Are They the Same? Exploring Visual Correspondence Shortcomings of Multimodal LLMs

Yikang Zhou[1*] Tao Zhang[1*] Shilin Xu[3] Shihao Chen[1] Qianyu Zhou[5] Yunhai Tong[3]
Shunping Ji[1†] Jiangning Zhang[4] Lu Qi[1] Xiangtai Li[2‡]
[1]Wuhan University [2]Bytedance Seed [3]Peking University [4] Zhejiang University [5] SJTU
{zhouyik,zhang_tao,jishunping}@whu.edu.cn, xiangtai94@gmail.com
https://zhouyiks.github.io/projects/CoLVA/

A. More Experiment Result

Ablation studies in more detailed results. Here, we present the detailed results of the main ablation experiments, as shown in Tab. 1. The table includes the overall accuracy and accuracy across eight different match types. Our method significantly improves accuracy over a strong baseline (45.83 vs. 32.38) across six match types. The improvement is less pronounced for the size (SZ) match type, where accuracy is approaching saturation (76.62 vs. 74.03).

CoLVA on the other base model. We combine CoLVA into Qwen2VL and test it on several general benchmarks, as shown in Tab. 2. CoLVA still works better.

Analysis on Different Match Types. From detailed results of Tab. 4, MLLMs work better in matching based on object size (SZ), shape (SP), and textual or LOGO markers (TM). These three types require focusing solely on the object itself, indicating that current MLLMs possess proficient object-level perception and understanding. In contrast, MLLMs find it more challenging to match based on object relative position (RP), object orientation and movement (OO), and binding relationships with other objects (BR). These require MLLMs to understand the interrelationships between objects and infer information that remains invariant across time and space.

CoLVA Failure Cases Analysis. We have observed that CoLVA tends to fail when performing matching in densely populated object scenarios, as illustrated in Fig. 1. One reason for this is that CoLVA is prone to hallucinations regarding the query object in multi-object, multi-image contexts. For instance, in the left example of Fig. 1, CoLVA correctly identifies the query object as a player. However, in the second image, it mistakenly hallucinates object-7, which is actually a horse, as the matched player. Additionally, in multi-view scenarios, CoLVA is susceptible to incorrectly matching another object based on partial information of the query object from a single viewpoint.

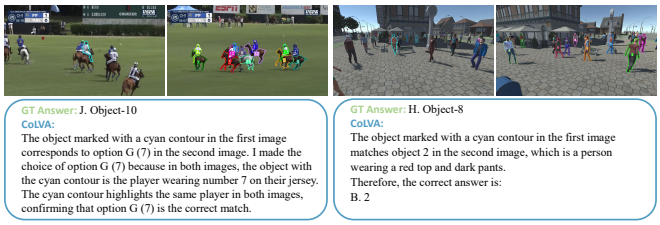

Figure 1. The failure cases of CoLVA on MMVM benchmark. CoLVA tends to fail when performing matching in densely populated object scenarios.

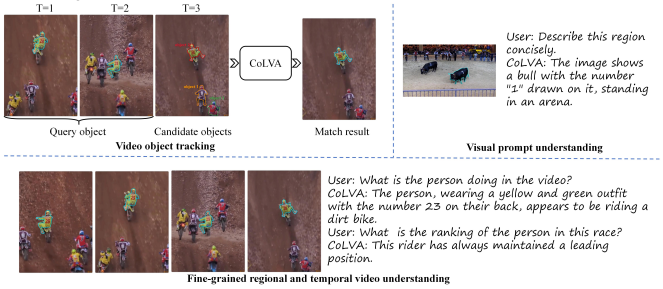

Figure 2. Potential real-world applications of CoLVA

B. More information on MMVM Benchmark

The MMVM benchmark is composed of the validation split from the video segmentation datasets (790 samples) and **manually collected** internet videos (720 samples). Additionally, the benchmark is **not** generated using the automated annotation pipeline employed for the training set, as it only requires matching results without the need for reasoning processes.

We categorize the 790 samples as the in-domain part, and the 720 samples as the out-domain part. Tab. 3 displays the test results of several methods on these two parts, which revealing that our CoLVA model achieves a significant gain in the out-domain segment (41.67 vs 13.89), thereby demonstrating its robust generalization capability.

