OpenVision 2 : A Family of *Generative Pretrained* Visual Encoders for Multimodal Learning

Supplementary Material

1. Additional Experimental Results

Table 1. **Generalization across Qwen2.5 [17] backbones.**

LLM Backbone	Vision Encoder	Param	Res.	TextVQA	ChartQA	OCRBench	MME	SEED	SQA	GQA	POPE
Qwen2.5-0.5B-Instruct	OpenVision 2	0.5B	224	41.5	11.0	257	1045/294	55.0	58.5	53.8	86.3
Qwen2.5-1.5B-Instruct	OpenVision 2	1.5B	224	52.5	12.7	292	1401/276	67.8	69.1	59.4	86.5
Qwen2.5-7B-Instruct	OpenAI-CLIP-L [14]	7B	224	54.5	15.3	167	1540/348	55.4	68.6	59.4	84.8
Qwen2.5-7B-Instruct	**OpenVision 2**	**7B**	**224**	**58.1**	**15.8**	**325**	**1533/320**	**66.5**	**69.5**	**62.9**	**87.2**
Llama3-8B [13]	**OpenVision 2**	**8B**	**224**	**59.0**	**13.7**	**327**	**1460/312**	**69.3**	**76.5**	**62.6**	**87.1**

Table 2. **Linear probing on ImageNet-1K and Places365.**

Model	Data	Objective	IN-1K	Places
OpenAI CLIP-L [14]	WIT-400M	Contrastive	82.6	57.2
OpenCLIP-L [5]	DC-1B (Raw) [3]	Contrastive	84.6	58.2
OpenVision 2-L	**DC-1B (ReCap) [7]**	**Generative**	**85.2**	**60.8**

1.1. Generalization Across LLM Backbones

To further evaluate cross-architecture generalization beyond LLaVA-style settings, we integrate OpenVision 2 with the Qwen2.5 family [17]. As shown in Table 1, performance scales consistently with the language model size from 0.5B to 7B. Under the same Qwen2.5-7B backbone, OpenVision 2 substantially outperforms OpenAI-CLIP-L on several multimodal benchmarks, including OCRBench [10] and TextVQA [15], demonstrating robust transfer across LLM backbones.

1.2. Linear Probing on General Vision Tasks

We additionally evaluate OpenVision 2 using linear probing on ImageNet-1K [1] and Places365 [18]. As shown in Table 2, OpenVision 2 improves over OpenCLIP-L on both object recognition and scene recognition using the same underlying image collection. The improvement on Places365 is particularly notable, suggesting that the generative objective with synthetic captions yields strong general-purpose visual representations across diverse domains.

1.3. Architecture and Efficiency Details

OpenVision 2 improves training efficiency through both structural simplification and reduced computation. Compared with the original OpenVision formulation, OpenVision 2 removes the text encoder entirely and retains only an image encoder together with a text decoder. This eliminates the forward and backward cost associated with contrastive text encoding.

In addition, approximately 65% of visual tokens are dropped before being fed into the decoder, which shortens the visual prefix and reduces the quadratic self-attention cost in the decoder. The decoder itself is a standard Transformer initialized from scratch, with scale matched to the vision encoder. It uses simple token concatenation and autoregressive generation, without cross-attention or additional fusion modules.

Unless otherwise specified, the reported pretraining efficiency numbers correspond to the vision encoder pretraining stage only, excluding the downstream LLaVA instruction-tuning stage, which is identical across compared models. Runtime is measured using wall-clock time under the same number of training steps.

1.4. Comparison with Concurrent Methods

We additionally compare OpenVision 2 against SigLIP 2 [16] under the LLaVA-1.5 [9] setting using the SO400m-14-224 vision backbone. As shown in Table 3, OpenVision 2 outperforms SigLIP 2 on SEED [6], TextVQA [15], GQA [4], ChartQA [12], POPE [8], and MME [2]-Cognition, while remaining competitive on OCR [10], ScienceQA [11], and MME [2]-Perception. These results further support that a simple generative-only pretraining recipe can produce strong visual encoders for multimodal transfer.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[2] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li,

Table 3. **Comparison with SigLIP 2 on LLaVA-1.5 using the SO400m-14-224 vision backbone.**

Method	SEED	TextVQA	GQA	OCR	ScienceQA	MME (perception/cognition)	ChartQA	POPE
SigLIP2	69.3	59.1	61.8	**350**	**77.0**	**1483**/265	12.8	87.0
OpenVision2	**69.5**	**59.4**	**62.4**	335	75.4	1433/**314**	**13.9**	**87.1**

Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1

[3] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 1

[4] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[5] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. *github*, 2021. 1

[6] 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

[7] Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024. 1

[8] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 1

[9] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. 1

[10] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), 2024. 1

[11] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1

[12] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, 2022. Association for Computational Linguistics. 1

[13] Meta LLaMA Team. Introducing Meta Llama 3: The most capable openly available LLM to date, 2024. 1

[14] 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 *ICML*, 2021. 1

[15] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 1

[16] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 1

[17] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 1

[18] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1