Supplementary Material

A. Ablation on Combination Design

To optimize the performance when inclusion of multi-grained learnable queries, we ablate on different design choice for the combination as shown in Table 1. From the experiment results, we observe that both the order of the granularity and the combination method matters. From the first row and the second row, there is a performance drop when the finer-grained learnable queries are placed nearer the instruction text and farer from the images, which is intuitive as images are much more fine-grain than high-level text in terms of semantics. Therefore, placing finer-grained query nearer the images can have a better transition from low-level to high-level understanding.

We also try not concatenating the learnable queries of different granularity together, but align them from the same fine-grained sequence before and after average pooling according to the unflattened sequence, which is shown in the third row of Table 1. The performance also draws a little compared to concatenate different learnable queries together.

And finally, we choose concatenate learnable queries with stride size 3 and 4 together, which achieves highest score according to the experiments.

B. Effect of Learnable Queries with Different Granularity

We demonstrate more results on the inclusion of learnable queries with different granualarity on different sub-tasks of MME [1] in Table 2 and Table 3. According to the results, we can find that finer-grained learnable queries (with patch size 2) can benefit more on the tasks requiring fine-grained recognition like OCR, numerical calculation and code reasoning. While coarse learnable queries will have an advantage on high-level understanding tasks, like color, posters, or artwork recognition and commonsense reasoning.

C. Visualization

C.1. Visualization Overall Results with Radar Chart

We visualize the performance of encoder-free MLLMs and the selected top 2 encoder-based MLLMs in Figure 1. It is obvious to see that BREEN performs comparable with a strong encoder-based baseline, which is Cambrian.

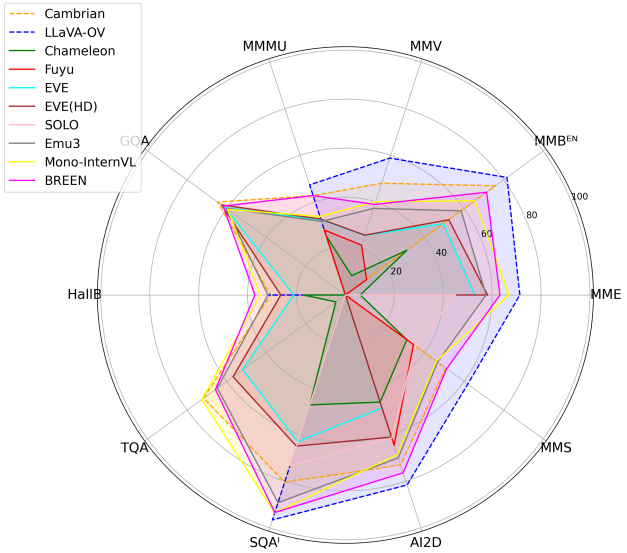

Figure 1. Comparison of the performance of BREEN with encoder-based and encoder-free models.

C.2. Visualization of Attention Weights

We visualize the attention weights from the critical output tokens to all the image tokens and the learnable query tokens in Figure 2. The first row of each example is the visualization for image tokens and the second row is for learnable queries. We can observe from the visualization that it is hard for attention weights to focus on the desired area in the first layer, they just distribute randomly with a higher average score. And we find that the attention to image tokens will refine to a specific area in the last layer, while still attending to a wider area in the middle layer. On the other hand, the attention scores to the learnable query can actively and correctly identify the important areas in the middle layer, but fade out in the last year. It is likely to be explained by the different granularity of the learnable query and image tokens.

Stride Size	Align Method	SQA^I	TQA	GQA	MME
4 & 2	Concat	62.8	41.0	50.8	1278.1
2 & 4	Concat	63.2	41.2	51.2	1290.4
2 & 4	AvgPool	63.1	40.9	51.0	1262.7
3 & 4	Concat	63.7	41.4	51.7	1333.4

Table 1. Ablation study of BREEN on the combination design for learnable queries with different granularity.

Stride/Patch Size	existence	count	position	color	posters	celebrity	scene	landmark	artwork	ocr	Total
2 (12×12)	173.3	88.3	78.3	101.6	56.5	47.1	139.0	82.5	85.8	67.5	919.9
3 (8×8)	175.0	88.3	101.7	111.7	62.9	62.6	138.3	104.3	86.8	57.5	989.0
4 (6×6)	175.0	90.0	98.3	145.0	68.4	47.4	134.3	85.3	90.8	57.5	991.8

Table 2. Sub-task performance on MME-P.

D. Limitation and Future Plan

While BREEN presents a significant step toward data-efficient encoder-free multimodal learning, several limitations remain, providing avenues for future research.