*Equal contribution. †Corresponding author. ‡Project leader.

Table 1. The effectiveness of our methods and MMVM data with detailed results. Data denotes using the combination of MMVM data and LLaVA SFT data. OCL denotes object-level contrastive learning. VE denotes fine-grained vision expert. IA denotes instruction augmentation. OA denotes the overall accuracy.

Data	OCL	VE	IA	OA	CL	SP	TM	SZ	RP	OO	BR	OM
				17.62	14.73	34.48	17.76	15.58	10.28	24.00	31.25	21.30
✓				32.38	25.04	24.14	32.71	74.03	19.00	35.20	43.18	36.57
✓	✓			34.05	25.78	26.77	31.97	75.01	22.32	35.29	42.98	37.51
✓		✓		32.25	24.22	27.59	31.78	68.83	19.14	35.20	40.34	39.35
✓	✓	✓		40.45	33.72	44.85	39.37	75.33	30.00	48.00	38.65	44.78
✓	✓	✓	✓	45.83	38.30	31.03	41.12	76.62	41.71	51.20	39.77	46.76

Table 2. The impact of Qwen2VL-CoLVA on general benchmarks.

MLLM	CoLVA	MME perception	MME reasoning	POPE Overall	BLINK Overall
Qwen2VL-2B	×	1471.10	404.64	86.83	44.50
	✓	1540.14	418.57	88.01	46.98

Table 3. The split of MMVM benchmark.

Method	Total	In-domain split	Out-domain split
GPT4o	42.65	46.46	38.47
InternVL2-4B	17.62	21.01	13.89
CoLVA-4B	49.87	57.22	41.67

C. Potential real-world applications of CoLVA

Object matching is fundamental to many real-world applications, such as video object tracking, re-identification (ReID), multi-image visual question answering (VQA), and video VQA. Our CoLVA also integrates visual prompt understanding capabilities. In Fig. 2, we showcase several real-world applications.

D. More Implementation Details

More training details. Our model comprises three components: a pre-trained MLLM InternVL2-4B [2], a fine-grained vision expert RADIO [20], and a RADIO adapter. We adopt Xtuner [3] codebase to implement our method. We maintain the original architecture of both InternVL2-4B and RADIO, while the RADIO adapter is implemented using a two-layer MLP. Our training includes two stages: pre-training and supervised fine-tuning (SFT). We freeze the MLLM and RADIO during the pre-train stage, focusing solely on training the RADIO Adapter. During the SFT stage, we freeze the RADIO, the RADIO adapter, and all components of InternVL2-4B except the LLM. The LLM of the MLLM is trained by applying LoRA [5].

During the pre-training phase, we sample 500k images with segmentation labels from SA1B [6]. For each image, we apply augmentations such as Crop, Resize, Flip, and Rotation to simulate a pseudo video. We then sample two frames from this pseudo video to serve as our training samples. Taking InternVL2 [2] as the base model and RADIO [20] as the vision expert, we input one image into the InternVL2 visual encoder and the other into RADIO. When selecting the (anchor, positive, negatives) triplet, the anchor is chosen from the image features output by RADIO, while the positive and negatives are selected from the image features output by the InternVL2 visual encoder. We perform full training from scratch on the RADIO adapter using only object-level contrastive loss.

In the fine-tuning phase, we apply instruction augmentation to the original 220k MMVM data samples using object-level representations. Consequently, we utilize a total of 440k MMVM data samples during fine-tuning. When using Qwen2VL [25] as the base model, to reduce sequence length and decrease computational resource requirements, we scale the long edge of all images to 1024 pixels and pad the short edge to 1024 pixels.

Inference details. When performing inference on the MMVM benchmark, we integrate CoLVA into the MLLMs. For inference on general VQA benchmarks, we maintain the MLLMs' original architecture and load the LLM parameters trained with CoLVA.

E. More visualization results

More PCA visualizations. In Fig. 3, we present additional PCA visualizations. The results reveal that the matched target (represented by a red dot) and other candidate objects (represented by blue dots) are clustered together, while being distant from the query object (represented by a red star). This clustering pattern makes it challenging for InternVL2 to distinguish the correct object. In contrast, our CoLVA brings the matched target and the query object closer together while distancing them from other candidate objects. This indicates that our CoLVA has learned fine-grained and discriminative visual features, which are beneficial for visual matching tasks.

