

Janus-Pro : Unified Multimodal Understanding and Generation with **Data and Model Scaling**

Introduction

- Separate multimodal understanding and generation tasks
 - Since the representations requires for these two tasks differ, Janus proposes decoupling visual encoding.

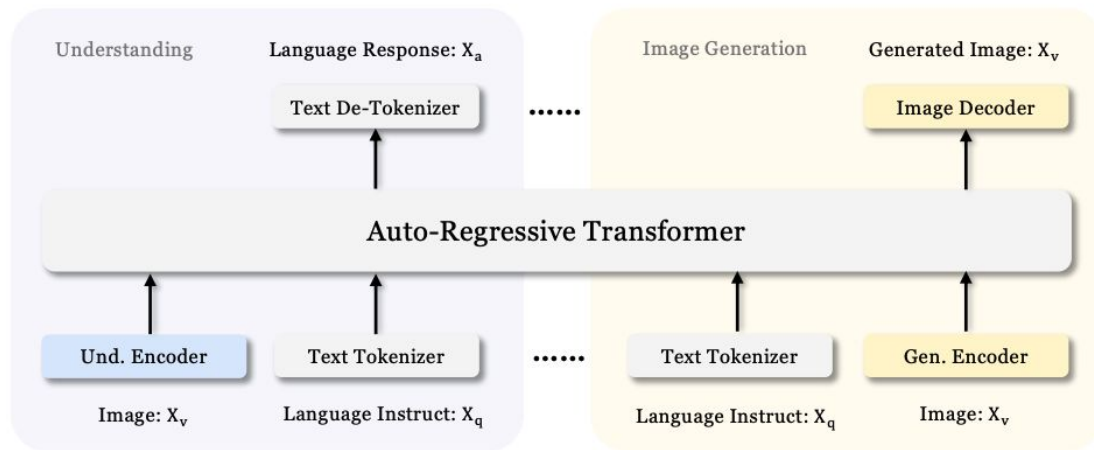
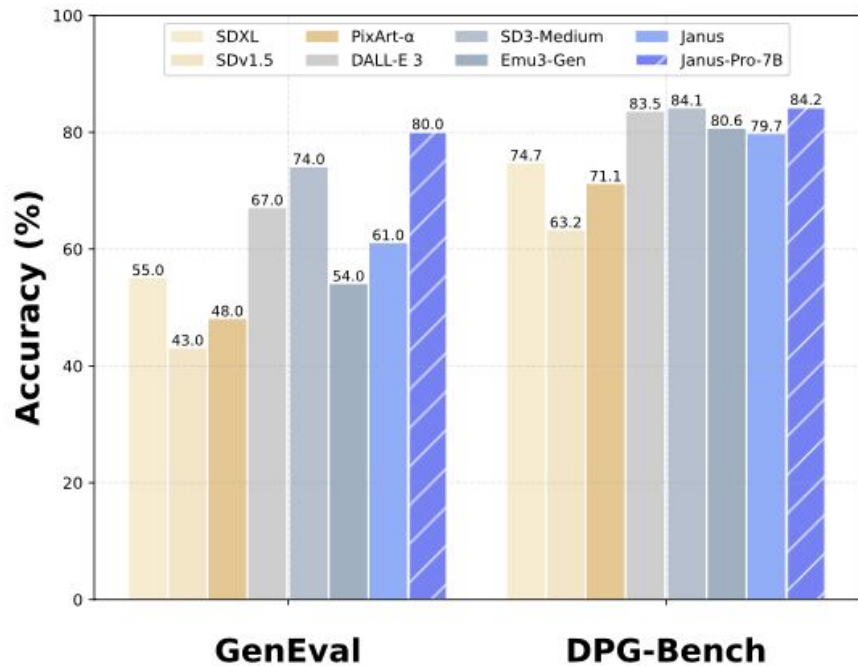
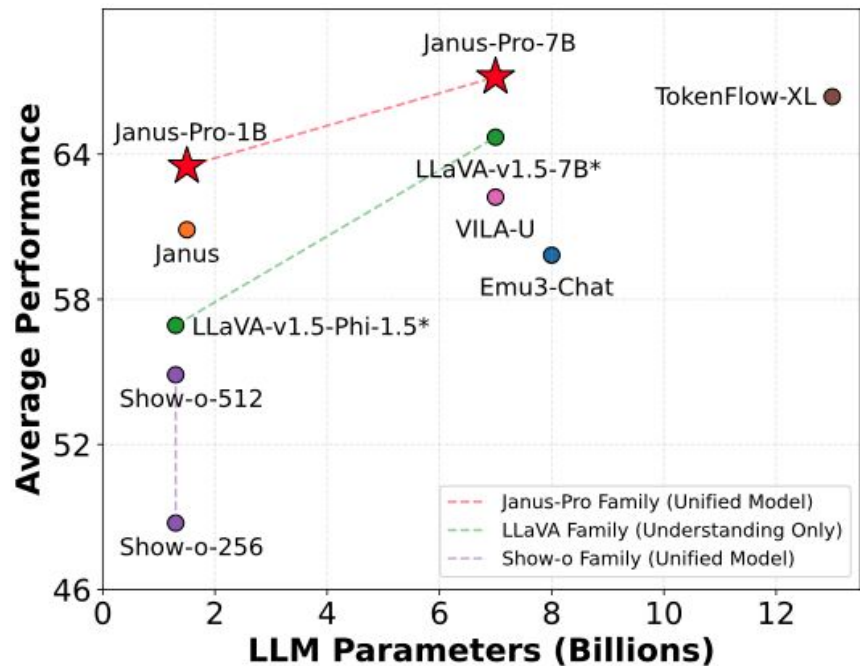


