



Janus: 解耦视觉编码以实现统一的多模态理解与生成

Chengyue Wu^{1,2} Xiaokang Chen^{1,*†} Zhiyu Wu^{1,3} Yiyang Ma^{1,3} Xingchao Liu¹ Zizheng Pan¹
Wen Liu¹ Zhenda Xie¹ Xingkai Yu¹ Chong Ruan¹ Ping Luo^{2,*}

¹DeepSeek-AI ²The University of Hong Kong ³Peking University

†: 项目负责人 *: 通讯作者

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

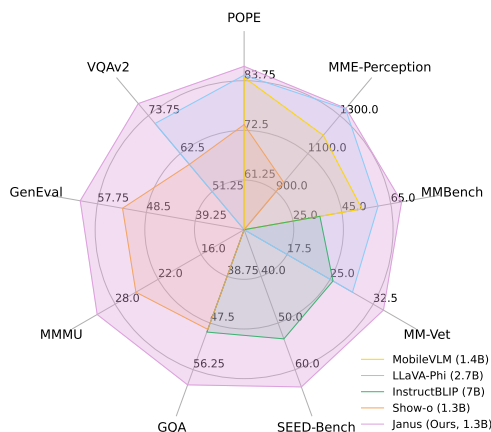
Abstract

本文介绍了 **Janus**，这是一个统一多模态理解与生成的自回归框架。先前的研究通常依赖单一的视觉编码器来处理这两项任务（例如 Chameleon）。然而，由于多模态理解与生成对信息粒度的要求不同，这种方法可能导致次优性能，尤其是在多模态理解方面。为解决这一问题，我们将视觉编码解耦为独立的路径，同时仍利用单一统一的 Transformer 架构进行处理。这种解耦不仅缓解了视觉编码器在理解与生成任务中的角色冲突，还增强了框架的灵活性。例如，多模态理解与生成组件均可独立选择最适合其自身的编码方法。实验表明，Janus 超越了以往的统一模型，并达到或超过了专用任务模型的性能。Janus 的简洁性、高灵活性和有效性使其成为下一代统一多模态模型的有力候选者。

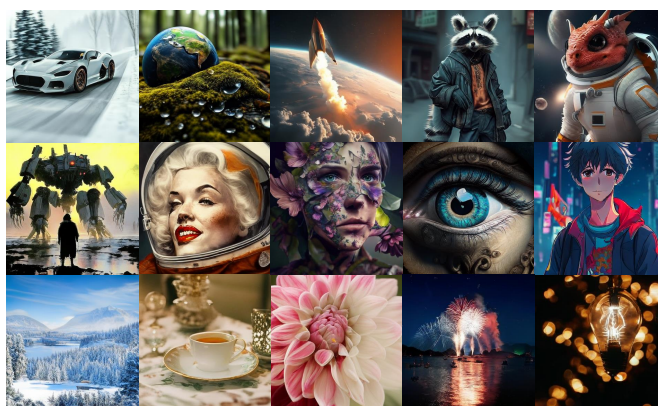
1. 引言

近年来，多模态大模型在理解与生成领域均取得了显著进展 [20, 51]。在多模态理解领域，研究人员遵循 LLaVA [51] 的设计思路，利用视觉编码器作为桥梁，使大语言模型 (LLMs) 能够理解图像。在视觉生成领域，基于扩散的方法 [9, 20, 20, 67] 取得了显著成功。最近，一些工作探索了用于视觉生成的自回归方法 [73, 79]，其性能已可与扩散模型相媲美。为了构建更强大、更通用的多模态模型，研究人员试图将多模态理解与生成任务相结合 [75, 77, 94]。例如，部分研究尝试将多模态理解模型与预训练扩散模型相连接 [27, 28, 75]。例如，Emu [75] 将 LLM 的输出作为预训练扩散模型的条件，随后依赖扩散模型生成图像。然而，严格来说，这种方法不能被视为真正的统一模型，因为视觉生成功能由外部扩散模型处理，而多模态 LLM 本身缺乏直接生成图像的能力。

其他方法 [77, 85, 86, 94] 采用单一 Transformer 来统一多模态理解与生成任务，这不仅提升了视觉生成的指令遵循能力，还激发了潜在的涌现能力，并减少了模型冗余。此类方法通常使用单一的视觉编码器来处理这两项任务的输入。然而，多模态理解与生成任务所需的表征存在显著差异。在多模态理解任务中，视觉编码器的目的是提取高层语义信息（例如图像中的物体类别或视觉属性）。理解任务的输出不仅涉及从图像中提取信息，还涉及复杂的语义推理。因此，视觉编码器表征的粒度主要侧重于高维语义表征。相比之下，在视觉生成任务中，重点在于生成局



(a) 基准测试性能。



(b) 视觉生成结果。

图 1 | Janus 的多模态理解与视觉生成结果。Janus 的性能优于以往最先进的统一多模态模型以及部分专用多模态理解模型，同时展现出强大的视觉生成能力。图像分辨率为 384×384 。建议在屏幕上查看。

部细节并保持图像的全局一致性。在此场景下，表征需要一种低维编码，以能够表达细粒度的空间结构和纹理细节。将这两项任务的表征统一在同一空间内会导致冲突与权衡。因此，现有的多模态理解与生成统一模型往往在多模态理解性能上做出妥协，明显落后于最先进的多模态理解模型。我们将在消融研究中进一步探讨这一问题。

为解决这一问题，我们提出了 **Janus**¹，这是一个解耦多模态理解与生成视觉编码的统一多模态框架。具体而言，我们引入了两条独立的视觉编码路径：一条用于多模态理解，另一条用于多模态生成，二者由相同的 Transformer 架构统一处理。该方法带来两大主要优势：(1) Janus 缓解了多模态理解与生成因粒度需求不同而产生的冲突，消除了在选择视觉编码器时需在两项任务间进行权衡的必要性。(2) Janus 具有灵活性和可扩展性。解耦后，理解与生成任务均可采用各自领域最先进的编码技术。此外，Janus 未来有望兼容更多输入类型，例如点云、脑电 (EEG) 信号或音频数据，其中独立编码器可提取特征，随后由统一的 Transformer 进行处理。

¹在罗马神话中，Janus 是二元与过渡之神，以其两面朝向相反方向的形象象征着矛盾力量的共存。类似地，我们的模型捕捉到了视觉任务之间的内在张力：理解需要抽象的高层语义表征，而生成则需要具体、详细的信息。通过将这两个过程解耦为专用编码器，我们的系统映射了 Janus 的双重特性，在统一架构内化解了这种张力。

据我们所知，我们是首个在统一多模态理解与生成框架中强调解耦视觉编码重要性的工作。实验结果表明，在多模态理解与生成基准测试上，Janus 超越了参数量相当的现有统一模型，达到了最先进的水平。值得注意的是，Janus 甚至优于部分参数量大得多的专用任务模型（图 1）。具体而言，在多模态理解基准 MMBench [54]、SEED-Bench [42] 和 POPE [48] 上，Janus (1.3B) 分别取得了 69.4、63.7 和 87.0 的分数，优于 LLaVA-v1.5 (7B) [50] 和 Qwen-VL-Chat (7B) [3]。在视觉生成基准 MSCOCO-30K [11] 和 GenEval [30] 上，Janus 取得了 8.53 的 FID 分数和 61% 的准确率，超越了 DALL-E 2 [66] 和 SDXL [62] 等文生图生成模型。我们相信，Janus 的强劲性能，结合其高灵活性与可扩展性，使其成为下一代统一多模态模型的有力候选者。

