# JanusFlow: Harmonizing Autoregression and Rectified Flow for Unified Multimodal Understanding and Generation

Supplementary Material

## A. Further Implementation Details

JanusFlow builds upon an enhanced version[1] of DeepSeek-LLM (1.3B) [1, 18]. The generation encoder $g_{enc}$ comprises a $2\times2$ patchify layer followed by two ConvNeXt [27] blocks and a linear layer. The generation decoder $g_{dec}$ combines two ConvNeXt blocks, a pixel-shuffle layer to upsample the feature map, and a linear layer. Our SigLIP encoder contains $\sim$ 300M parameters. $g_{enc}$ and $g_{dec}$ are light-weight modules, containing $\sim$ 70M parameters in total. We employ an exponential moving average (EMA) with a ratio of 0.99 to ensure training stability.

For data preprocessing, we deal with understanding and generation data differently. For understanding tasks, we maintain all image information by resizing the long side to the target size and padding the image to squares. For generation tasks, we resize the short side to the target size and apply random square cropping to avoid padding artifacts. During training, multiple sequences are packed to form a single sequence of length $4,096$ for training efficiency. Our implementation is based on the HAI-LLM platform [7] using PyTorch [22]. Training was conducted on NVIDIA A100 GPUs, with each model requiring $\sim 1,600$ A100 GPU days.

The datasets used in the pre-training stage for understanding include DetailedCaption [4], SAM [9], arXivQA[14], DenseFusion-1M [15], MMSci[16], PixelProse [24], re-captioned LAION-Aesthetics [3], re-captioned Open Images V4 [11], ShareGPT4V [2], WikiHow [10] and WIT [25]. The datasets used in the pre-training stage for generation include re-captioned LAION-Aesthetics [3], DALL-E 3 1M [5], SAM [9], Open Images V4 [11], Megalith-10M [19], YFCC-15M [20], PixelProse[24] and JourneyDB [26].

## B. Performance Analysis of the 256 Resolution Model

We trained our model at two resolutions: $256 \times 256$ and $384 \times 384$. The main paper presents results from the $384 \times 384$ model as our primary results. Here, we provide a comprehensive evaluation of the $256 \times 256$ model's performance. The visual understanding performances are presented in Tab. 1. The generation capabilities are evaluated using GenEval [6], DPG-Benchmark [8], and MJHQ FID-30k [13], with results shown in Tab. 2 and 3. We also provide the sub-task scores in these tables.

As expected, the $256 \times 256$ model shows slightly lower performance compared to the $384 \times 384$ model on visual understanding metrics due to its reduced resolution. Interestingly, however, the $256 \times 256$ model outperforms its higher-resolution counterpart on GenEval and DPG-Bench - benchmarks specifically designed to evaluate instruction following capabilities and semantic accuracy. This superior performance on semantic tasks can be attributed to the model's better control over lower-resolution images, where reduced visual complexity allows for more precise semantic manipulation.

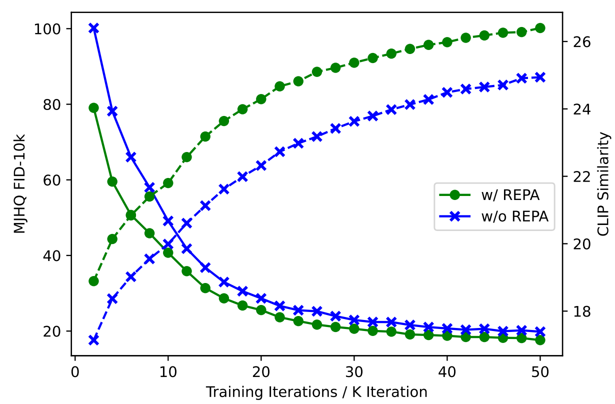



Figure 1. The FID and CLIP similarity during the first $50,000$ iterations.

## C. Details of REPA Ablation

We provide the FID and CLIP similarity of the first 50,000 training iterations of the pre-train stage in Fig. 1 with and without representation alignment regularization. The gap between the two models demonstrates the benefits of using representation alignment regularization.

