OpenVision: A Fully-Open, Cost-Effective Family of Advanced Vision Encoders for Multimodal Learning

Xianhang Li* Yanqing Liu* Haoqin Tu Cihang Xie
University of California, Santa Cruz

Project Page: https://ucsc-vlaa.github.io/OpenVision
Model Training: https://github.com/UCSC-VLAA/OpenVision
Model Zoo: click me

1. Appendix

1.1. Ablation w.r.t. Input Resolutions of the Vision Encoder

We present model performance under the LLaVA 1.5 setting with varied input resolutions in Table 1. We draw a similar conclusion as the findings from the LLaVA-Next setting: a higher resolution into the vision encoder during training always help boost model performance on vision-language benchmarks.

Table 1. Ablation study on our OpenVision visual encoder with different input resolutions resulting from the three-stage training pipeline, evaluated under the LLaVA-1.5 setting.

Res.	Text VQA	Chart QA	OCR.	MME	SEED	MMVet	SQA	GQA	POPE
84×84	50.4	12.1	231	1372/290	63.5	28.8	76.6	58.8	83.9
224×224	57.7	13.9	315	1487/317	69.5	35.2	73.6	62.9	86.4
336×336	61.2	15.7	339	1525/315	70.5	36.2	75.1	63.7	87.2

1.2. Visual Encoder Configuration

We present detailed visual encoder configurations in Table 2. We demonstrate the flexibility of our approach by scaling OpenVision up/down and varying patch size for different application scenarios, and by showcasing its adaptability even with very small language models.

Table 2. **Visual encoder configurations** used in our paper.

Model Size	Patch Size	Layers	Width	Heads	#Params (M)
Tiny	16 or 8	12	192	3	5
Small	16 or 8	12	384	6	22
Base	16 or 8	12	768	12	86
Large	14	24	1024	16	303
SoViT-400M [1]	14	27	1152	16	412
Huge	14	32	1280	16	631

1.3. Ablation w.r.t. Learning Rate

We also conduct comprehensive ablations w.r.t. the learning rate and other hyper-parameters during VLLM training. In Table 3 shows that a mid-range learning rate setting of 5×10^{-5} (Stage 2 ViT) and 5×10^{-4} (Stage 3 LLM) achieves the best overall scores—*TextVQA* 33.2, *MME* 743/212, *POPE* 85.0 — whereas overly lower or higher rates degrade accuracy. Careful hyperparameter tuning is essential to maximize performance for practical and extensible multimodal pipelines.

*Equal contribution.

References

[1] Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting vit in shape: Scaling laws for compute-optimal model design. *Advances in Neural Information Processing Systems*, 36:16406–16425, 2023. 1

Table 3. Ablation study on the Stage 2 & Stage 3's learning rate. Results show that both contribute to better performance across multimodal benchmarks.

Stage 2	Stage 3 LLM	Stage 3 ViT	Text VQA	Chart QA	OCR	MME	SEED	MMVet	SQA	GQA	POPE
1e-5	1e-5		26.5	10.5	136	618/242	26.2	16.7	36.5	38.6	72.7
1e-5			32.8	10.2	171	806/213	48.7	16.7	37.8	54.2	85.4
3e-5			33.2	10.6	173	759/215	47.8	17.0	38.1	54.9	84.7
5e-5			33.2	10.3	194	743/212	48.8	15.8	38.2	54.2	85.0
7e-5			32.6	10.2	184	845/205	42.0	14.4	32.8	54.2	85.7
1e-4	5e-4	***Frozen***	32.5	9.4	165	734/211	48.1	14.4	37.8	53.1	85.4
3e-4			29.2	9.2	149	649/205	44.5	14.1	35.5	50.5	83.3
5e-4			25.4	9.8	86	684/205	27.1	10.7	34.9	49.4	81.1
7e-4			23.0	9.2	22	812/210	28.0	14.4	35.0	47.5	79.3
1e-3			22.5	9.2	20	656/206	27.5	11.2	34.3	44.7	77.5
	1e-5		26.1	10.0	147	672/221	24.8	15.8	34.1	39.4	78.2
	3e-5		29.1	10.0	178	769/259	26.9	16.0	35.6	44.2	80.7
	5e-5		29.6	10.0	176	797/240	27.3	15.8	35.7	46.3	82.1
	7e-5		30.4	10.1	185	836/235	27.2	13.9	35.3	47.7	83.3
5e-5	1e-4	***Frozen***	31.7	9.8	185	876/260	27.4	13.9	35.5	9.4	84.3
	3e-4		32.8	10.4	198	717/210	44.5	13.3	36.9	53.2	84.7
	5e-4		33.2	10.3	194	743/212	48.8	15.8	38.2	54.2	85.0
	7e-4		32.4	10.3	191	793/215	49.5	15.1	35.9	54.8	86.3
	1e-3		32.1	10.8	202	808/247	50.2	15.3	31.6	55.4	85.5
5e-5	5e-4	1e-6	21.7	9.3	32	705/223	27.2	12.5	34.9	46.2	79.23
		5e-6	21.7	9.3	31	706/223	27.3	12.8	34.8	46.1	79.2