**Enhancing Large Language Model with Composite Learning and Cross-Attention for Cross-Modal Query Understanding**

1. **Introduction**

Recently, Large Language Models (LLMs) [1] have created tremendous advancements in Natural Language Processing (NLP) tasks, demonstrating their incredible ability to comprehend and generate human language. However, a distinct problem has emerged with the advent of multimodal information, which is data that spans multiple modalities, including text, image, and audio. In order to meet the growing need for cross-modal query understanding, the traditional scope of LLMs is limited to textual data. However, real-world applications necessitate the combination of multimodal data to generate highly comprehensive and insightful responses. Integrating massive and multimodal datasets that are frequently noisy and heterogeneous becomes challenging. Hence, optimally balancing the accuracy of each modality and the requirement for cross-modal generalization during LLM training to comprehend textual and non-textual data. Furthermore, integrating various modalities necessitates substantial resources for training and inference, which is challenging for cross-modal query understanding without increasing computational complexity. As a result, improving LLMs for multimodal interpretation is crucial. Multimodal LLM [2] have surfaced, building on existing capabilities and extending beyond text to encompass several modalities. Advances in LLMs and multimodal LLMs [3] have recently demonstrated tremendous potential in multimodal task handling. Furthermore, Multimodal pretraining enforced the new standards in the real-time industry by consistently improving performance across various downstream activities. Traditional LLM models have been developed to process and comprehend textual input; however, recent works have concentrated on enhancing the existing systems to handle multimodal data, including audio, video, and images. In order to customize the LLM for the non-textual modalities, LLMs need to be integrated with different mechanisms and strategies [4, 5]. For example, a Convolutional Neural Network (CNN) is paired with the LLM to combine image modality with textual modality and to associate late fusion or cross-attention strategies. By late fusion, separate models can produce independent representations of the modalities combined to produce a response or prediction after the process.

Conversely, cross-attention methods allow the model to flexibly focus on pertinent information across modalities, resulting in a more comprehensive understanding of the data. To further enhance LLMs for multimodal tasks, advanced architectural adaptations are implemented [6, 7], such as the shared embedding space in which textual, visual, and audio data are all mapped into a unified vector space, thereby enabling the model to align and correlate data from different modalities effectively. Also, multimodal pretraining has been introduced in which the LLM is pre-trained on large-scale datasets that combine text with other modalities, such as image captions or video descriptions. Accordingly, the multimodal LLMs [8] enable the decision-making model to learn relationships across modalities during training, improving its performance in cross-modal tasks such as visual question answering or image captioning systems. Thus, enhancing LLM is essential to handle complex, multimodal queries with greater accuracy and relevance.

1. **Significance of the Research**

Improving LLMs in cross-modal query understanding is primarily driven by the possibility of utilizing the capabilities of robust AI systems in various industries. For instance, to deliver precise diagnoses, multimodal systems in the healthcare industry evaluate medical textual records in addition to radiological scan images. Similarly, the models assist customers in e-commerce product searches by fusing textual descriptions with visual elements. Efficiently managing cross-modal queries paves the way for intelligent systems to respond to users' demands in a contextually relevant and intuitive manner. Although improving LLMs for cross-modal query understanding becomes advantageous, substantial obstacles exist, such as the synchronization of several modalities due to the unique features and patterns. It is challenging for a single model to understand text and image characteristics simultaneously. Hence, architectural enhancement is essential for the multimodal LLM [9], such as cross-attention mechanisms and multimodal encoders with contextual alignment to ensure the coherent representation of multimodal data.

* 1. **Potential Applications**

The task of cross-modal query understanding has been involved in various real-time applications [10], which are listed as follows.

