OpenMAIA: a Multimodal Automated Interpretability Agent based on open-source models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Interpreting the internal mechanisms of large neural networks remains a main challenge for trustworthy AI. Recent works such as *MAIA* (a Multimodal Automated Interpretability Agent) have shown that agent-based systems can iteratively generate and test hypotheses about neuron function without the need for human intervention, which offers a scalable solution for mechanistic interpretability. However, these agent-based systems rely on closed-source APIs, limiting reproducibility and access. To address this, we introduce *OpenMAIA*, an open-source implementation of *MAIA* that replaces its main components with open-source models. To this end, we experiment with two state-of-the-art multimodal Large Language Models (LLMs) (Gemma-3-27B, Mistral-Small-3.2-24B) as backbone models, and update the agent's interpretability toolset with open-source models. Following the neuron description evaluation protocol established in the original *MAIA* paper, applied across multiple vision backbones and synthetic neurons, *OpenMAIA* achieves predictive accuracy comparable to Claude Sonnet 4 while converging more efficiently. These results demonstrate that competitive, agent-based interpretability can be achieved with a fully open stack, providing a practical and reproducible foundation for community-driven research.

1 Introduction

Understanding the functional role of individual neurons within deep neural networks remains a central challenge in mechanistic interpretability. While a variety of tools and techniques have been developed to support these analyses, they often rely heavily on manual inspection and expert intuition, limiting their scalability and reproducibility. To address these limitations, recent works such as *MAIA* (a Multimodal Automated Interpretability Agent) [1] have explored the use of autonomous agents for interpreting vision systems and neuron-level behaviors. In particular, *MAIA* combines multimodal reasoning with tools used in an iterative loop, enabling high-quality neuron explanations and systematic detection of spurious features, offering the potential to automate and systematize the interpretation procedure. However, *MAIA* strongly rely on commercial platforms and closed-source APIs, which restricts transparency, limits customization, includes overhead costs, and introduce barriers to their widespread use.

In this work, we introduce *OpenMAIA*, an open-source implementation of *MAIA* that retains the agentic experimentation loop of *MAIA* while replacing closed components with state-of-the-art open multimodal LLMs. We experiment with Gemma-3-27B and Mistral-Small-3.2-24B [2, 3, 4], and provide updated open-source-based tools for image generation, editing, and summarization. By grounding the framework in openly available resources, *OpenMAIA* facilitates access and experimentation, enabling deeper mechanistic investigations than closed APIs permit. Figure 1 illustrates a simplified trace of the system: beginning with dataset exemplars, the agent proposes hypotheses, de-

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

signs targeted experiments, and iteratively refines them until producing a natural-language description of the neuron behavior.

Our evaluation on the neuron description tasks introduced in *MAIA* shows that open backbones achieve competitive neuron explanations compared to Claude-Sonnet-4 [5], while converging more efficiently in shorter dialogues. Our results empirically demonstrate that state-of-the-art agent-based interpretability is feasible with a fully open stack, providing a foundation for transparent, scalable, and community-driven research in settings where interpretability is critical.

