



Janus-Pro: 具有数据和模型缩放的统一的多模式理解和代

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan

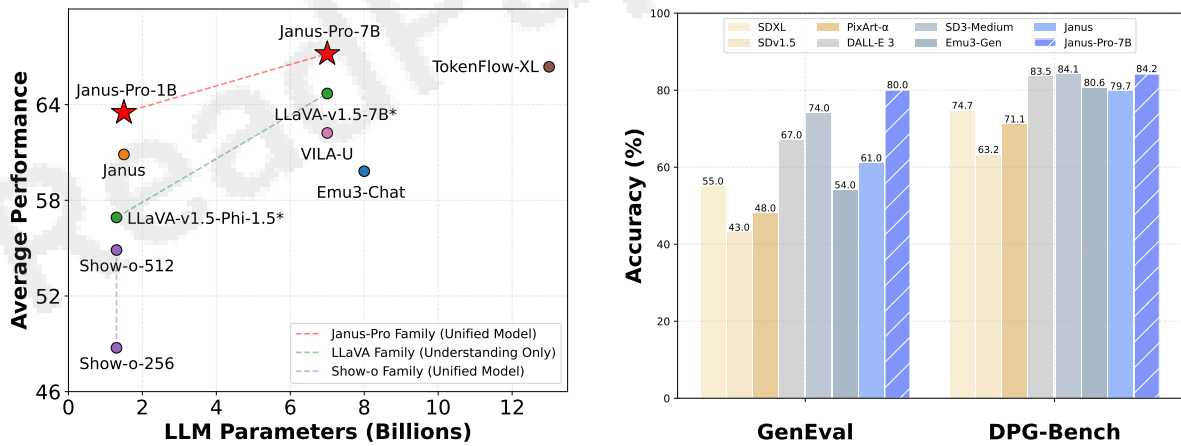
DeepSeek-AI

项目页面: <https://github.com/deepseek-ai/Janus>

Abstract

在这项工作中，我们介绍了先前作品Janus的高级版本Janus-Pro。从特定的角度来看，Janus-Pro合并了（1）优化的训练策略，（2）扩展的训练数据，以及（3）扩展到更大的模型大小。通过这些改进，Janus-Pro在多模式的理解和文本对图像遵循的CAPA命令方面都取得了重要的影响，同时也增强了文本对图像生成的稳定性。我们希望这项作品威尔斯特（Willinspire）在该领域进行进一步探索。代码和模型公开可用。

一、简介



(a) 四个多模式理解基准测试的平均性能。 (b) 以指导性基准为基准的性能的性能。

图1 | 多模式的理解和视觉产生来自我们的Janus-Pro。用于多

模态理解，我们平均教皇，MME感知，GQA和MMMU的准确性。MMME感知的分数除以20，以扩展为[0, 100]。对于视觉生成，我们评估了两个指导台式台式，Geneval和DPG基座的绩效。总体而言，Janus-Pro优于预防的最先进的统一多模式模型以及某些特定于任务的模型。最佳查看的屏幕。

on screen.



图2 | Janus-Pro与其前身Janus之间的文本到图像生成的比较。Janus-Pro为短提示提供了更稳定的输出，具有提高的视觉质量，更丰富的细节以及生成简单文本的能力。图像分辨率为 384×384 。屏幕上最佳浏览。

统一的多模式理解和生成模型的最新进展是Havedend示意的显着进步[30, 40, 45, 46, 48, 50, 54, 55]。这些方法已得到证实，以增强视觉生成任务中的指导跟踪功能，同时重新探讨模型冗余。这些方法中的大多数都利用相同的视觉编码器来处理多模式理解和生成任务。由于对这两个任务的表示形式有所不同，因此这通常会导致多模量质疑的次优性能。为了解决这个问题，Janus [46]提出了解耦的视觉编码，ThatalLeviata多模式理解与发电任务之间的冲突，在这两个任务中都能达到excellent绩效。