* Captioning
* Autonomous Driving
* Healthcare Diagnostics
* Surveillance and Security
* Multimodal Search Engines
* Visual Question Answering (VQA)
* E-learning

1. **Literature Survey**

**Enhanced LLM Approaches for Multimodal Data:** The work in [11] develops a Multimodal Large Language Model (MLLM), referred to as KOSMOS-2, that jointly learns the object descriptions in terms of bounding boxes and grounding texts in the visual environment. To train the LLM, it builds large-scale Grounded Image-Text (GRIT) pairs with the assistance of multimodal corpora. Also, it demonstrates its versatility in fusing language with visual information and includes multimodal grounding, producing referring expressions, perception-language tasks, and general language tasks. The work in [12] introduced a unique training paradigm that uses modularized learning to give large language models (LLMs) multimodal capabilities. In order to accommodate many modalities and facilitate collaboration between unimodal and multimodal skills, this paradigm incorporates a foundation LLM, a visual knowledge module, and a visual abstractor module. Two steps comprise the training process: first, a frozen LLM is used to train the abstractor and visual knowledge modules to align text and images. During the second stage, the visual knowledge module is locked, and the low-rank adaptation (LoRA) module on the LLM and abstractor is refined using both language-only and multimodal datasets. This method maintains and even improves the generating capabilities of the LLM.

In order to bridge the gap between language and sensory impressions, the work [13] presents the idea of embodied language models, which include continuous real-world sensor modalities in language models. Multimodal sentences are used as the input of PaLM-E. The phrases integrate textual, visual, and continuous state estimation input encodings. The encodings are trained end-to-end to execute various embodied activities, including captioning, visual question answering, and sequential robotic manipulation planning, in conjunction with a pretrained large language model. The work in [14] presents the idea of federated example selection to improve AI models' multimodal reasoning capabilities by addressing critical issues with data privacy, diversity, and representation. This method incorporates a federated example selection algorithm into the open-source Llama large language model (LLM), resulting in significant performance gains on a range of multimodal tasks. Within a federated learning framework, the methodology encompasses an all-encompassing system that guarantees safe data gathering, algorithmic design, and iterative model improvement. By examining it from a new representation learning angle, the work in [15] addresses the problem of hallucinations in MLLMs. It employs the hallucinogenic material as complex negative examples to include contrastive learning in MLLMs to overcome the research constraints in multimodal learning. By efficiently minimizing hallucinations and improving cross-modal representation alignment, it helps to represent non-hallucinating text and related visual samples.

The work in [16] integrates visual knowledge at two levels to improve multimodal large language models (MLLMs). Firstly, it presents a step-by-step integration of fine-grained spatially aware visual knowledge via a vision aggregator that works with vision-language (VL) tasks at the area level. It devises a stage-wise instruction-tuning technique employing a mixture of adapters to resolve the conflict between image-level and region-level VL tasks, enabling mutual reinforcement between both tasks. Second, it incorporates a variety of picture tags within the model's text instruction, so utilizing soft prompting of high-level semantic visual evidence. The work in [17] improves vision-and-language models' ability to respond to inquiries requiring outside information. The method focuses on enhancing the models by including an external knowledge source comprising multimodal documents instead of merely depending on innovative structures and adapters. Wiki-LLaVA gains an advantage over other models in handling dialogues that need comprehension of both visual and external knowledge by this strategy to generate more accurate and contextually enhanced responses. The work in [18] presents Multimodal Composition Learning (MCL) to improve the synchronization and fusion of visual and linguistic inputs. Multimodal-Context Retrieval (MC-Ret) attempts to retrieve contextually pertinent knowledge from the inputs. Multimodal-Context captioning (MC-Cap) is intended to produce correct textual descriptions from integrated visual and linguistic inputs. By enhancing the accuracy and caliber of text generation and visual representation, the exercises aid frozen LLMs in comprehending multimodal settings more effectively.

