Supplementary Material for *CoVFT: Context-aware Visual Fine-tuning for Multimodal Large Language Models*

Nan Zhou[1,2] Huiqun Wang[1,2] Yaoyan Zheng[1,2] Di Huang[1,2*]
[1]State Key Laboratory of Complex and Critical Software Environment, Beihang University
[2]School of Computer Science and Engineering, Beihang University
{zhounan0431,hqwangscse,yaoyanzheng,dhuang}@buaa.edu.cn

This supplementary document provides additional experimental results and extended analyses that complement the main paper. The content is organized as follows:

- **Section A** presents extended ablation studies of CoVFT, including additional design choices and hyperparameters.
- **Section B** presents extended analyzes on the gradient conflict, context-aware counterparts, learnable parameters and computational overhead.
- **Section C** provides visualizations of contextual vectors and multimodal samples.
- **Section D** reports the complete benchmark results for all evaluated tasks, extending the averaged results shown in the main paper.

A. Extended Ablation Studies

In this section, we provide additional analyses of key design choices in CoVFT, focusing on the placement of CoMoE layers and the number of experts per CoMoE block.

Placement of CoMoE layers. We evaluate the impact of different CoMoE placements within the vision encoder in Table A. Since only the first 23 ViT blocks participate in the LLaVA-1.5 forward pass [3], we index them from 0 to 22 and examine four representative configurations: inserting CoMoE into *all* eligible layers (0–22), into the *first half* (0–10), into the *second half* (11–22), and into progressively deeper subsets of layers. From Table A, we observe that introducing CoMoE consistently improves performance over the baseline (*N/A*), which corresponds to full fine-tuning without CoMoE, demonstrating its overall effectiveness. Furthermore, deeper CoMoE placement generally yields larger gains under comparable layer budgets. This trend aligns with the observation in Fig. 1(c), where visual preference conflicts become increasingly pronounced in deeper layers of the vision encoder. We additionally evaluate several reduced-layer variants and find that placing CoMoE in layers 11–22 delivers the most consistent improvements across benchmarks. Consequently, we adopt this configuration as the default setting for CoVFT.

Table A. Ablation on the placement of CoMoE layers within the vision encoder. "Start" and "End" denote the layer range where CoMoE is inserted. "Num." is the total number of CoMoE layers applied. The baseline (N/A) corresponds to full fine-tuning without CoMoE.

Start	End	Num.	General	Know.&OCR	Vision	Mean (%)
N/A			66.69	61.29	52.17	59.29
0	22	23	66.72	61.43	54.32	60.23
0	10	11	66.75	61.43	52.13	59.33
11	22	12	**67.04**	**61.93**	**55.81**	**61.08**
15	22	8	66.96	62.10	52.87	59.87
19	22	4	66.78	61.70	53.50	59.98

Table B. Ablation on the number of experts in each CoMoE block.

Num.	General	Know.&OCR	Vision	Mean (%)
2	66.24	61.63	55.16	60.47
4	**67.04**	**61.93**	55.81	**61.08**
8	66.12	61.73	**56.03**	60.82

Number of experts. Table B reports an ablation on the number of experts in each CoMoE block. Using four experts yields the best mean performance among the tested configurations. We note that the optimal number of experts is related to the scale and diversity of training data [6]. The hyperparameters in Table B are determined under an instruction-tuning set of roughly 665K samples. Based on existing MoE scaling observations [4, 6], we expect that larger-scale training may benefit from increasing the number of experts, and adjusting expert width according to empirical scaling rules may further improve performance.

To further investigate the optimal MoE capacity under

*Corresponding author.

different levels of training data diversity, we partition the LLaVA-1.5 instruction data into 15 task types using Qwen-Plus, based on perceptual focus, semantic granularity, and reasoning requirements, and then ablate the number of experts. The results are shown in Table C. As task diversity increases (3→15), the optimal number of experts also increases (2→4→8), indicating that the optimal MoE capacity scales with data diversity. This trend suggests that CoMoE allocates more specialized expert subspaces to accommodate heterogeneous visual preferences induced by increasingly diverse multimodal contexts.

Table C. Vision-centric task performance under different levels of task diversity and numbers of experts. **Bold** indicates the best configuration for each diversity level.