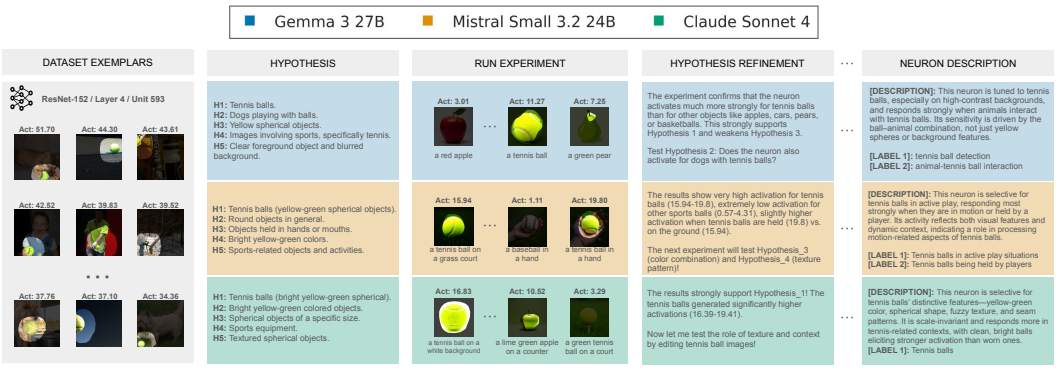

Figure 1: **Simplified experiment for a single neuron across agents.** Illustrative trace for *ResNet-152 (layer 4, unit 593)*: agents start from dataset exemplars, propose hypotheses, and run targeted image generation/editing experiments. Iterative refinement reveals selective features (e.g., tennis ball color, context, interaction), concluding with a natural-language description and concise labels.

2 Related Work

Tool-Based / Automated Neuron Labeling. Network Dissection [6] addressed automatic CNN interpretability by aligning individual units with a predefined set of concepts based on activation–mask overlap. While it provided a useful benchmark, its analysis was limited to a fixed, closed vocabulary. Building on the same idea, some works have taken advantage of CLIP's open vocabulary space to improve neuron labeling. CLIP-Dissect [7] used the pre-trained CLIP model to assign open-vocabulary labels to neurons without retraining or manual annotation, offering an efficient and scalable approach to interpretability. FALCON [8] extended this line of work by enhancing label fidelity through counterfactual analysis and spatial filtering, though its outputs remain short phrases rather than richer descriptions. In contrast, MILAN [9] moved beyond tag interpretability by producing open-ended, natural-language descriptions of individual neurons. It does so by training a captioning model on top-activating image regions (neuron exemplars), using crowd-sourced human annotations. More recently, Describe-and-Dissect [10] leveraged pretrained models (BLIP for captioning, GPT-3.5 for summarization, and synthetic image generation for validation)[11, 12] in a three-stage pipeline, providing concise natural-language explanations.

Agent-Based Interpretability. Early work by Bills et al. [13] used GPT-4 to generate natural-language explanations for neurons and test them through activation prediction, introducing a quantitative self-consistency measure. However, this framework was essentially single-pass and did not support iterative hypothesis refinement. Schwettmann et al. [14] advanced this paradigm with the FIND benchmark, where a language-based interpretability agent iteratively proposed, tested, and revised hypotheses on synthetic functions with known ground truth. Iteration improved descriptive accuracy but often failed to capture fine-grained behaviors. Most recently, *MAIA* [1] generalizes this agentic approach to the multimodal setting, equipping a pretrained multimodal LLM with tools for exemplar selection, synthetic image generation, targeted editing, and summarization. Through iterative experimentation, *MAIA* not only produces expert-level neuron explanations and interventions on spurious features but also extends to broader interpretability tasks such as bias detection and failure mode analysis.

Recent Progress in Open Multimodal LLMs. Over the last two years, open-source multimodal large language models have made rapid progress. Early efforts such as BLIP-2 and LLaVA showed that linking vision encoders with LLMs could yield surprisingly capable visual assistants[15, 16],

while projects like OpenFlamingo and Otter scaled this approach to multi-image and multi-turn reasoning[17, 18, 19]. More recently, families such as Mistral 3 (Small, Medium, Large) and Gemma 3 (1B–27B) have demonstrated strong multimodal reasoning and tool-use abilities, narrowing the gap with proprietary models (GPT-4o, Claude Sonnet 4)[20] while remaining efficient enough for consumer-grade hardware. To our knowledge, prior interpretability frameworks have not yet leveraged these open multimodal backbones, and *OpenMAIA* is the first automated, agent-based interpretability framework that is fully implemented on top of open-source models.

3 OpenMAIA Framework

We introduce OpenMAIA, which is based on the agentic iterative experimentation paradigm of MAIA [1], but is fully grounded in open-source models. We experiment with Gemma-3-27B and Mistral-Small-3.2-24B for the backbone multimodal LLMs, and update the tool suite introduced in MAIA with state-of-the-art open-source models.