Figure 3 | **Architecture of our Janus-Pro.** We decouple visual encoding for multimodal understanding and visual generation. “Und. Encoder” and “Gen. Encoder” are abbreviations for “Understanding Encoder” and “Generation Encoder”, respectively. Best viewed on screen.

Introduction

- Model Scaling
 - As a pioneering model, Janus is validated at the 1B parameter scale.
 - Due to the limited amount of training data and small model capacity, it shows certain shortcomings, such as suboptimal performance and unstable generation.



(a) Average performance on four multimodal understanding benchmarks.

(b) Performance on instruction-following benchmarks for text-to-image generation.

Method

- Architecture
 - Multimodal Understanding - SigLIP(Vision Encoder)
 - Features are flattened from a 2-D grid into a 1-D sequence
 - an understanding adaptor is used to map these image features into the input space of the LLM
 - Visual Generation tasks - VQ Tokenizer
 - After the ID Sequences is flattened into 1-D, we use a generation adaptor to map the codebook embeddings corresponding to each ID into the input space of the LLM
 - Then concatenate these feature sequences to form a multimodal feature sequence, which is fed into the LLM for processing

Method

- Optimized Training Strategy
 - Three-stage training process
 - Stage 1
 - focuses on training the adaptors and the image head.
 - trains on ImageNet data, text-to-image generation (text - category names)
 - $\frac{2}{3}$ training steps of Stage 2
 - Longer Training in Stage 1
 - We increase the training steps in Stage I, allowing sufficient training on the ImageNet dataset. Our findings reveals that even with the LLM parameters fixed, the model could effectively model pixel dependence and generate reasonable images based on category names.