# Experts	# Task types→	3	6	9	12	15
Full ft.		45.49	47.66	49.01	51.79	52.17
2		**48.09**	**49.55**	51.09	53.14	55.16
4		47.36	48.71	**52.27**	**54.72**	55.81
8		46.94	48.35	52.01	54.50	**56.03**

B. Extended Analysis

Gradient conflict. In Fig. 1(a) of the main body, we analyze visual preference conflict from the perspective of parameter distance. Here, we further provide an analysis from the gradient perspective. Specifically, we compute the cosine similarity between each gradient update and the dominant gradient direction during training under the standard mixed-task setting. The results are shown in Fig. A. Full fine-tuning exhibits frequent near-orthogonal or even opposing gradients, indicating update conflicts across heterogeneous instructions. In contrast, CoVFT substantially increases the mean gradient similarity (0.076→0.189) while reducing the standard deviation (0.112→0.051), suggesting that it improves gradient alignment and stabilizes optimization.

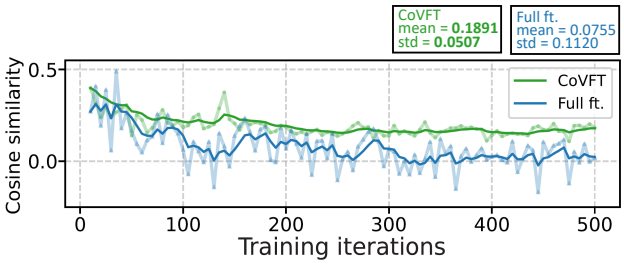

Figure A. Cosine similarity between each gradient update and the dominant gradient direction during training under the standard instruction-tuning setting.

Comparison with context-aware counterparts. To further compare with existing context-aware counterparts, we re-implement QA-ViT [2] and Q-MoE [8] within LLaVA-1.5. CoVFT achieves **61.08**% mean accuracy, outperforming QA-ViT (59.12%) and Q-MoE (59.29%) across 12 benchmarks, demonstrating the advantage of our method for context-aware visual adaptation in MLLMs.

Parameter-matched comparison. We compare different methods under matched trainable parameter budgets. With 4 CoMoE layers, CoVFT matches LoRA (r=512) in trainable parameters (6.85B vs. 6.86B) while achieving better overall performance (59.98 vs. 59.20), especially on vision-centric tasks (53.50 vs. 52.67). When increasing the number of CoMoE layers to 12, CoVFT continues to scale effectively, improving overall performance from 59.98 to 61.08 and vision-centric performance from 53.50 to 55.81. In contrast, existing PETL methods are more difficult to scale in this manner, highlighting the superior scalability and effectiveness of CoVFT under similar parameter budgets.

Table D. Comparison of diverse VFT methods under different trainable parameters. N denotes the number of CoMoE layers.

Method	G.	K.	V.	Mean (%)	Param. (B)
Freeze	66.23	61.20	51.71	58.93	6.76
LoRA (r=8)	65.93	60.86	52.45	59.04	6.76
LoRA (r=512)	66.09	60.90	52.67	59.20	6.86
CoVFT (N=4)	**66.78**	**61.70**	**53.50**	**59.98**	6.85
Full fine-tuning	66.69	61.29	52.17	59.29	7.06
CoVFT (N=12)	**67.04**	**61.93**	**55.81**	**61.08**	7.24

Computational overhead. We analyze the computational overhead of CoVFT and compare it with two widely used VFT baselines: *Freeze* and *Full fine-tuning*. All pre-training and instruction-tuning experiments are conducted on 8×NVIDIA H100 GPUs, with batch sizes identical to those used in LLaVA-1.5. For inference, we evaluate each method on a single NVIDIA H100 GPU with a batch size of 1.