More challenging test cases of our MMVM. Here, we present more examples from the MMVM benchmark,

which features diverse scenes and presents significant challenges, as illustrated in Fig. 4. In particular, our MMVM contains extremely small objects.

F. Further Discussion

Future works. We have argued the fine-grained visual perception and logical reasoning ability of MLLMs in the main paper. We give a more detailed description here.

The former means the MLLMs must understand various scale objects well, where detailed information, such as object parts, remote objects, and thin objects, play a critical role in perception. Thus, equipping MLLMs with dense perception ability and visual prompts [7, 11, 21, 30, 31] is needed.

The latter means that MLLMs must have instance-aware understanding and can perform visual comparisons [19]. With this ability, MLLMs can distinguish various objects and perform visual reasoning. This is why we adopt contrastive loss during the pre-training stage.

In addition, automatically collecting more high-quality supervised fine-tuning data is another way to boost MLLMs.

Board impact. Our works explore one fundamental limitation of current SOTA MLLMs: visual correspondence shortcomings. We present a new benchmark: MMVM, a training dataset, and a new training framework, CoLVA, to improve the visual correspondence in MLLM models. Our work will raise the attention of visual correspondence in MLLM design and inspire research on cross-image VQA tasks and fine-grained VQA tasks.

Table 4. More MMVM Benchmark results. Accuracy is the metric, and the overall accuracy is computed across all 1,510 evaluation samples. The accuracy for each of the eight match types is calculated separately on their respective samples. The full term of the match type abbreviation can be found in the main text. For MLLMs that only support single-image input, we simply concatenate all the images vertically into one image and then input it.

