ROSE: Rotate Your Large Language Model to See

Supplementary Material

Overview

In this supplementary material, we provide following items:
- (Sec.1) Preliminary experiment details.
- (Sec.2) Training details.
- (Sec.3) Evaluation details.

1. Details of Preliminary Experiments

To empirically determine which component of pretrained parameters serves as the primary carrier of semantic knowledge in LLMs, we conduct a preliminary experiment that disentangles the contributions of direction and magnitude in their interactions with input tokens. The experiment is designed as follows.

Given a pretrained LLM, each linear projection $W = [w_1, \ldots, w_n] \in \mathbb{R}^{d\times n}$ interacts with an input token $x \in \mathbb{R}^d$ through its column vectors:

$$z = W^\top x = [\, w_1^\top x,\ w_2^\top x,\ \ldots,\ w_n^\top x \,]^\top. \tag{1}$$

For each vector w_i, the interaction can be further written as:

$$z_i = w_i^\top x = \|w_i\|\,\|x\|\,\cos(\theta_i), \tag{2}$$

where θ_i denotes the angle between w_i and x. This expression reveals that the influence of parameter w_i on the input x can be decomposed into two parts: first projecting x onto the direction of w_i captured by $\cos(\theta_i)$, and then scaling this projection by the magnitude $\|w_i\|$. These two components together reflect the patterns learned by w_i during pretraining. To understand their respective roles, we aim to disentangle and ablate their individual contributions. Specifically, we examine the following three settings:

- **Keep Direction**. We remove the scaling effect induced by the vector magnitude and retain only the directional relationship between w_i and x, setting $z_i = \cos(\theta_i)$. This corresponds to normalizing all parameter vectors onto the unit hypersphere, preserving only their geometric orientation in the semantic space.
- **Keep Magnitude**. We eliminate the influence of direction by enforcing $z_i = \|w_i\|\,\|x\|$, effectively assuming that all w_i are perfectly aligned with x. In this setting, each vector contributes solely through its scaling effect while its orientation is ignored.
- **Add Noise**. We directly perturb each w_i by adding Gaussian noise $\mathcal{N}(0, 1)$, thereby destroying both its directional and magnitude information. This serves as a control condition in which the semantic structure encoded in the pretrained parameters is disrupted.

Table 1. **The training details of ROSE.**

Config	Stage 1	Stage 2	Stage 3
LLM backbone	Qwen2.5-7B [17]		
ViT backbone	SigLIP2-So400m-384 [15]		
Global batch size	2048	2048	256
Batch size per GPU	32	32	4
Accumulated steps	1	1	1
DeepSpeed zero stage	2	2	3
Learning rate	1×10^{-4}	5×10^{-5}	2×10^{-5}
Learning rate schedule	cosine decay		
Warmup ratio	0.01		
Weight decay	0		
Epoch	1		
Optimizer	AdamW		
Precision	bf16		

We then iterate through the LLM layer by layer. For the i-th layer, we apply one of the three intervention strategies to all linear projections within that layer while keeping the rest of the network unchanged. We then evaluate the modified model on the MMLU [4] benchmark to obtain the corresponding accuracy. Repeating this process across all layers produces layer-wise accuracies, from which the Figure 2 of the main manuscript are derived.

2. Training Details

The overall training process adopts a three-stage paradigm, initially involving the S1. Perceptual Pretraining, S2. Semantic Pretraining, and S3. Self-supervised Finetuning. Table 1 presents the details of this three-stage training for **ROSE**.

3. Evaluation Details

3.1. Details About the Benchmarks

We conduct a comprehensive evaluation of **ROSE**, including both multimodal benchmarks and NLP benchmarks.

Multimodal Benchmarks. We conduct experiments across 12 widely recognized multimodal benchmarks, including MMBench [9], MMMU [18], SEED-Image [7], MME [3], MMStar [1], TextVQA [14], ChartQA [10], DocVQA [11], InfoVQA [12], AI2D [6], RealWorldQA [16], and GQA [5]. These benchmarks span a broad spectrum of multimodal tasks.

Table 2. **Summary of the evaluation multimodal benchmarks.** Prompts are mostly borrowed from LMMs-Eval [20].

