SPATIAL REASONING IS NOT A FREE LUNCH: A CONTROLLED STUDY ON LLAVA

Nahid Alam[1,†], Leema Krishna Murali[1,5,†], Siddhant Bharadwaj[2], Patrick Liu[3,†], Timothy Chung[4,1], Drishti Sharma[1], Akshata A.[1,†], Kranthi Kiran[1,6], Wesley Tam[6,†], Bala Krishna S Vegesna[7]

[1]Cohere Labs Community, [2]Indian Institute of Science, Bangalore, [3]UIUC, [4]Imperial College London, [5]Eisai Inc., [6]EleutherAI, [7]Georgia Institute of Technology

† Work done as part of the EleutherAI SOAR Program
nahid.m.alam@gmail.com

ABSTRACT

Vision-language models (VLMs) have advanced rapidly, yet they still struggle with basic spatial reasoning. Despite strong performance on general benchmarks, modern VLMs remain brittle at understanding 2D spatial relationships such as relative position, layout, and counting. We argue that this failure is not merely a data problem, but is closely tied to dominant design choices in current VLM pipelines: reliance on CLIP-style image encoders and the flattening of images into 1D token sequences with 1D positional encoding. We present a controlled diagnostic study within the LLaVA framework to isolate how these choices affect spatial grounding. We evaluate frontier models and LLaVA variants on a suite of spatial benchmarks, comparing CLIP-based encoders against alternatives trained with denser or generative objectives, as well as variants augmented with 2D positional encoding. Our results show consistent spatial performance gaps across models, and indicate that encoder objectives and positional structure shape spatial behavior, but do not fully resolve it.

1 BACKGROUND AND MOTIVATION

Modern vision-language models (VLMs) almost universally rely on large pre-trained image encoders such as CLIP and SigLIP (Dosovitskiy et al., 2021; Radford et al., 2021; Sun et al., 2023; Oquab et al., 2024; Zhai et al., 2023; Tschannen et al., 2025). These encoders are trained primarily to align global image representations with text and are then integrated into systems such as Flamingo, LLaVA, BLIP-2, KOSMOS, Florence-2, and Molmo (Alayrac et al., 2022; Liu et al., 2023c;b; Li et al., 2023; Peng et al., 2023; Pan et al., 2023; Xiao et al., 2024; Deitke et al., 2024). While this paradigm has driven progress on captioning and VQA, it does not explicitly optimize for structured spatial representations. Prior work shows that CLIP-style encoders emphasize semantic alignment while underperforming on fine-grained and spatially grounded tasks (Tong et al., 2024b; Anis et al., 2025). Recent encoders introduce denser or generative objectives (Oquab et al., 2024; Maninis et al., 2025; Tschannen et al., 2025; Fini et al., 2024), but their impact on spatial grounding in VLMs remains underexplored.

Beyond the encoder, multimodal alignment introduces a second structural bottleneck. Most VLMs flatten images into 1D token sequences before applying 1D rotary positional encoding (Su et al., 2023), collapsing 2D structure during fusion. Recent analyses argue that this design undermines spatial reasoning even when strong visual features are available (Zhang et al., 2025a;b). While Qwen2-VL introduces multimodal rotary embeddings that preserve height and width information (Wang et al., 2024), systematic evidence isolating the role of positional structure remains limited.

At the evaluation level, spatial reasoning is rarely treated as a first-class capability. Foundational VLMs are primarily reported on general benchmarks, while spatial understanding is often omitted (Alayrac et al., 2022; Liu et al., 2023c; Li et al., 2023; Xiao et al., 2024). Meanwhile, several bench-

marks and datasets now explicitly target spatial reasoning, including MMVP, CV-Bench, GQA, VSR, TopViewRS, TallyQA, and CountBenchQA (Tong et al., 2024b;a; Hudson & Manning, 2019; Liu et al., 2023a; Li et al., 2024b; Acharya et al., 2019; Beyer et al., 2024), as well as training-driven efforts such as RoboSpatial, SpatialVLM, MM-Spatial, and SpatialRGPT (Song et al., 2025; Chen et al., 2024; Daxberger et al., 2025; Cheng et al., 2024). However, these efforts often conflate data, scale, and architecture, making it difficult to isolate which design factors shape spatial behavior.

In this work, we focus on static 2D spatial reasoning and present a controlled diagnostic study within the LLaVA framework. We evaluate frontier VLMs and systematically vary two under-examined design dimensions: the image encoder objective and the positional structure used during multimodal alignment, introducing 2D rotary positional encoding. This setting allows us to isolate how these factors influence spatial grounding and to assess the extent to which they account for observed spatial failures.

