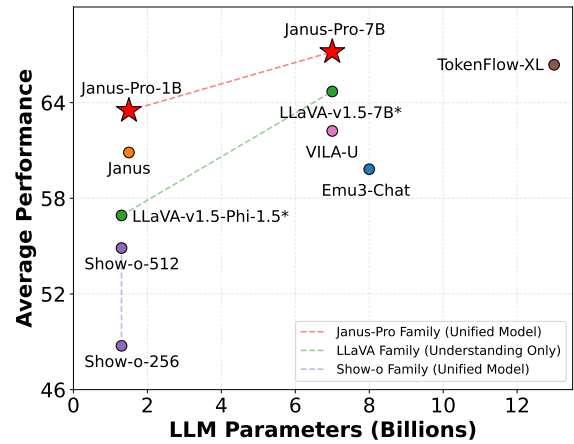


Janus-Pro

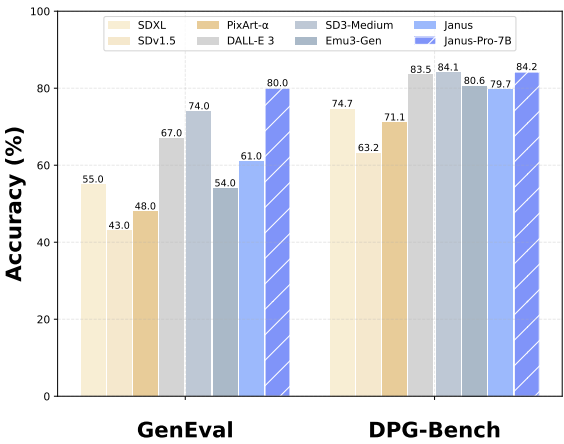
Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan  
DeepSeek-AI  
<https://github.com/deepseek-ai/Janus>

Abstract

Janus-Pro 1 Janus 2 Janus-Pro 3 CAPA  
Willinspire



a



b

1 | Janus-Pro  
MME GQA MMMU MMME  
[0 100] Janus-Pro  
Geneval DPG 20

on screen.

Janus	Janus-Pro-7B	Janus	Janus-Pro-7B	Janus	Janus-Pro-7B
-------	--------------	-------	--------------	-------	--------------

The face of a beautiful girl.	A steaming cup of coffee on a wooden table.	A glass of red wine on a reflective surface.
-------------------------------	---	--

" Hello"

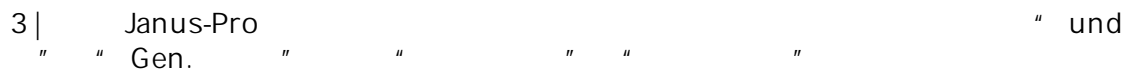
2   Janus-Pro	Janus	Janus-Pro
384x 384		

	Havedend	[30 40 45 46
48 50 54 55]		

Janus [46]	ThatalLeviate
excellent	

	Janus 1B	
	Janus-Pro Janus-Pro	Janus 1B
7B		

	Janus-Pro	
	Janus-Pro- B	
MMBench [29] 79.2		Janus [46]
(69.4) TokenFlow [34] (68.9)	MetaMorph [42] (75.2)	
GenEval [14] Janus-Pro- B 0.80	Janus [46] (0.61)	DALL-E
3 (0.67) Stable Di usion 3 Medium [11] 0.74		



2.1.

Rawinputs  
siglip [53]

Figure 1 illustrates the comparison of VQ and LLM-based methods. The diagram is organized into two main columns: VQ (left) and LLM (right). The VQ column shows a sequence of components: ID, LLM, and LLM. The LLM column shows a sequence of components: ID, LLM, and LLM. The LLM column further branches into 2-D and 1-D, with 2-D leading to LLM and 1-D leading to ID. The LLM column also includes CodeBook embeddings and embeddings. The LLM column includes The entire and Randomly components.

Janus

[4]  
sam an sas

66. 67

• I I I Imagenet  
 LLM  
 II Imagenet  
 to-to-to-to-to-to-to-to-to-to-to-to-to-to-to-to-to-to-to-  
 II

7 3 3 10 5 1 4

iMpred

2.3

Janus

•  
 [49] 9000  
 [31]  
 [20]

Meme Gealpent

•

Janus-Pro

7200

[43]