**LLM-assisted Cross-Modal Learning Approaches:** The work [19] introduces a Large Language Model-guided Moment Retrieval (LMR), which utilizes the extensive knowledge that LLMs possess to improve cross-modal alignment and video context representation, ultimately leading to more accurate moment localization in films. The context enhancement method LMR presents uses LLMs to produce target-related context semantics. The semantics are combined with visual elements to produce more recognizable video representations. In order to enhance Visual Question Answering (VQA) by integrating contextual information, the work [20] presents a prompting technique called Causal Context Generation (Causal-CoG). In order to provide more accurate answers to inquiries, this method asks multimodal language models (MLMs) to produce text descriptions of images. It also investigates the causality perspective on the impact of context, presenting a causality filtering technique that chooses samples when contextual information significantly improves the VQA performance. The work in [21] provides a multi-turn interactive picture retrieval system that allows users to improve their queries based on relevance feedback. In order to enhance the quality of text-based inquiries, a vision-language model (VLM) for image captioning is incorporated. To further improve and rectify errors in the expanded text queries, an LLM denoiser is utilized.

By adopting LLMs as virtual knowledge bases, the work [22] presents an end-to-end generative framework for multimodal information retrieval. Using database searches, pertinent documents are retrieved, and knowledge hints are generated from queries. It utilizes a knowledge-guided generation technique during decoding to improve clue generation, and it introduces object-aware prefix-tuning to align visual features with LLMs for cross-modal learning. The CCF-LLM framework is introduced by the work in [23, which employs the user-item interactions in the hybrid prompts that combine signals for collaboration and semantic understanding. Also, to efficiently combine latent embeddings from various modalities, the system exploits an attentive cross-modal fusion technique. The CCF-LLM framework improves suggestion quality by utilizing collaborative filtering and semantic understanding to produce recommendations that are more precise and pertinent to the context. The work in [24] presents a generative cross-modal retrieval framework. Initially, it teaches the MLLM to generate the corresponding identifier based on a given textual query, and then it learns the association between images and their identifiers. The MLLM is assigned unique identifier strings to represent images. Distinct from typical discriminative approaches, it presents a new paradigm in cross-modal retrieval, emphasizing image memorization within MLLMs for more efficient retrieval. The work in [25] introduces a Visual Question Answering (VQA) framework called Teach Prompt Learning (TPL), which is based on prompt-based learning and a cross-modal approach for deep semantic understanding. It applies prompt-based learning to combine language models and sophisticated visual encoders, improving the integration of visual content and verbal semantic reasoning. By utilizing continuous vectors that function as context words learned from the input, TPL broadens the semantic domain of the model and enhances its understanding of visual concepts.

1. **Gap Analysis and Drawbacks of the Existing Systems**

Even though different strategies for the multimodal data have enhanced LLMs, the cross-modal query understanding still confronts several shortcomings listed below.

* Aligning the heterogeneous representations of various modalities, such as text, images, and videos, is challenging in cross-modal learning.
* LLM relies on the modalities in the generic contexts for understanding; hence, exploring dynamic contextual adaptation remains a gap.
* Multimodal LLMs often lack the ability to generalize across different domains, particularly in cross-modal tasks, due to the expensive retraining for domain adaptation.
* Traditional multimodal LLMs often struggle with learning fine-grained textual and visual features for understanding complex queries due to the lack of interpreting the intricate relationships in the cross-modal queries.
* Supervised learning-based cross-modal understanding paradigms often require large annotated datasets for the interpreted multimodal inputs, which is infeasible and leads to the mapping of non-contextual modalities.

1. **Problem Statement**

In the task of cross-modal query understanding, interpreting complicated queries effectively is complex, and it involves reasoning and integrating across modalities such as text, images, audio, and video. Also, with the inherent heterogeneity in representations across multiple modalities, grounding multimodal inputs is critical; hence, resolving ambiguous and noisy data in user queries becomes challenging. Traditional LLMs cannot contextually understand the visual and textual data due to their inability to handle the fine-grained interpretable features in a particular domain, leading to contextual multimodal alignment issues. Due to the limited multimodal reasoning capabilities in the LLMs, bias, and inconsistencies in the multimodal training data affect performance. In addition, real-time multimodal applications often deal with scalability and efficiency constraints while handling the generalization and temporal dynamics. Hence, to contextually understand the cross-modal queries involving multiple modalities, enhancing the LLM to generate accurate and pertinent outputs is essential, improving the user experience and satisfaction with the real-time interactive applications.

