A Study on Multi-modal Interaction in Vision Large Language Models

Houjing Wei[1] 趙 羽風[1] Yuting Shi[1] Naoya Inoue[1,2]
[1] 北陸先端科学技術大学院大学 [2]RIKEN
{houjing,yfzhao,s2210096,naoya-i}@jaist.ac.jp

Abstract

Vision **L**arge **L**anguage **M**odels (VLLMs) usually take input as a concatenation of image token embeddings and text token embeddings and conduct causal modeling. Based on observations, this paper hypothesizes that intensive multimodal interactions happen in the mid-to-late layers. To verify, we apply cosine similarity measurement and norm-based attention analysis. Our experiments indicate that in the mid-to-late layers of LM decoder, there is a rise in inter-modal similarity and gradual accumulation in attention allocation to visual tokens, suggesting a four-phase inference dynamics against the LM layers, including **I)** Alignment, **II)** Intra-modal Encoding, **III)** Inter-modal Encoding, and **IV)** Output Preparation.

1 Introduction

Recently, instruction-tuned **L**anguage **M**odels (LMs) have demonstrated remarkable performance on cross-modal tasks when incorporated with other modalities, mainly vision [1, 2, 3, 4, 5, 6]. These VLLMs extend the instruction-following capability of LMs for handling multiple modalities, e.g., a concatenation of image tokens and text tokens, exhibiting impressive abilities, such as drafting stories based on images and building a website based on the hand-sketched image. Given their surprising achievements, how these models bridge the modality gap to enable information transition between image tokens and text tokens is still underexplored. In paper [7], the authors identify multimodal neurons in Transformer MLP layers and translate them into semantically related text. Another work in [8] indicates that LMs account for modeling domain-specific visual attributes while fine-tuning the cross-modal projector does not enhance such capability. Recent work explores some specific aspects of the inner workings of VLLMs via a mechanistic interpretation lens, such as localization and evolution of object-centric visual tokens, storage and transfer of multi-modal knowledge, and cross-modal information flow across LM decoder layers [9, 10, 11]. Although these works provide perspective insights on the inner dynamics of VLLMs, to the best of our knowledge, the magnitude of cross-modal interactions along LMs' layers remains unexplored, leading to our main research question: *How does multi-modal interaction evolve along the layers of the LM decoder in VLLMs?*

To answer this question, we first examine whether image tokens can be translated into linguistic semantics during the language modeling computation. Experiment results show that the visual representations are refined towards the embedding of interpretable tokens in the LM vocabulary space, even though VLLMs are not explicitly pre-trained for next-token prediction. We conjecture such a phenomenon might originate from multi-modal interaction, that is, the multimodal interaction leads to this refinement. Then, we propose to investigate the multi-modal interaction dynamics using similarity metrics and norm-based attention analysis. Specifically, we first investigate the magnitude of contextualization [12] to characterize the cross-modal dynamics along LMs' layers. Our experiments reveal a phase diagram of multimodal contextualization as shown in Fig. 1, suggesting that as inputs pass through successive layers of the Transformer-based decoder, a four-phase multimodal contextualization appears (Fig. 3). In addition, a norm-based attention analysis is conducted to visualize such multimodal interaction along LM decoder layers. This analysis reveals two patterns during model inference: gradual attention accumulation against Transformer layers and stronger attention focusing on specific tokens.

