

21 A Formal Description of the BIFROST-1 MLLM Architecture

22 As illustrated in ??, we initialize the visual generation branch from the pretrained MLLM by creating
 23 a trainable copy of the MLP and attention QKV projection layers [16]. The only component randomly
 24 initialized is the vision head, which is a simple linear projection layer. By reusing the majority of
 25 parameters from the pretrained MLLM, we avoid the costly process of realigning image embeddings.

26 Specifically, we use **blue** color to denote the original *frozen* MLLM-specific modules, which handle
 27 language modeling (**LM**) and image understanding (**Img-U**) tasks. These modules remain unchanged
 28 during training. In addition, we use **yellow** color to represent the newly introduced *trainable*
 29 modules for the image generation task (**Img-G**), which are initialized as trainable copies of the
 30 corresponding blue modules.

31 During the image generation training process, we first obtain input hidden states for each task.
 32 Text inputs x^{Text} are projected through a linear embedding layer to produce $h_{\text{in}}^{\text{Text}}$. Images used for
 33 image understanding ($x^{\text{Img-U}}$) and image generation ($x^{\text{Img-G}}$) are both passed through the frozen
 34 MLLM-embedded visual encoder \mathcal{E}^{Und} , producing hidden states $h_{\text{in}}^{\text{Img-U}}$ and $h_{\text{in}}^{\text{Img-G}}$ respectively:

$$h_{\text{in}}^{\text{Text}} = \text{Linear}_{\text{Text}}(x^{\text{Text}}) \quad h_{\text{in}}^{\text{Img-U}} = \mathcal{E}^{\text{Und}}(x^{\text{Img-U}}) \quad h_{\text{in}}^{\text{Img-G}} = \mathcal{E}^{\text{Und}}(x^{\text{Img-G}})$$

35 During the image generation inference process, $h_{\text{in}}^{\text{Img-G}}$ is initialized from 2D learnable mask tokens.

36 *By using patch-level CLIP image embeddings that are natively aligned with the MLLM’s visual*
 37 *encoder to represent visual signals in vision generation tasks, we eliminate the need for any additional*
 38 *alignment between the visual generation representation and the MLLM.*

39 For attention processing, we use the frozen MLLM attention layers to compute Q, K, and V matrices
 40 for the text and image understanding tokens $h_{\text{in}}^{\text{Text}}$ and $h_{\text{in}}^{\text{Img-U}}$, and use the newly added visual
 41 generation branch to process the image generation tokens $h_{\text{in}}^{\text{Img-G}}$:

$$h_Q^{\text{Text}}, h_K^{\text{Text}}, h_V^{\text{Text}} = \text{QKV}_{\text{MLLM}}(h_{\text{in}}^{\text{Text}}) \quad h_Q^{\text{Img-U}}, h_K^{\text{Img-U}}, h_V^{\text{Img-U}} = \text{QKV}_{\text{MLLM}}(h_{\text{in}}^{\text{Img-U}})$$

42

$$h_Q^{\text{Img-G}}, h_K^{\text{Img-G}}, h_V^{\text{Img-G}} = \text{QKV}_{\text{Img-G}}(h_{\text{in}}^{\text{Img-G}})$$

For language modeling and image understanding tasks, we replicate the standard MLLM attention structure by attending over their respective modalities only. For image generation task, we enable cross-module attention, allowing image generation queries to attend jointly over all token types. Specifically:

$$\begin{aligned} h_O^{\text{Text}} &= \text{O}_{\text{MLLM}}(\text{Attn}(h_Q^{\text{Text}}, [h_K^{\text{Text}} \circ h_K^{\text{Img-U}}], [h_V^{\text{Text}} \circ h_V^{\text{Img-U}}])) \\ h_O^{\text{Img-U}} &= \text{O}_{\text{MLLM}}(\text{Attn}(h_Q^{\text{Img-U}}, [h_K^{\text{Img-U}} \circ h_K^{\text{Text}}], [h_V^{\text{Img-U}} \circ h_V^{\text{Text}}])) \\ h_O^{\text{Img-G}} &= \text{O}_{\text{Img-G}}(\text{Attn}(h_Q^{\text{Img-G}}, [h_K^{\text{Img-G}} \circ h_K^{\text{Img-U}} \circ h_K^{\text{Text}}], [h_V^{\text{Img-G}} \circ h_V^{\text{Img-U}} \circ h_V^{\text{Text}}])) \end{aligned}$$

43 where \circ denotes concatenation, and $O(\cdot)$ denotes linear output projection layer. We apply a causal
 44 mask to text and image understanding tokens, and a bidirectional mask to image generation tokens.

