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

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

DeepSeek-AI

Project Page: https://github.com/deepseek-ai/Janus

Abstract

In this work, we introduce **Janus-Pro**, an advanced version of the previous work Janus. Specifically, Janus-Pro incorporates (1) an optimized training strategy, (2) expanded training data, and (3) scaling to larger model size. With these improvements, Janus-Pro achieves significant advancements in both multimodal understanding and text-to-image instruction-following capabilities, while also enhancing the stability of text-to-image generation. We hope this work will inspire further exploration in the field. Code and models are publicly available.

1. Introduction
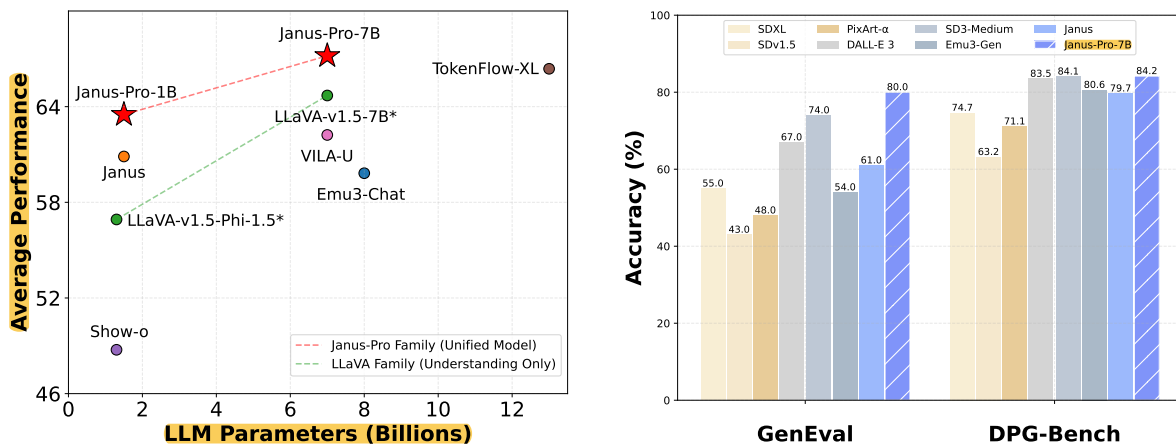
(a) Average performance on four multimodal understanding benchmarks.

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

Figure 1 | **Multimodal understanding and visual generation results from our Janus-Pro**. For multimodal understand, we average the accuracy of POPE, MME-Perception, GQA, and MMMU. The scores of MME-Perception are divided by 20 to scale to [0, 100]. For visual generation, we evaluate the performance on two instruction-following benchamrks, GenEval and DPG-Bench. Overall, Janus-Pro outperforms the previous state-of-the-art unified multimodal models as well as some task-specific models. Best viewed 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.

A minimalist photo of an orange tangerine with a green stem and leaves, symbolizing prosperity, sitting on a red silk cloth during Chinese New Year.　　A clear image of a blackboard with a clean, dark green surface and the word 'Hello' written precisely and legibly in the center with bold, white chalk letters.　　Capture a close-up shot of a vibrant sunflower in full bloom, with a honeybee perched on its petals, its delicate wings catching the sunlight.

Figure 2 | **Comparison of text-to-image generation between Janus-Pro and its predecessor, Janus.** Janus-Pro delivers more stable outputs for short prompts, with improved visual quality, richer details, and the ability to generate simple text. The image resolution is 384 × 384. Best viewed on screen.

Recent advancements in unified multimodal understanding and generation models have demonstrated significant progress [30, 40, 45, 46, 48, 50, 54, 55]. These approaches have been proven to enhance the instruction-following capabilities in visual generation tasks while reducing model redundancy. Most of these methods utilize the same visual encoder to process inputs for both multimodal understanding and generation tasks. Since the representations required for these two tasks differ, this often results in suboptimal performance in multimodal understanding. To address this issue, Janus [46] proposes decoupling visual encoding, which alleviates the conflict between multimodal understanding and generation tasks, achieving excellent performance in both tasks.