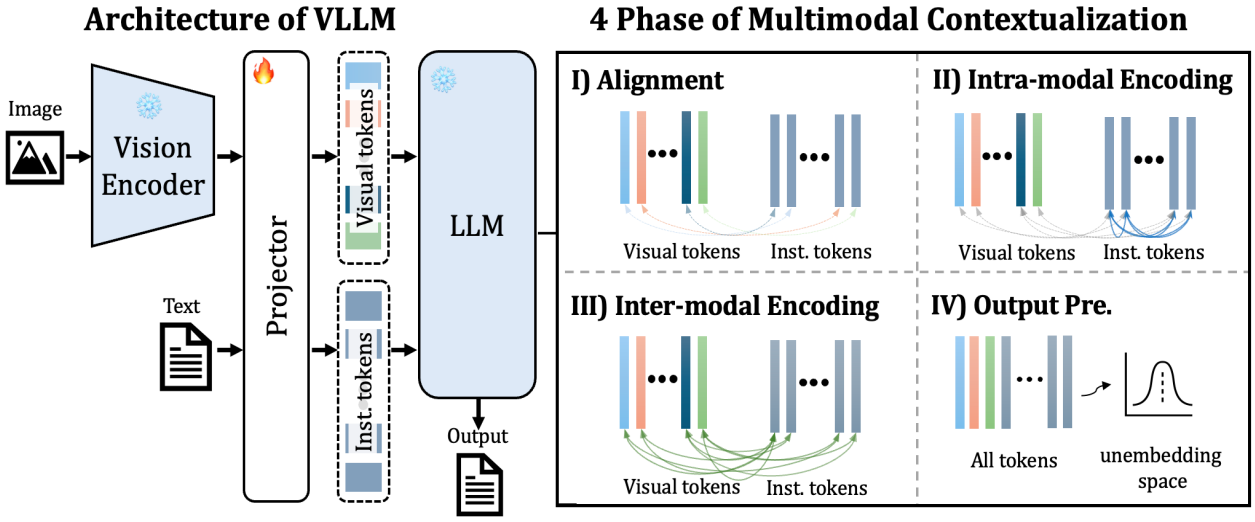

Figure 1 A four-phase diagram of feed-forward dynamics of LMs in VLLMs. **I) Alignment** of two different feature spaces occurs. **II) Intra-modal Encoding** is enhanced while cross-modal encoding is inhibited. **III) Inter-modal Encoding** appears and strengthens. **IV) Output Preparation** requires hidden states to be aligned toward output embedding space.

2 Preliminary Observation: Projecting Visual Tokens into Vocabulary Space

This section demonstrates to what extent visual tokens can be converted into linguistic concepts represented in the language vocabulary space. We use LogitLens [13] technique to project intermediate representations of visual tokens into the LM vocabulary space by multiplying them with the unembedding matrix.

Specifically, we extract the hidden representations of 32 visual tokens at each layer of LM in InstructBLIP[1], then decode them into language words. We define a visual hidden state as being decoded correctly if its decoded word matches the ground-truth caption. Therefore, precision indicates how many correctly decoded words overlap with all decoded words, while recall refers to the proportion of correctly decoded words to ground-truth caption words.

Precision and recall are computed and plotted in Fig. 2. It generally presents a continuously rising tendency in mid-to-late layers, indicating the intermediate representations of visual tokens are progressively morphed into linguistic forms that match with correct ground-truth captions. In the lower layers (near embedding space), both precision and recall are nearly negligible, suggesting that raw image tokens tend to produce irrelevant word distributions. In the mid-to-late layers (from around 10th), both lines continuously climb, reflecting an ongoing process of refinement where the visual token representations become more semantically coupled with the textual domain. Around the deepest layers (after 30th), we observe a slight variability between precision and recall, indicating a possible reduction in correctly decoded words.

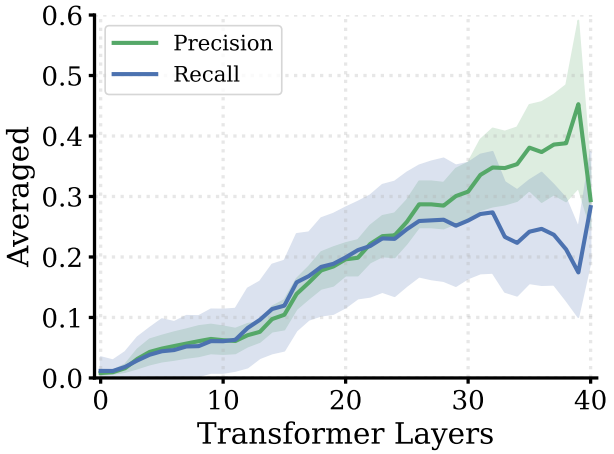

Figure 2 Averaged precision and recall for decoded words of visual tokens along Transformer layers. Results are computed and averaged on COCO validation set and Winoground dataset, from which we randomly choose 400 image-caption pairs. Shaded regions around each curve represent the standard deviation across multiple data samples.

Overall, this result shows that representation from the vision modality can be directedly decoded into nature language, and in mid-to-late layers they are decoded more correctly. This leads to our hypothesis that the intensive inter-modal interaction happen in those mid-to-late layers, during which intermediate representations of visual tokens are successively shaped to the most likely linguistic tokens.