## D. Analysis of CFG Factor and Sampling Steps

We investigate the impact of two key generation parameters: the Classifier-Free Guidance (CFG) factor and the number of sampling steps. While our main results use $w = 2$ for CFG and 30 sampling steps to calculate FID, here we present a comprehensive analysis of these hyperparameters. Fig. 2 shows the effect of varying CFG factors while main-

[1] This version has been demonstrated to possess better performance on multiple-choice benchmarks (e.g., MMBench [17] and SEED Bench [12]). Our preliminary experiments suggest that it has minimal impact on the quality of visual generation.

Table 1. Results on visual understanding tasks.

| Model | LLM Params | POPE↑ | MME-P↑ | MMB$_{dev}$ ↑ | SEED↑ | VQAv2$_{test}$↑ | GQA↑ | MM-Vet↑ |
|---|---|---|---|---|---|---|---|---|
| **JanusFlow 256** | 1.3B | 85.3 | 1203.0 | 71.9 | 67.6 | 76.3 | 58.4 | 27.4 |
| **JanusFlow 384** | 1.3B | 88.0 | 1333.1 | 74.9 | 70.5 | 79.8 | 60.3 | 30.9 |

Table 2. Results on GenEval [6].

| Method | LLM Params | Single Obj. | Two Obj. | Count. | Colors | Pos. | Color Attri. | Overall↑ |
|---|---|---|---|---|---|---|---|---|
| **JanusFlow 256** | 1.3B | 0.98 | 0.73 | 0.54 | 0.83 | 0.63 | 0.53 | 0.70 |
| **JanusFlow 384** | 1.3B | 0.97 | 0.59 | 0.45 | 0.83 | 0.53 | 0.42 | 0.63 |

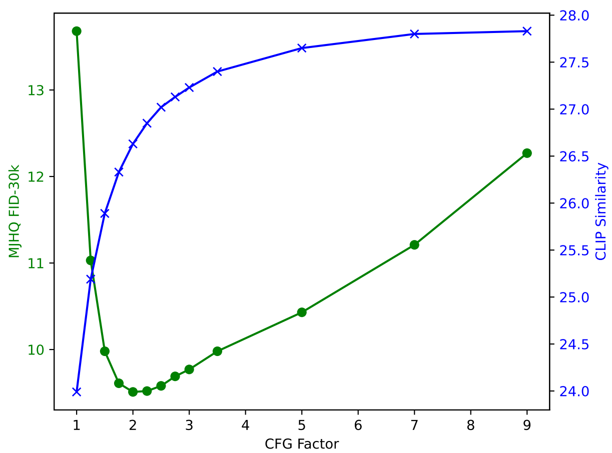



Figure 2. Results of varying CFG Factors

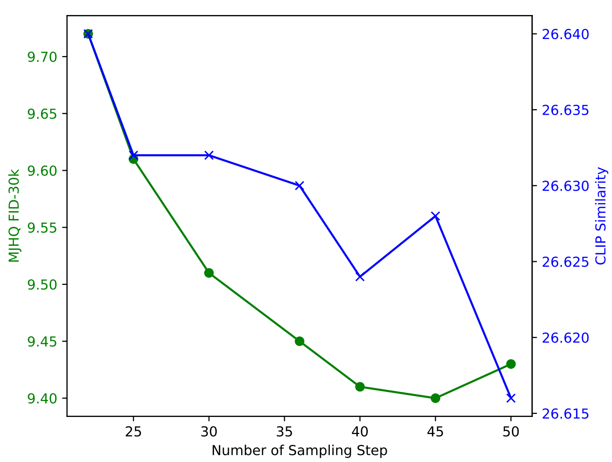



Figure 3. Results of Varying Numbers of Sampling Steps

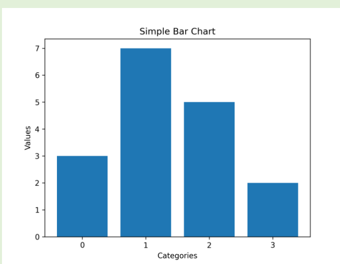
taining 30 sampling steps. The results reveal an optimal CFG value for FID scores, while CLIP [23] similarity continues to improve with increasing CFG values, consistent with findings from previous work [21]. Fig. 3 demonstrates the impact of different sampling steps while maintaining a CFG factor of 2. The number of sampling steps shows relatively minor influence on performance. Our choice of 30 steps in the main paper represents a balance between generation quality and computational efficiency.