2. 相关工作

2.1. 视觉生成

视觉生成是一个快速发展的领域，它将自然语言处理的概念与 Transformer 架构的进展相结合。受语言处理领域成功经验的启发，自回归模型利用 Transformer 来预测离散视觉标记（码本 ID）的序列 [24, 65, 75]。这些模型对视觉数据进行标记化，并采用类似于 GPT 风格 [64] 的预测方法。此外，掩码预测模型 [7, 8] 借鉴了 BERT 风格 [19] 的掩码方法，通过预测视觉输入中被掩码的部分来提高合成效率，并且已被应用于视频生成 [89]。与此同时，连续扩散模型在视觉生成方面展现了令人印象深刻的能力 [33, 67, 71]，它们通过概率视角进行生成，从而与离散方法形成互补。

2.2. 多模态理解

多模态大语言模型（MLLMs）融合了文本与图像 [6, 80, 81]。通过利用预训练的大语言模型（LLMs），MLLMs [1, 2, 12, 51, 55, 82, 95] 展现出强大的多模态信息理解与处理能力。近期的研究探索了将预训练扩散模型与 MLLMs 结合以辅助图像生成的方法 [27, 29, 36, 75, 76]。这些方法属于工具利用范畴，即利用扩散模型根据 MLLM 输出的条件来生成图像，而 MLLM 本身并不具备直接进行视觉生成的能力。此外，整个系统的生成能力通常受限于外部扩散模型，导致其性能往往不如直接单独使用扩散模型 [27, 75]。

2.3. 统一的多模态理解与生成

统一的多模态理解与生成模型被认为在促进跨模态无缝推理与生成方面具有强大能力 [77, 94]。此类模型的传统方法通常对理解和生成任务共用单一的视觉表示，无论其是基于自回归（AR）模型 [77, 85] 还是扩散模型 [86, 94]。例如，Chameleon [77] 采用 VQ Tokenizer 对图像进行编码，以同时服务于多模态理解与生成。然而，这种做法可能导致次优结果，因为视觉编码器可能在理解与生成的需求之间面临权衡。相比之下，我们的 Janus 能够显式地解耦用于理解和生成的视觉表示，认识到不同任务可能需要不同层次的信息。

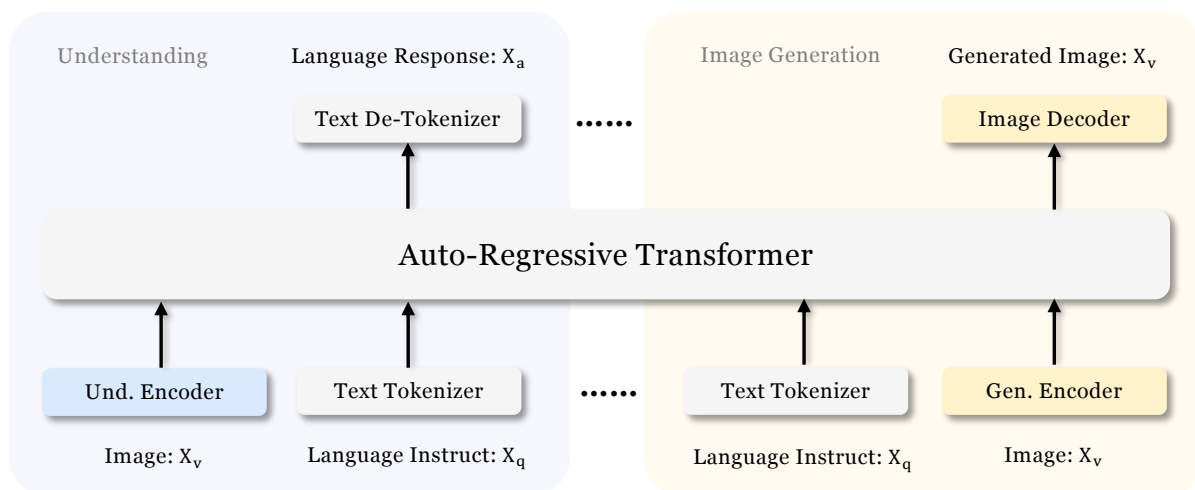


图 2 | **Janus 的架构**。与以往通常假设视觉理解与生成需要相同视觉编码器的方法 [77, 85] 不同，我们的 Janus 将视觉理解与视觉生成的视觉编码过程解耦。“Und. Encoder”和“Gen. Encoder”分别是“Understanding Encoder”（理解编码器）和“Generation Encoder”（生成编码器）的缩写。彩色显示效果更佳。

3. Janus：一个简单、统一且灵活的多模态框架

3.1. 架构

Janus 的架构如图 2 所示。针对纯文本理解、多模态理解和视觉生成任务，我们采用独立的编码方法将原始输入转换为特征，随后由一个统一的自回归 Transformer 进行处理。具体而言，对于文本理解，我们使用大语言模型（LLM）内置的分词器将文本转换为离散 ID，并获取每个 ID 对应的特征表示。对于多模态理解，我们使用 SigLIP [92] 编码器从图像中提取高维语义特征。这些特征从 2-D 网格展平为 1-D 序列，并通过一个理解适配器将这些图像特征映射到 LLM 的输入空间中。对于视觉生成任务，我们使用来自 [73] 的 VQ 分词器将图像转换为离散 ID。在将 ID 序列展平为 1-D 后，我们使用一个生成适配器将每个 ID 对应的码本嵌入映射到 LLM 的输入空间中。随后，我们将这些特征序列拼接起来，形成一个多模态特征序列，并输入到 LLM 中进行处理。LLM 内置的预测头用于纯文本理解和多模态理解任务中的文本预测，而视觉生成任务中的图像预测则使用一个随机初始化的预测头。整个模型遵循自回归框架，无需专门设计的注意力掩码。

3.2. 训练流程

Janus 的训练分为三个阶段，如图 3 所示。具体细节如下。

阶段一：训练适配器与图像预测头。该阶段的主要目标是在嵌入空间内建立视觉元素与语言元素之间的概念联系，使 LLM 能够理解图像中展示的实体，并具备初步的视觉生成能力。在此阶段，我们保持视觉编码器和 LLM 的参数冻结，仅更新理解适配器、生成适配器和图像预测头中的可训练参数。

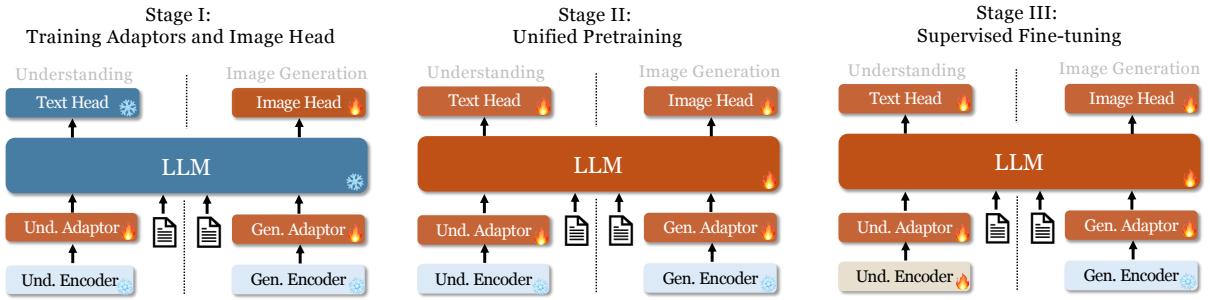


图 3 | 我们的 Janus 采用三阶段训练流程。图中使用火焰符号/雪花符号表示模块更新/不更新其参数。

阶段二：统一预训练。在此阶段，我们使用多模态语料进行统一预训练，使 Janus 能够同时学习多模态理解与生成能力。我们解冻 LLM，并利用所有类型的训练数据：纯文本数据、多模态理解数据和视觉生成数据。受 PixArt [9] 的启发，我们首先使用 ImageNet-1k 进行简单的视觉生成训练，以帮助模型掌握基本的像素依赖关系。随后，我们使用通用的文本到图像数据来增强模型的开放域视觉生成能力。