1. **The Scope of the Research**

In the real world, the improved LLM enhances multimodal image-text applications, such as the caption generation, and Visual Question Answering (VQA), to accomplish cross-modal query processing precisely. Combining, processing, and reasoning across several modalities, such as text and images, enhanced LLM enables highly complex interpretation and contextual grounding of multimodalities. The enhanced multimodal LLM with the inputs of image-text data can potentially improve tasks, such as image captioning and visual question answering by integrating visual and textual information and investigating sophisticated cross-modal retrieval strategies. Research boundaries include examining the handling of complex multimodal data, creating innovative architectures that improve feature alignment, and ensuring the generalization across multiple modalities without requiring much retraining. Interpreting the generative AI model in the text query enrichment benefits the multimodal LLM's understanding of cross-modal inputs. Thus, this research aims to improve the performance of multimodal applications that utilize text and image modalities, allowing contextually-aware and coherent applications in various domains.

1. **Research Questions**

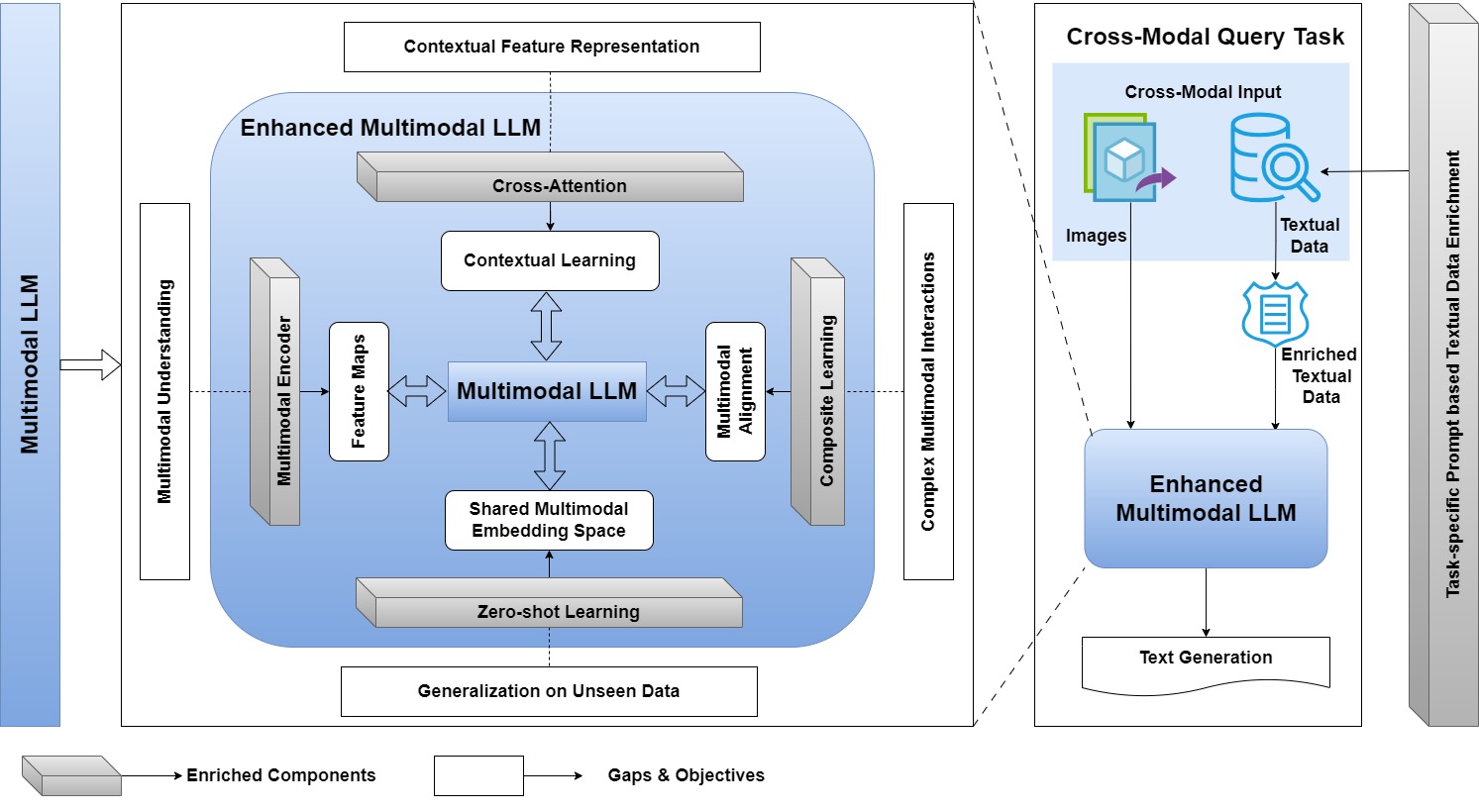
* How does the research work handle the ambiguity and noise in the multimodal queries to improve the LLM response?
* How do the multimodal alignment and representation strategies enhance cross-modal query interpretation by the LLM?
* Which optimization method in the research efficiently integrates the multimodal inputs to interpret cross-modal queries accurately?
* How does the research work adaptively and automatically to enhance the ability of LLM to identify the different combinations and contexts of modalities in the user queries?
* Which method effectively extracts the contextually relevant features from the LLM inputs to identify user intention across various modalities?
* How does the research customize the LLM from its pretrained generalized cross-modal knowledge with the interactive outputs?