The computational cost comparison among *Freeze*, *Full FT*, and *CoVFT* is reported in Table E. Compared with the freeze setting and full fine-tuning, CoVFT introduces 18 minutes in total training time (approximately 3.8% overhead relative to Full FT), adds ∼10ms latency during inference, and incurs a moderate increase in peak GPU memory usage (13.5% over Full FT). The frozen BERT encoder incurs around 1.14ms inference latency per sample. Importantly, this slight computational and memory overhead yields substantial accuracy gains across 12 multimodal benchmarks and delivers markedly improved stability, offering a significantly better trade-off among performance, robustness, and efficiency. This demonstrates that context-aware visual adaptation is both practical and cost-effective for modern MLLMs. Moreover, CoVFT remains fully compatible with existing MoE-parallelization and expert shard-

ing techniques [9], which can further reduce training, inference, and memory costs, suggesting that its efficiency can be improved even further in future deployments.

C. Visualizations

In the main paper (Fig. 3(a)), we visualize the contextual vectors extracted from the CVE module, showing clear semantic grouping aligned with distinct visual preference patterns. Here, we extend this analysis by presenting the full clustering results in Fig. B, together with representative samples from each cluster.

Visualization procedure. We randomly sample 5,000 multimodal instruction–image pairs from the LLaVA-665K instruction-tuning dataset and extract their contextual vectors from the penultimate vision encoder layer (i.e., the features actually fed to the LLM). These vectors are grouped using k-means [1] clustering, projected into two dimensions using PCA [5] for visualization, and we select the four samples nearest to each cluster centroid to illustrate the dominant characteristics of each cluster.

Observed cluster patterns. As shown in Fig. B, the contextual vectors exhibit a clear task-dependent organization: clusters naturally align with distinct types of multimodal reasoning despite the absence of explicit task labels. For example, Clusters 0 and 8 correspond to *relationship reasoning* involving structural or relational understanding across objects; Clusters 1 and 2 capture *region captioning*, focusing on localized box-level descriptions; and Clusters 3, 4, and 5 represent *region grounding*, mapping textual phrases to spatial coordinates. Other clusters display more heterogeneous behaviors: Cluster 7 leans toward *visual entity recognition*, including OCR and attribute-based identification; Cluster 9 frequently mixes grounding and captioning; while Cluster 6 spans diverse queries without a dominant pattern. These emergent structures indicate that CVE effectively organizes the multimodal space into semantically meaningful subregions, capturing latent task semantics purely from contextual signals. Building upon these structured contextual embeddings, CoMoE further routes samples to expert pathways that better align the vision encoder's updates with multimodal instruction-following objectives while reducing conflicting optimization signals.

Discussion. While CVE provides an effective mechanism for capturing contextual signals, the clustering results suggest room for further refinement. Some clusters (e.g., Cluster 6) exhibit coarse or mixed semantics, suggesting that contextual modeling could be made more discriminative. Conversely, overly fine-grained contextual partitioning may also hinder expert-level knowledge sharing. Balancing fine-grained modeling of visual preferences with the need for shared representation learning, as well as understanding how longer and more complex textual contexts shape these clusters, offers promising directions for future research.

D. Full Results

For reasons of space, the main paper reports the averaged results for several groups of benchmarks. In this section, we provide the complete per-task results corresponding to all experiments in the main text[1]. Specifically, Table F presents the full results for the design choices of the contextual vector and routing strategy (Table 2 in the main paper). Table G reports the per-task results under different data scales (Figure 4 in the main paper). Table H and Table I provide the detailed results for the ablations on the placement of CoMoE layers and the number of experts, respectively, corresponding to Tables A and B of the supplementary material.

References

[1] Michael R Anderberg. *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*. Academic press, 2014. 3, 4

[2] Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. Question aware vision transformer for multimodal reasoning. In *CVPR*, 2024. 2

[3] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 1

[4] Jan Ludziejewski, Maciej Pióro, Jakub Krajewski, Maciej Stefaniak, Michał Krutul, Jan Małaśnicki, Marek Cygan, Piotr Sankowski, Kamil Adamczewski, Piotr Miłoś, et al. Joint moe scaling laws: Mixture of experts can be memory efficient. *arXiv preprint arXiv:2502.05172*, 2025. 1

[5] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 1901. 3, 4

[6] Changxin Tian, Kunlong Chen, Jia Liu, Ziqi Liu, Zhiqiang Zhang, and Jun Zhou. Towards greater leverage: Scaling laws for efficient mixture-of-experts language models. *arXiv preprint arXiv:2507.17702*, 2025. 1

[7] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024. 3

[8] Hanzi Wang, Jiamin Ren, Yifeng Ding, Lei Ren, Huixing Jiang, Wei Chen, Fangxiang Feng, and Xiaojie Wang. Q-moe: Connector for mllms with text-driven routing. In *ACM MM*, 2024. 2

[1] Following [7], we divide the MME Perception score by 20 to have the same scale as other benchmarks before averaging.