## E. Additional Qualitative Results

Additional qualitative examples for both understanding and generation tasks are presented in Fig. 4 and Fig. 5, respectively. The understanding examples demonstrate JanusFlow's diverse capabilities, including code generation, person identification, character recognition, and visual reasoning. For image generation, our model exhibits strong performance in both visual quality and semantic alignment with input prompts.

Table 3. Results on DPG-Bench [8] and MJHQ FID-30k [13].

| Method | DPG-Bench↑ | | | | | | MJHQ FID-30k↓ |
|---|---|---|---|---|---|---|---|
| | Global | Entity | Attribute | Relation | Other | Overall | |
| **JanusFlow 256** | 91.20 | 88.83 | 88.00 | 87.60 | 89.53 | 81.23 | 12.70 |
| **JanusFlow 384** | 87.03 | 87.31 | 87.39 | 89.79 | 88.10 | 80.09 | 9.51 |



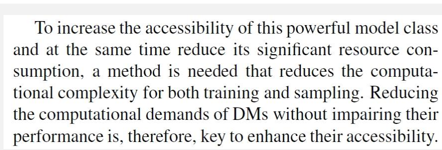

**User:** Please write a python code to draw this plot.

**JanusFlow:** Here is the Python code to draw this plot:

```python
import matplotlib.pyplot as plt

# Data
values = [3, 7, 5, 2]

# Create the bar chart
plt.figure(figsize=(10, 6))
plt.bar(range(len(values)), values)

# Title
plt.title('Simple Bar Chart')

# Labels
plt.xlabel('Categories')
plt.ylabel('Values')

# Show the plot
plt.show()
```

**User:** Explain why this image is funny.

**JanusFlow:** This image is likely to be a humorous or artistic rendition of the famous painting "The Mona Lisa" with a dog instead of a human. The dog's expression and pose are similar to the iconic pose of the original painting, which adds a playful and unexpected twist to the artwork.



To increase the accessibility of this powerful model class and at the same time reduce its significant resource consumption, a method is needed that reduces the computational complexity for both training and sampling. Reducing the computational demands of DMs without impairing their performance is, therefore, key to enhance their accessibility.

**User:** What are the texts in this passage?

**JanusFlow:** To increase the accessibility of this powerful model class and at the same time reduce its significant resource consumption, a method is needed that reduces the computational complexity for both training and sampling. Reducing the computational demands of DMs without impairing their performance is, therefore, key to enhance their accessibility.

**User:** Who is this in the figure?

**JanusFlow:** George W. Bush.

Figure 4. More multimodal understanding cases.

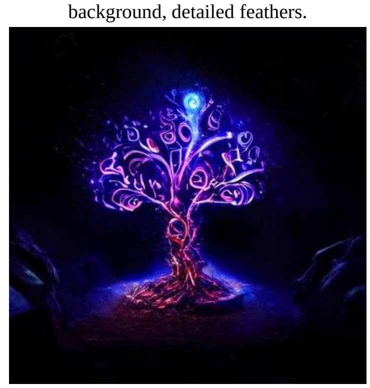
Massive cathedral church, battle between Heaven and hell, church on fire, 8k hyper real ultra sharp renaissance by Francisco Goya.

A handsome 24-year-old boy in the middle with sky color background wearing eye glasses, it's super detailed with anime style.

Happy dreamy owl monster sitting on a tree branch, colorful glittering particles, forest background, detailed feathers.

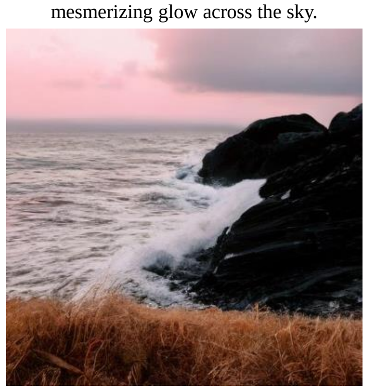
A man wearing Fedora hat with mafia style, realistic photography, intricate details, magical lighting, vibrant background, complex textures, rich colors, realistic style, front-facing view.