2 Experimental Setup

2.1 Methods

All experiments are conducted within the LLaVA (Liu et al., 2023c) framework. We construct controlled LLaVA-1.5 (7B) variants by swapping the image encoder while holding the language backbone and training pipeline fixed. Specifically, we compare CLIP, SigLIP, SigLIP2, and AIMv2, and evaluate each with and without 2D rotary positional encoding (2D-RoPE).

Unlike standard 1D RoPE, 2D-RoPE encodes both horizontal and vertical patch indices and is applied to query and key projections during multimodal attention. This preserves explicit 2D structure during image–text fusion and allows us to isolate the effect of positional structure on spatial grounding.

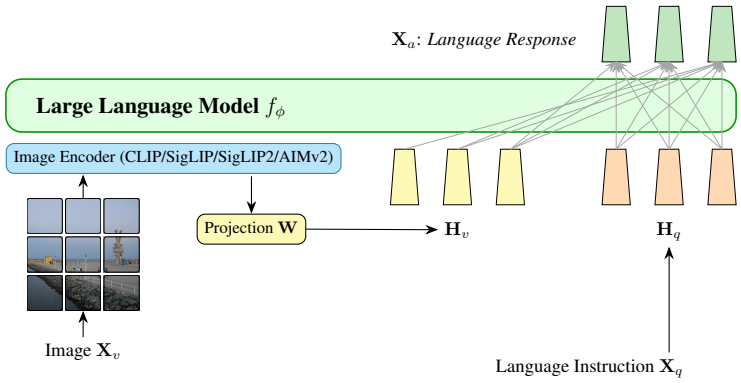

Figure 1: Our experimental approach with LLaVA Framework Liu et al. (2023c) that compares the performance of different image encoders and 2D-RoPE variants.

2.2 Training

Models are trained using the standard two-stage LLaVA recipe: projection pretraining followed by full instruction tuning, using the original LLaVA datasets. Images are resized to 256×256. Pretraining updates only the projection layer, while instruction tuning updates all parameters. Full training and optimization details are provided in the appendix.

3 Results

We evaluate frontier multimodal models and controlled LLaVA-1.5 (7B) variants across spatial reasoning benchmarks. All encoder-swapped models and their 2D-RoPE counterparts are trained on the same 7B LLaVA backbone for controlled comparison. Frontier models between 2B and 8B parameters include LLaVA-NeXT (Liu et al., 2024), LLaVA-OneVision-qwen2-7B-ov-hf (Li et al., 2024a),

Table 1: Comparison of frontier models and LLaVA variants across spatial understanding benchmarks. Values underlined indicate the best-performing frontier model; values in **bold** indicate the best-performing LLaVA variant.

Models	MMVP	CV-Bench 2D Overall	TallyQA	GQA Overall	VSR	Top-ViewRS	Count-BenchQA
LLaVA-NeXT	0.667	0.606	0.733	63.786	63.994	0.409	0.515
LLaVA-OneVision	0.767	0.730	0.797	62.140	77.741	0.414	0.823
Qwen2.5-VL	0.770	0.754	0.800	60.391	89.116	0.456	0.891
SmolVLM2	0.687	0.577	0.729	50.574	71.277	0.416	0.692
Gemma3-4b-it	0.708	0.659	0.525	31.277	55.074	0.334	0.713
PaliGemma	0.667	0.624	0.794	62.570	65.139	0.322	0.674
Molmo	0.753	0.728	0.808	55.295	76.432	0.323	0.858
LLaVA v1.5	0.577	0.490	0.707	33.225	55.810	0.384	0.468
LLaVA-2D-RoPE	0.513	0.443	0.654	34.433	57.201	0.283	0.290
LLaVA-SigLIP	0.433	0.412	0.672	25.648	54.910	0.349	0.581
LLaVA-SigLIP-2D-RoPE	0.507	0.425	0.616	**38.448**	57.692	0.295	0.483
LLaVA-SigLIP2	0.427	0.442	0.684	23.970	52.701	**0.371**	0.532
LLaVA-SigLIP2-2D-RoPE	0.480	0.415	0.646	34.560	56.465	0.330	0.402
LLaVA-AIMv2	0.513	**0.466**	**0.710**	32.541	56.219	0.339	**0.739**
LLaVA-AIMv2-2D-RoPE	**0.560**	0.432	0.690	32.342	**60.311**	0.338	0.719

Qwen2.5-VL-8B (Bai et al., 2025), SmolVLM2-2.2B-Instruct (Marafioti et al., 2025), Gemma3-4b-it (Team et al., 2025), PaliGemma2-3b-mix-448 (Beyer et al., 2024), and Molmo-7B-D-0924 (Deitke et al., 2024).