Method

- Optimized Training Strategy
 - Three-stage training process
 - Stage 2
 - handles unified pretraining, during which all components(except the understanding generation encoder), has their parameters update
 - Focused Training in Stage 2
 - Drop ImageNet data, utilize normal text-to-image data to generate images
 - enables stage 2 to utilize the text-to-image data more efficiently, resulting in improved training efficiency and overall performance

Method

- Optimized Training Strategy
 - Three-stage training process
 - Stage 3
 - supervised fine-tuning
 - unlocking the understanding encoder (not generation encoder)
 - data portion
 - changing multimodal data, pure text data, text-to-image data from 7:3:10 to 5:1:4
 - By slightly reducing the proportion of text-to-image data, we observe that this adjustment allows us to maintain strong visual generation capabilities while achieving improved multimodal understanding performance.

Method

- Data Scaling
 - Multimodal Understanding
 - For stage 2
 - DeepSeek-VL2 + 90M samples
 - 90M samples
 - image caption datasets(e.g.YFCC)
 - table char and document understanding(e.g. Doc-matrix)
 - For stage 3
 - stage 2 + MEME + Chinese conversational data + ...
 - Visual Generation
 - 72 million samples = Janus data + synthetic data
 - <https://huggingface.co/datasets/vivym/midjourney-prompts>

Method

- Model Scaling

	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

Experiments

- Implementation details

- understanding vision encoder
 - SigLIP-Large-Patch16-384
- generation vision encoder
 - codebook size 16384
- understanding and generation adaptor
 - two-layer MLPs
- note that for Stage II, we employ an early stopping strategy, halting at 270K steps
- All images are resized to 384x384
- For multimodal understanding data, we resize the long side of the image and pad the short side with the background color (RGB: 127, 127, 127) to reach 384
- For visual generation data, the short side is resized to 384, and the long side is cropped to 384.
- The whole training process took about 9/14 days on a cluster of 16/32 nodes for 1.5B/7B model, each equipped with 8 Nvidia A100 (40GB) GPUs

	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

Experiments

- Evaluation

Table 3 | **Comparison with state-of-the-arts on multimodal understanding benchmarks.** “Und.” and “Gen.” denote “understanding” and “generation”, respectively. Models using external pretrained diffusion model are marked with [†].

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

Experiments

- Evaluation

Table 4 | **Evaluation of text-to-image generation ability on GenEval benchmark.** “Und.” and “Gen.” denote “understanding” and “generation”, respectively. Models using external pretrained diffusion model are marked with [†].

Type	Method	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall [†]
<i>Gen. Only</i>	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
<i>Und. and Gen.</i>	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

Experiments

- Evaluation

Table 5 | **Performances on DPG-Bench.** The methods in this table are all generation-specific models except Janus and 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- 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

Exper

● Eve

Image Description



Describe this scene in detail.



This image depicts a **stunning coastal landscape** featuring a series of **tall, rugged rock** formations emerging from the ocean. The rocks are layered with sedimentary rock, giving them a distinct stratified appearance. The largest formation is prominently positioned in the foreground, with smaller ones receding into the distance. The ocean waves crash against the base of these formations, creating **white foam and a dynamic sense of movement**.

The **coastline is sandy**, with a narrow strip of beach visible on the right side of the image. The beach is bordered by steep cliffs, which rise dramatically from the shoreline. The cliffs are covered in patches of green vegetation, adding a touch of color to the otherwise earthy tones of the rocks.

The sky above is **overcast**, with thick clouds that suggest an impending storm or recent rainfall. The light is diffused, casting a soft, muted glow over the entire scene. Despite the cloudy sky, there is a sense of tranquility and natural beauty in this coastal landscape.

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？

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 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.

Ex



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.



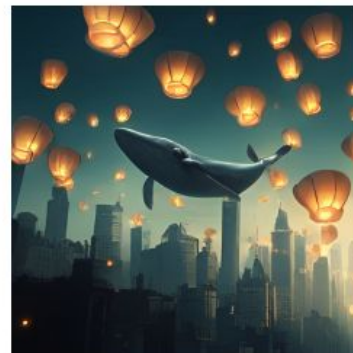
An ancient stone bridge arching over a crystal-clear mountain stream, surrounded by lush greenery.



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