作为一个开创性的模型，Janus 在 1B 参数尺度上得到了验证。然而，由于训练数据量有限且模型容量相对较小，它表现出一定的缺点，例如短提示图像生成性能不佳、文本到图像生成质量不稳定等。在本文中，我们介绍了 Janus-Pro，这是 Janus 的增强版本，它融合了三个维度的改进：训练策略、数据和模型大小。Janus-Pro系列包括两种型号尺寸：1B和7B，展示了视觉编码解码方法的可扩展性。

我们在多个基准上对 Janus-Pro 进行了评估，结果揭示了其卓越的多模态理解能力，并显着提高了文本到图像的指令跟踪性能。具体来说，Janus-Pro-7B 在多模态理解基准 MMBench [29] 上取得了 79.2 的分数，超越了最先进的统一多模态模型，如 Janus [46] (69.4)、TokenFlow [34] (68.9) 和 MetaMorph [42] (75.2)。此外，在文本到图像指令跟踪排行榜 GenEval [14] 中，Janus-Pro-7B 得分为 0.80，优于 Janus [46] (0.61)、DALL-E 3 (0.67) 和 Stable Diffusion 3 Medium [11] (0.74)。

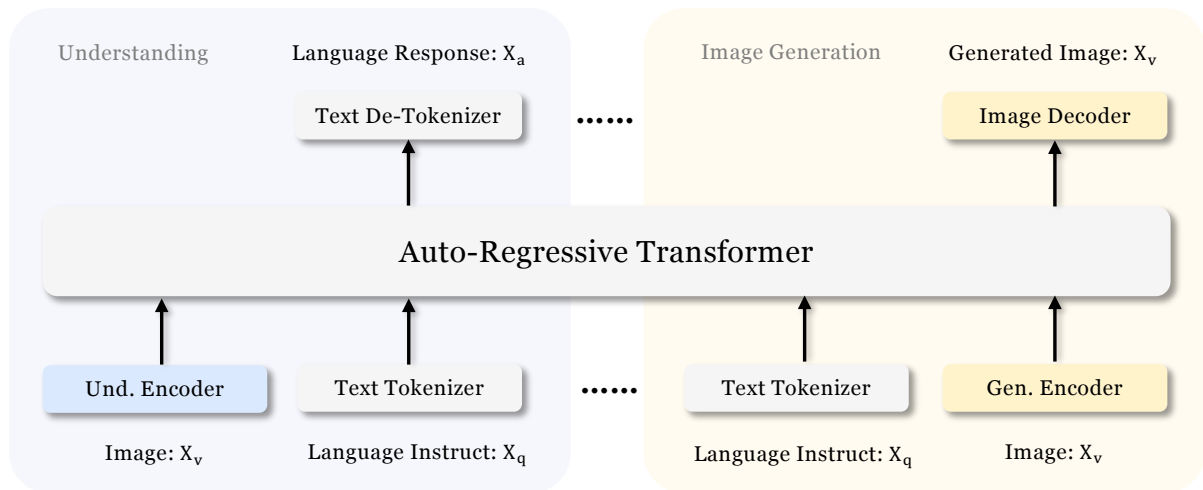


图3 |我们的Janus-Pro的建筑。我们将视觉编码解码用于多模式的不良和视觉产生。“und.编码器”和“Gen.编码器”分别是“理解编码器”和“生成编码器”的缩写。最好在屏幕上查看。

2. 方法

2.1. 建筑学

Janus-Pro的架构如图3所示，与Janus [46]相同。整体体系结构的核心原理是将视觉编码解码，以进行多模式验证和发电。我们应用独立的编码方法将Rawinputs转换为功能，然后由统一自回归变压器处理。表型的理解，我们使用siglip [53]编码器从图像中提取高维语义表。将这些特征从2-D网格扁平化为1-D序列，并使用弥漫的适配器将这些图像特征映射到LLM的输入空间中。前言生成任务，我们使用[38]的VQ令牌将图像转换为离散ID。在将ID序列扁平为1-D之后，我们使用一代适配器将与每个ID相对应的CodeBookembeddings映射到每个ID中LLM。然后，我们会产生特征序列，形成多模式特征序列，随后将其用于加工的LLM。除了LLM中的内置预测头外，我们还利用Arandomly初始化的预测头来进行视觉生成任务中的图像预测。Theentire模型遵守自回归框架。