As a pioneering model, Janus is validated at the 1B parameter scale. However, due to the limited amount of training data and the relatively small model capacity, it exhibits certain shortcomings, such as suboptimal performance on short prompts image generation and unstable text-to-image generation quality. In this paper, we introduce Janus-Pro, an enhanced version of Janus that incorporates improvements across three dimensions: training strategies, data, and model size. The Janus-Pro series includes two model sizes: 1B and 7B, demonstrating scalability of the visual encoding decoding method.

We evaluate Janus-Pro on multiple benchmarks, and the results reveal its superior multimodal understanding capabilities and significantly improved text-to-image instruction-following performance. Specifically, Janus-Pro-7B achieved a score of 79.2 on the multimodal understanding benchmark MMBench [29], surpassing state-of-the-art unified multimodal models such as Janus [46] (69.4), TokenFlow [34] (68.9) and MetaMorph [42] (75.2). Additionally, in the text-to-image instruction-following leaderboard GenEval [14], Janus-Pro-7B scores 0.80, outperforming Janus [46] (0.61), DALL-E 3 (0.67), and Stable Diffusion 3 Medium [11] (0.74).
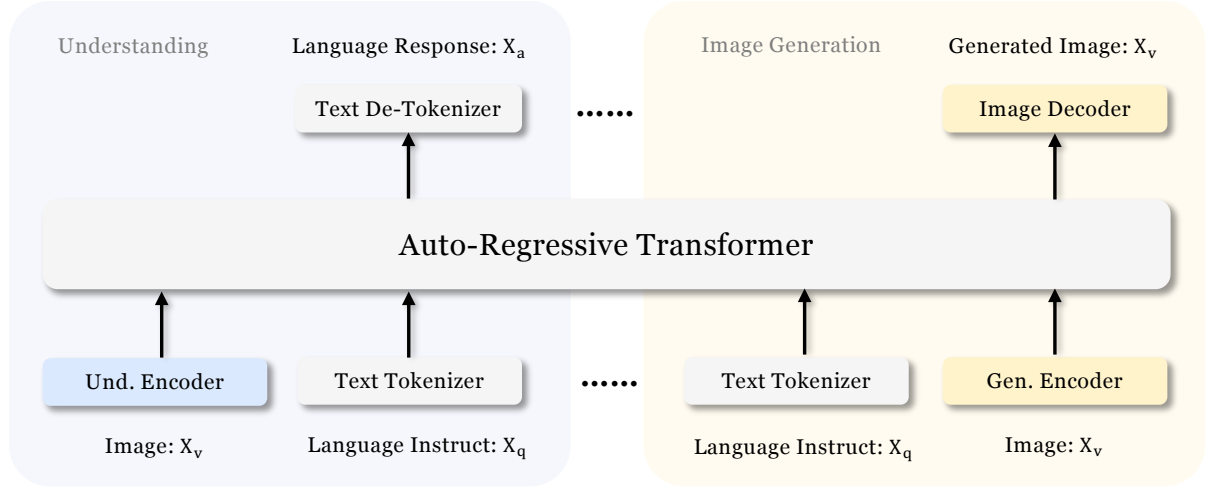
2

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.

2. Method

2.1. Architecture

The architecture of Janus-Pro is shown in Figure 3, which is the same as Janus [46]. The core design principle of the overall architecture is to decouple visual encoding for multimodal understanding and generation. We apply independent encoding methods to convert the raw inputs into features, which are then processed by an unified autoregressive transformer. For multimodal understanding, we use the SigLIP [53] encoder to extract high-dimensional semantic features from images. These features are flattened from a 2-D grid into a 1-D sequence, and an understanding adaptor is used to map these image features into the input space of the LLM. For visual generation tasks, we use the VQ tokenizer from [38] to convert images into discrete IDs. After the ID sequence 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. We then concatenate these feature sequences to form a multimodal feature sequence, which is subsequently fed into the LLM for processing. Apart from the built-in prediction head in the LLM, we also utilize a randomly initialized prediction head for image predictions in the visual generation task. The entire model adheres to an autoregressive framework.