阶段三：监督微调。在此阶段，我们使用指令微调数据对预训练模型进行微调，以增强其指令遵循和对话能力。我们微调除生成编码器之外的所有参数。我们专注于监督答案部分，同时对系统和用户提示进行掩码处理。为确保 Janus 在多模态理解和生成方面的熟练度，我们不为特定任务微调单独的模型。相反，我们混合使用纯文本对话数据、多模态理解数据和视觉生成数据，确保模型在各种场景下的通用性。

3.3. 训练目标

Janus 是一个自回归模型，在训练过程中我们直接采用交叉熵损失：

$$\mathcal{L} = - \sum_{i=1} \log P_{\theta}(x_i | x_{<i}) \quad (1)$$

其中， $P(\cdot | \cdot)$ 表示由 Janus 的权重 θ 建模的条件概率。对于纯文本理解和多模态理解任务，我们在文本序列上计算损失。对于视觉生成任务，我们仅在图像序列上计算损失。为了保持设计简洁，我们未对不同任务分配不同的损失权重。

3.4. 推理

在推理阶段，我们的模型采用下一个 token 预测方法。对于纯文本理解和多模态理解，我们遵循标准做法，从预测分布中依次采样 token。对于图像生成，我们采用无分类器引导 (CFG)²，这与先前的工作 [8, 26, 73] 类似。具体而言，对于每个 token，logit l_g 的计算公式为： $l_g = l_u + s(l_c - l_u)$ ，

²在训练期间，我们以 10% 的概率将文生图数据中的文本条件替换为 pad token，从而使模型具备无条件视觉生成能力。

其中 l_c 为条件 logit, l_u 为无条件 logit, s 为无分类器引导的缩放系数。在后续的评估中, s 的默认值设为 5。

3.5. 可能的扩展

需要指出的是, 我们的设计为理解和生成分别配备了独立的编码器, 结构直观且易于扩展。

多模态理解。 (1) 对于多模态理解组件, 可以选择更强的视觉编码器, 而无需担心该编码器是否具备处理视觉生成任务的能力, 例如 EVA-CLIP [74]、InternViT [13] 等。(2) 为了处理高分辨率图像, 可以采用动态高分辨率技术 [50]。这使得模型能够适配任意分辨率, 而无需对 ViT 进行位置嵌入插值。为了节省计算成本, 还可以进一步压缩 Token, 例如使用像素重排 (pixel shuffle) 操作 [12]。

视觉生成。 (1) 对于视觉生成, 可以选择更细粒度的编码器, 以便在编码后保留更多的图像细节, 例如 MoVQGAN [93]。(2) 可以采用专为视觉生成设计的损失函数, 例如扩散损失 (diffusion loss) [46]。(3) 在视觉生成过程中, 可以结合自回归 (因果注意力) 与并行 (双向注意力) 方法, 以减少视觉生成过程中的累积误差 [79]。

支持额外模态。 Janus 直观的架构使其能够轻松集成额外的编码器, 从而支持多种模态, 如 3D 点云 [53]、触觉 [88] 和脑电图 (EEG) [4]。这使得 Janus 具备成为更强大的多模态通用模型的潜力。

4. 实验

在本节中, 我们展示了一系列全面的实验, 旨在评估我们的方法在多种视觉理解与生成任务上的性能。我们首先详细介绍实验设置, 包括模型架构、训练数据集和评估基准。接着, 我们报告 Janus 的性能, 并将其与其他最先进模型在多模态理解与生成基准上的表现进行比较。我们还进行了广泛的消融实验, 以验证所提方法的有效性。最后, 我们提供了一些定性结果。

4.1. 实现细节

在我们的实验中, 我们使用最大支持序列长度为 4096 的 DeepSeek-LLM (1.3B) [5] 作为基础语言模型。对于理解任务中使用的视觉编码器, 我们选择 SigLIP-Large-Patch16-384 [92]。生成编码器的码本大小为 16,384, 并将图像下采样 16 倍。理解适配器和生成适配器均为两层 MLP。各阶段的详细超参数见表 1。所有图像均调整大小为 384×384 像素。对于多模态理解数据, 我们调整图像的长边, 并用背景色 (RGB: 127, 127, 127) 填充短边至 384。对于视觉生成数据, 我们将短边调整至 384, 并将长边裁剪至 384。我们在训练过程中使用序列打包以提高训练效率。我们在单个训练步骤中按照指定比例混合所有数据类型。我们的 Janus 使用 HAI-LLM [32] 进行训练和评估, 这是一个基于 PyTorch 构建的轻量级高效分布式训练框架。整个训练过程在由 16 个节点组成的集群上耗时 7 天, 每个节点配备 8 块 Nvidia A100 (40GB) GPU。

表 1 | Janus 的详细超参数。数据比例指多模态理解数据、纯文本数据和视觉生成数据的比例。

超参数	阶段 1	阶段 2	阶段 3
学习率	1.0×10^{-3}	1×10^{-4}	2.0×10^{-5}
学习率调度器	Cosine	Constant	Constant
权重衰减	0.0	0.0	0.1
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	300	5,000	0
训练步数	10,000	180,000	24,000
批次大小	256	512	256
数据比例	1 : 0 : 1	2 : 3 : 5	7 : 3 : 10

4.2. 数据设置

在本节中，我们详细介绍了预训练和有监督微调数据集的构成。

阶段 I。我们使用的数据集包含来自 ShareGPT4V [10] 的 125 万张图文配对数据，用于多模态理解；以及来自 ImageNet-1k [18] 的约 120 万个样本，用于视觉生成。ShareGPT4V 数据的格式为“<image><text>”。ImageNet 数据使用类别名称组织成文生图数据格式：“<category_name><image>”。此处，“<>”符号表示占位符。

阶段 II。我们将数据组织为以下几类。(1) 纯文本数据。我们使用来自 DeepSeek-LLM [5] 的预训练文本语料库。(2) 图文交错数据。我们使用 WikiHow [39] 和 WIT [72] 数据集。(3) 图像描述数据。我们使用来自 [17, 18, 23, 38, 40, 45, 47, 49, 70] 的图像。其中，我们采用开源多模态模型对 [17, 40] 中的图像进行重新描述。图像描述数据被格式化为问答对，例如：“<image>Describe the image in detail.<caption>”。(4) 表格与图表数据。我们使用来自 DeepSeek-VL [55] 的相应表格和图表数据。数据格式为“<question><answer>”。(5) 视觉生成数据。我们利用来自多个数据集的图像-描述对，包括 [17, 38, 40, 57, 58, 60, 63, 70]，以及 200 万条内部数据。对于来自 [38, 70] 的图像，我们根据美学评分和图像尺寸进行过滤，最终保留 20%。在训练过程中，我们以 25% 的概率随机仅使用描述的第一句话，以鼓励模型发展出对简短描述的强生成能力。ImageNet 样本 [18] 仅在前 12 万步训练中出现，而其他数据集的图像则出现在后 6 万步中。正如 [9] 所建议的，这种方法有助于模型先学习基本的像素依赖关系，然后再过渡到更复杂的场景理解。视觉生成数据提供的格式为：“<caption><image>”。

阶段 III。对于文本理解，我们使用来自 [43] 的数据。对于多模态理解，我们使用来自 [31, 34, 35, 43, 56, 69] 的指令微调数据。对于视觉生成，我们使用来自 [17, 60, 70] 的图文对（阶段 II 的子集）以及 400 万条内部数据。我们采用以下格式进行指令微调：“User:<Input Message> \n Assistant: <Response>”。对于多轮对话，我们重复此格式来组织数据。

4.3. 评估设置

多模态理解。为了评估多模态理解能力，我们在广泛认可的基于图像的视觉语言基准上对我们的模型进行评估，这些基准包括 VQAv2 [31]、GQA [35]、POPE [48]、MME [25]、SEED [42]、