2.2. 优化的培训策略

贾努斯（Janus）的先前版本采用了三阶段的培训过程。阶段我将重点放在适配器和图像头上。第二阶段处理统一的预审查，在此期间，除了理解编码器和生成编码器外，所有组件都具有其参数。第三阶段是通过在训练期间进一步解锁理解编码器的参数，在第二阶段进行了微调。该培训策略具有某些意义。在第二阶段，Janus将文本对图像功能的培训分为两个部分，以下两个部分[4]。第一部分使用ImageNet [9]数据进行训练，使用图像类别samansas提示文本到图像生成，目的是建模像素依赖性。第三部分训练正常的文本形象数据。在实施过程中，第二阶段的文本对图像培训步骤中有66.67%分配给第一部分。但是，进一步

表 1 | Janus-Pro 的架构配置。我们列出了该架构的超参数。

	Janus-Pro-1B	Janus-Pro-7B
Vocabulary size	100K	100K
Embedding size	2048	4096
Context Window	4096	4096
#Attention heads	16	32
#Layers	24	30

表 2 |用于训练 Janus-Pro 的详细超参数。数据比例是指多模态理解数据、纯文本数据、视觉生成数据的比例。

	Janus-Pro-1B			Janus-Pro-7B		
Hyperparameters	Stage 1	Stage 2	Stage 3	Stage 1	Stage 2	Stage 3
Learning rate	1.0×10^{-3}	1.0×10^{-4}	4.0×10^{-5}	1.0×10^{-3}	1.0×10^{-4}	4.0×10^{-5}
LR scheduler	Constant	Constant	Constant	Constant	Constant	Constant
Weight decay	0.0	0.0	0.0	0.0	0.0	0.0
Gradient clip	1.0	1.0	1.0	1.0	1.0	1.0
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)			AdamW ($\beta_1 = 0.9, \beta_2 = 0.95$)		
Warm-up steps	600	5000	0	600	5000	0
Training steps	20K	360K	80K	20K	360K	40K
Batch size	256	512	128	256	512	128
Data Ratio	1:0:3	2:3:5	5:1:4	1:0:3	2:3:5	5:1:4

3. 实验

3.1. 实施细节

在我们的实验中，我们利用DeepSeek-llm（1.5b和7b）[3]，最大支持的序列长度为4096作为基本语言模型。对于理解任务中使用的视觉编码器，我们选择siglip-large-patch16-384 [53]。该一代编码器具有16、384和下图像的代码手册，构成了16倍。理解适配器和生成适配器都是两层MLP。表2中提供了每个阶段的详细超参数。请注意，对于第二阶段，我们采用了早期停止策略，Haltingat 270k步骤。所有图像都调整为 384×384 像素。对于多模式理解数据，我们调整了图像的长度大小，并用背景颜色（RGB: 127,127, 127）在短侧达到384。裁剪为384。我们在训练过程中使用序列包装来提高训练效率。我们根据单个训练步骤中的指定比率混合所有数据类型。我们的Janus-Prois使用HAI-LLM [15]进行了训练和评估，这是一个在Pytorch顶部建造的轻巧有效的分布式培训框架。整个训练过程大约需要9/14天的1.5b/7b型号的16/32个节点，每个节点配备了8个NVIDIA A100（40GB）GPU。

3.2. 评估设置

多模式理解。为了评估多模式的理解能力，我们评估了我们的模型广泛认可的基于图像的视觉语言基准，其中包括GQA

表 3 | 与多模态理解基准的最先进技术进行比较。“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用 † 标记。