2.2. Optimized Training Strategy

The previous version of Janus employs a three-stage training process. Stage I focuses on training the adaptors and the image head. Stage II handles unified pretraining, during which all components except the understanding encoder and the generation encoder has their parameters updated. Stage III is supervised fine-tuning, building upon Stage II by further unlocking the parameters of the understanding encoder during training. This training strategy has certain issues. In Stage II, Janus divides the training for text-to-image capabilities into two parts following PixArt [4]. The first part trains on ImageNet [9] data, using image category names as prompts for text-to-image generation, with the goal of modeling pixel dependence. The second part trains on normal text-to-image data. During implementation, 66.67% of the text-to-image training steps in Stage II are allocated to the first part. However, through further

experimentation, we find that this strategy is suboptimal and lead to significant computational inefficiency.

To address this issue, we make two modifications.

- **Longer Training in Stage I**: 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.
- **Focused Training in Stage II**: In Stage II, we drop ImageNet data and directly utilize normal text-to-image data to train the model to generate images based on dense descriptions. This redesigned approach enables Stage II to utilize the text-to-image data more efficiently, resulting in improved training efficiency and overall performance.

We also adjust the data ratio in Stage III supervised fine-tuning process across different types of datasets, changing the proportion of multimodal data, pure text data, and 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.

2.3. Data Scaling

We scale up the training data used for Janus in both multimodal understanding and visual generation aspects.

- **Multimodal Understanding**. For the Stage II pretraining data, we refer to DeepSeek-VL2 [49] and add approximately 90 million samples. These include image caption datasets (e.g., YFCC [31]), as well as data for table, chart, and document understanding (e.g., Docmatix [20]). For the Stage III supervised fine-tuning data, we also incorporate additional datasets from DeepSeek-VL2, such as MEME understanding, Chinese conversational data, and datasets aimed at enhancing dialogue experiences. These additions significantly expanded the model's capabilities, enriching its ability to handle diverse tasks while improving the overall conversational experience.
- **Visual Generation**. We observe that the real-world data used in the previous version of Janus lacks quality and contains significant noise, which often leads to instability in text-to-image generation, resulting in aesthetically poor outputs. In Janus-Pro, we incorporate approximately 72 million samples of synthetic aesthetic data, bringing the ratio of real to synthetic data to 1:1 during the unified pretraining stage. The prompts for these synthetic data samples are publicly available, such as those in [43]. Experiments demonstrat that the model converges faster when trained on synthetic data, and the resulting text-to-image outputs are not only more stable but also exhibit significantly improved aesthetic quality.

2.4. Model Scaling

The previous version of Janus validates the effectiveness of visual encoding decoupling using a 1.5B LLM. In Janus-Pro, we scaled the model up to 7B, with the hyperparameters of both the 1.5B and 7B LLMs detailed in Table 1. We observe that when utilizing a larger-scale LLM, the convergence speed of losses for both multimodal understanding and visual generation improved significantly compared to the smaller model. This finding further validates the strong scalability of this approach.

Still need larger models

Table 1 | **Architectural configuration for Janus-Pro**. We list the hyperparameters of the architecture.

	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

Table 2 | **Detailed hyperparameters for training Janus-Pro**. Data ratio refers to the ratio of multimodal understanding data, pure text data, and visual generation data.

	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. Experiments

3.1. Implementation Details

In our experiments, we utilize DeepSeek-LLM (1.5B and 7B) [3] with a maximum supported sequence length of 4096 as the base language model. For the vision encoder used in understanding tasks, we select SigLIP-Large-Patch16-384 [53]. The generation encoder has a codebook of size 16,384 and downsamples images by a factor of 16. Both the understanding adaptor and the generation adaptor are two-layer MLPs. The detailed hyperparameters for each stage are provided in Table 2. All images are resized to 384×384 pixels. 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. We use sequence packing during training to improve training efficiency. We mix all data types according to the specified ratios in a single training step. Our Janus is trained and evaluated using HAI-LLM [15], which is a lightweight and efficient distributed training framework built on top of PyTorch. The whole training process took about 7/14 days on a cluster of 16/32 nodes for 1.5B/7B model, each equipped with 8 Nvidia A100 (40GB) GPUs.