3.1 Frontier models still fail on spatial reasoning

Table 1 shows that Qwen2.5-VL is the strongest frontier model overall, leading on CV-Bench 2D Overall, MMVP, VSR, TopViewRS, and CountBenchQA, while LLaVA-NeXT leads frontier models on GQA Overall and Molmo leads on TallyQA. Despite these gains, spatial performance remains uneven across tasks and models, indicating that spatial grounding is not consistently captured by general-purpose training and scaling.

3.2 Encoder choice dominates spatial performance in controlled LLaVA variants

Within the controlled LLaVA setting, we observe that the choice of image encoder strongly shapes spatial behavior. LLaVA-AIMv2 yields the most consistent improvements over the CLIP-based LLaVA baseline, achieving the best LLaVA scores on CV-Bench 2D Overall, TallyQA, and CountBenchQA. LLaVA-AIMv2-2D-RoPE further improves MMVP and achieves the best LLaVA score on VSR. In contrast, SigLIP- and SigLIP2-based variants show more fragmented gains, with LLaVA-SigLIP-2D-RoPE performing best on GQA Overall and LLaVA-SigLIP2 performing best on TopViewRS.

Overall, these results suggest that spatial failures in VLMs are strongly coupled to the visual backbone: encoders optimized primarily for global image–text alignment do not reliably support spatial reasoning, while encoders trained with denser or generative objectives can improve spatial outcomes within the same multimodal framework.

3.3 Qualitative localization reveals missing spatial grounding

Quantitative results suggest that encoder choice strongly affects spatial behavior. To further examine this, we qualitatively compare LLaVA-SigLIP2 and LLaVA-AIMv2 on object localization prompts. Figure 2 shows representative examples where models are asked to localize visually grounded regions and output bounding boxes.

(a) SigLIP2

(b) AIMv2

(c) Prompt: *Locate the cup that contains green liquid. Provide the bounding boxes.*

(d) SigLIP2

(e) AIMv2

(f) Prompt: *Locate the pieces of green onions in the ramen bowl. Provide the bounding box coordinates given the image size is 550×550.*

Figure 2: Object localization in LLaVA-SigLIP2 vs. LLaVA-AIMv2. AIMv2 yields tighter and more spatially aligned localizations, while SigLIP2 often produces imprecise or misaligned boxes.

Figure 3: Example image from LLaVA-Bench (In-the-Wild) (Liu et al., 2023c).

Table 2: Model outputs for the prompt *Are the chopsticks to the left or right of the bowl?* on image shown in Figure 3.

Model	Output
LLaVA-v1.5	The chopsticks are to the right of the bowl.
LLaVA-2D-RoPE	The chopsticks are to the right of the bowl.
LLaVA-SigLIP	The chopsticks are to the right of the bowl.
LLaVA-SigLIP-2D-RoPE	The chopsticks are to the right of the bowl.
LLaVA-SigLIP2	The chopsticks are to the right of the bowl.
LLaVA-SigLIP2-2D-RoPE	Right
LLaVA-AIMv2	The chopsticks are to the right of the bowl.
LLaVA-AIMv2-2D-RoPE	The chopsticks are to the right of the bowl.
Qwen2.5-VL	The chopsticks are to the right of the bowl.
Gemma3-4b-it	**The chopsticks are to the left of the bowl.**

Across examples, LLaVA-AIMv2 produces tighter and more accurate localizations, while LLaVA-SigLIP2 frequently outputs imprecise or spatially misaligned boxes. These failures are consistent with the quantitative trends on CV-Bench and VSR, and suggest that encoders trained with dense or generative supervision better preserve spatial detail needed for grounded perception. In contrast, encoders optimized primarily for global alignment struggle to support region-level reasoning, even when paired with the same language model and training pipeline.

3.4 2D POSITIONAL STRUCTURE HELPS, BUT DOES NOT RESOLVE SPATIAL GROUNDING

We also evaluate the impact of 2D-RoPE across encoder variants. Improvements are mixed: 2D-RoPE helps in some settings (e.g., AIMv2 on MMVP and VSR; SigLIP on GQA Overall), but degrades others (e.g., CV-Bench 2D Overall for AIMv2, and several tasks for the CLIP baseline). These results indicate that preserving 2D positional structure alone is not sufficient; spatial grounding depends jointly on the visual features learned by the encoder and the positional structure used during multimodal fusion.

4 CONCLUSION

Despite rapid progress in VLMs, spatial reasoning remains fragile and inconsistent. Even frontier models vary widely across spatial benchmarks, and strong performance on general multimodal tasks does not reliably translate into spatial grounding. In a controlled LLaVA setting, we show that architectural choices—particularly the image encoder objective and multimodal positional structure shape spatial behavior. Encoders trained with denser or generative supervision improve spatial performance, while 2D positional structure alone is insufficient. Overall, our results suggest that spatial

reasoning is an under-addressed design dimension in modern VLMs. We hope this diagnostic study encourages treating spatial representation as a first-class concern in VLM design and evaluation.