Type	Model	# LLM Params	POPE↑	MME-P↑	MMB↑	SEED↑	GQA↑	MMMU↑	MM-Vet↑
Und. Only	LLaVA-v1.5-Phi-1.5 [50]	1.3B	84.1	1128.0	-	-	56.5	30.7	-
	MobileVLM [6]	1.4B	84.5	1196.2	53.2	-	56.1	-	-
	MobileVLM-V2 [7]	1.4B	84.3	1302.8	57.7	-	59.3	-	-
	MobileVLM [6]	2.7B	84.9	1288.9	59.6	-	59.0	-	-
	MobileVLM-V2 [7]	2.7B	84.7	1440.5	63.2	-	61.1	-	-
	LLaVA-Phi [56]	2.7B	85.0	1335.1	59.8	-	-	-	28.9
	LLaVA [27]	7B	76.3	809.6	38.7	33.5	-	-	25.5
	LLaVA-v1.5 [26]	7B	85.9	1510.7	64.3	58.6	62.0	35.4	31.1
	InstructBLIP [8]	7B	-	-	36.0	53.4	49.2	-	26.2
	Qwen-VL-Chat [1]	7B	-	1487.5	60.6	58.2	57.5	-	-
	IDEFICS-9B [19]	8B	-	-	48.2	-	38.4	-	-
	Emu3-Chat [45]	8B	85.2	1244	58.5	68.2	60.3	31.6	37.2
	InstructBLIP [8]	13B	78.9	1212.8	-	-	49.5	-	25.6
Und. and Gen.	DreamLLM† [10]	7B	-	-	-	-	-	-	36.6
	LaVIT† [18]	7B	-	-	-	-	46.8	-	-
	MetaMorph† [42]	8B	-	-	75.2	71.8	-	-	-
	Emu† [39]	13B	-	-	-	-	-	-	-
	NExT-GPT† [47]	13B	-	-	-	-	-	-	-
	Show-o-256 [50]	1.3B	73.8	948.4	-	-	48.7	25.1	-
	Show-o-512 [50]	1.3B	80.0	1097.2	-	-	58.0	26.7	-
	D-Dit [24]	2.0B	84.0	1124.7	-	-	59.2	-	-
	Gemini-Nano-1 [41]	1.8B	-	-	-	-	-	26.3	-
	ILLUME [44]	7B	88.5	1445.3	65.1	72.9	-	38.2	37.0
	TokenFlow-XL [34]	13B	86.8	1545.9	68.9	68.7	62.7	38.7	40.7
	LWM [28]	7B	75.2	-	-	-	44.8	-	9.6
	VILA-U [48]	7B	85.8	1401.8	-	59.0	60.8	-	33.5
	Chameleon [40]	7B	-	-	-	-	-	22.4	8.3
	Janus	1.5B	87.0	1338.0	69.4	63.7	59.1	30.5	34.3
	Janus-Pro-1B	1.5B	86.2	1444.0	75.5	68.3	59.3	36.3	39.8
	Janus-Pro-7B	7B	87.4	1567.1	79.2	72.1	62.0	41.0	50.0

[17], POPE [23], MME [12], SEED [21], MMB [29], MM-Vet [51], and MMMU [52].

视觉生成。为了评估视觉产生能力，我们使用Geneval [14]和DPG基础[16]。Geneval是文本到图像生成的具有挑战性的基准，它通过提供对其组成能力的详细分析来设计视觉生成模型的全面生成能力。DPG基础（密集提示 GraphBenchmark）是一个全面的数据集，该数据集由1065个冗长，密集的提示组成，设计了文本对图像模型的复杂语义对齐功能。

3.3.与最先进技术的比较

多模式理解性能。我们将所提出的方法与表 3 中最先进的统一模型和仅理解模型进行比较。Janus-Pro 取得了总体最佳结果。这可以归因于多模态理解和生成的视觉编码的解耦，减轻了这两个任务之间的冲突。与尺寸明显更大的型号相比，Janus-Pro 仍然具有很强的竞争力。例如，Janus-Pro-7B 在除 GQA 之外的所有基准测试中均优于 TokenFlow-XL (13B)。