1. **Aims and Objectives**

* To enhance the LLM for an efficient and accurate multimodal interpretation across various heterogeneities, ambiguities, and noises, even in unseen data.
* To develop the cross-modal query understanding paradigm with the enriched semantic and contextual representation of the LLM among the multimodal inputs.
* To design the adaptive feature representation for multimodal inputs to personalize the LLM in understanding the user inputs from the generative outputs.

1. **Proposed Methodology**

This work introduces a unique LLM architecture that integrates text and image modality with the enhanced LLM in cross-modal query processing. In order to improve the capability of learning multiple modalities and create richer representations of complex interactions, this work adopts the composite learning advantages in the multimodal LLM architecture. Also, the cross-attention mechanism enables dynamic alignment between modalities, whereas zero-shot learning ensures higher generalization of unseen domains or data, achieving improved interpretability and accuracy. Furthermore, task-specific prompt-based enrichment of input text data personalizes capturing multimodal feature maps with contextual and semantic information, thereby improving multimodal comprehension and response generation. Figure 1 illustrates the enhanced LLM and its application in the cross-modal query understanding task.



**Figure 1: Illustration of Enhanced LLM**

**Enhanced Multimodal LLM:** Besides the input encoder, feature fusion, shared embedding space, and decoder in the traditional multimodal LLMs, the enhanced multimodal LLM comprises cross-attention, zero-shot learning, and composite learning components.

**Multimodal Encoder:** In the multimodal LLM, the multimodal encoder is essential to learn the feature maps from each modality instead of early fusing the input modalities in which feature representation and encoders vary with the enhancement over the modalities.

**Cross-Attention:** In multimodal LLM, cross-attention is an essential process that allows for dynamic alignment between multiple modalities, including text and images. It improves the comprehension of intricate relationships across the modalities by enabling it to concentrate on pertinent aspects from one modality while considering context from another. Cross-attention generates coherent and contextually appropriate outputs by computing the attention scores to weigh feature importance.

**Zero-shot Learning:** In multimodal LLM, the model can generalize across modalities and tasks without requiring specialized training through the contextually shared multimodal embedding space. Learning the joint representations captures the interactions between modalities, such as text and images. This approach eliminates the need for large labeled datasets and prior knowledge to provide predictions and answers for the unseen data or domain.