45 For the MLP layers, we follow the same branching: text and image-understanding tokens $h_{\text{in}}^{\text{Text}}$ and
 46 $h_{\text{in}}^{\text{Img-U}}$ are passed through the frozen MLLM MLP, while image generation tokens $h_{\text{in}}^{\text{Img-G}}$ use the
 47 trainable MLP from the generation branch:

$$h_{\text{MLP}}^{\text{Text}} = \text{MLP}_{\text{MLLM}}(h_O^{\text{Text}}) \quad h_{\text{MLP}}^{\text{Img-U}} = \text{MLP}_{\text{MLLM}}(h_O^{\text{Img-U}}) \quad h_{\text{MLP}}^{\text{Img-G}} = \text{MLP}_{\text{Img-G}}(h_O^{\text{Img-G}})$$

48 Finally, task-specific heads convert the hidden states to output predictions. For language modeling
 49 and image understanding, we apply a linear projection to $h_{\text{MLP}}^{\text{Text}}$, and for image generation, we project
 50 $h_{\text{MLP}}^{\text{Img-G}}$ using the vision generation head:

$$h_{\text{out}}^{\text{Text}} = \text{TextHead}(h_{\text{MLP}}^{\text{Text}}) \quad h_{\text{out}}^{\text{Img-G}} = \text{VisionHead}(h_{\text{MLP}}^{\text{Img-G}})$$

Table 1: Comparison with state-of-the-arts on multimodal generation benchmarks.

Method	Base (M)LLM	Training Steps×Batch Size	COCO FID↓	MJHQ FID↓	GenEval↑	DPG-Bench↑
EMU [18]	LLaMA 13B	-	11.66	-	-	-
DreamLLM [5]	Vicuna 7B	-	8.46	-	-	-
Chameleon [14]	From Scratch 7B	-	26.74	-	0.39	-
Show-o-512 [22]	Phi-1.5 1.3B	-	9.24	15.18	0.68	-
VILA-U [21]	LLaMA-2 7B	-	-	7.69	-	-
EMU3 [18]	From Scratch 7B	-	12.80	-	0.66	80.60
MetaMorph [15]	LLaMA-3 8B	-	11.8	-	-	-
MetaQuery-L [12]	Qwen2.5-VL 3B	-	8.87	6.35	0.78	81.10
MetaQuery-XL [12]	Qwen2.5-VL 7B	200M	8.69	6.02	0.80	82.05
TokenFlow-XL [6]	Qwen-2.5 14B	-	-	-	0.63	73.38
Transfusion [23]	From Scratch 7B	-	8.70	-	0.63	-
LMFusion [13]	LLaVA-Next 8B	-	8.20	-	-	-
Janus [20]	DeepSeek-LLM 1.5B	100M	8.53	10.10	0.61	-
JanusFlow [11]	DeepSeek-LLM 1.5B	211M	-	9.51	0.63	80.09
JanusPro-1B [4]	DeepSeek-LLM 1.5B	200M	-	14.33	0.73	82.63
JanusPro-7B [4]	DeepSeek-LLM 7B	194M	-	13.48	0.80	84.19
BIFROST-1 (Ours)	Qwen2.5-VL 3B	9M	23.02	15.24	0.61	76.41

B Comparison with SoTAs on Multimodal Generation Benchmarks

Table 1 compares BIFROST-1 with other unified models on image generation benchmarks, including COCO and MJHQ for visual quality and GenEval and DPG-Bench for prompt following ability. As we can see, our model trained on only 9M image-text pairs for 1 epoch matches the performance with baselines trained with much higher compute, including Janus on GenEval benchmark, and outperforms TokenFlow-XL on DPG-Bench. In addition, we would like to highlight that the FID scores on the COCO and MJHQ datasets are heavily influenced by the choice of diffusion model. Diffusion models fine-tuned on aesthetic datasets (*e.g.*, FLUX.1-dev) typically achieve worse FID scores compared to models that have not undergone extensive aesthetic fine-tuning (*e.g.*, SD1.5, as used in MetaQuery).

C Broader Impacts

BIFROST-1 is motivated by the fact that training a unified multimodal generation and understanding model that can perform native generation with high visual quality usually requires huge computational cost. By bridging pretrained MLLM with pretrained diffusion models, training BIFROST-1 can be significantly faster. Therefore, we believe that our work can be a strong contribution to efficient unified model training. While our framework can benefit numerous applications in image generation, similar to other image generation frameworks, it can also be used for potentially harmful purposes (*e.g.*, creating false information or misleading images). Therefore, it should be used with caution in real-world applications.