表 2 | 在多模态理解基准上与最先进模型的对比。“Und.” 和 “Gen.” 分别表示 “理解” 和 “生成”。使用外部预训练扩散模型的模型标记为 †。

类型	模型	LLM 参数量	POPE↑	MME-P↑	MMB↑	SEED↑	VQAv2 _(test) ↑	GQA↑	MMMUp↑	MM-Vet↑
仅理解	LLaVA-v1.5-Phi-1.5 [86]	1.3B	84.1	1128.0	-	-	75.3	56.5	30.7	-
	MobileVLM [14]	1.4B	84.5	1196.2	53.2	-	-	56.1	-	-
	MobileVLM-V2 [15]	1.4B	84.3	1302.8	57.7	-	-	59.3	-	-
	MobileVLM [14]	2.7B	84.9	1288.9	59.6	-	-	59.0	-	-
	MobileVLM-V2 [15]	2.7B	84.7	1440.5	63.2	-	-	61.1	-	-
	LLaVA-Phi [96]	2.7B	85.0	1335.1	59.8	-	71.4	-	-	28.9
	LLaVA [51]	7B	76.3	809.6	38.7	33.5	-	-	-	25.5
	LLaVA-v1.5 [50]	7B	85.9	1510.7	64.3	58.6	78.5	62.0	35.4	31.1
	InstructBLIP [16]	7B	-	-	36.0	53.4	-	49.2	-	26.2
	Qwen-VL-Chat [3]	7B	-	1487.5	60.6	58.2	78.2	57.5	-	-
	IDEFICS-9B [41]	8B	-	-	48.2	-	50.9	38.4	-	-
	Emu3-Chat [83]	8B	85.2	-	58.5	68.2	75.1	60.3	31.6	-
	InstructBLIP [16]	13B	78.9	1212.8	-	-	-	49.5	-	25.6
	理解与生成	DreamLLM† [21]	7B	-	-	-	-	72.9	-	-
LaVIT† [36]		7B	-	-	-	-	66.0	46.8	-	-
Emu† [75]		13B	-	-	-	-	52.0	-	-	-
NExT-GPT† [84]		13B	-	-	-	-	66.7	-	-	-
-----		-----	-----	-----	-----	-----	-----	-----	-----	-----
Show-o [86]		1.3B	73.8	948.4	-	-	59.3	48.7	25.1	-
Gemini-Nano-1 [78]		1.8B	-	-	-	-	62.7	-	26.3	-
LWM [52]		7B	75.2	-	-	-	55.8	44.8	-	9.6
VILA-U [85]		7B	85.8	1401.8	-	59.0	79.4	60.8	-	33.5
Chameleon [77]		7B	-	-	-	-	-	-	22.4	8.3
Janus (Ours)		1.3B	87.0	1338.0	69.4	63.7	77.3	59.1	30.5	34.3

MMB [54]、MM-Vet [90] 和 MMMU [91]。

视觉生成。为了评估视觉生成能力，我们使用 MSCOCO-30K [11]、MJHQ-30K [44] 和 GenEval [30] 基准。MSCOCO-30K 和 MJHQ-30K 采用 Fréchet Inception Distance (FID) 指标，将生成的图像与 3 万张高质量图像进行比较，以指示图像生成的整体效果。GenEval 是一个具有挑战性的基准，旨在通过对模型组合能力进行详细的实例级分析，来反映视觉生成模型的综合生成能力。

4.4. 与现有最优方法的对比

多模态理解性能。我们在表 2 中将所提方法与最先进的统一模型及仅理解模型进行了比较。Janus 在同等规模的模型中取得了整体最佳的结果。具体而言，与之前最佳的统一模型 Show-o [86] 相比，我们在 MME 和 GQA 数据集上的性能分别提升了 41% (949 → 1338) 和 30% (48.7 → 59.1)。这归功于 Janus 解耦了多模态理解与生成的视觉编码，从而缓解了这两项任务之间的冲突。与规模显著更大的模型相比，Janus 依然保持高度的竞争力。例如，Janus 在 POPE、MMbench、SEED Bench 和 MM-Vet 等多个数据集上均优于 LLaVA-v1.5 (7B)。

视觉生成性能。我们报告了 Janus 在 GenEval、COCO-30K 和 MJHQ-30K 基准测试上的视觉生成性能。如表 3 所示，我们的 Janus 在 GenEval 上取得了 61% 的整体准确率，优于之前最佳的统一模型 Show-o (53%) 以及一些流行的仅生成方法，例如 SDXL (55%) 和 DALL-E 2 (52%)。

表 3 | 在 GenEval 基准上对文生图生成能力的评估。“Und.” 和 “Gen.” 分别表示 “理解” 和 “生成”。使用外部预训练扩散模型的模型标记为 †。

类型	方法	参数量	单对象	双对象	计数	颜色	位置	颜色属性	整体 ↑
仅生成	LlamaGen [73]	0.8B	0.71	0.34	0.21	0.58	0.07	0.04	0.32
	LDM [67]	1.4B	0.92	0.29	0.23	0.70	0.02	0.05	0.37
	SDv1.5 [67]	0.9B	0.97	0.38	0.35	0.76	0.04	0.06	0.43
	PixArt- α [9]	0.6B	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1 [67]	0.9B	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2 [66]	6.5B	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	Emu3-Gen [83]	8B	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	SDXL [62]	2.6B	0.98	0.74	0.39	0.85	0.15	0.23	0.55
理解与生成	SEED-X [†] [29]	17B	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	Show-o [86]	1.3B	0.95	0.52	0.49	0.82	0.11	0.28	0.53
	LWM [52]	7B	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	Chameleon [77]	34B	-	-	-	-	-	-	0.39
	Janus (Ours)	1.3B	0.97	0.68	0.30	0.84	0.46	0.42	0.61

这表明我们的方法具有更好的指令遵循能力。如表 4 所示，Janus 在 COCO-30K 和 MJHQ-30K 基准测试上分别取得了 8.53 和 10.10 的 FID 分数，超越了统一模型 Show-o 和 LWM，并与一些知名的仅生成方法相比展现出具有竞争力的性能。这证明了 Janus 生成的图像质量良好，并凸显了其在视觉生成方面的潜力。

表 4 | 在 MSCOCO-30K 和 MJHQ-30K 基准上对文本到图像生成能力的评估。“Und.” 和 “Gen.” 分别表示 “理解” 和 “生成”。使用外部预训练扩散模型的模型标记为 †。

类型	模型	参数量	COCO-30K↓	MJHQ-30K↓
仅生成	DALL·E [65]	12B	27.50	-
	GLIDE [59]	5B	12.24	-
	LDM [67]	1.4B	12.64	-
	DALL·E 2 [66]	6.5B	10.39	-
	SDv1.5 [67]	0.9B	9.62	-
	GigaGAN [37]	0.9B	9.09	-
	PixArt- α [9]	0.6B	7.32	-
	Imagen [68]	34B	7.27	-
	RAPHAEL [87]	3B	6.61	-
理解与生成	Emu [†] [75]	13B	11.66	-
	NEXT-GPT [†] [84]	13B	11.28	-
	SEED-X [†] [29]	17B	14.99	-
	Show-o [86]	1.3B	9.24	15.18
	LWM [52]	7B	12.68	17.77
	VILA-U (256) [85]	7B	-	12.81
	VILA-U (384) [85]	7B	-	7.69
	Janus (Ours)	1.3B	8.53	10.10

4.5. 消融实验

我们精心设计消融实验以验证 Janus 设计概念的有效性。首先，我们设计实验以验证解耦视觉编码的重要性和优势。其次，我们研究统一训练对多模态理解或视觉生成等单项任务的影响。结果列于表 5。

表 5 | **消融实验**。我们验证了解耦视觉编码的有效性，并比较了统一训练与特定任务训练。“Und.”、“Gen.” 和 “SE. Tokenizer” 分别表示 “理解”、“生成” 和 “语义分词器”。