表 4 | 在 GenEval 基准上评估文本到图像的生成能力。“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用 † 标记。


Type	Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall↑
Gen. Only	LlamaGen [38]	0.71	0.34	0.21	0.58	0.07	0.04	0.32
	LDM [37]	0.92	0.29	0.23	0.70	0.02	0.05	0.37
	SDv1.5 [37]	0.97	0.38	0.35	0.76	0.04	0.06	0.43
	PixArt- α [4]	0.98	0.50	0.44	0.80	0.08	0.07	0.48
	SDv2.1 [37]	0.98	0.51	0.44	0.85	0.07	0.17	0.50
	DALL-E 2 [35]	0.94	0.66	0.49	0.77	0.10	0.19	0.52
	Emu3-Gen [45]	0.98	0.71	0.34	0.81	0.17	0.21	0.54
	SDXL [32]	0.98	0.74	0.39	0.85	0.15	0.23	0.55
	DALL-E 3 [2]	0.96	0.87	0.47	0.83	0.43	0.45	0.67
Und. and Gen.	SD3-Medium [11]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
	SEED-X† [13]	0.97	0.58	0.26	0.80	0.19	0.14	0.49
	Show-o [50]	0.95	0.52	0.49	0.82	0.11	0.28	0.53
	D-DiT [24]	0.97	0.80	0.54	0.76	0.32	0.50	0.65
	LWM [28]	0.93	0.41	0.46	0.79	0.09	0.15	0.47
	Transfusion [55]	-	-	-	-	-	-	0.63
	ILLUME [44]	0.99	0.86	0.45	0.71	0.39	0.28	0.61
	TokenFlow-XL [28]	0.95	0.60	0.41	0.81	0.16	0.24	0.55
	Chameleon [40]	-	-	-	-	-	-	0.39
	Janus [46]	0.97	0.68	0.30	0.84	0.46	0.42	0.61
	Janus-Pro-1B	0.98	0.82	0.51	0.89	0.65	0.56	0.73
	Janus-Pro-7B	0.99	0.89	0.59	0.90	0.79	0.66	0.80

表 5 | DPG-Bench 上的表演。此表中的方法都是除 Janus 和 Janus-Pro 之外的特定代模型。

Method	Global	Entity	Attribute	Relation	Other	Overall↑
SDv1.5 [36]	74.63	74.23	75.39	73.49	67.81	63.18
PixArt- α [4]	74.97	79.32	78.60	82.57	76.96	71.11
Lumina-Next [57]	82.82	88.65	86.44	80.53	81.82	74.63
SDXL [33]	83.27	82.43	80.91	86.76	80.41	74.65
Playground v2.5 [22]	83.06	82.59	81.20	84.08	83.50	75.47
Hunyuan-DiT [25]	84.59	80.59	88.01	74.36	86.41	78.87
PixArt- Σ [5]	86.89	82.89	88.94	86.59	87.68	80.54
Emu3-Gen [45]	85.21	86.68	86.84	90.22	83.15	80.60
DALL-E 3 [2]	90.97	89.61	88.39	90.58	89.83	83.50
SD3-Medium [11]	87.90	91.01	88.83	80.70	88.68	84.08
Janus	82.33	87.38	87.70	85.46	86.41	79.68
Janus-Pro-1B	87.58	88.63	88.17	88.98	88.30	82.63
Janus-Pro-7B	86.90	88.90	89.40	89.32	89.48	84.19

