Multi-Layer Visual Feature Fusion in Multimodal LLMs: Methods, Analysis, and Best Practices

Supplementary Material

1. Visualization of Cross-Attention Modules

Fig. 1 illustrates the architectures of different Internal Modular Fusion used in our experiments. The "Gated Xatten" module, adapted from [1], is introduced as a new component in these architectures. Based on where the "Gated Xatten" module is inserted, we identify three distinct architectures: pre-cross, post-cross, and parallel. These architectures are further compared in terms of their efficiency in fusing multi-layer visual features, as detailed in Tab. 2 in the main paper.

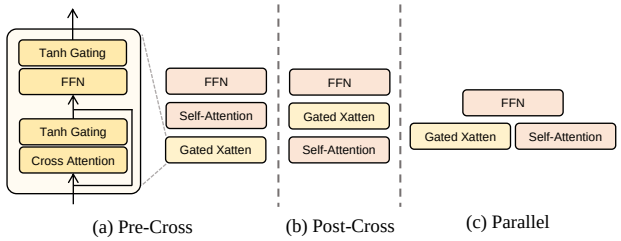

Figure 1. Architecture Comparisons between three Current Internal Modular Fusion Strategies.

2. Training Datasets and Evaluation Benchmarks

2.1. Training Datasets

For the training data, we utilize three main datasets:

1. The first dataset is used for pretraining. This dataset comprises a subset of 558K LAION-CC-SBU [22, 23, 26] image-text pairs with BLIP-generated captions [11], which is the same as the first stage of the LLaVA-1.5 [13] pre-training.
2. The second and third datasets are used for instruction tuning. The second dataset, which is the primary fine-tuning dataset used in most of our experiments, consists of a 665K instruction-following data summarized by LLaVA-1.5. The third part, derived from Cambrian-1 [29], builds upon the 665K instruction-following dataset of LLaVA-1.5 by adding a small number of OCR and chart data. The detailed composition of these datasets can be found in Tab. 1.

2.2. Evaluation Benchmarks

To conduct a comprehensive evaluation of MLLMs across different configurations, we have prepared seven distinct benchmarks, categorized into four types: *General*, *OCR*, *CV-Centric*, and *Hallucination*. The capabilities evaluated by each category are as follows:

- **General**: Evaluates the general capabilities of multimodal models, including cognition and perception. Benchmarks in this category include: GQA [6], MM-Bench (MMB) [16], and MME [4], which is further divided into MME Cognition (MME^C) and MME Perception (MME^P).
- **OCR**: Evaluates the model's performance in text recognition and understanding tasks. Benchmarks in this category include: TextVQA [28] and OCRBench [15].
- **CV-Centric**: This category focuses on better evaluating visual representations in an integrated multimodal setting. Benchmarks in this category include: CV-Bench [29], which is further divided into CV-Bench^{2D} and CV-Bench^{3D}.
- **Hallucination**: Evaluates the model's ability to generate accurate and truthful information, avoiding hallucinations. The benchmark in this category is: POPE [12].

Table 1. The mixture detail of fine-tuning dataset for LLaVA-1.5 665K and Cambrian-1 737K.

Data	Size
LLaVA [14]	158K
+ ShareGPT [25]	40K
+ VQAv2 [5]	83K
+ GQA [6]	72K
+ OKVQA [19]	9K
+ OCRVQA [21]	80K
+ A-OKVQA [24]	66K
+ TextCaps [27]	22K
+ RefCOCO [8, 18]	48K
+ VG [10]	86K
LLaVA-1.5 Total	665K
+ AI2D [9]	16K
+ DocVQA [20]	15K
+ DVQA [7]	13K
Cambrian-1 Total	737K

3. More Detail about Results

In Section 5, we present the performance differences of various fusion strategies under the *Triple* layer selection set when dealing with different data scales and model components, as shown in Fig. 5 and Fig. 6 in the main paper. To provide a more comprehensive understanding of the performance differences, we include the complete evaluation results in Tab. 2 and Tab. 3.

Table 2. Comparison on Different Training Datasets. **Note:** E, I, D, and M represent External Fusion, Internal Fusion, Direct Fusion, and Modular Fusion, respectively.

FT	PT+IT	General				OCR		CV-Centric		Hallu	Avg.
		GQA	MMB	MME^C	MME^P	TextVQA	OCRBench	$CVBench^{2D}$	$CVBench^{3D}$	POPE	
E + D	558k + 332k	56.04	49.14	224	1126	34.56	266	43.96	50.00	85.91	47.83
	558k + 665k	59.19	53.78	238	1141	38.35	256	42.05	50.50	86.33	49.18
	558k + 737k	59.73	52.84	208	1202	39.21	285	41.53	50.58	86.71	49.47
E + M	558k + 332k	54.90	50.77	243	1055	34.16	250	45.53	51.58	86.01	47.90
	558k + 665k	58.43	52.66	225	1173	36.25	262	45.78	52.83	86.03	49.44
	558k + 737k	58.82	51.11	241	1211	37.19	280	44.11	51.92	86.78	49.85
I + D	558k + 332k	55.06	52.14	238	1046	33.74	219	48.77	55.83	86.04	48.39
	558k + 665k	58.59	47.47	223	1207	36.24	255	41.87	53.08	85.87	48.54
	558k + 737k	58.09	49.66	244	1188	37.05	272	43.52	50.50	86.13	49.56
I + M	558k + 332k	52.26	43.56	234	1027	31.12	236	43.01	48.08	84.96	45.24
	558k + 665k	57.56	49.66	212	1163	34.06	255	38.66	47.42	84.69	46.91
	558k + 737k	58.09	51.11	241	1172	35.08	272	46.96	47.75	86.09	49.00

Table 3. Comparison on Different MLLM Components.