Benchmark	Response formatting prompts
MMB	Answer with the option's letter from the given choices.
MMMU	Answer with the option's letter from the given choices.
SEED	Answer with the option's letter from the given choices.
MME	Answer the question using a single word or phrase.
MMStar	Answer with the option's letter from the given choices.
TextVQA	Answer the question using a single word or phrase.
ChartQA	Answer the question with a single word.
DocVQA	Answer the question using a single word or phrase.
InfoVQA	Answer the question using a single word or phrase.
AI2D	Answer with the option's letter from the given choices.
RWQA	Answer the question with a single word.
MathVista	Answer with the option's letter from the given choices.

NLP Benchmarks. We conduct experiments across 5 widely recognized benchmarks, including MMLU [4], Hellaswag [19], ARC-C [2], TruthfulQA [8], and Winogrande [13]. They cover multiple knowledge dimensions and domain focuses.

3.2. Evaluation Protocol

We adopt LMMs-Eval as our evaluation toolkit. For evaluation prompts, we provide a thorough examination of all evaluation benchmarks utilized in this paper in Table 2. For model efficiency, the FLOPs and latency are calculated using the DeepSpeed toolkit on a single NVIDIA H20 GPU without any engineering acceleration techniques.

3.3. Pseudocode

To provide a comprehensive understanding of our method, we present the pseudocode of **ROSE** in Algorithm 1.

Algorithm 1 Forward Process of ROSE

Require: Input Image I, Input Text Sequence x, Pretrained LLM Weights Θ, Rotation Layer Index List $\mathcal{H}$
Ensure: Output Logits y

Stage 1: Visual Encoding
$v \leftarrow \text{VisualEncoder}(I)$ ▷ Extract visual tokens $v \in \mathbb{R}^{L_v \times d_v}$
Stage 2: LLM Forward with Parameter Rotation
$h \leftarrow \text{WordEmbedding}(x)$
for each layer $l = 1$ to L **do**
  $\mathcal{W}_l \leftarrow \{W_q, W_k, W_v, W_o, W_{up}, W_{down}, W_{gate}\}$ ▷ Weights in layer l
  for each weight matrix $W \in \mathcal{W}_l$ **do**
    if $l \in \mathcal{H}$ **then**
      $R_v \leftarrow \text{VRMG}(v, W, r)$
      $W' \leftarrow R_v W$ ▷ Inject visual info via rotation
    else
      $W' \leftarrow W$
    end if
  end for
  ▷ Standard Transformer computation with rotated weights
  $h_{attn} \leftarrow \text{Self-Attn}(h, W'_q, W'_k, W'_v, W'_o)$
  $h \leftarrow h + h_{attn}$
  $h_{ffn} \leftarrow \text{FFN}(h, W'_{up}, W'_{down}, W'_{gate})$
  $h \leftarrow h + h_{ffn}$
end for
$y \leftarrow \text{LanguageHead}(h)$
return y

function $\text{VRMG}(v, W, r)$
  Initialize list of sub-blocks $\mathcal{B} \leftarrow []$
  Learnable queries $Q = \{q_1, \ldots, q_r\}$ specific to W
  for $i = 1$ to r **do**
    $q'_i \leftarrow \text{CrossAttention}(q_i, v, v)$ ▷ Aggregate visual info
    $t_i \leftarrow \text{Linear}(q'_i)$ ▷ Map to flat parameters
    $T_v^i \leftarrow \text{Reshape}(t_i)$ ▷ Reshape to lower-triangular form
    $P_v^i \leftarrow T_v^i - (T_v^i)^\top$ ▷ Construct skew-symmetric matrix
    $R_v^i \leftarrow (I + P_v^i)(I - P_v^i)^{-1}$ ▷ Cayley Orthogonalization
    $\mathcal{B}.\text{append}(R_v^i)$
  end for
  $R_v \leftarrow \text{BlockDiagonal}(\mathcal{B})$ ▷ Construct sparse orthogonal matrix
  return R_v
end function

References

[1] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024. 1

[2] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. 2

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

[4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 1, 2

[5] 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

[6] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016. 1

[7] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 1

[8] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252, 2022. 2

[9] 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, 2024. 1

[10] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022. 1

[11] 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

[12] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 1

[13] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64 (9):99–106, 2021. 2

[14] 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

[15] 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, 2025. 1

[16] X.ai. Grok-1.5 vision preview. https://x.ai/blog/grok-1.5v, 2024. 1

[17] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1

[18] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567, 2024. 1

[19] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800, 2019. 2

[20] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, 2025. 2