3 Methodology

We investigate multi-modal interaction by implementing the following two approaches, i.e., cosine similarity measurements and layer-wise attention analysis, aiming to obtain a comprehensive view of how visual and linguistic representations interact and evolve within the Transformer-based LM decoder in VLLMs.

Contextualization as Interaction Magnitude. Inspired by [12], we use cosine similarity as a measurement of contextuality to explore how hidden states from two different representation spaces interact in LMs. In detail, let $v_i^{(l)}$ and $w_j^{(l)}$ denote the hidden state vectors of tokens i and j, respectively. The average cosine similarity for the hidden states at each layer l in LMs is thus defined as follows:

$$s^{(l)} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \cos\left(v_i^{(l)}, w_j^{(l)}\right), \tag{1}$$

where m and n indicate the number of tokens in two sets. Inter-modal similarity is computed by choosing $v_i^{(l)}$ from vision tokens and $w_j^{(l)}$ from text tokens. Intra-modal similarity is computed by ensuring $v_i^{(l)}$ and $w_j^{(l)}$ from the same modality (e.g., both from vision or both from text). Higher similarity suggests that the two sets of vectors occupy closely related subspaces in the representation space, indicating that they may encode similar features.

Visualization via Norm-based Attention. To investigate how the multimodal information interacts via multi-head self-attention mechanism, considering the faithfulness problem of attention score as an explanation [14, 15, 16], we use norm-based attention proposed by [17], which uses the norm of multi-head attention's output transformation to scale the attention score to investigate linguistic capabilities of Transformer. By taking the magnitudes of transformed vectors into consideration, this norm-based attention analysis provides a relatively faithful interpretation of the contribution of the input vector to the output.

For a more detailed experimental setup about this section, we recommend that readers refer to Appendix §A.1.

4 Multimodal Inference Dynamics in VLLMs

This section investigates the multimodal inference dynamics along Transformer layers in VLLMs.

§4.1 reveals that multimodal interaction evolves as the Transformer layer goes deeper, introducing our finding of a four-phase multimodal interaction pattern during the feed-forward calculation. Besides, attention analysis is conducted to visualize the multimodal interaction along LM layers. §4.2 demonstrates that intensive attention progressively emerges in the mid-to-late layers, suggesting a consistent correlation with the finding of similarity-based multi-modal interaction pattern.

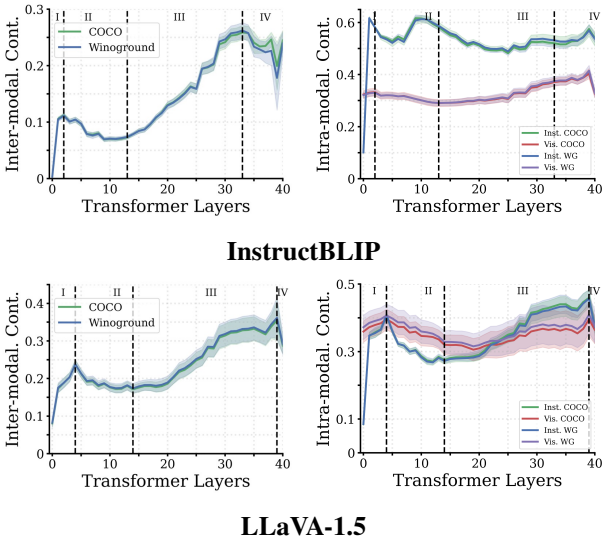

Figure 3 Inter-modal and intra-modal contextualization against layer depth, demonstrating our proposed four-phase inference dynamics. A higher value indicates stronger interaction. Similarity values are averaged over randomly chosen 600 images for each dataset. Shaded regions show standard deviations over randomly sampled images.

4.1 Four-phase Multimodal Contextualization

As described in equation (1), we calculate inter-modal similarity as well as intra-modal similarity using hidden states at each layer of LM decoder in VLLMs.