视觉生成性能。我们报告了 GenEval 和 DPG-Bench 上的视觉生成性能。如表 4 所示，我们的 Janus-Pro-7B 在 GenEval 上获得了 80% 的总体准确率，这优于所有其他统一或仅生成的方法，例如 Transfusion [55] (63%)SD3-Medium (74%) 和达尔-E 3 (67%)。这表明我们的方法具有更好的指令跟踪能力。如表 5 所示，Janus-Pro 在 DPG-Bench 上的得分为 84.19，超过了所有其他方法。这表明 Janus-Pro 擅长遵循密集指令来生成文本到图像。

3.4.定性结果我们在图 4 中展示了多模态理解的结果。Janus-Pro 在处理来自不同上下文的输入时表现出令人印象深刻的理解能力，展示了其强大的功能。我们还在图4的下半部分展示了一些文本到图像的生成结果。Janus-Pro-7B生成的图像非常逼真，尽管分辨率只有 384×384 ，但它们仍然包含很多细节。对于富有想象力和创造力的场景，Janus-Pro-7B 可以准确捕捉提示中的语义信息，生成逻辑清晰、连贯的图像。



详细描述一下这个场景。


这张图片描绘了令人惊叹的海岸景观，其中包括一系列从海洋中升起的高大、崎岖的岩层。这些岩石由沉积岩分层而成，呈现出明显的分层外观。最大的结构位于前景的显着位置，较小的结构则退到远处。海浪撞击这些地层的底部，产生白色泡沫和动态的运动感。

海岸线是沙质的，图像右侧可以看到一条狭窄的海滩。海滩的边缘是陡峭的悬崖，从海岸线急剧上升。悬崖上覆盖着一片片绿色植被，为岩石原本的泥土色调增添了一抹色彩。

头顶的天空阴沉沉的，厚厚的云层预示着即将到来的暴风雨或最近的降雨。光线被扩散，在整个场景上投射出柔和、柔和的光芒。尽管天空多云，这片海岸风景却给人一种宁静和自然之美的感觉。


Landmark Recognition

这张图片展示的是中国杭州的西湖，特别是著名的三潭印月岛。图片中的建筑是三潭印月的亭子，周围是平静的湖水，远处是连绵的山脉。西湖是中国著名的旅游景点之一，以其美丽的自然景观和丰富的历史文化而闻名。



猜猜这是哪里？

Text Recognition




What's written on the large black board?

The large black board has the following text:

"Serving Soul since Twenty Twelve"

Additionally, there are smaller sections with more detailed information, but the main heading is clearly visible.


General Knowledge




Can you introduce the background story of this cake?

This cake is themed around the classic cartoon character **Tom and Jerry**. The cake features the characters Tom (the cat) and Jerry (the mouse) in a playful setting, with a backdrop that includes the words "Tom & Jerry". The cake itself is decorated with fondant shaped like cheese, which is a recurring theme in the cartoon series. The design also includes a small fondant house with a door, adding to the whimsical and playful theme of the cake.


Text-To-Image Generation




A golden retriever lying peacefully on a wooden porch, with autumn leaves scattered around.




A young woman with freckles wearing a straw hat, standing in a golden wheat field.




A single drop of water clinging to a green leaf, with sunlight creating a faint rainbow prism.




An ancient stone bridge arching over a crystal-clear mountain stream, surrounded by lush greenery.




A glowing crystal ball floating above a sandstone table in the middle of a desert at sunset.



A tiny galaxy contained inside a glass bottle, glowing brightly against a dark velvet cloth.



A giant whale flying through a city skyline, surrounded by floating glowing lanterns.



Astronaut in a jungle, cold color palette, muted colors, detailed, 8k

图4 |多模式理解和视觉生成能力的定性结果。模型是Janus-Pro-7b，视觉生成的图像输出分辨率为 384×384 .最best在屏幕上。88

4. 结论

本文从训练策略、数据、模型大小三个方面介绍了Janus的改进。这些增强功能在多模式理解和文本到图像指令跟踪能力方面取得了显著进步。然而，Janus-Pro仍然有一定的局限性。在多模态理解方面，输入分辨率限制为 384×384 ，这影响了其在OCR等细粒度任务中的性能。对于文本到图像的生成，低分辨率加上视觉分词器引入的重建损失，导致图像虽然语义内容丰富，但仍然缺乏精细细节。例如，占据有限图像空间的小面部区域可能会出现-详细。提高图像分辨率可以缓解这些问题。

References

- [1] 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.
- [2] 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. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] R. Dong, C. Han, Y. Peng, Z. Qi, Z. Ge, J. Yang, L. Zhao, J. Sun, H. Zhou, H. Wei, et al. Dream-llm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499, 2023.