实验编号	视觉编码器	训练任务	POPE↑	MMB↑	SEED↑	MMMU↑	COCO-FID↓
A	VQ Tokenizer	理解 + 生成	60.1	35.0	34.9	24.7	8.72
B	SE. Tokenizer	理解 + 生成	82.4	52.7	54.9	26.6	7.11
C	SE. Tokenizer	理解	83.9	62.1	60.8	27.5	-
D	SigLIP + VQ (Ours)	理解 + 生成	87.0	69.4	63.7	30.5	8.53
E	SigLIP	理解	85.9	70.6	64.8	28.8	-
F	VQ Tokenizer	生成	-	-	-	-	8.92

基线构建。遵循先前的工作 [77]，我们选择一个 VQ tokenizer [73] 来为多模态理解和生成任务编码图像，作为基线 (Exp-A)。考虑到 Exp-A 中的 VQ tokenizer 在提取语义信息方面可能较弱，从而使其在多模态理解上的效果不佳，我们还构建了一个更强的基线 Exp-B。我们采用 SigLIP 来蒸馏一个增强的语义 tokenizer³，它既能从图像中提取高层语义信息，又具备将图像转换为离散 ID 的能力，这与 [85] 中的方法类似。语义 tokenizer 的详细信息可在附录 A.1 中找到。

解耦视觉编码的影响。(1) 从 Exp-A 的结果来看，我们发现该模型在视觉生成基准上取得了令人满意的性能 (在 COCO 上 FID 为 8.72)。然而，Exp-A 与我们的模型 (Exp-D) 在理解基准上存在显著差距。(2) 将 Exp-B 与 Exp-A 进行比较时，结果显示多模态理解能力有明显提升，尽管与我们的方法相比仍有较大差距。在视觉生成方面，Exp-B 的表现优于 Exp-D。我们推测这有两个可能的原因。首先，语义 tokenizer 生成的离散 ID 具有更强的语义连贯性，为 LLM 提供了更合理的预测目标。其次，Exp-B 中的视觉编码器参数量显著多于 Exp-D 中的 Gen. 编码器。(3) 为了探究使用单一视觉编码器是否会导致多模态理解与生成之间的权衡，我们基于 Exp-B 进一步设计了 Exp-C，该实验仅专注于多模态理解训练。Exp-C 的多模态理解能力显著优于 Exp-B。这表明 Exp-B 中的视觉编码器在多模态理解与生成之间做出了权衡，最终牺牲了其多模态理解能力。上述实验阐明了解耦视觉编码的重要性。

统一模型 vs. 纯理解与纯生成。我们比较了统一训练 (Exp-D) 与纯理解 (Exp-E) 和纯生成 (Exp-F) 训练的性能。对于纯理解训练，我们省略了视觉生成数据；对于纯生成训练，我们排除了理解数据。请注意，统一训练和纯理解训练在理解部分遵循相同的步骤。同样，统一训练和纯生成训练在视觉生成部分也遵循相同的步骤。实验结果表明，统一训练的性能与仅针对理解或仅针对视觉生成训练的性能相当。这证明我们的模型 Janus 能够在融入强大生成能力的同时，对多模态理解性能的影响降至最低。

4.6. 定性结果

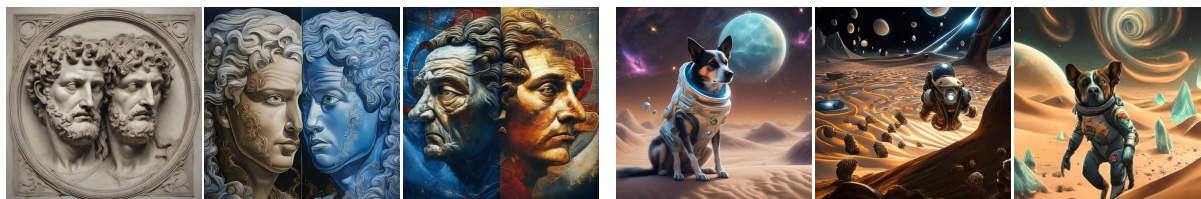
视觉生成可视化结果。图 4 提供了我们的模型与基于扩散的模型 (如 SDXL [62]) 以及自回归模型 LlamaGen [73] 之间的定性对比。结果表明，我们的模型在视觉生成方面展现出更优越的指

³该语义 tokenizer 仅在消融实验中作为更强的基线使用。为简化起见，我们在主实验中使用了普通的 VQ tokenizer [73]。



A close-up high-contrast photo of Sydney Opera House sitting next to Eiffel tower, under a blue night sky of roiling energy, exploding yellow stars, and radiating swirls of blue.

A wise old owl with golden plumage perched on a luminous crystal tree in a magical forest. Radiant fireflies swirl around while ethereal mist rolls through the trees, illuminated by swirls of iridescent moonlight and glistening emerald leaves.



A detailed portrait of the Roman god Janus, featuring his two faces looking in opposite directions. One face appears aged, with deep-set wrinkles and a wise, contemplative expression, while the other face is youthful, exuding vigor and curiosity. His hair is styled in flowing curls, framing both faces with a sense of divine symmetry. The artwork is rich in contrasting colors, with the left side dominated by cold blues and silvers, symbolizing winter and reflection, and the right side awash with warm golds and reds, representing spring and renewal. The background is a celestial tapestry, adorned with stars and symbolic motifs of time and passage.

A brave dog wearing a futuristic space suit, exploring an alien planet amidst swirling dunes of stardust and meteor showers. The landscape is dotted with glowing crystal formations and ethereal terraforms, creating a surreal environment in which swirling vortexes in the sky depict the endless dance of distant galaxies.

图 4 | 与 LlamaGen 和 SDXL 的视觉生成定性对比. Janus 生成的图像与用户提示词表现出更好的一致性。SDXL、LlamaGen 和我们的方法生成的图像分辨率分别为 1024×1024 、 512×512 和 384×384 。建议在屏幕上查看。

令遵循能力，能够准确捕捉用户提示词中的大部分细节。这表明统一模型在视觉生成领域具有巨大潜力。更多可视化结果可在附录 B 中找到。

梗图 (MEME) 上的多模态理解。图 5 展示了 Janus 的多模态理解能力的定性结果，并与 Chameleon [77] 和 Show-o [86] 进行了对比。Janus 能够准确解读图像配文，并捕捉到梗图所传达的情感。相比之下，Chameleon 和 Show-o 在准确识别图像中的文本方面均存在困难。此外，Chameleon 未能识别出梗图中的物体，而 Show-o 则错误判断了狗的颜色。这些例子表明，与 Chameleon 和 Show-o 使用的共享编码器相比，解耦的视觉编码器显著提升了 Janus 的细粒度多模态理解能力。更多多模态理解示例可在附录 B 中找到。

5. 结论

在本文中，我们介绍了 Janus，一个简单、统一且可扩展的多模态理解与生成模型。Janus 的核心思想是将多模态理解与生成中的视觉编码进行解耦，这有助于缓解理解与生成任务对视觉编码器不同需求所引发的冲突。大量实验证明了 Janus 的有效性与领先性能。值得注意的是，Janus 具有高度的灵活性且易于扩展。除了在多模态理解与生成方面具有巨大的提升潜力外，Janus 还易于扩展以纳入更多的输入模态。上述优势表明，Janus 有望为下一代多模态通用模型的开发提供启发。

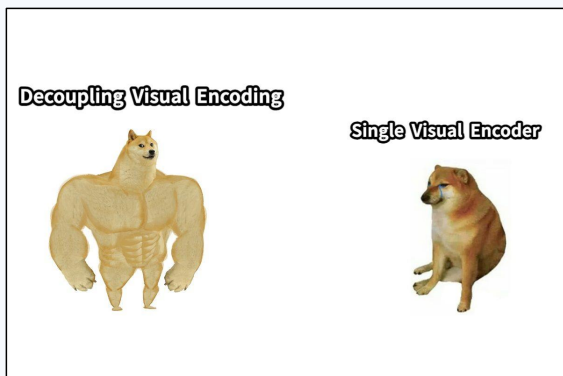


User: What is the meaning of this meme from the top panel to the bottom?