Adaptive Query Selection for Task-Specific Needs BREEN employs a dual-granularity learnable query mechanism, allowing the model to leverage both fine-grained and coarse-grained representations for different multimodal tasks. However, the current approach lacks an explicit mechanism for dynamically selecting or weighting queries based on the task context. Future work could explore adaptive query selection strategies, such as attention-based weighting or reinforcement learning-based query routing, to optimize query utilization based on task requirements.

Efficient Token Utilization and Sequence Compression One limitation of BREEN is that the addition of learnable queries increases the token sequence length, leading to higher computational costs during training and inference. Although learnable queries improve vision-language alignment, their fixed length may introduce redundancy, especially for tasks with varying levels of visual detail. To address this, future work could explore mechanisms for adaptive query compression, where the model dynamically selects or merges learnable queries based on task requirements. Additionally, compressing image tokens alongside learnable queries could further enhance efficiency while retaining critical visual information.

Exploring Stronger Vision Encoder Teachers BREEN distills visual knowledge from CLIP [2] to enhance alignment in an encoder-free setting. However, recent advancements in vision encoders, such as SigLIP and SigLIP2 [3, 4], have demonstrated superior visual representations with stronger semantic understanding. Future work could explore using these more powerful vision encoders as teachers to further improve the quality of learnable queries, potentially leading to better multimodal reasoning and more robust performance across diverse tasks.

References

[1] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv: 2306.13394*, 2023. 1

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2

[3] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2

[4] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. 2

Stride/Patch Size	common sense reasoning	numerical calculation	text translation	code reasoning	Total
2 (12 × 12)	70.0	50.0	95.0	87.5	302.5
3 (8 × 8)	72.1	42.5	110.0	80.0	304.64
4 (6 × 6)	83.6	47.5	72.5	55.0	258.57

Table 3. Sub-task performance on MME-C.

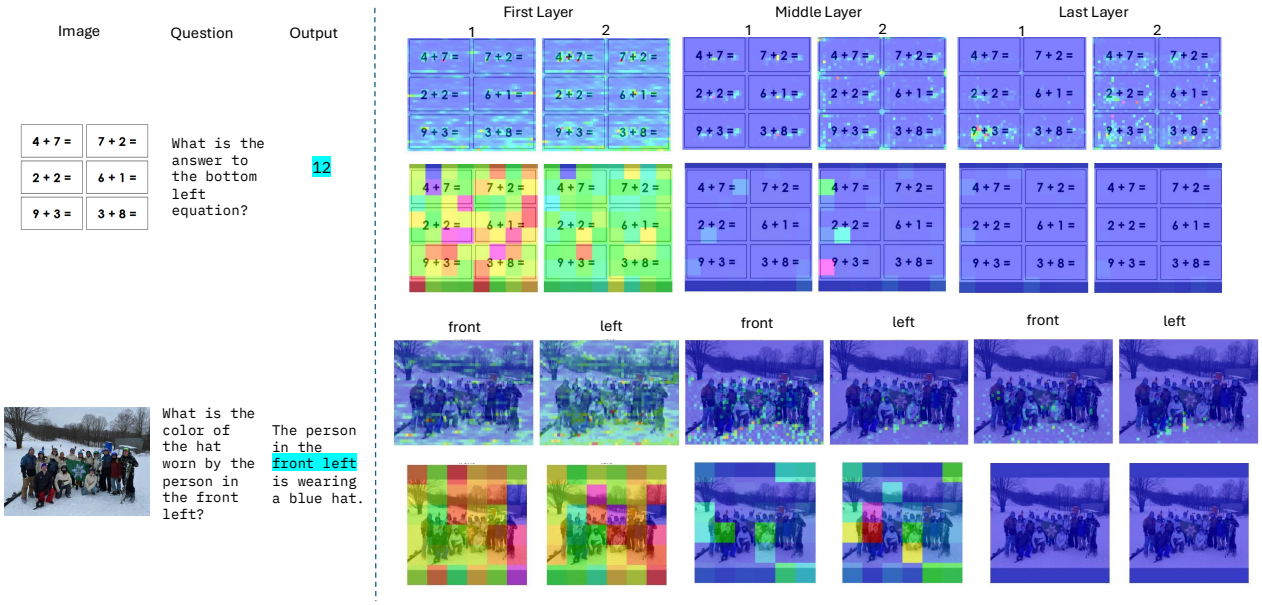

Figure 2. The output attention visualization to image tokens and learnable query tokens in different layers.