
- [104] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. In Proc. Int’l Conf. Machine Learning, 2024.
- [105] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In Proc. IEEE Int’l Conf. Computer Vision and Pattern Recognition, 2024.
- [106] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In Proc. IEEE Int. Conf. Comput. Vision, 2023.

- [107] C. Zhao, Y. Song, W. Wang, H. Feng, E. Ding, Y. Sun, X. Xiao, and J. Wang. MonoFormer: One transformer for both diffusion and autoregression. [arXiv preprint arXiv:2409.16280](#), 2024.
- [108] 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. [arXiv preprint arXiv:2408.11039](#), 2024.
- [109] Y. Zhu, M. Zhu, N. Liu, Z. Ou, X. Mou, and J. Tang. LLaVA-Phi: Efficient multi-modal assistant with small language model. [arXiv preprint arXiv:2401.02330](#), 2024.
- [110] L. Zhuo, R. Du, H. Xiao, Y. Li, D. Liu, R. Huang, W. Liu, L. Zhao, F.-Y. Wang, Z. Ma, et al. Lumina-Next: Making Lumina-T2X stronger and faster with Next-DiT. [arXiv preprint arXiv:2406.18583](#), 2024.

## 附录

### A. 256 分辨率模型的性能分析

我们在两个分辨率下训练了我们的模型：256×256 和 384×384。主论文将 384×384 模型的结果作为我们的主要结果进行展示。在此，我们提供对 256×256 模型性能的综合评估。视觉理解性能结果如表 1 所示。生成能力使用 GenEval [28]、DPG-Benchmark [34] 和 MJHQ FID-30k [48] 进行评估，结果如表 2 和 3 所示。

表 1 | 视觉理解任务的结果。

模型	LLM 参数量	POPE↑	MME-P↑	MMB <sub>dev</sub> ↑	SEED↑	VQA <sub>v2<sub>test</sub></sub> ↑	GQA↑	MM-Vet↑
JanusFlow 256	1.3B	85.3	1203.0	71.9	67.6	76.3	58.4	27.4
JanusFlow 384	1.3B	88.0	1333.1	74.9	70.5	79.8	60.3	30.9

表 2 | GenEval [28] 上的结果。

方法	LLM 参数量	单物体	双物体	计数	颜色	位置	颜色属性	总体 ↑
JanusFlow 256	1.3B	0.98	0.73	0.54	0.83	0.63	0.53	0.70
JanusFlow 384	1.3B	0.97	0.59	0.45	0.83	0.53	0.42	0.63

表 3 | DPG-Bench [34] 和 MJHQ FID-30k [48] 上的结果。

方法	DPG-Bench↑						MJHQ FID-30k↓
	全局	实体	属性	关系	其他	总体	
JanusFlow 256	91.20	88.83	88.00	87.60	89.53	81.23	12.70
JanusFlow 384	87.03	87.31	87.39	89.79	88.10	80.09	9.51

正如预期，由于分辨率较低，256×256 模型在视觉理解指标上的表现略低于 384×384 模型。然而有趣的是，256×256 模型在 GenEval 和 DPG-Bench 上的表现优于其高分辨率对应模型——这两个基准专门用于评估指令遵循能力和语义准确性。在语义任务上的这种优越表现可归因于模型对低分辨率图像具有更好的控制力，较低的视觉复杂性使得更精确的语义操控成为可能。

### B. 数据集详情

理解阶段预训练使用的数据集包括 DetailedCaption [20]、SAM [41]、arXivQA [50]、DenseFusion-1M [51]、MMSci [53]、PixelProse [82]、重新配文的 LAION-Aesthetics [16]、重新配文的 Open Images V4 [43]、ShareGPT4V [11]、WikiHow [42] 和 WIT [84]。生成阶段预训练使用的数据集包括重新配文的 LAION-Aesthetics [16]、DALL-E 3 1M [21]、SAM [41]、Open Images V4 [43]、Megalith-10M [68]、YFCC-15M [71]、PixelProse [82] 和 JourneyDB [85]。