Model Size	Method	Overall	CL	SP	TM	SZ	RP	OO	BR	OM
∼4B	InternVL2-2B [2]	9.87	9.66	6.90	10.28	10.39	8.28	11.20	10.80	8.80
	xGen-MM-v1.5-4B [26]	13.50	10.47	17.24	18.69	25.97	6.71	19.20	17.61	16.20
	VILA1.5-3B[14]	15.36	10.96	6.89	19.62	29.87	9.57	20.80	19.30	18.98
	Qwen2-VL-2B-Instruct [25]	15.69	13.42	20.69	17.75	31.16	9.57	22.40	18.75	16.67
	Ovis1.6-Llama3.2-3B [18]	16.62	13.09	20.69	20.56	33.77	9.28	22.40	21.59	20.83
	DeepSeek-VL-1.3B [17]	16.82	12.60	13.79	18.69	37.66	10.43	22.40	21.59	17.59
	InternVL2-4B [2]	17.62	14.73	34.48	17.76	15.58	10.28	24.00	31.25	21.30
4B∼13B	Chameleon-7B [23]	10.07	9.49	17.24	14.95	11.69	6.86	9.60	13.07	10.65
	Cambrian-13B [24]	10.72	9.32	6.89	9.34	23.37	6.28	16.00	15.34	7.87
	Mini-Gemini-7B-HD [12]	13.18	10.80	10.34	14.95	25.97	8.28	14.40	18.18	13.89
	LLaVA-NEXT-13B [16]	13.77	8.35	10.34	10.28	22.08	7.57	22.4	22.73	18.52
	LLaVA1.5-13B [15]	14.04	11.78	13.79	14.02	31.17	7.57	20.00	18.18	14.35
	MiniCPM-V2.5-8B [27]	14.11	10.80	17.24	13.08	31.17	6.28	24.00	20.45	17.13
	Monkey-7B [13]	14.43	13.09	6.89	14.01	31.16	7.85	17.60	18.18	15.74
	VILA1.5-13B [14]	14.70	13.91	13.79	13.08	36.36	7.57	22.40	17.04	15.74
	Slime-13B [32]	14.83	11.29	6.89	16.82	32.46	9.00	18.40	21.02	17.59
	mPLUG-Owl3-7B [28]	16.22	14.07	20.68	16.82	31.16	8.57	20.80	20.45	19.90
	InternVL2-8B [2]	16.89	13.58	20.69	22.43	24.68	11.57	24.00	23.30	18.52
	VITA-8*7B [4]	17.42	14.57	13.79	23.36	29.87	10.57	24.80	22.16	20.37
	DeepSeek-VL-7b [17]	17.68	14.24	17.24	20.56	35.06	10.00	22.40	25.00	23.61
	Ovis1.6-Gemma2-9B [18]	17.75	17.68	17.24	15.89	32.47	12.14	20.00	19.32	18.98
	LLaVA-Next-Interleave-7B [10]	19.34	15.88	41.38	15.89	41.56	10.71	19.20	23.30	27.78
	LLaVA-OneVision-ov-7B [9]	20.92	16.69	17.24	25.23	31.16	14.28	22.40	30.68	25.92
	Qwen2-VL-7B-Instruct [25]	27.48	24.87	37.93	30.84	62.33	17.85	28.00	28.97	31.94
13B∼40B	Yi-VL-34B [29]	11.26	9.49	17.24	18.69	12.99	7.57	9.60	15.34	11.57
	Eagle-X5-34B-Chat [22]	13.84	10.47	13.79	13.08	27.27	7.86	23.20	18.18	14.81
	LLaVA-Next-34B [16]	15.03	11.29	20.69	16.82	32.47	8.71	21.6	19.89	17.13
	VILA1.5-40B [14]	15.36	14.73	20.69	14.95	36.36	5.00	22.40	18.18	17.13
	InternVL2-40B [2]	26.03	24.88	41.38	33.64	42.86	16.86	31.20	31.82	31.02
40B∼	Idefics-80B-instruct [8]	13.58	11.13	13.79	14.95	24.68	7.00	20.80	17.61	13.89
	InternVL2-76B [2]	25.83	24.06	31.03	30.84	40.26	19.28	31.20	30.11	31.02
	LLaVA-OneVision-ov-72B [9]	29.34	28.48	34.48	26.17	55.84	21.14	28.00	34.66	32.41
	InternVL2.5-78B [1]	36.42	35.02	37.93	38.32	58.44	25.86	38.40	39.20	43.98
	Qwen2-VL-72B-Instruct [25]	38.08	37.64	44.83	42.06	64.94	32.28	36.00	35.80	39.81
Unkown	Claude3-5V-Sonnet	40.20	34.21	41.38	56.07	77.92	34.86	40.00	32.39	40.28
	GeminiPro1-5	40.73	36.00	44.83	44.86	74.02	35.14	44.80	38.07	38.42
	GPT4o-20240806	42.65	39.28	**65.52**	**60.75**	67.53	32.28	44.00	43.18	50.00
2B	CoLVA-Qwen2VL-2B (Ours)	47.48	40.92	31.03	47.66	68.83	**50.57**	49.60	33.52	38.42
4B	CoLVA-InternVL2-4B (Ours)	49.80	**43.21**	41.38	45.79	77.92	44.43	**53.60**	44.89	**53.24**
7B	CoLVA-Qwen2VL-7B (Ours)	**51.06**	42.72	37.93	49.53	**80.52**	46.43	52.80	**47.73**	49.54