5 FUTURE WORK

Our study focused on static, 2D images, benchmarks, and encoder variants within the LLaVA framework. This work can extend to 3D spatial reasoning in conjunction with the dynamic environment. Another potential extension can be on SigLIP2 with NaFlex. The flexible resolution image preprocessing of NaFlex mitigates the information loss observed in fixed-resolution encoders. In terms of visual backbones, incorporating DINOv2 in LLaVA is left out for future work. Future work will analyze why 2D-RoPE sometimes degrades performance by adding attention-based diagnostics and testing for potential conflicts with the LLM's positional encoding. We will introduce spatial probing of encoder and fused representations to identify what spatial information is present and where it is lost in the pipeline. Another area is to explicitly study image resolution as a confounder by running controlled experiments beyond the current 256×256 setting and reporting its independent effect on spatial reasoning.

6 ACKNOWLEDGMENT

We thank EleutherAI for their generous GPU support. We also thank our mentor, Jekaterina Novikova, for asking critical questions that strengthened this project.

REFERENCES

Manoj Acharya, Kushal Kafle, and Christopher Kanan. TallyQA: Answering Complex Counting Questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8076–8084, 2019. doi: 10.1609/aaai.v33i01.33018076. URL https://ojs.aaai.org/index.php/AAAI/article/view/4815.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Ahmad Mustafa Anis, Hasnain Ali, and Saquib Sarfraz. On the limitations of vision-language models in understanding image transforms, 2025. URL https://arxiv.org/abs/2503.09837.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, 2025. URL https://arxiv.org/abs/2502.13923.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024. URL https://arxiv.org/abs/2407.07726.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities, 2024. URL https://arxiv.org/abs/2401.12168.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. SpatialRGPT: Grounded Spatial Reasoning in Vision Language Models, 2024. URL https://arxiv.org/abs/2406.01584.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and Peter Grasch. MM-Spatial: Exploring 3D Spatial Understanding in Multimodal LLMs, 2025. URL https://arxiv.org/abs/2503.13111.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models. *arXiv preprint arXiv:2409.17146*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander T Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua M. Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders, 2024. URL https://arxiv.org/abs/2411.14402.

Drew A. Hudson and Christopher 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 (CVPR)*, pp. 6700–6709, 2019. doi:

10.1109/CVPR.2019.00686. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer, 2024a. URL https://arxiv.org/abs/2408.03326.

Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topviewrs: Vision-language models as top-view spatial reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1786–1807, Miami, Florida, USA, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.106. URL https://aclanthology.org/2024.emnlp-main.106/.

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 *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual Spatial Reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023a. doi: 10.1162/tacl_a_00566. URL https://aclanthology.org/2023.tacl-1.37/.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, 2023c.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, Dan Gnanapragasam, Mojtaba Seyedhosseini, Howard Zhou, and Andre Araujo. TIPS: Text-Image Pretraining with Spatial awareness, 2025. URL https://arxiv.org/abs/2410.16512.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-G: Generating Images in Context with Multimodal Large Language Models. *ArXiv*, abs/2310.02992, 2023.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding Multimodal Large Language Models to the World. *arXiv preprint arXiv:2306.14824*, 2023.

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 *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics, 2025. URL https://arxiv.org/abs/2411.16537.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. URL https://arxiv.org/abs/2104.09864.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023. URL https://arxiv.org/abs/2303.15389.

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.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. In *Advances in Neural Information Processing Systems*, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9ee3a664ccfeabc0da16ac6f1f1cfe59-Paper-Conference.pdf.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024b. URL https://arxiv.org/abs/2401.06209.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features, 2025. URL https://arxiv.org/abs/2502.14786.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4818–4829, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Huanyu Zhang, Chengzu Li, Wenshan Wu, Shaoguang Mao, Yifan Zhang, Haochen Tian, Ivan Vulić, Zhang Zhang, Liang Wang, Tieniu Tan, and Furu Wei. Scaling and beyond: Advancing spatial reasoning in mllms requires new recipes, 2025a. URL https://arxiv.org/abs/2504.15037.

Wanyue Zhang, Yibin Huang, Yangbin Xu, JingJing Huang, Helu Zhi, Shuo Ren, Wang Xu, and Jiajun Zhang. Why do mllms struggle with spatial understanding? a systematic analysis from data to architecture, 2025b. URL https://arxiv.org/abs/2509.02359.