Fig. 3 generally exhibits an upward trajectory, showing a consistent trend with the observation from LogitLens; meanwhile, four distinct monotonic intervals are observed. Based on this monotonicity depicted, we introduce our findings of four-phase inference dynamics against LM layers, which coincide with Fig. 1: **I) Alignment**, during which an early alignment between two modalities occurs. **II) Intra-modal Encoding**, within which intra-modal similarity is significantly higher than that of inter-modal similarity (Fig. 3), indicating the model starts encoding visual tokens and text tokens separately. **III) Inter-modal Encoding** shows a swift rise in inter-modal similarity (Fig. 3), indicating an incremental inter-modal interaction. **IV) Output preparation** presents the global reduction in inter-modal similarity, suggesting the model shifts focus away

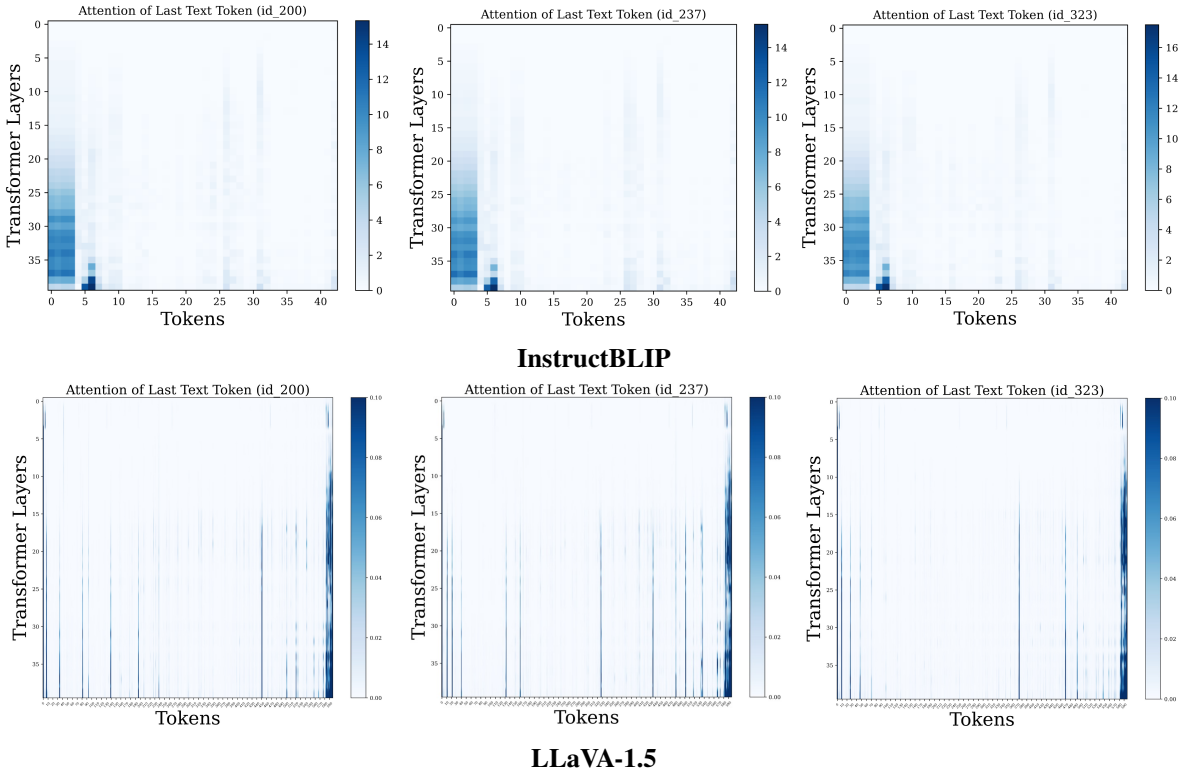

Figure 4 Qualitative analysis of norm-based attention results (id_200 refers to a randomly chosen image). Heatmaps showcase norm-based attention of the last text token (left 4) and the last vision token (right 4) to its preceding tokens along Transformer layers. The color intensity (moving from light to dark) indicates the magnitude of attention paid to each token.

from multimodal interaction.

4.2 Visualization of Multimodal Interaction

To further examine our findings regarding the four-phase inference dynamics between image tokens and text tokens, we employ norm-based attention analysis to elucidate how the attention allocation between the two modalities changes against LM decoder layers in VLLMs.