3.2. Evaluation Setup

Multimodal Understanding. To assess multimodal understanding capabilities, we evaluate our model on widely recognized image-based vision-language benchmarks, which include GQA

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 [50]	1.3B	73.8	948.4	-	-	48.7	25.1	-
	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].

Visual Generation. For evaluating visual generation capabilities, we use GenEval [14] and DPG-Bench [16]. GenEval is a challenging benchmark for image-to-text generation, designed to reflect the comprehensive generative abilities of visual generation models by offering a detailed instance-level analysis of their compositional capabilities. DPG-Bench (Dense Prompt Graph Benchmark) is a comprehensive dataset consisting of 1065 lengthy, dense prompts, designed to assess the intricate semantic alignment capabilities of text-to-image models.

3.3. Comparison with State-of-the-arts

Multimodal Understanding Performance. We compare the proposed method with state-of-the-art unified models and understanding-only models in Table 3. Janus-Pro achieves the overall best results. This can be attributed to decoupling the visual encoding for multimodal understanding and generation, mitigating the conflict between these two tasks. When compared to models with significantly larger sizes, Janus-Pro remains highly competitive. For instance, Janus-Pro-7B outperforms TokenFlow-XL (13B) on all benchmarks except GQA.

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↑
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
	SD3-Medium [11]	0.99	0.94	0.72	0.89	0.33	0.60	0.74
Und. and Gen.	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

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
Janus	82.33	87.38	87.70	85.46	86.41	79.68
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

Visual Generation Performance. We report visual generation performance on GenEval and DPG-Bench. As shown in Table 4, our Janus-Pro-7B obtains 80% overall accuracy on GenEval, which outperforms all the other unified or generation-only methods, e.g., Transfusion [55] (63%) SD3-Medium (74%) and DALL-E 3 (67%). This demonstrates that our approach has better instruction-following capabilities. As shown in Table 5, Janus-Pro achieves a score of 84.19 on DPG-Bench, surpassing all other methods. This demonstrates that Janus-Pro excels in following dense instructions for text-to-image generation.