D Safeguards

BIFROST-1 is built upon pretrained MLLM (*i.e.*, Qwen2.5-VL) and diffusion models (*i.e.*, FLUX.1-dev) with strong safeguards, and trained on publically available image datasets (*i.e.*, MSCOCO and SA1B) that removes unsafe concepts. Therefore, our model avoids the high risk for misuse.

E Limitations

Note that BIFROST-1 is designed as a bridging method that connects existing MLLMs with diffusion-based image generation models. As such, its performance, output quality, and potential visual artifacts are inherently influenced by the capabilities and limitations of the underlying backbone models it relies on. For instance, if the diffusion model used as the visual backbone struggles with generating complex, rare, or previously unseen scenes and objects, then BIFROST-1, which builds upon this foundation, may also exhibit suboptimal image generation results. This dependency highlights the importance of selecting strong and well-generalized base models when applying BIFROST-1 to real-world or open-domain generation tasks.

83 **F License**

84 We use standard licenses from the community and provide the following links to the licenses for the
85 datasets, codes, and models that we used in this paper. For further information, please refer to the
86 specific link.

- 87 PyTorch [1]: [BSD-style](#)
- 88 HuggingFace Transformers [19]: [Apache License 2.0](#)
- 89 HuggingFace Diffusers [17]: [Apache License 2.0](#)
- 90 FLUX.1-dev [8]: [Non-Commercial License](#)
- 91 Qwen2.5-VL [2]: [Non-Commercial License](#)
- 92 MSCOCO dataset [10]: [CC BY 4.0](#)
- 93 CC12M dataset [3]: [Permissive Custom License](#)
- 94 SA1B dataset [7]: [SA-1B Dataset Research License](#)
- 95 MJHQ30k dataset [9]: [Playground v2 Community License](#)



Colosseum at night, beautifully illuminated, 20 minutes after sunset, starry night, dramatic, emotional, saturated, hyper realistic, professional, award winning, fine art

A high end residential swimming pool with the Atlanta skyline serving as a dynamic backdrop, a state-of-the-art drone camera and feature a chic outdoor bar, a luxury spa area, and stunning poolside lighting

Starry night blended into realistic walkway, in the style of Vincent van Gogh, oil on canvas, watercolor painting, ambient occlusion

Medieval fantasy heroic universe, natural underground cave, unreal 5 engine

Modern Egypt with flying spaceships, energy generating pyramids, giants, rivers in the sky, gold houses. Ultra Realistic, high detailed, 8k, bright colors, high contrast



A black lion in futuristic shades wearing neon sunglasses, in the style of Dan Mumford, 32k uhd, Valentin Rekunenko, laser segall, mashup of styles, 1990s, solarpunk



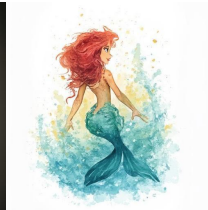
Steampunk style owl singing and holding a camera ultra detailed in complete white background



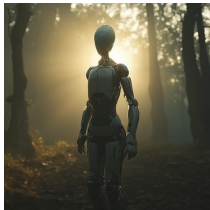
A cute baby tiger in a tuxedo playing the violin, a cute character, Disney style, full body



A gorilla, dressed as Karate player, by Annie Leibovitz



The little mermaid Disney watercolor



Cinematic, realistic robot humanoid walking in a forest forest background, 16K, ultra realistic, 116, v5, dramatic looking, lens flare, evening sky



Spiderman closeup on fire, realistic, 4k



Futuristic Sonic the Hedgehog, Sega, aviator sunglasses, dark backdrop



Fantastic image of a wizard who can't understand a book of magic spell



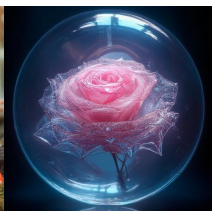
Background of blue and orange flames with a white horse rearing in foreground facing an orange Yamaha MT09



Pumpkin in a different color, in the style of Mark Henson, detailed background elements, Victor Nizovtsev, dark indigo, cute and quirky, leaf patterns, sculpted



A magic mushroom house in a fantasy world, 3d



Lifelike, mystical, fantasy iridescent rose in a glass dome, like the rose from Beauty and the Beast, with water droplets. Hundreds of tiny fantasy details, beautiful design, high contrast, sharp focus



Cherry tree on the surface of the moon



Detailed painting of magnolias on a tree by Vincent Van Gogh, Lisa Frank and J.C. Leyendecker, solid color background, soft colors, contrasting colors



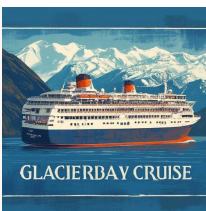
A 3d head of broccoli wearing sunglasses on a white background



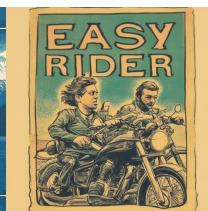
Round logo, transparent, of red dragon, television, explosion



Logo minimalist girl unicorn



Retro travel poster of Westerdam cruise ship in Glacier Bay National Park



Create a vintage poster of Easy Rider the movie

Figure 1: Visualization examples from MJHQ30k dataset.