**Composite Learning:** To generate richer representations for multimodal data, LLM adopts composite learning that combines data from several modalities, including text and images. By capturing intricate relationships and dependencies between modalities it improves the LLM in taking advantage of each data type without compromising its contextual significance. In particular, composite learning allows the model to learn from many inputs simultaneously, improving the learning of semantic and contextual knowledge and performance on tasks like visual question answering and image captioning.

**Task-specific Prompt-based Textual Data Enrichmen**t: In addition to multimodal LLM components enhancement at the architecture level, this work focuses on customizing input textual data or queries with the knowledge of image modality to enhance the ability of comprehension and answer generation. It is accomplished by task-specific prompt engineering, offering the pertinent contextual and semantic information to prompts. Thus, the proposed approach produces highly accurate and relevant outputs in various cross-modal applications, providing logical and appropriate responses in a particular context through the enhanced LLM.

1. **Experimental Software Requirements and Performance Evaluation**

To implement the proposed model, the experimental system executes the proposed algorithm using Python programming language associated with Keras, scikit-learn, opencv, pillow, and sentence-transformer libraries. The Python software version 3.8 is implemented on the Ubuntu 16.04 64-bit machine to conduct the experiments. In the cross-modal query understanding, image captioning and visual query answering tasks are widely evaluated with the Flickr30K and COCO benchmark datasets and Visual Question Answering (VQA) v2.0 and Visual Genome datasets, respectively.

In the image captioning task, the Flickr30 dataset [26] comprises 31,783 images associated with an average number of 5 human-annotated textual descriptions. In contrast, the MS-COCO dataset [27] contains 330,000 images with five annotated sentences for each image. In the VQA task, the VQAv2 dataset [28] comprises 265,016 images with a minimum of three textual questions for each image and ten ground truth answers for each question. In contrast, the Visual Genome dataset [29] contains 101,174 images with 1.7 million question-answer pairs and 17 average textual questions for each image.

**Performance Metrics:**

This work employs metrics to assess the enhanced LLM in the cross-modal query understanding application, such as the precision @n, recall @n, and BLEU score.

**Precision @n**: It measures relevant results within the top-n outcomes observed from the images for a text query.

**Recall @n**: It measures the total retrieved relevant results within the top-n outcomes observed from the images for a text query.

**BLEU:** It measures text quality in terms of captions or descriptions generated by the model, referencing ground truth text based on the n-gram overlap.

1. **Conclusion**

This work suggested a research contribution for enhancing multimodal LLM to handle the cross-modal query understanding tasks. The enhanced multimodal LLM comprises cross-attention, composite learning, and zero-shot learning. Also, task-specific prompt engineering enriches the textual input for the multimodal LLM, improving the performance of LLM in cross-modal learning. Thus, the enhanced LLM ensured the provision of coherent, context-aware outputs in various multimodal applications, such as visual question answering and image captioning.