Multimodal LLM Backbone. We adopt Gemma-3-27B and Mistral-Small-3.2-24B as OpenMAIA's multimodal backbones due to their validated strength in agentic tool use and vision–language reasoning. Independent evaluations show that both models excel despite their moderate size: in the LMSYS Vision Arena (Aug 2025), Gemma-3-27B-IT and Mistral-3.2 ranked #12 and #23 (Elo $\approx$1162 and 1135), surpassing larger open VLMs [21]. On VLM@school (multilingual visual QA), Mistral 24B (42.9%) and Gemma 3 27B (40.2%) outperformed size-based expectations [22], while Gemma 3 achieved the lowest mean error on ORBIT object-property reasoning [23]. For agentic integration, both models support structured function calling, with Mistral 3.2 independently verified to improve API reliability in AWS Bedrock trials and aligning with the BFCL benchmark standard[24, 25]. Together, these results establish them as efficient, reproducible, and high-performing choices for open multimodal interpretability agents.

Tools. We updated the tool suite introduced in MAIA [1], replacing closed components with state-of-the-art open-source components. Below we provide the details of the updated tools.

Image Generation. MAIA generated images for hypothesis testing with Stable Diffusion v1.5[26] and reported an ablation where DALL·E 3 [27] improved performance. It also documented prompt-following failures for SD-v1.5 that occasionally confused the agent's conclusions [1]. Motivated by these limitations, we adopt FLUX.1 [dev], a 12B rectified-flow transformer released as open weights, which we deploy via Diffusers with 4-bit NF4 quantization. This configuration substantially reduces VRAM while maintaining practically useful visual quality, as shown in Diffusers' quantization guide and model card comparisons; we use it to keep the agent loop interactive on a single high-memory GPU [28, 29, 30].

Image Editing. MAIA performed text-guided edits with InstructPix2Pix, but noted frequent failures on precise or negative edits (e.g., removing objects or assigning attributes correctly) [31, 1]. We replace this with FLUX.1 Kontext [dev], an open-weights model that unifies generation and editing and is explicitly optimized for local edits, character/style consistency, and multi-turn workflows. Its technical report introduces KontextBench and reports strong single-turn quality and multi-turn consistency across local/global edits, reference character/style, and typography, aligning well with OpenMAIA's causal intervention loop [32, 33].

Image Summarization. MAIA used one of the first versions of GPT-4 with vision capabilities (GPT-4V) to (i) describe highlighted regions per image and (ii) summarize common content across exemplars, to mitigate confirmation bias [1]. For openness and semantic consistency with the agent, we instead reuse our backbones (Gemma 3 27B, Mistral Small 3.2). As we previously discussed, their strong performance on vision–language benchmarks makes them suitable for this task.

4 Evaluation

To test *OpenMAIA*, we follow the same quantitative evaluation protocol as in *MAIA*. Concretely, we evaluated *OpenMAIA* on the neuron description task across four vision backbones: ResNet-152, DINO ViT-S/8, CLIP RN50, and synthetic neurons with known ground truth—using two open multimodal LLM backbones (Gemma-3-27B and Mistral-Small-3.2-24B). For each condition, we sample neurons uniformly across layers, yielding 384 evaluations per backbone. For each model,