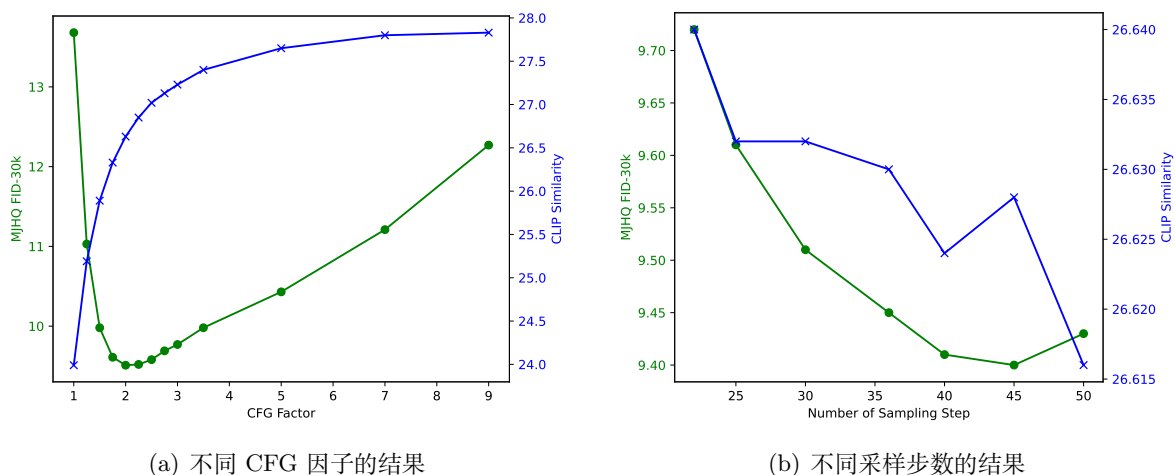


图 1 | 不同 CFG 因子与采样步数的结果. 在图 (a) 中, 采样步数设置为 30。在图 (b) 中, CFG 因子设置为 2。

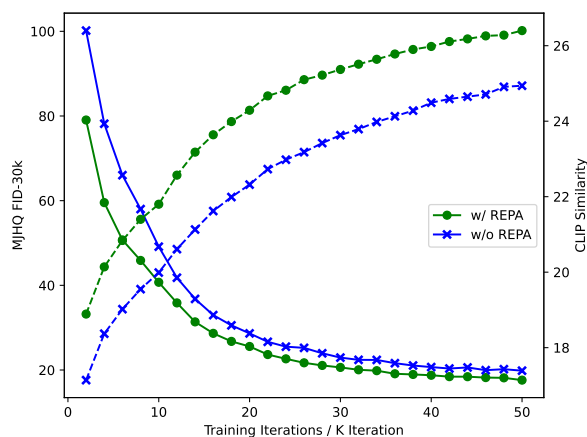


图 2 | 前 50,000 次迭代过程中的 FID 与 CLIP 相似度。

### C. CFG 因子与采样步数分析

我们研究了两个关键生成参数的影响：无分类器引导（Classifier-Free Guidance, CFG）因子和采样步数。尽管我们的主要结果在计算 FID 时使用了  $w = 2$  的 CFG 因子和 30 步采样，但在此我们对这些超参数进行了全面分析。图 1(a) 展示了在保持 30 步采样的情况下，改变 CFG 因子的效果。结果表明，FID 分数存在一个最优的 CFG 值，而 CLIP [75] 相似度则随着 CFG 值的增加而持续提升，这与先前工作的发现一致 [73]。图 1(b) 展示了在保持 CFG 因子为 2 的情况下，不同采样步数的影响。采样步数对性能的影响相对较小。我们在主文中选择 30 步是在生成质量与计算效率之间取得平衡的结果。

## D. REPA 消融实验细节

我们在图 2 中提供了预训练阶段前 50,000 次训练迭代中，使用与不使用表示对齐正则化 (representation alignment regularization) 的 FID 和 CLIP 相似度。两种模型之间的差距证明了使用表示对齐正则化的优势。