3.4. Qualitative Results

We present results on multimodal understanding in Figure 4. Janus-Pro exhibits impressive comprehension abilities when handling inputs from various contexts, showcasing its powerful capabilities. We also present some text-to-image generation results in the lower part of Figure 4. The images generated by Janus-Pro-7B are highly realistic, and despite having a resolution of only 384×384, they still contain a lot of details. For imaginative and creative scenes, Janus-Pro-7B accurately captures the semantic information from the prompts, producing well-reasoned and coherent images.
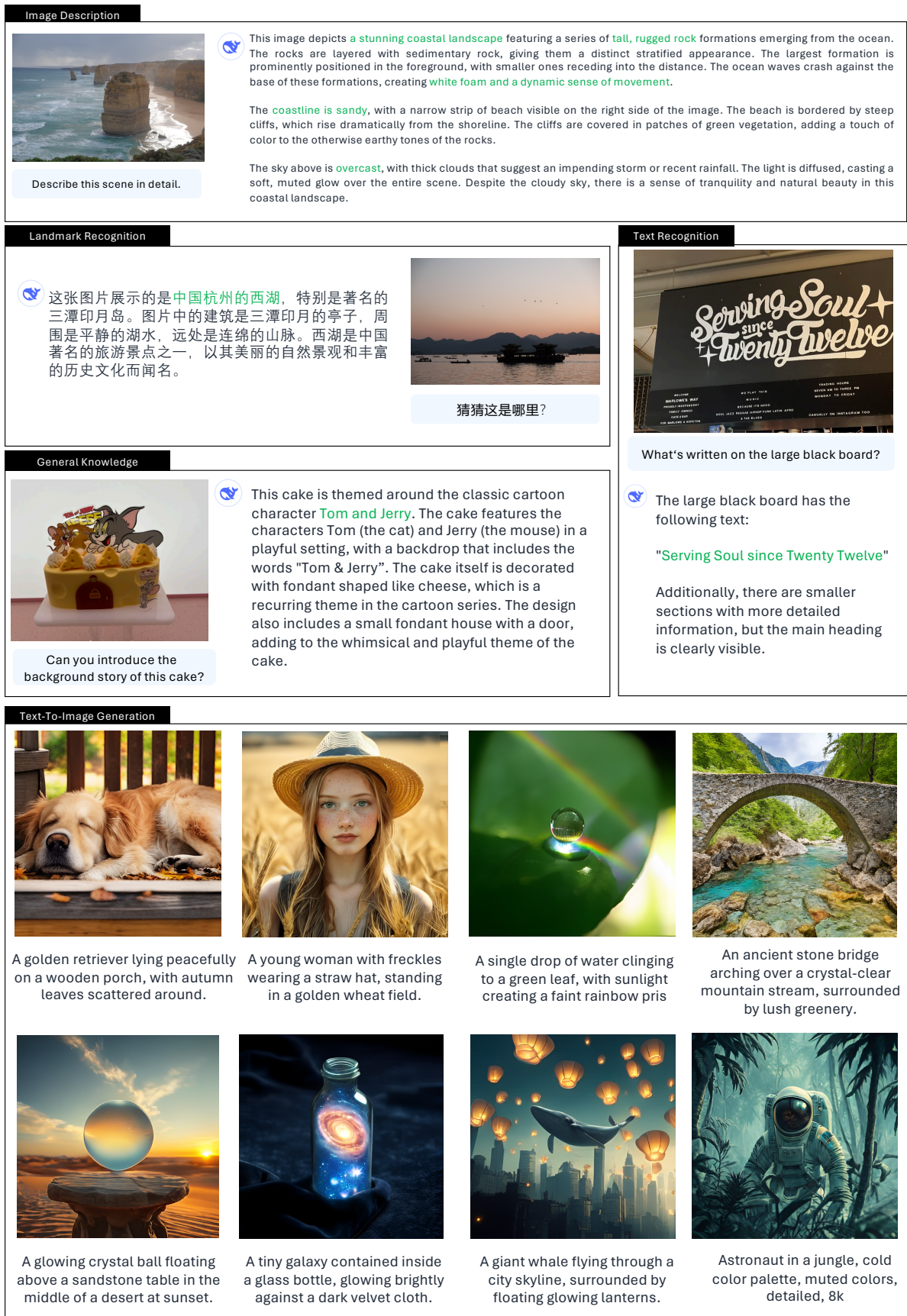
Image Description

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

猜猜这是哪里?

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。

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 pris

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

Figure 4 | **Qualitative results of multimodal understanding and visual generation capability**. The model is Janus-Pro-7B and the image output resolution of visual generation is 384×384. Best viewed on screen.

4. Conclusion

This paper introduces improvements to Janus from three aspects: training strategy, data, and model size. These enhancements have led to significant advancements in both multimodal understanding and text-to-image instruction-following capabilities. However, Janus-Pro still has certain limitations. In terms of multimodal understanding, the input resolution is limited to 384×384, which affects its performance in fine-grained tasks such as OCR. For text-to-image generation, the low resolution, combined with reconstruction losses introduced by the vision tokenizer, results in images that, while rich in semantic content, still lack fine details. For example, small facial regions occupying limited image space may appear under-detailed. Increasing the image resolution could mitigate these issues.

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. 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.

[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 state-of-the-art visual language model, 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. Team. Chameleon: Mixed-modal early-fusion foundation models. 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] 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.

[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.