References

- [1] J. Ansel, E. Yang, H. He, N. Gimelshein, A. Jain, M. Voznesensky, B. Bao, P. Bell, D. Berard, E. Burovski, G. Chauhan, A. Chourdia, W. Constable, A. Desmaison, Z. DeVito, E. Ellison, W. Feng, J. Gong, M. Gschwind, B. Hirsh, S. Huang, K. Kalambarkar, L. Kirsch, M. Lazos, M. Lezcano, Y. Liang, J. Liang, Y. Lu, C. Luk, B. Maher, Y. Pan, C. Puhersch, M. Reso, M. Saroufim, M. Y. Siraichi, H. Suk, M. Suo, P. Tillet, E. Wang, X. Wang, W. Wen, S. Zhang, X. Zhao, K. Zhou, R. Zou, A. Mathews, G. Chanan, P. Wu, and S. Chintala. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, Apr. 2024.
- [2] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [3] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.
- [4] X. Chen, Z. Wu, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, and C. Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.
- [5] R. Dong, C. Han, Y. Peng, Z. Qi, Z. Ge, J. Yang, L. Zhao, J. Sun, H. Zhou, H. Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*, 2023.
- [6] M. Geyer, O. Bar-Tal, S. Bagon, and T. Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *The Twelfth International Conference on Learning Representations*, 2023.
- [7] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.
- [8] B. F. Labs. Flux.1-dev. <https://huggingface.co/black-forest-labs/FLUX.1-dev>, 2024.
- [9] 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.
- [10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.
- [11] 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.
- [12] X. Pan, S. N. Shukla, A. Singh, Z. Zhao, S. K. Mishra, J. Wang, Z. Xu, J. Chen, K. Li, F. Juefei-Xu, J. Hou, and S. Xie. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.
- [13] W. Shi, X. Han, C. Zhou, W. Liang, X. V. Lin, L. Zettlemoyer, and L. Yu. Lmfusion: Adapting pretrained language models for multimodal generation, 2025.
- [14] C. Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

- 144 [15] S. Tong, D. Fan, J. Zhu, Y. Xiong, X. Chen, K. Sinha, M. Rabbat, Y. LeCun, S. Xie, and Z. Liu.
145 Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint*
146 *arXiv:2412.14164*, 2024.
- 147 [16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and
148 I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*,
149 volume 30. Curran Associates, Inc., 2017.
- 150 [17] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, and
151 T. Wolf. Diffusers: State-of-the-art diffusion models. [https://github.com/huggingface/](https://github.com/huggingface/diffusers)
152 *diffusers*, 2022.
- 153 [18] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu, et al.
154 Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.
- 155 [19] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf,
156 M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L.
157 Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Transformers: State-of-the-art natural
158 language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural*
159 *Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association
160 for Computational Linguistics.
- 161 [20] C. Wu, X. Chen, Z. Wu, Y. Ma, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, C. Ruan, et al. Janus:
162 Decoupling visual encoding for unified multimodal understanding and generation. *arXiv*
163 *preprint arXiv:2410.13848*, 2024.
- 164 [21] Y. Wu, Z. Zhang, J. Chen, H. Tang, D. Li, Y. Fang, L. Zhu, E. Xie, H. Yin, L. Yi, et al. Vila-u:
165 a unified foundation model integrating visual understanding and generation. *arXiv preprint*
166 *arXiv:2409.04429*, 2024.
- 167 [22] J. Xie, W. Mao, Z. Bai, D. J. Zhang, W. Wang, K. Q. Lin, Y. Gu, Z. Chen, Z. Yang, and M. Z.
168 Shou. Show-o: One single transformer to unify multimodal understanding and generation.
169 *arXiv preprint arXiv:2408.12528*, 2024.
- 170 [23] C. Zhou, L. Yu, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, J. Kahn, X. Ma, L. Zettlemoyer,
171 and O. Levy. Transfusion: Predict the next token and diffuse images with one multi-modal
172 model, 2024.