A vivid depiction of the Northern Lights dancing above the snow-covered mountains in Iceland, casting a mesmerizing glow across the sky.

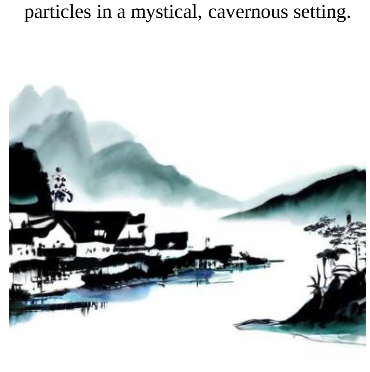


A dark, high-contrast render of a psychedelic Tree of Life glowing brilliantly, illuminating swirling dust particles in a mystical, cavernous setting.

The image features a mushroom growing on grassy ground amidst fallen leaves. Their caps are light brownish-white with visible gills underneath; the stems appear dark and sturdy. In the background, there's an out-of-focus scene that includes greenery and possibly some structures or trees shrouded by mist or fog, giving it a serene yet slightly eerie atmosphere. This photograph employs shallow depth of field to emphasize the mushrooms while blurring the surroundings for artistic effect.



The image captures a vast ocean view at either sunrise or sunset, with soft pink hues near the horizon blending into darker clouds above. Waves crash against rugged black rocks on the right, where water flows down onto smaller stones below. In the foreground, dry grass contrasts with the smooth sea surface. The scene feels tranquil but also reveals the raw power of nature through the interaction between the dynamic waves and the solid land.



A serene Chinese ink painting depicts a tranquil mountain village. Simple homes nestle at the foot of misty peaks, while a gentle river winds through the village. Bamboo and pine trees dot the landscape. The minimalist brushstrokes reflect a harmonious relationship between nature and human life, capturing the peaceful essence of the scene with elegant simplicity.

Figure 5. More text-to-image generation results.

## References


[1] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. DeepSeek LLM: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954, 2024. 1

[2] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793, 2023. 1

[3] dclure. LAION-Aesthetics-UMAP, 2022. 1

[4] echo840. Detailed caption, 2023. 1

[5] Ben Egan, Alex Redden, XWAVE, and SilentAntagonist. DALLE-3 1 million+ high quality captions, 2024. 1

[6] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2024. 1, 2

[7] High-flyer. HAI-LLM: Efficient and lightweight training tool for large models, 2023. 1

[8] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with llm for enhanced semantic alignment. arXiv preprint arXiv:2403.05135, 2024. 1, 3

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In Proc. IEEE Int. Conf. Comput. Vision, 2023. 1

[10] Mahnaz Koupaee and William Yang Wang. WikiHow: A large scale text summarization dataset. arXiv preprint arXiv:1810.09305, 2018. 1

[11] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. Int'l Journal of Computer Vision, 2020. 1

[12] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125, 2023. 1

[13] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2.5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245, 2024. 1, 3

[14] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arXiv: A dataset for improving scientific comprehension of large vision-language models. In Annual Meeting of the Association for Computational Linguistics, 2024. 1

[15] Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. DenseFusion-1M: Merging vision experts for comprehensive multimodal perception. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2024. 1

[16] Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, et al. MMSci: A multimodal multi-discipline dataset for phd-level scientific comprehension. In AI for Accelerated Materials Design, 2024. 1

[17] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In Proc. European Conf. Computer Vision, 2024. 1

[18] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. DeepSeek-VL: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525, 2024. 1

[19] madebyollin. Megalith-10M, 2024. 1

[20] mehdidc. YFCC-15M, 2024. 1

[21] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In Proc. Int'l Conf. Learning Representations, 2024. 2

[22] PyTorch-Contributors. PyTorch, 2024. 1

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In Proc. Int'l Conf. Machine Learning, 2021. 2

[24] Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions. arXiv preprint arXiv:2406.10328, 2024. 1

[25] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In Proc. ACM SIGIR Conf. Research and Develop. in Info. Retrieval, 2021. 1

[26] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. JourneyDB: A benchmark for generative image understanding. In Proc. Annu. Conf. Neural Inf. Process. Systems, 2024. 1

[27] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt v2: Co-designing and scaling ConvNets with masked autoencoders. In Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, 2023. 1