**References**

1. Hadi, M. U., Al Tashi, Q., Shah, A., Qureshi, R., Muneer, A., Irfan, M., ... & Shah, M. (2024). Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.
2. Huang, D., Yan, C., Li, Q., & Peng, X. (2024). From Large Language Models to Large Multimodal Models: A Literature Review. *Applied Sciences*, *14*(12), 5068.
3. Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., ... & Cucchiara, R. (2024). The (r) evolution of multimodal large language models: A survey. *arXiv preprint arXiv:2402.12451*.
4. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., & Shou, M. Z. (2024). Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.
5. Qin, Z., Chen, D., Zhang, W., Yao, L., Huang, Y., Ding, B., ... & Deng, S. (2024). The Synergy between Data and Multimodal Large Language Models: A Survey from Co-Development Perspective. *arXiv preprint arXiv:2407.08583*.
6. Wang, J., Jiang, H., Liu, Y., Ma, C., Zhang, X., Pan, Y., ... & Zhang, S. (2024). A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks. *arXiv preprint arXiv:2408.01319*.
7. Zhang, D., Yu, Y., Li, C., Dong, J., Su, D., Chu, C., & Yu, D. (2024). Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.
8. Wang, Y., Chen, W., Han, X., Lin, X., Zhao, H., Liu, Y., ... & Yang, H. (2024). Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.
9. Bai, T., Liang, H., Wan, B., Yang, L., Li, B., Wang, Y., ... & Zhang, W. (2024). A Survey of Multimodal Large Language Model from A Data-centric Perspective. *arXiv preprint arXiv:2405.16640*.
10. Zhou, R. (2024). Advanced Embedding Techniques in Multimodal Retrieval Augmented Generation A Comprehensive Study on Cross Modal AI Applications. *Journal of Computing and Electronic Information Management*, *13*(3), 16-22.
11. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., & Wei, F. (2023). Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.
12. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., ... & Zhou, J. (2023). mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.
13. Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., ... & Florence, P. (2023). Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.
14. Fawcett, B., Ashworth, F., & Dunbar, H. (2024). Improving multimodal reasoning in large language models via federated example selection.
15. Jiang, C., Xu, H., Dong, M., Chen, J., Ye, W., Yan, M., ... & Zhang, S. (2024). Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 27036-27046).
16. Chen, G., Shen, L., Shao, R., Deng, X., & Nie, L. (2024). Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 26540-26550).
17. Caffagni, D., Cocchi, F., Moratelli, N., Sarto, S., Cornia, M., Baraldi, L., & Cucchiara, R. (2024). Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1818-1826).
18. Li, W., Fan, H., Wong, Y., Yang, Y., & Kankanhalli, M. Improving Context Understanding in Multimodal Large Language Models via Multimodal Composition Learning. In *Forty-first International Conference on Machine Learning*.
19. Liu, W., Miao, B., Cao, J., Zhu, X., Liu, B., Nasim, M., & Mian, A. (2024). Context-Enhanced Video Moment Retrieval with Large Language Models. arXiv preprint arXiv:2405.12540.
20. Zhao, S., Li, Z., Lu, Y., Yuille, A., & Wang, Y. (2024). Causal-CoG: A Causal-Effect Look at Context Generation for Boosting Multimodal Language Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13342-13351).
21. Zhu, H., Huang, J. H., Rudinac, S., & Kanoulas, E. (2024, May). Enhancing Interactive Image Retrieval With Query Rewriting Using Large Language Models and Vision Language Models. In Proceedings of the 2024 International Conference on Multimedia Retrieval (pp. 978-987).
22. Long, X., Zeng, J., Meng, F., Ma, Z., Zhang, K., Zhou, B., & Zhou, J. (2024, March). Generative multimodal knowledge retrieval with large language models. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 17, pp. 18733-18741).
23. Liu, Z., Zhang, H., Dong, K., & Fang, Y. (2024). Collaborative Cross-modal Fusion with Large Language Model for Recommendation. arXiv preprint arXiv:2408.08564.
24. Li, Y., Wang, W., Qu, L., Nie, L., Li, W., & Chua, T. S. (2024). Generative cross-modal retrieval: Memorizing images in multimodal language models for retrieval and beyond. arXiv preprint arXiv:2402.10805.
25. Zhu, S., Peng, S., & Chen, S. (2024, April). Enhancing Visual Question Answering with Prompt-based Learning: A Cross-modal Approach for Deep Semantic Understanding. In Proceedings of the International Conference on Algorithms, Software Engineering, and Network Security (pp. 713-717).
26. Flickr30K Dataset, Available Online at: <https://datasets.activeloop.ai/docs/ml/datasets/flickr30k-dataset/>, Accessed On September, 2024
27. MS-COCO Dataset, Available Online at: <https://paperswithcode.com/dataset/coco-captions>, Accessed On September, 2024
28. VQAv2 Dataset, Available Online at: <https://paperswithcode.com/dataset/visual-question-answering-v2-0>, Accessed On September, 2024
29. Visual Genome Dataset, Available Online at: <https://paperswithcode.com/dataset/visual-genome>, Accessed On September, 2024