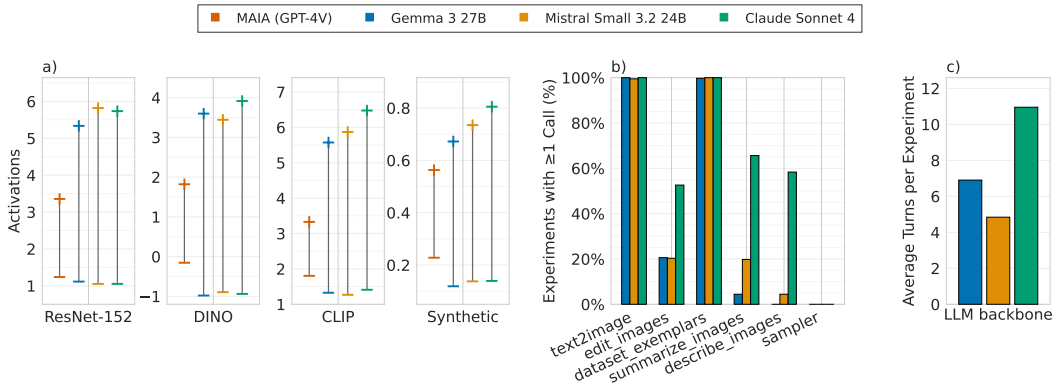

Figure 2: **Comparison of LLM backbones. (a) Predictive activation evaluation.** For each *vision backbone* (ResNet-152, DINO ViT-S/8, CLIP RN50, Synthetic neurons) we aggregate activations produced under each *LLM backbone* evaluation. The vertical segment per axis spans the mean of the positive and neutral activations; the "+" marker denotes the mean positive activation and the "−" marker the mean neutral activation. *MAIA* score is used only as a *reference*. **(b) Tool usage.** For each LLM backbone, bars show the percentage of experiments that invoked each tool at least once. **(c) Conversation length.** Average number of turns per experiment for each LLM backbone.

quality is assessed by whether the obtained labels can reliably distinguish positive from neutral generations in terms of neuron activations. In addition to these open models, we also report results for Claude-Sonnet-4 as a strong closed-source baseline. Finally, we include the original *MAIA* (GPT-4V) numbers as a reference: while not directly comparable due to differences in the tool suite, they help contextualize the impact of our open-source toolchain and illustrate the evolution of agent-based interpretability in terms of performance.

Figure 2a shows that both Gemma-3-27B and Mistral-Small-3.2-24B achieve strong predictive gaps between positive and neutral generations, on par with Claude-Sonnet-4 and clearly outperforming the original *MAIA* baseline (GPT-4V). This indicates that fully open backbones and tools can produce competitive explanations without relying on proprietary APIs.

Tool usage. As shown in Figure 2b, Gemma-3-27B and Mistral-Small-3.2-24B rely primarily on dataset exemplars and image generation, with occasional use of image editing, but make little use of image summarization or description tools. In contrast, Claude-Sonnet-4 employs a more diverse tool mix, including frequent calls to summarization modules. Interestingly, this diversity in tool usage correlates with overall performance: Claude Sonnet achieves the strongest predictive scores, followed by Mistral and then Gemma, mirroring the ordering of tool diversity. This suggests that a broader use of the available tools may contribute to more accurate neuron explanations.

Efficiency. Figure 2c reports conversation length. Mistral-Small-3.2-24B reached conclusions in the fewest experimental iterations ($\sim$5 on average), while Gemma-3-27B required slightly longer dialogues ($\sim$7 turns). Claude-Sonnet-4, by contrast, often exceeded 10 turns. This suggests that *OpenMAIA*'s open backbones tend to converge more quickly, which is desirable for efficiency. At the same time, the longer trajectories of Claude-Sonnet-4 may reflect a more cautious style, repeatedly validating hypotheses before committing to a final description.

5 Conclusion

We introduced OpenMAIA, the first open-source interpretability agent, which is based on MAIA. OpenMAIA includes two open-source LLM backbones (Gemma 3 27B, Mistral Small 3.2), and an interpretability tools that is also based in open-source models. Our evaluation shows that the open-source LLM backbones achieve competitive results on neuron descriptions when compared to Claude-Sonnet-4, while converging more efficiently, with fewer experimental iterations. By eliminating dependence on closed-source APIs, OpenMAIA facilitates reproducible interpretations of model behavior, addressing the need for scalable, open, and free of cost framework in multimodal interpretability, which has been highlighted in recent surveys [34].

References

[1] Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent. In *Forty-first International Conference on Machine Learning*, 2024.

[2] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL `https://arxiv.org/abs/2503.19786`.

[3] Mistral AI. Mistral small 3.1 (mistral-small-2503). `https://mistral.ai/news/mistral-small-3-1`, 2025. Apache 2.0 license; released March 17, 2025.

[4] Mistral AI. Mistral small 3.2 (mistral-small-3.2-24b-instruct-2506). `https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506`, 2025. Apache 2.0 license; released June 2025. Minor update of Mistral Small 3.1.