## E. 额外的定性结果

理解任务和生成任务的额外定性示例分别展示在图 3 和图 4 中。理解示例展示了 JanusFlow 的多样化能力，包括代码生成、人物识别、字符识别和视觉推理。在图像生成方面，我们的模型在视觉质量和与输入提示的语义对齐方面均表现出强劲的性能。

**Simple Bar Chart**

Categories	Values
0	3
1	7
2	5
3	2

**User:** Please write a python code to draw this plot.

**JanusFlow:** Here is the Python code to draw this plot:

```
```python
import matplotlib.pyplot as plt

# Data
values = [3, 7, 5, 2]

# Create the bar chart
plt.figure(figsize=(10, 6))
plt.bar(range(len(values)), values)

# Title
plt.title('Simple Bar Chart')

# Labels
plt.xlabel('Categories')
plt.ylabel('Values')

# Show the plot
plt.show()
```
```

**User:** Explain why this image is funny.

**JanusFlow:** This image is likely to be a humorous or artistic rendition of the famous painting "The Mona Lisa" with a dog instead of a human. The dog's expression and pose are similar to the iconic pose of the original painting, which adds a playful and unexpected twist to the artwork.

**User:** What are the texts in this passage?

**JanusFlow:** To increase the accessibility of this powerful model class and at the same time reduce its significant resource consumption, a method is needed that reduces the computational complexity for both training and sampling. Reducing the computational demands of DMs without impairing their performance is, therefore, key to enhance their accessibility.

**User:** Who is this in the figure?

**JanusFlow:** George W. Bush.

图 3 | 更多多模态理解案例。



Massive cathedral church, battle between Heaven and hell, church on fire, 8k hyper real ultra sharp renaissance by Francisco Goya.



A handsome 24-year-old boy in the middle with sky color background wearing eye glasses, it's super detailed with anime style.



Happy dreamy owl monster sitting on a tree branch, colorful glittering particles, forest background, detailed feathers.



A man wearing Fedora hat with mafia style, realistic photography, intricate details, magical lighting, vibrant background, complex textures, rich colors, realistic style, front-facing view.



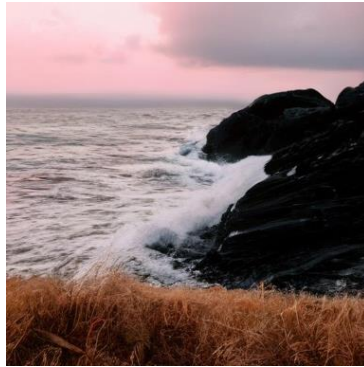
A vivid depiction of the Northern Lights dancing above the snow-covered mountains in Iceland, casting a mesmerizing glow across the sky.



A dark, high-contrast render of a psychedelic Tree of Life glowing brilliantly, illuminating swirling dust particles in a mystical, cavernous setting.



The image features a mushroom growing on grassy ground amidst fallen leaves. Their caps are light brownish-white with visible gills underneath; the stems appear dark and sturdy. In the background, there's an out-of-focus scene that includes greenery and possibly some structures or trees shrouded by mist or fog, giving it a serene yet slightly eerie atmosphere. This photograph employs shallow depth of field to emphasize the mushrooms while blurring the surroundings for artistic effect.



The image captures a vast ocean view at either sunrise or sunset, with soft pink hues near the horizon blending into darker clouds above. Waves crash against rugged black rocks on the right, where water flows down onto smaller stones below. In the foreground, dry grass contrasts with the smooth sea surface. The scene feels tranquil but also reveals the raw power of nature through the interaction between the dynamic waves and the solid land.



A serene Chinese ink painting depicts a tranquil mountain village. Simple homes nestle at the foot of misty peaks, while a gentle river winds through the village. Bamboo and pine trees dot the landscape. The minimalist brushstrokes reflect a harmonious relationship between nature and human life, capturing the peaceful essence of the scene with elegant simplicity.

图 4 | 更多文生图生成结果。