Janus (Ours): The meme is a humorous take on the concept of a **Nobel Prize in Physics**. The top panel shows a man in a red jacket with his hand up, seemingly confused or surprised, while the bottom panel shows the same man with a smile on his face, seemingly pleased or proud, as if he has just won the Nobel Prize. The meme is a play on the idea that the man in the top panel might be **confused or surprised by the idea of winning a Nobel Prize in Physics**, while the man in the bottom panel seems to be genuinely happy about it.

Chameleon: The meme is a visual representation of a conversation between two people, with the top panel featuring the text "I'm not arguing, I'm just explaining why I'm right" and the bottom panel featuring the text "No, you're not." The meme is using a technique called "irony" to convey the idea that even though one person is presenting themselves as non-argumentative, they are actually being argumentative. The use of the phrase "I'm not arguing, I'm just explaining why I'm right" is a common tactic used by people who are trying to avoid an argument but are actually engaging in one. The meme is poking fun at this tactic by showing the person's actions and words as being in direct contradiction to their claim.

Show-o: The meaning of the meme from the top panel to the bottom is that it is a humorous and light-hearted depiction of a man wearing a yellow hat and a red jacket, with the caption "I'm not a fan of the sun." This suggests that the man is not a fan of the sun, and he is using the sun as a metaphor for his feelings or emotions. The meme is likely meant to be humorous and entertaining, as it combines a serious subject with a lighthearted and playful approach.



User: Explain this meme.

Janus (Ours): The meme is comparing two different types of neural networks used in machine learning. The first type is called "Decoupling Visual Encoding," and the second type is called "Single Visual Encoder." The meme is trying to convey that the first type is **more powerful and can handle more complex tasks**, while the second type is **simpler and less powerful**.

Chameleon: This image shows a cartoon of a pig and a bear side-by-side, with the text "When you're a vegan and your friends still order sausage and bacon" above them. The image is a humorous take on the challenges of being a vegan in a non-vegan environment, where one's friends or family may not share their dietary preferences.

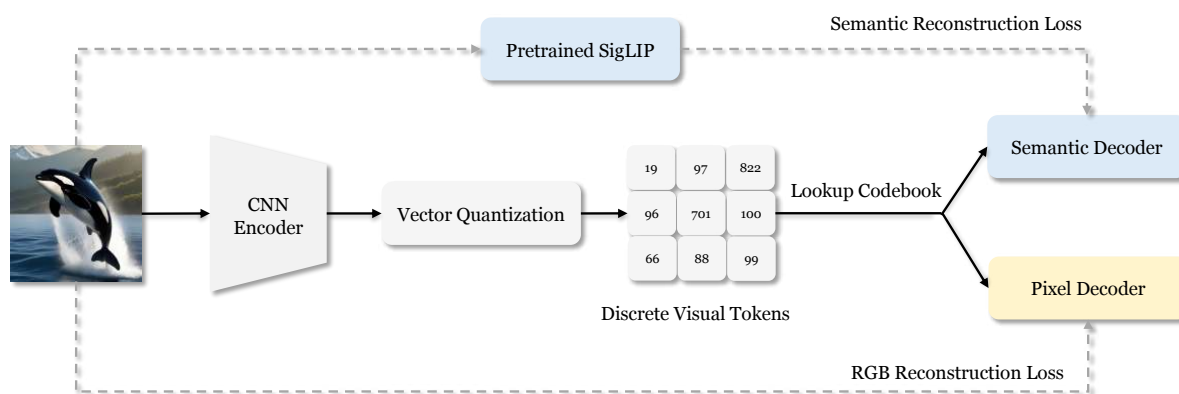
Show-o: The image features two dogs, **one brown and one white**, standing side by side. The brown dog is positioned on the left side of the image, while the white dog is on the right. Both dogs are looking at the camera, seemingly posing for a picture. The brown dog is slightly larger than the white dog. The image is captioned with the phrase "I'm not a dog, but I'm a dog." This humorous statement is accompanied by the image of the two dogs, emphasizing their ...

图 5 | 幽默梗图上的多模态理解定性结果。我们将模型的回复与 Chameleon-7B [77] 和 Show-o [86] 进行了对比。我们在回复中强调了关键点。屏幕观看效果最佳。

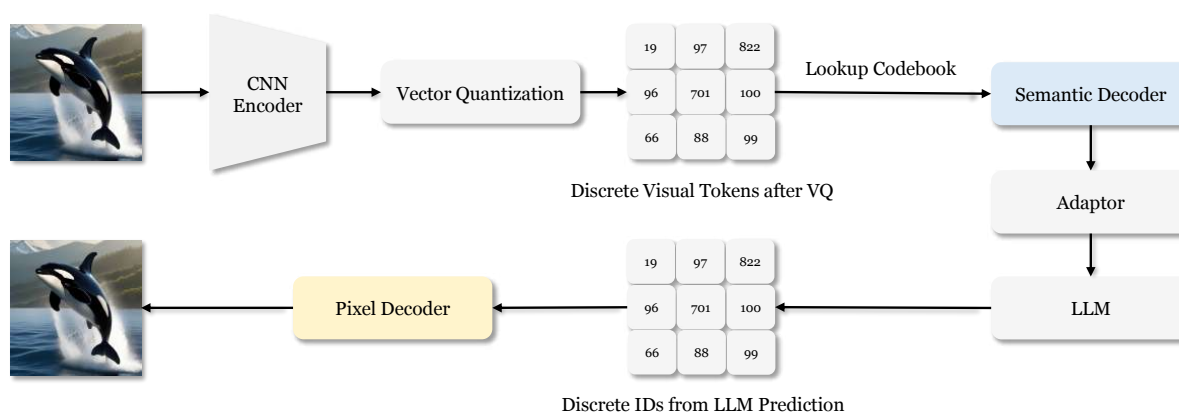
附录

A. 消融实验中提及的语义分词器细节

A.1. 语义分词器的架构



(a) Architecture of Semantic Tokenizer



(b) Architecture of LLM with Semantic Tokenizer Integration

图 6 | 语义分词器的架构与使用方式。(a) 训练语义分词器时使用的架构。我们使用预训练的 SigLIP [92] 来监督语义信息的重建，同时使用原始图像来监督 RGB 值的重建。(b) 将大语言模型 (LLM) 与语义解码器集成。语义解码器输出具有高层语义的连续特征，这些特征经过适配器 (adaptor) 处理后，作为 LLM 的输入。请注意，语义分词器仅在消融实验中使用，未在主实验中使用。

我们基于 [73] 中提出的分词器架构构建了语义分词器，其下采样率为 16。除了原有的 CNN 像素解码器外，我们在向量量化 (Vector Quantization) 之后添加了一个额外的语义解码器分支，如图 6 (a) 所示。该语义解码器是一个 12 层的 ViT [22]，包含 12 个注意力头，隐藏层维度为 768。对于语义解码器，我们使用因果注意力掩码 (causal attention mask)，以便在与 LLM 集成时促进下一个 token 的预测。

A.2. 训练

训练流程。语义分词器采用两阶段方式从头训练。在第一阶段，我们在 ImageNet-1k [18] 数据集上训练模型 40 个 epoch。在第二阶段，我们在 5000 万张图像上对模型进行 1 个 epoch 的微调。这些图像来源于 Janus 预训练过程中使用的视觉生成数据。我们使用 $1e-4$ 的固定学习率和 128 的批次大小 (batch size)。

训练损失。语义分词器的训练损失由两部分组成。一方面，我们采用 [73] 中描述的 RGB 重建损失。另一方面，我们使用 SigLIP-Large-Patch16-384 作为教师模型，来监督语义解码器的语义特征重建结果。我们采用了 BEiT-v2 [61] 中的损失函数。具体而言，我们最大化语义解码器预测的语义特征与 SigLIP 输出之间的余弦相似度。语义重建损失的权重设置为 0.25。