To this end, we extract norm-based attention results from two representative VLLMs, i.e., InstructBLIP and LLaVA-1.5, and plot the attention heatmaps for 100 images that were randomly selected from COCO and Winoground datasets, respectively. After manually inspecting them, we found almost identical pattern holds. We thus showcase the norm-based attention heatmaps of three images (id_200, id_237, id_323) for qualitative analysis.

Overall, Fig. 4 illustrates the norm-based attention of the last text token constantly increasing from the middle layers. Meanwhile, it reveals that image tokens at different positions receive varying degrees of saliency assignment. This observation holds for both models despite their architectural differences. In detail, first, a gradual accumulation of attention degree against layers is observed. In the early layers, attention tends to be more dispersed. As the model proceeds to the middle and deeper layers, we observe stronger attention allocation, especially more focused attention on several specific tokens. This pattern suggests that crucial cross-modal interaction intensifies in those mid-to-late layers. Second, a noticeable disparity in attention across image tokens is observed. Certain tokens are attended to much more strongly than others. One possible conjecture for explaining this phenomenon could be that these image tokens are presumably tied to semantically rich regions within the image, thus providing critical clues for accurate textual predictions.

5 Conclusion

This paper proposes to utilize contextualization as a measurement to explain multimodal interaction in LMs of VLLMs. By incorporating other investigation methods, i.e. norm-based attention, our extensive experiments indicate the multimodal interaction dynamics during the model's feed-forward pass.

Ackknowledgement

This work is supported by the Nakajima Foundation.

References

[1] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: towards general-purpose vision-language models with instruction tuning. In **Proceedings of the 37th International Conference on Neural Information Processing Systems**, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

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

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

[4] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In **The Twelfth International Conference on Learning Representations**, 2024.

[5] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning, 2023.

[6] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. **arXiv preprint arXiv:2209.15162**, 2022.

[7] Sarah Schwettmann, Neil Chowdhury, Samuel Klein, David Bau, and Antonio Torralba. Multimodal neurons in pretrained text-only transformers, 2023.

[8] Gaurav Verma, Minje Choi, Kartik Sharma, Jamelle Watson-Daniels, Sejoon Oh, and Srijan Kumar. Cross-modal projection in multimodal llms doesn't really project visual attributes to textual space. 2024.

[9] Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models, 2024.

[10] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. **arXiv preprint arXiv:2410.07149**, 2024.

[11] Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models, 2024.

[12] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 55–65, 2019.

[13] nostalgebraist. logit lens on non-gpt2 models. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, 2021.

[14] Kevin Clark. What does bert look at? an analysis of bert's attention. **arXiv preprint arXiv:1906.04341**, 2019.

[15] Sofia Serrano and Noah A Smith. Is attention interpretable? **arXiv preprint arXiv:1906.03731**, 2019.

[16] Sarthak Jain and Byron C. Wallace. Attention is not explanation, 2019.

[17] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 7057–7075, Online, November 2020. Association for Computational Linguistics.

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[19] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015.

[20] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality, 2022.

A Appendix

A.1 Experimental Settings

VLLMs usually take image patches and text tokens as input and generate text as output. Based on how visual features are mapped into the language embedding space, they can be broadly categorized into:1) Models employing cross-attention mechanisms to enable interaction between vision encoder's outputs and the language embedding space for extracting task-relevant image features (e.g., Flamingo, BLIP family); 2) Models using projection layers to map the vision encoder's outputs directly into the language embedding space (e.g., Mini-GPT4, LLaVA family).

Models. We conduct experiments on two representative VLLMs. 1) InstructBLIP [1], which is extended from BLIP-2 [18], introducing an instruction-aware Query Transformer to extract task-relevant image features tailored to the given textual instruction. 2) LLaVA-1.5 [3] apply an MLP projection as the cross-modal connector on top of CLIP vision encoder, establishing new SOTA baselines across 11 VL benchmarks.

Datasets. For evaluation, we use COCO captions validation set [19] and Winoground dataset [20]. COCO is a commonly used image-caption dataset that contains 164K images, each annotated with five captions. Winoground is a carefully handcrafted probing dataset, comprising 400 items, each including two pairs of images and corresponding captions.

Other Details. For the similarity experiment, we randomly select 600 images respectively from Winoground and COCO caption validation set. For the attention analysis experiment, we manually examine 100 randomly selected image instances from two datasets.