themodel

2.4

Janus

Janus-Pro

1.5B LLM

7B 1  
 LLM

1.5B 7B LLM

1 | Janus-Pro

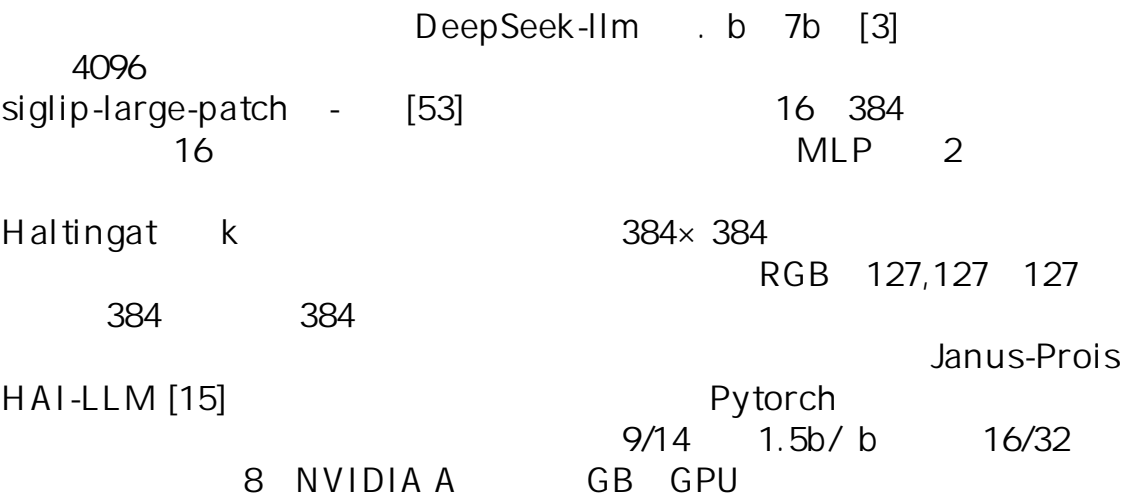
	Janus-Pro-1B	Janus-Pro-7B
Vocabulary size	100K	100K
Embedding size	2048	4096
Context Window	4096	4096
#Attention heads	16	32
#Layers	24	30

2 | Janus-Pro

	Janus-Pro-1B						Janus-Pro-7B					
Hyperparameters	Stage 1		Stage 2		Stage 3		Stage 1		Stage 2		Stage 3	
Learning rate	1.0	10 <sup>-3</sup>	1.0	10 <sup>-4</sup>	4.0	10 <sup>-5</sup>	1.0	10 <sup>-3</sup>	1.0	10 <sup>-4</sup>	4.0	10 <sup>-5</sup>
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 ( $v_1 = 0.9, v_2 = 0.95$ )						AdamW ( $v_1 = 0.9, v_2 = 0.95$ )					
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.

3.1.



3.2.

GQA

3 | " Und." " Gen."  
" " " " †

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 <sup>y</sup> [10]	7B	-	-	-	-	-	-	36.6
	LaVIT <sup>y</sup> [18]	7B	-	-	-	-	46.8	-	-
	MetaMorph <sup>y</sup> [42]	8B	-	-	75.2	71.8	-	-	-
	Emu <sup>y</sup> [39]	13B	-	-	-	-	-	-	-
	NExT-GPT <sup>y</sup> [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].

Geneval [14] DPG [16]  
Geneval  
DPG  
GraphBenchmark 1065

3.3.

3  
Janus-Pro  
Janus-Pro  
TokenFlow-XL ( B)  
Janus-Pro- B GQA

4 | GenEval

" " " "

" Und." " Gen."

†

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-U [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
	SD3-Medium [11]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
Und. and Gen.	SEED-X <sup>y</sup> [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

5 | DPG-Bench

Janus

Janus-Pro

Method	Global	Entity	Attribute	Relation	Other	Overall"
SDv1.5 [36]	74.63	74.23	75.39	73.49	67.81	63.18
PixArt-U [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- [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

GenEval

DPG-Bench

4

Janus-Pro- B

GenEval

80%

Transfusion [55] ( %)

SD -Medium (74%)

5

-E 3 (67%)

Janus-Pro

DPG-Bench

84.19

Janus-Pro

3. 4.

4

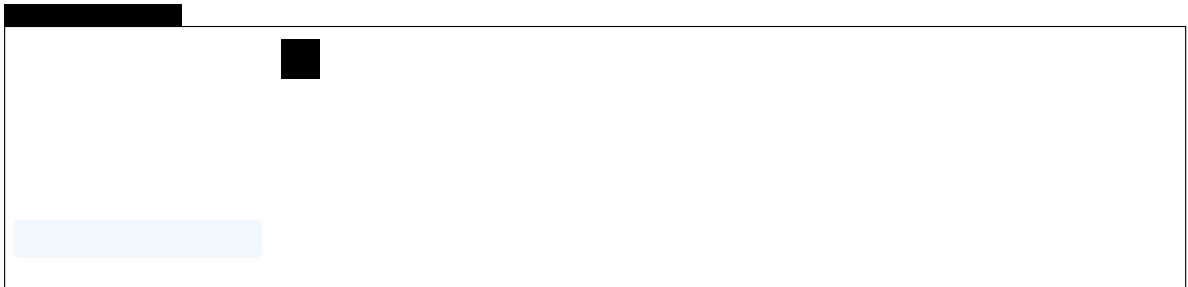
Janus-Pro

4

Janus-Pro- B

384× 384

Janus-Pro- B



A)-6( ).<4B"2&\*-/S/&-

! " # \$ % & ' ( ) \* + , ' - . / 0 1 ( 2 3 '  
4 5 6 7 8 9 # \$ ) ' : ; ( 4 5 6 7 ' < = / >  
? ( @ A ' . B / C D ( E F ' G H 9 - . ( ) \*  
2 3 ' I J K L M N / O P Q R ' S T K U V W X  
' Y Z [ \ ] ^ 3 9

! ! " # \$ % !