- [11] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, Y. Marek, and R. Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL <https://arxiv.org/abs/2403.03206>.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] High-flyer. Hai-llm: Efficient and lightweight training tool for large models, 2023. URL <https://www.high-flyer.cn/en/blog/hai-llm>.
- [16] 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.
- [17] D.A.哈德森和C.D.曼宁。Gqa: 用于现实世界视觉推理和组合问答的新数据集。IEEE/CVF 计算机视觉和模式识别会议记录, 第 6700-6709 页, 2019 年。—
- [18] 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.
- [19] 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 最先进的视觉语言模型, 2023。URL <https://huggingface.co/blog/idefics>。
- [20] H. Laurençon, A. Marafioti, V. Sanh, and L. Tronchon. Building and better understanding vision-language models: insights and future directions., 2024.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] Z. Li, H. Li, Y. Shi, A. B. Farimani, Y. Kluger, L. Yang, and P. Wang. Dual diffusion for unified image generation and understanding. arXiv preprint arXiv:2501.00289, 2024.
- [25] 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.

- [26] 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.
 - [27] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024.
 - [28] 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.
 - [29] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mm-bench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281, 2023.
 - [30] Y. Ma, X. Liu, X. Chen, W. Liu, C. Wu, Z. Wu, Z. Pan, Z. Xie, H. Zhang, X. Yu, L. Zhao, Y. Wang, J. Liu, and C. Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2024.
 - [31] mehdidc. Yfcc-huggingface. <https://huggingface.co/datasets/mehdidc/yfcc15m>, 2024.
 - [32] 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.
 - [33] 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. 2024.
 - [34] L. Qu, H. Zhang, Y. Liu, X. Wang, Y. Jiang, Y. Gao, H. Ye, D. K. Du, Z. Yuan, and X. Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. arXiv preprint arXiv:2412.03069, 2024.
 - [35] 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.
 - [36] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. 2022.
 - [37] 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.
 - [38] 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.
 - [39] 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.
 - [40] C.团队。变色龙：混合模式的早期融合基础模型。 Arxiv Preprintarxiv: 2405.09818, 2024。
-
- [41] 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.

- [42] S. Tong, D. Fan, J. Zhu, Y. Xiong, X. Chen, K. Sinha, M. Rabbat, Y. LeCun, S. Xie, and Z. Liu. Metamorph: Multimodal understanding and generation via instruction tuning. arXiv preprint arXiv:2412.14164, 2024.
- [43] Vivym. Midjourney prompts dataset. <https://huggingface.co/datasets/vivym/midjourney-prompts>, 2023. Accessed: [Insert Date of Access, e.g., 2023-10-15].
- [44] C. Wang, G. Lu, J. Yang, R. Huang, J. Han, L. Hou, W. Zhang, and H. Xu. Illume: Illuminating your llms to see, draw, and self-enhance. arXiv preprint arXiv:2412.06673, 2024.
- [45] 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.
- [46] 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.
- [47] 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.
- [48] 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.
- [49] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai, H. Gao, Y. Ma, C. Wu, B. Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302, 2024.
- [50] 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.
- [51] 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.
- [52] 多模式的多模式理解和推理基准的专家AGI。在IEEE/CVF计算机愿景和盖特恩认可会议论文集, 第9556–9567页, 2024年。
-
- [53] 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.
- [54] 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.
- [55] 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.
- [56] 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.

- [57] 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.