A.3. 与 LLM 的集成

我们在图 6 (b) 中展示了语义分词器与 LLM 的集成方式。图像首先通过 CNN 编码器、向量量化和语义解码器转换为连续特征。随后，LLM 处理这些特征并生成图像 ID 的预测结果。最后，像素解码器将这些离散 ID 转换回 RGB 值。

B. 额外的定性结果

更多文生图可视化结果。我们在图 7 中展示了更多的文生图结果。显然，Janus 能够生成高质量且紧密遵循给定提示词的图像。我们进一步探索了模型的多语言文生图能力，如图 8 所示。令人惊喜的是，尽管我们的训练数据仅包含英文文生图数据，Janus 仍然能够处理其他语言的文生图任务。我们将这种多语言能力归因于原始大语言模型的内在特性。LLM 首先将各种语言映射到统一的语义空间中，使得 Janus 无需额外训练即可自然地执行文生图任务。

更多多模态理解结果。图 9 展示了更多多模态理解的结果。在处理来自不同上下文的输入时，Janus 展现出了令人印象深刻的理解能力，彰显了其强大的性能。



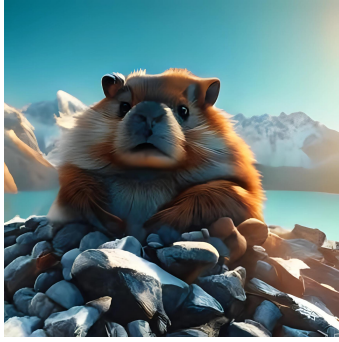
a young woman, looks like mix of Lana Del Rey and grimes. flowing cool colored hair, marbled, iridescent, shoujo manga, pre-raphaelite, k-pop, gilded, pearl, spun silk, clouds, ghost, glowing jellyfish, billowing gossamer cloth, Alexander McQueen, handmade lace, floral embroidery, snakeskin, dramatic lighting



Real photo of a cup of hot steaming coffee and a brass vase with a large bouquet of spring flowers by an old oak window at sunrise, fine details, rich colors taken with a nikon z6 camera and a nikon nikkor lens with 50 f5.6 iso 100 and a shutter speed of 1400 knot. UHD dtm HDR 8k



Portrait of a beautiful, curvaceous, Pirate princess goddess babe, red hair, intricate ornate costume, Caribbean background + outdoors + Ocean, painted by ArtGerm, Alphonse Mucha, Roberto Ferri, Ross Tran, Pixar, low angle shot, digital painting, cinematic rim lighting, Unreal Engine 5, 8K



a cute fluffy chubby marmot sunbathing on a pile of rocks, snow mountains background, turquoise glacier lake afar, clear blue sky, highly detailed, golden hour, natural light, octane render, unreal engine



epic 3d portrait of white King Kong wearing mech armor made of black crystals, golden ornate around the armor, symmetrical body, hyperrealistic, intricate details, shiny, cinematic, unreal engine, artstation, octane render.



Tiny cute adorable mouse dressed as a king in a castle, anthropomorphic, Jean-Baptiste Monge, soft cinematic lighting, 8k, intricate details, portrait, Pixar style character, old fashioned movie style



a panda that has been cybernetically enhanced more cybernetics3d 4k unreal engine chaos 20



A stunning princess from kabul in red, white traditional clothing, blue eyes, brown hair.



The ultimate wrist watch watch time machine, super advanced technology, holographic display, intricate mechanism.



Tiny cute adorable fluffy baby raccoon with knitted blue scarf leaning at a table in a medieval pub holding a coffee cup, anthropomorphic, Jean-Baptiste Monge, soft cinematic lighting, 8k, intricate details, portrait, Pixar style character, old fashioned movie style



Architectural parametric pavilion made from wood and glass, with organic cavities, surrounded by a beautiful forest. Dramatic scene, photorealistic, hyperrealistic, raytracing reflections, 8k hd, intricate detail in the style of Frank Lloyd Wright



Beautiful surreal symbolism the mesmerizing vision of a Cleopatra Queen of Egypt, full body, 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 composing, authentic, professional by Rorianai style 36k s1000

图 7 | 更多文生图生成结果。为了获得更好的可视化效果，我们将图像上采样至 1024 × 1024。

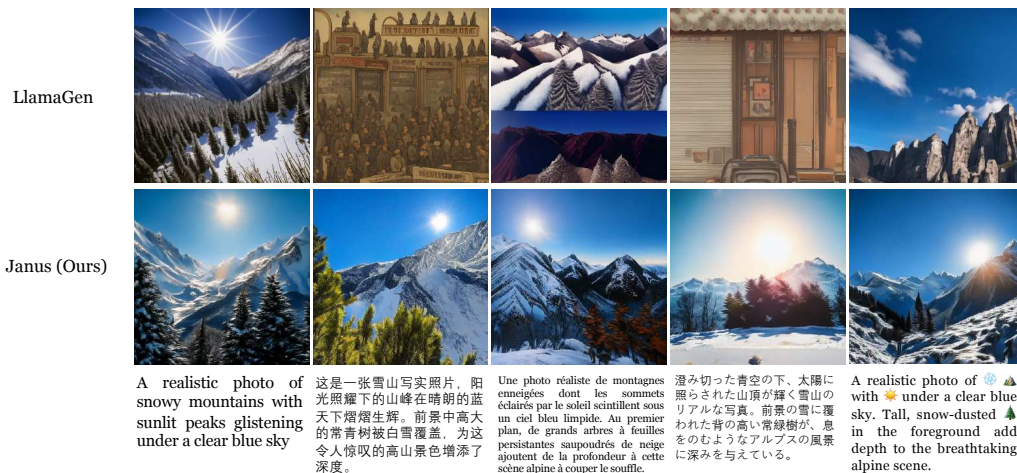


图 8 | 与 LlamaGen [73] 对比的多语言文生图生成样本。请注意，我们在训练中仅使用了英文文生图数据，这是模型涌现出的一种能力。提示词中使用的语言从左到右依次为：英语、中文、法语、日语以及带有表情符号的英语。

Prompt
Generate the latex code of this formula.

$$A_n = a_0 \left[1 + \frac{3}{4} \sum_{k=1}^n \left(\frac{4}{9} \right)^k \right]$$

Response

$$A_n = a_0 \left[1 + \frac{3}{4} \sum_{k=1}^n \left(\frac{4}{9} \right)^k \right]$$

Prompt
What algorithm does the image depict?

Response
The image depicts a CLIP (Contrastive Language-Image Pre-training) model. The CLIP model is a type of pre-training algorithm that combines the power of language and vision to learn a unified representation of the world. The image encoder processes the input image, and the text encoder processes the text associated with the image. The output of the text encoder is then used to predict the correct pair of images, which is a key component of the CLIP model.

Prompt
Generate the matplotlib.pyplot code for this plot.

Response
The image you've provided is a pie chart that represents the distribution of different types of animals based on their frequency. The chart is divided into four segments, each representing a different type of animal and its corresponding percentage. Here's the code to generate the pie chart:

```
python
import matplotlib.pyplot as plt
# Data
animals = [
    {"name": "Hogs", "frequency": 30.0},
    {"name": "Frog", "frequency": 15.0},
    {"name": "Logs", "frequency": 10.0},
    {"name": "Dogs", "frequency": 45.0}
]
# Create the pie chart
plt.pie(animals, labels=animals, autopct='%1.1f%%',
startangle=90)
# Set the title and labels
plt.title("Distribution of Animals")
plt.title.set_text("Frequency of Animals")
# Show the plot
plt.show()
```

Prompt
Why is this painting interesting?

Response
This image is an interesting piece of art because it combines a well-known painting, the "Mona Lisa," with a dog. The dog is depicted in the same pose as the Mona Lisa, but with a different expression and a different background. The painting is a unique blend of art styles and subjects, making it stand out.