[5] Anthropic. System card for claude opus 4 and claude sonnet 4. Technical report, Anthropic, 2025. Introduces Claude Sonnet 4, describes model capabilities, safety testing, and deployment under AI Safety Level 2.

[6] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[7] Tuomas Oikarinen and Tsui-Wei Weng. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *The Eleventh International Conference on Learning Representations*, 2023.

[8] Neha Kalibhat, Shweta Bhardwaj, Bayan Bruss, Hamed Firooz, Maziar Sanjabi, and Soheil Feizi. Identifying interpretable subspaces in image representations. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[9] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *International Conference on Learning Representations*, 2021.

[10] Nicholas Bai, Rahul A Iyer, Tuomas Oikarinen, Akshay Kulkarni, and Tsui-Wei Weng. Interpreting neurons in deep vision networks with language models. *TMLR*, 2025.

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. preprint; published January 28, 2022.

[12] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, ..., and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. Introduces GPT-3, released May 2020.

[13] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. 2023.

[14] Sarah Schwettmann, Tamar Shaham, Joanna Materzynska, Neil Chowdhury, Shuang Li, Jacob Andreas, David Bau, and Antonio Torralba. Find: A function description benchmark for evaluating interpretability methods. *Advances in Neural Information Processing Systems*, 36: 75688–75715, 2023.

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=w0H2xGHlkw`.

[17] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[18] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022.

[19] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.

[20] OpenAI. Gpt-4o technical report, 2024. `https://openai.com/index/gpt-4o`.

[21] Christopher Chou, Lisa Dunlap, Koki Mashita, Krishna Mandal, Trevor Darrell, Ion Stoica, Joseph E. Gonzalez, and Wei-Lin Chiang. Visionarena: 230k real world user-vlm conversations with preference labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3877–3887, June 2025.

[22] René Peinl and Vincent Tischler. Vlm@school – evaluation of ai image understanding on german middle-school knowledge. arXiv preprint arXiv:2506.11604, 2025. Evaluates 13 open-weight VLMs on German-language school-level visual QA; Mistral 24B scored 42.86

[23] Abhishek Kolari, Mohammadhossein Khojasteh, Yifan Jiang, Floris den Hengst, and Filip Ilievski. Orbit: An object property reasoning benchmark for visual inference tasks. *arXiv preprint*, arXiv:2508.10956, 2025. Systematic benchmark of object-property reasoning; best VLMs achieve only 40

[24] Amazon Web Services. Mistral small 3.2 24b instruct (v2506) on amazon bedrock and sagemaker jumpstart. AWS ML Blog (July 29, 2025), 2025. Independent report of improved function-calling reliability and image-text reasoning capabilities in Mistral 3.2.

[25] Berkeley Function-Calling Leaderboard Team. Berkeley function-calling leaderboard (bfcl). BFCL online benchmark (ICML 2025), 2025. Standardized benchmark for structured tool/function-calling ability in LLMs.

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[27] OpenAI. Dall·e 3, 2023. `https://openai.com/dall-e-3`.

[28] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[29] Hugging Face. Quantizing diffusers models — 4-bit nf4 inference. Hugging Face blog, 2025. Guidelines and evaluation of quantization strategies (NF4) for diffusion models in Diffusers framework.

[30] Black Forest Labs. 4-bit nf4 quantization of flux.1 [dev]. Hugging Face model card documentation, 2025. Demonstrates effective deployment of FLUX.1 [dev] under 4-bit NF4 quantization while maintaining visual quality.

[31] Thomas Brooks et al. Instructpix2pix: Learning to follow image editing instructions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. Text-guided image editing model; often struggles with localized compositional edits.

[32] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL `https://arxiv.org/abs/2506.15742`.

[33] Black Forest Labs. Flux.1 kontext [dev] — 12 b open-weight model for image editing. Hugging Face model card, 2025. Developer-focused open-weight version with local/global editing and multi-turn robustness.

[34] Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, et al. A survey on mechanistic interpretability for multi-modal foundation models. *arXiv preprint arXiv:2502.17516*, 2025.