Table E. Comparison of computational overhead across different VFT strategies. Training time is reported in hours and minutes; inference time is averaged per sample.

Method	Pre-training	Instruction-tuning	Total	Inference	Peak GPU memory
Freeze	2h 38m	4h 28m	7h 6m	77.04ms	41,352MB
Full ft.	2h 38m	5h 14m	7h 52m	77.04ms	44,036MB
CoVFT	2h 38m	5h 32m	8h 10m	88.17ms	49,974MB

Figure B. PCA [5] visualization of 5,000 contextual vectors extracted from the CVE module. Vectors are grouped using k-means [1], and for each cluster, four nearest samples to the cluster centroid are shown to illustrate representative task patterns.

Table F. Per task results by various ablation settings in Table 2 of the main body.

	General				Knowledge & OCR			Vision-centric					Mean (%)			
Ablation	MME^{P}	MMB^{en}	MMB^{cn}	GQA	SQA^{I}	AI2D	TextVQA	MMVP	RWQA	COCO	ADE	Omni3D	**G.**	**K.**	**V.**	**Avg.**
w.r.t. contextual vector																
N/A	1510.2	67.44	59.88	63.92	67.67	56.51	59.70	27.33	55.82	65.71	51.97	60.00	66.69	61.29	52.17	59.29
Image-only	1492.2	68.56	61.17	62.04	68.96	56.80	59.32	26.00	56.99	66.71	55.13	61.00	66.60	61.69	53.17	59.77
Text-only	1505.6	69.07	60.22	62.77	68.72	57.35	59.52	30.66	56.99	68.82	57.35	59.83	66.84	61.86	54.73	60.55
Concat[I, T]	1498.6	68.90	60.48	62.82	69.16	56.67	59.54	31.33	55.82	68.70	55.29	61.67	66.78	61.79	54.56	60.44
CVE	1525.2	68.13	60.40	63.37	69.51	56.64	59.64	36.67	57.52	66.96	56.08	61.83	67.04	61.93	55.81	61.08
w.r.t. routing strategy																
Uniform	1478.8	67.61	59.28	63.89	69.26	56.80	59.20	28.67	57.65	66.21	51.03	61.67	66.18	61.75	53.05	59.60
Sparse@2	1478.8	69.07	60.57	62.91	69.91	56.67	60.15	30.00	56.21	67.20	51.82	62.75	66.63	61.78	53.60	60.10
Dense	1525.2	68.13	60.40	63.37	69.51	56.64	59.64	36.67	57.52	66.96	56.08	61.83	67.04	61.93	55.81	61.08

Table G. Per task results by various ablation settings in Figure 4 of the main body.