图 9 | 更多多模态理解结果。Janus 具备强大的多模态理解能力，能够处理来自各种上下文的输入，例如科学图表、艺术作品图像、LaTeX 公式图像等。

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] Anthropic. The claude 3 model family: Opus, sonnet, haiku. <https://www.anthropic.com>, 2024.
- [3] 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.
- [4] Y. Bai, X. Wang, Y.-p. Cao, Y. Ge, C. Yuan, and Y. Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals. [arXiv preprint arXiv:2306.16934](#), 2023.
- [5] 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.
- [6] T. B. Brown. Language models are few-shot learners. [arXiv preprint arXiv:2005.14165](#), 2020.
- [7] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman. Maskgit: Masked generative image transformer. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition](#), pages 11315–11325, 2022.
- [8] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. [arXiv preprint arXiv:2301.00704](#), 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] 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.
- [11] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. [arXiv preprint arXiv:1504.00325](#), 2015.
- [12] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. [arXiv preprint arXiv:2404.16821](#), 2024.

- [13] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24185–24198, 2024.
- [14] 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.
- [15] 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.
- [16] 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, 2023.
- [17] dclure. Laion-aesthetics-umap. <https://huggingface.co/datasets/dclure/laion-aesthetics-12m-umap>, 2022.
- [18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009.
- [19] J. Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
- [20] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34:8780–8794, 2021.
- [21] 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. arXiv preprint arXiv:2309.11499, 2023.
- [22] A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
- [23] Echo840. Detailed caption dataset. https://huggingface.co/datasets/echo840/Detailed_Caption, 2023.
- [24] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12873–12883, 2021.

- [25] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394, 2023.
- [26] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, and Y. Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In European Conference on Computer Vision, pages 89–106. Springer, 2022.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] D. Ghosh, H. Hajishirzi, and L. Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems, 36, 2024.
- [31] 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 Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6904–6913, 2017.
- [32] High-flyer. Hai-llm: Efficient and lightweight training tool for large models, 2023. URL <https://www.high-flyer.cn/en/blog/hai-llm>.
- [33] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 33:6840–6851, 2020.
- [34] Y.-C. Hsiao, F. Zubach, M. Wang, et al. Screenqa: Large-scale question-answer pairs over mobile app screenshots. arXiv preprint arXiv:2209.08199, 2022.
- [35] D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6700–6709, 2019.
- [36] Y. Jin, K. Xu, L. Chen, C. Liao, J. Tan, B. Chen, C. Lei, A. Liu, C. Song, X. Lei, et al. Unified language-vision pretraining with dynamic discrete visual tokenization. arXiv preprint arXiv:2309.04669, 2023.

- [37] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10124–10134, 2023.
- [38] 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 Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015–4026, 2023.
- [39] M. Koupaei and W. Y. Wang. Wikihow: A large scale text summarization dataset. arXiv preprint arXiv:1810.09305, 2018.
- [40] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallocci, A. Kolesnikov, T. Duerig, and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. IJCV, 2020.
- [41] H. Laurençon, D. van Strien, S. Bekman, L. Tronchon, L. Saulnier, T. Wang, S. Karamcheti, A. Singh, G. Pistilli, Y. Jernite, and et al. Introducing idefics: An open reproduction of state-of-the-art visual language model, 2023. URL <https://huggingface.co/blog/idefics>.
- [42] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan. Seed-bench: Benchmarking multi-modal llms with generative comprehension. arXiv preprint arXiv:2307.16125, 2023.
- [43] 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.
- [44] 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.
- [45] 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. arXiv preprint arXiv:2403.00231, 2024.
- [46] T. Li, Y. Tian, H. Li, M. Deng, and K. He. Autoregressive image generation without vector quantization. arXiv preprint arXiv:2406.11838, 2024.
- [47] X. Li, F. Zhang, H. Diao, Y. Wang, X. Wang, and L.-Y. Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. arXiv preprint arXiv:2407.08303, 2024.
- [48] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen. Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355, 2023.

- [49] 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. arXiv preprint arXiv:2407.04903, 2024.
- [50] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26296–26306, 2024.
- [51] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024.
- [52] 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.
- [53] M. Liu, R. Shi, K. Kuang, Y. Zhu, X. Li, S. Han, H. Cai, F. Porikli, and H. Su. Openshape: Scaling up 3d shape representation towards open-world understanding. Advances in neural information processing systems, 36, 2024.
- [54] 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? arXiv preprint arXiv:2307.06281, 2023.
- [55] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, Y. Sun, et al. Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525, 2024.
- [56] 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. arXiv preprint arXiv:2110.13214, 2021.
- [57] madebyollin. Megalith-huggingface. <https://huggingface.co/datasets/madebyollin/megalith-10m>, 2024.
- [58] mehdidc. Yfcc-huggingface. <https://huggingface.co/datasets/mehdidc/yfcc15m>, 2024.
- [59] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741, 2021.
- [60] J. Pan, K. Sun, Y. Ge, H. Li, H. Duan, X. Wu, R. Zhang, A. Zhou, Z. Qin, Y. Wang, J. Dai, Y. Qiao, and H. Li. Journeydb: A benchmark for generative image understanding, 2023.
- [61] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. arXiv preprint arXiv:2208.06366, 2022.

- [62] 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. arXiv preprint arXiv:2307.01952, 2023.
- [63] ProGamerGov. Dalle3-high-quality-captions. <https://huggingface.co/datasets/ProGamerGov/synthetic-dataset-1m-dalle3-high-quality-captions>, 2024.
- [64] A. Radford. Improving language understanding by generative pre-training. 2018.
- [65] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In International conference on machine learning, pages 8821–8831. Pmlr, 2021.
- [66] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 1(2):3, 2022.
- [67] 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, pages 10684–10695, 2022.
- [68] 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. Advances in neural information processing systems, 35:36479–36494, 2022.
- [69] S. Shah, A. Mishra, N. Yadati, and P. P. Talukdar. Kvqa: Knowledge-aware visual question answering. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 8876–8884, 2019.
- [70] 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.
- [71] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020.
- [72] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, pages 2443–2449, 2021.
- [73] 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.

- [74] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao. Eva-clip: Improved training techniques for clip at scale. [arXiv preprint arXiv:2303.15389](#), 2023.
- [75] 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. [arXiv preprint arXiv:2307.05222](#), 2023.
- [76] 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](#), pages 14398–14409, 2024.
- [77] C. Team. Chameleon: Mixed-modal early-fusion foundation models. [arXiv preprint arXiv:2405.09818](#), 2024.
- [78] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. [arXiv preprint arXiv:2312.11805](#), 2023.
- [79] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. [arXiv preprint arXiv:2404.02905](#), 2024.
- [80] 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.
- [81] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batta, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. [arXiv preprint arXiv:2307.09288](#), 2023.
- [82] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. [Advances in Neural Information Processing Systems](#), 36, 2024.
- [83] 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.
- [84] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua. Next-gpt: Any-to-any multimodal llm. [arXiv preprint arXiv:2309.05519](#), 2023.
- [85] 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.

- [86] 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.
- [87] Z. Xue, G. Song, Q. Guo, B. Liu, Z. Zong, Y. Liu, and P. Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. Advances in Neural Information Processing Systems, 36, 2024.
- [88] F. Yang, C. Ma, J. Zhang, J. Zhu, W. Yuan, and A. Owens. Touch and go: Learning from human-collected vision and touch. arXiv preprint arXiv:2211.12498, 2022.
- [89] L. Yu, Y. Cheng, K. Sohn, J. Lezama, H. Zhang, H. Chang, A. G. Hauptmann, M.-H. Yang, Y. Hao, I. Essa, et al. Magvit: Masked generative video transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10459–10469, 2023.
- [90] 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. arXiv preprint arXiv:2308.02490, 2023.
- [91] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9556–9567, 2024.
- [92] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11975–11986, 2023.
- [93] 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, 35: 23412–23425, 2022.
- [94] 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.
- [95] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592, 2023.
- [96] 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.