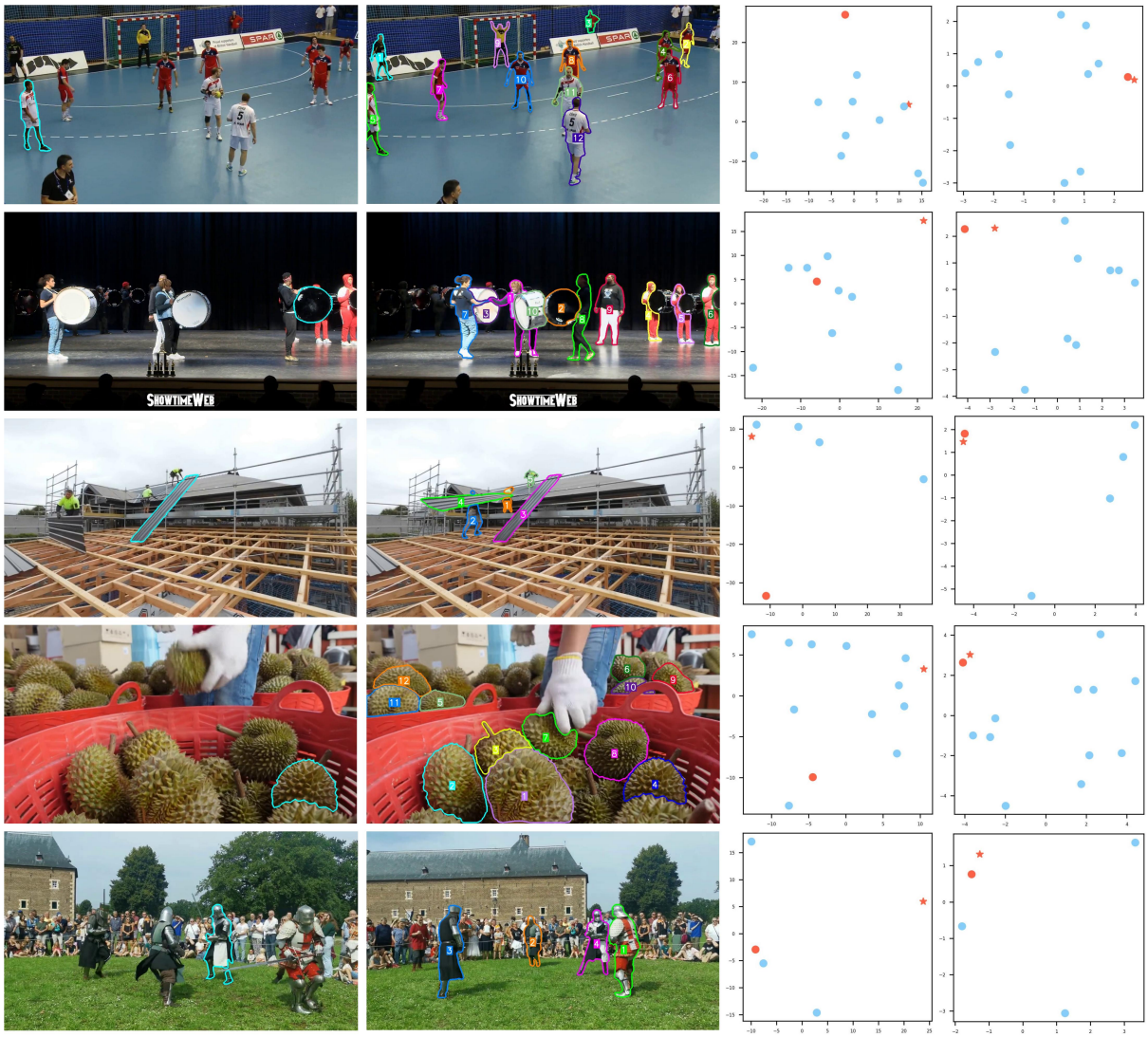

Figure 3. More PCA visualizations of learned object embeddings by InternVL2-4B and our CoLVA-4B. The object embeddings are obtained by applying average pooling to the visual tokens using mask annotations. The red star represents the query object in the first image. The red dot represents the matched target in the second image. The blues dots represent other candidates.

Figure 4. More challenging test cases of our MMVM benchmark, where each row shows cases of different match types.

References

[1] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 4

[2] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 2, 4

[3] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner, 2023. 2

[4] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024. 4

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[7] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 3

[8] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. In *NeurIPS*, 2024. 4

[9] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 4

[10] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 4

[11] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *CVPR*, 2024. 3

[12] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2023. 4

[13] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *CVPR*, 2024. 4

[14] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, 2024. 4

[15] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 4

[16] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 4

[17] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 4

[18] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024. 4

[19] Liang Peng, Junyuan Gao, Xinran Liu, Weihong Li, Shaohua Dong, Zhipeng Zhang, Heng Fan, and Libo Zhang. Vasttrack: Vast category visual object tracking. *arXiv preprint arXiv:2403.03493*, 2024. 3

[20] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *CVPR*, 2024. 2

[21] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[22] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024. 4

[23] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024. *arXiv preprint arXiv:2405.09818*, 2024. 4

[24] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 4

[25] 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. 2, 4

[26] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 4

[27] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu

Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 4

[28] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024. 4

[29] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024. 4

[30] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *CVPR*, 2024. 3

[31] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. In *NeurIPS*, 2024. 3

[32] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024. 4