! "#\$4B"2&\*-/S/&-

! "#\$%&'()\$\$+, '-, '\$'+'.#)/+°.#12°-#)34

! "#\$%&'()#\$%&'()\*+,-, &'-\$" &.\$/"#\$  
0, % 1 2 3 (\$/#4/5

67#°823(\$7, 9%. 23\*#! 1#3/:! 1#9B#6

; --2/2, 3&% 4/"#°\$&'#\$\$. =&%#°\$  
. #\*/2, 3. \$12/"\$=, '\$\$-#/82°#-\$  
230, '= &/2, 3<4) 9/\$/"#\$= &23°"#\$-23 (\$  
2.\$\*%#&'&%\$82. 2)##>

, "-.)1C-&>1"6\*\*

! "2.\$\*&+#2.\$/"#=#-\$&', 93-\$/"#\$°%&. 2\*\$\*&'/, , 3\$  
\*" &'&\*/#°\$!, =&3-?#°':>! "°\$\*&+#0#&/9°#.\$/"#\$  
\*" &'&\*/#°. \$!, =&°/"#\$\*&/A&3-?#°':&°/"#\$=, 9. #A&3\$&\$  
B%&: 09%. #/123 (<12/"\$&\$) &\*+-' , B\$/"&/23\*°9-#.\$/"#\$  
1, '-. \$6!, =&C\$°#°': D> ! "°\$\*&+#2/. #°02.\$-#\*, '&/#-\$  
12/"°0, 3-&3/\$. " &B#-°2+°\$\*"##. #<1 "2\*"2.\$&\$  
'#\*9'"23 (\$/"#=#23\$/"#\$\*&'/, , 3\$. #°2#.> \$! "°\$-# . 2(3\$  
&%, 23\*°9-#.\$&\$ . =&%00, 3-&3/\$', 9. #°12/"\$&\$-, , '<  
&--23 (\$/, \$/"#\$1 "2=. 2\*°%&3-\$B%&: 09%/"#=#\$, 0\$/"#\$  
\*&+>

5#, '6-7", \$)-371+'+'  
0#12/)-7, 3°&\$-)6'-8\$""&'1#2+4

! "#\$%&'()\*\*+, "-.)\$&-

; \$(, %-#3\$ #/'2#8°\$%23 (\$B#&\*#09%\$ ; \$, 93 (\$1, =&3\$12/"°0'#+%#-\$  
, 3\$&\$1, , -#3\$B, '""<12/"\$&9/9=3\$ 1#&'23 (\$&\$. /' &1\$' &/&\$. /&3-23 (\$  
%#&8#. \$. \*&//#°#-\$&', 93-> 23\$&\$ (\$, %-#3\$1 "°#&/02#%>

; \$. 23 (%#°-\$, B\$, 0\$1 &/#°\$\*°23 (23 (\$  
, \$&\$ ('##3\$°#&0<12/"\$. 93°2 (" /\$  
\*\*#&/23 (\$&0&23/\$°&23), 1\$B'2.

; 3\$&3\*2#3/\$. /, 3#°\$)'2- (\$\$  
&\*""23 (\$, 8#°\$&\$\*': . /&°\$\*%#&'\$  
=, 93/&23\$. /'°&=<4. 9'', 93-#-\$  
) : \$°9. ""\$ ('##3#':>

; \$(% 123 (\$\*': . /&% )&%00% &/23 (\$  
&), 8#°\$&\$. &3-. /, 3#°\$/\$&)%#°23\$/"#\$  
=2--%#\$, 0\$&\$-#.#' /\$&/\$. 93. #/>

; \$/23: \$(&%&4: \$\*, 3/&23#-°23. 2-#°\$  
&\$(\$&. . \$), //#<\$(% 123 (\$)'2 (" /% \$  
& (&23. /\$&\$-&'&+°B#°B#/\$°% /'">

; \$(2&3/\$1 "°&%#0% 23 (\$""', 9 ("°\$&\$  
\*2/: \$. +: °23#<4. 9'', 93-#-\$): \$  
0% &/23 (\$(\$% 123 (\$°&3/#'3.>

; . /', 3&9/°23\$&\$°93 (%#<4°, %-\$  
\*, % ' B&%#//#<4=9/#-\$\*, %' . <  
-#/&2°#-<4G+

Janus-Pro

384× 384

OCR

-

## 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-0: 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. C.D. Gqa IEEE/CVF 6700-6709 2019 —
- [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 idfics: An open reproduction of 2023 URL <https://huggingface.co/blog/idfics>
- [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. Arxiv Preprint arxiv 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] AGI      IEEE/CVF  
9556–9567      2024

- 
- [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.