	General				Knowledge & OCR			Vision-centric					Mean (%)			
Method	MME^{P}	MMB^{en}	MMB^{cn}	GQA	SQA^{I}	AI2D	TextVQA	MMVP	RWQA	COCO	ADE	Omni3D	**G.**	**K.**	**V.**	**Avg.**
100% data																
Freeze	1473.7	67.87	60.30	63.07	69.31	55.76	58.53	28.00	56.73	63.73	49.61	60.50	66.23	61.20	51.71	58.93
Full fine-tuning	1510.2	67.44	59.88	63.92	67.67	56.51	59.70	27.33	55.82	65.71	51.97	60.00	66.69	61.29	52.17	59.29
CoVFT	1525.2	68.13	60.40	63.37	69.51	56.64	59.64	36.67	57.52	66.96	56.08	61.83	67.04	61.93	55.81	61.08
50% data																
Freeze	1446.0	65.55	57.47	60.84	68.96	55.63	57.01	28.00	53.46	64.72	49.45	59.42	64.04	60.53	51.01	57.73
Full fine-tuning	1445.2	67.18	56.87	62.23	66.09	57.35	57.40	29.33	53.59	65.96	52.13	58.00	64.63	60.28	51.80	58.20
CoVFT	1492.0	67.7	58.08	61.25	68.32	56.83	58.07	26.00	55.29	66.96	49.92	58.17	65.41	61.07	51.27	58.43
25% data																
Freeze	1463.4	62.89	55.07	58.67	67.82	54.92	55.54	19.33	51.63	57.64	47.39	60.25	62.45	59.43	47.25	55.36
Full fine-tuning	1397.6	63.32	53.18	60.43	67.53	54.76	56.24	24.67	50.72	57.76	47.87	62.92	61.70	59.51	48.79	55.77
CoVFT	1467.0	64.35	54.90	59.91	67.13	55.60	56.93	23.33	53.73	57.27	47.39	59.83	63.13	59.89	48.31	56.14
12.5% data																
Freeze	1430.6	59.19	53.09	56.11	66.68	54.60	53.38	15.33	47.97	55.53	49.29	58.17	59.98	58.55	45.26	53.49
Full fine-tuning	1408.2	57.82	50.52	57.93	66.63	53.79	54.50	22.00	50.46	54.78	46.45	58.50	59.17	58.31	46.44	53.65
CoVFT	1382.2	59.11	50.69	57.98	66.34	53.95	55.37	21.33	49.93	59.38	48.18	59.50	59.22	58.55	47.66	54.24

[9] Zheng Zhang, Yaqi Xia, Hulin Wang, Donglin Yang, Chuang Hu, Xiaobo Zhou, and Dazhao Cheng. Mpmoe: Memory efficient moe for pre-trained models with adaptive pipeline parallelism. *IEEE Transactions on Parallel and Distributed Systems*, 35(6):998–1011, 2024. 3

Table H. Per task results by various ablation settings in Table A of the supplementary material.

			General				Knowledge & OCR			Vision-centric					Mean (%)			
Start	End	Num.	MME^{P}	MMB^{en}	MMB^{cn}	GQA	SQA^{I}	AI2D	TextVQA	MMVP	RWQA	COCO	ADE	Omni3D	**G.**	**K.**	**V.**	**Avg.**
	N/A		1510.2	67.44	59.88	63.92	67.67	56.51	59.70	27.33	55.82	65.71	51.97	60.00	66.69	61.29	52.17	59.29
0	22	23	1508.4	68.08	60.55	62.84	67.92	56.93	59.45	30.67	55.69	67.20	55.29	62.75	66.72	61.43	54.32	60.23
0	10	11	1501.4	68.04	60.65	63.22	68.62	56.99	58.68	27.33	56.86	62.11	52.34	62.00	66.75	61.43	52.13	59.33
11	22	12	1525.2	68.13	60.40	63.37	69.51	56.64	59.64	36.67	57.52	66.96	56.08	61.83	67.04	61.93	55.81	61.08
15	22	8	1518.0	68.21	60.45	63.29	69.66	56.77	59.88	31.33	56.60	64.57	54.50	57.33	66.96	62.10	52.87	59.87
19	22	4	1513.2	68.10	60.33	63.01	69.36	56.25	59.50	29.33	56.60	66.71	55.13	59.75	66.78	61.70	53.50	59.98

Table I. Per task results by various ablation settings in Table B of the supplementary material.

	General				Knowledge & OCR			Vision-centric					Mean (%)			
Num.	MME^{P}	MMB^{en}	MMB^{cn}	GQA	SQA^{I}	AI2D	TextVQA	MMVP	RWQA	COCO	ADE	Omni3D	**G.**	**K.**	**V.**	**Avg.**
2	1478.6	67.61	60.91	62.51	69.06	56.25	59.59	32.00	58.04	68.57	55.77	61.42	66.24	61.63	55.16	60.47
4	1525.2	68.13	60.40	63.37	69.51	56.64	59.64	36.67	57.52	66.96	56.08	61.83	67.04	61.93	55.81	61.08
8	1476.2	67.61	60.48	62.57	68.62	56.99	59.58	30.67	56.73	70.68	58.14	63.92	66.12	61.73	56.03	60.82