FT	Component	General				OCR		CV-Centric		Hallu	Avg.
		GQA	MMB	MME^C	MME^P	TextVQA	OCRBench	$CVBench^{2D}$	$CVBench^{3D}$	POPE	
E + D	Baseline	59.19	53.78	238	1141	38.35	256	42.05	50.50	86.33	49.18
	+SigLIP	60.69	53.26	219	1245	45.98	302	41.31	57.92	86.84	51.76
	+MobileLLaMA 2.7B	61.39	59.28	238	1293	42.54	280	45.50	55.17	87.41	52.63
I + D	Baseline	58.59	47.47	223	1207	36.24	255	41.87	53.08	85.87	48.54
	+SigLIP	59.32	54.73	230	1162	41.66	272	42.91	55.33	86.02	50.45
	+MobileLLaMA 2.7B	60.56	58.33	235	1283	39.94	266	39.06	54.42	86.62	51.01
I + M	Baseline	57.56	49.66	212	1163	34.06	255	38.66	47.42	84.69	46.91
	+SigLIP	50.82	45.45	262	1029	16.43	134	44.94	52.67	81.52	43.27
	+MobileLLaMA 2.7B	58.10	50.77	246	1233	33.09	246	46.72	50.00	86.22	49.10

We further explored different layer combinations and experimented with a larger language model (Vicuna 1.5 7B [3]). As shown in Tab. 4, fusing visual layer {3} alone yields lower performance than fusing visual layer {18}, indicating that the earlier layer (layer 3) has a lesser impact. Comparing the two fusion methods, external direct fusion shows greater performance gains (average score: 49.18 $\rightarrow$ 52.63 $\rightarrow$ 58.62) than internal direct fusion (average score: 48.54 $\rightarrow$ 51.01 $\rightarrow$ 54.37). Coupled with the weaker results on smaller datasets (see Fig. 5 in the main paper), this phenomenon suggests that internal fusion may disrupt intrinsic feature distributions within LLM. However, by more closely integrating ViT's object-focused features with the LLM's higher-level abstractions, internal fusion holds substantial theoretical promise. As illustrated in Fig. 2, the training loss for internal fusion gradually converges but continues to improve, implying that, with sufficient data, this approach has the possibility of outperforming simpler external fusion strategies.

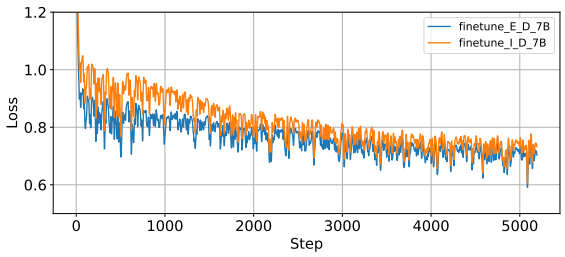

Figure 2. Loss curves for different fusion strategies.

Table 4. Comparison of performance across different configurations.

Model	Avg.	General	OCR	CV-Centric	Hallu
Mini-LLaVA	48.51	49.15	29.69	47.37	85.83
E+D {3}	48.26	49.24	29.53	46.17	86.01
E+D {3,18,23} Vicuna 1.5 7B	58.62	59.57	39.84	61.39	86.84
I+D {3,18,23} Vicuna 1.5 7B	54.37	57.85	30.93	55.84	84.40

4. Limitations

In this study, we examine model scaling by parameter and dataset size. Specifically, for LLMs, our experiments are limited to a maximum scale of 7B parameters. Similarly, for datasets, we constrain the scale to about 1M samples. While these limits are smaller compared to the largest LLMs and datasets available in the field, our exploration still holds significant value due to the practical relevance and commonality of such scales in many works [2, 17, 30].

References

[1] 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. 1

[2] Yue Cao, Yangzhou Liu, Zhe Chen, Guangchen Shi, Wenhai Wang, Danhuai Zhao, and Tong Lu. Mmfuser: Multimodal multi-layer feature fuser for fine-grained vision-language understanding. *arXiv preprint arXiv:2410.11829*, 2024. 3

[3] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 2

[4] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[6] 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*, pages 6700–6709, 2019. 1

[7] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018. 1

[8] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 1

[9] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 1

[10] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 1

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[12] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1

[13] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[15] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 1

[16] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 1

[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. 3

[18] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1

[19] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[20] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 1

[21] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 1

[22] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015. 1

[23] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1

[24] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 1

[25] ShareGPT. https://sharegpt.com/, 2023. 1

[26] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 1

[27] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 1

[28] 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

[29] 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. 1

[30] Huanjin Yao, Wenhao Wu, Taojiannan Yang, YuXin Song, Mengxi Zhang, Haocheng Feng, Yifan Sun, Zhiheng Li, Wanli Ouyang, and Jingdong Wang. Dense connector for mllms. *arXiv preprint arXiv:2405.13800*, 2024. 3