Journal of Organizational and End User Computing
Volume 36 • Issue 1

Enhancing Multimodal Understanding With LIUS:
A Novel Framework for Visual Question Answering in Digital Marketing

Chunlai Song, Department of Global Business, Kyungil University, South Korea*

ABSTRACT

VQA (visual question and answer) is the task of enabling a computer to generate accurate textual answers based on given images and related questions. It integrates computer vision and natural language processing and requires a model that is able to understand not only the image content but also the question in order to generate appropriate linguistic answers. However, current limitations in cross-modal understanding often result in models that struggle to accurately capture the complex relationships between images and questions, leading to inaccurate or ambiguous answers. This research aims to address this challenge through a multifaceted approach that combines the strengths of vision and language processing. By introducing the innovative LIUS framework, a specialized vision module was built to process image information and fuse features using multiple scales. The insights gained from this module are integrated with a "reasoning module" (LLM) to generate answers.

KEYWORDS

Digital Marketing, Feature Extraction and Fusion, Image Features, LLM, Text Information, Text-Image Matching, VQA

1. INTRODUCTION

Visual Question Answering (VQA) is an interdisciplinary research field that combines computer vision and natural language processing (NLP). Its goal is to develop intelligent systems capable of comprehending visual content and textual questions and generating accurate natural language answers. In the VQA task, a system is presented with an image and a question related to the image content. The ultimate objective is for the system to understand both visual and textual information, and generate accurate answers relevant to the questions posed (Akula et al., 2021). To achieve this goal, the system needs to extract meaningful features from the image, analyze grammar, recognize keywords, and grasp subtle contextual differences in the questions. It then effectively bridges the visual and textual cues to produce coherent and accurate responses.

At the current stage, the VQA task still faces challenges. For instance, research has shown that models might rely on spurious language correlations rather than multimodal reasoning. Asking what sport is associated with the VQA v1.0 dataset, the model might simply answer "tennis," achieving an

DOI: 10.4018/JOEUC.336276 *Corresponding Author

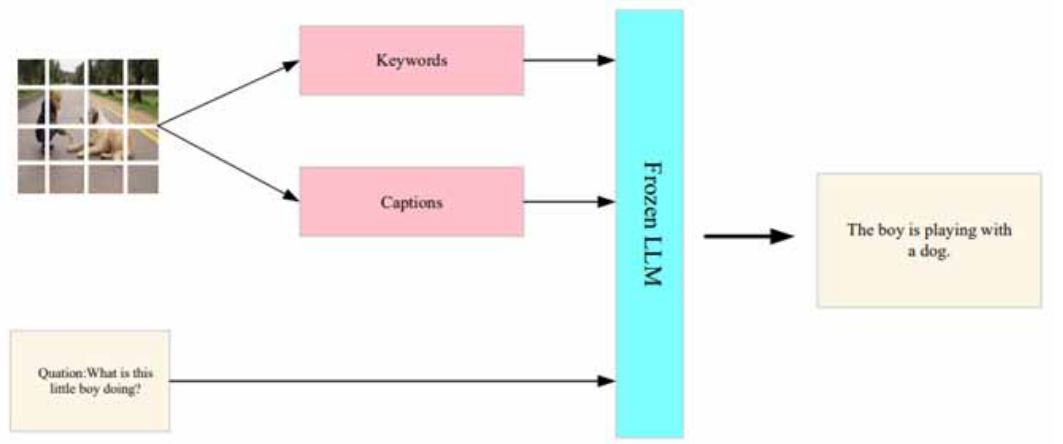
accuracy of around 40% (Anderson et al., 2018). Additionally, the computational cost associated with extra pre-training stages remains a challenge. For example, Flamingo introduces new cross-attention layers into the LLM (Language and Vision Model), incorporating visual features, and then pre-trains from scratch. This pre-training phase still requires more than 2 billion image-text pairs, covering around 43 million web pages, and takes approximately 15 days (Antol et al., 2015).

Moreover, in the process of handling image features, different scales and levels of information are often overlooked. Although Vision Transformers use self-attention mechanisms to model relationships between different positions in images and capture contextual information at a global scope, certain low-level image features like textures, edges, and colors can be disregarded. To address the aforementioned issues, as illustrated in Figure 1, a modular VQA system is proposed in this work, built on top of pre-trained LLMs. This offers three benefits. Firstly, it reduces deployment costs and simplifies the deployment process. Secondly, upgrading the LLM is straightforward. Additionally, image feature information is extracted from various angles and categorized into labels, attributes, actions, relationships, etc.

A key task of the LIUS model is feature extraction and fusion, which is closely related to the model's accuracy (Yang et al., 2016). Since features extracted by deep neural networks and those extracted by shallow neural networks carry different meanings, it is necessary to comprehensively consider multi-scale feature extraction and the establishment of connections between different positions in images through self-attention mechanisms. In previous research, Convolutional Neural Networks (CNNs) have been commonly used (Schwenk et al., 2022). CNNs process images hierarchically, gradually learning features ranging from low-level edge features to high-level object representations. In image processing, lower layers of convolutional layers can capture basic features like edges, corners, and textures. As the layers go deeper, convolutional layers start capturing more complex features like object contours, combinations of textures, and some simple shapes. Deeper layers can capture advanced object representations, including parts, compositions, and wholes, aiding accurate comprehension of image content in VQA tasks. However, with the increasing depth of convolutional layers, the issue of vanishing gradients arises during backpropagation. This phenomenon involves the gradual reduction of gradients, which can make the training process challenging and even hinder the network from learning deeper-level features.

Pre-trained models play a significant role in image feature extraction (Guo et al., 2023). Taking ResNet-152 as an example, this deep convolutional neural network has achieved remarkable success in the field of image recognition. It makes the network more amenable to training and optimization,

Figure 1. The LIUS framework



alleviating issues like vanishing and exploding gradients. However, deep-layer networks might lead to the learning of less useful information, thus affecting the model's generalization ability.

Furthermore, image pyramids and feature pyramids are commonly used strategies in VQA tasks to comprehensively capture multi-scale information within images (Lu et al., 2023). An image pyramid involves obtaining a series of images at different scales by scaling the input image. On the other hand, a feature pyramid entails extracting features at different levels of convolutional networks. The objective of these strategies is to capture details of objects, global context, and other significant features at different scales, thereby providing a more abundant source of information for VQA systems. However, due to the differences between these two pyramid mechanisms, it is challenging to capture all crucial information in a single model.

We propose the LIUS approach, which leverages Language and Vision Models (LLMs) as the "reasoning module," combined with an independent "vision module" setup. In the LIUS method, we first utilize a pre-trained vision module to extract rich visual information. To combine high-resolution features from shallow networks with high-level semantic information from deep networks, we introduce the feature pyramid attention Network (FPAN). This algorithm extracts features from deep networks, obtains features matching shallow-level ones through upsampling, and then employs attention mechanisms to weight these features for the input of the classification network. Subsequently, the fusion of these two sets of features is utilized as the output of the classification network. Additionally, we employ ALBEF to extract image information, a strategy that has demonstrated promising results in classification tasks. The integrated concept of LIUS enables us to harness the latest advancements in both computer vision and natural language processing, maximizing the strengths of these fields.

- The LINS method proposed in this study integrates the Language and Vision Model (LLM) as an "inference module" with an independent "visual module." This modality fusion effectively harnesses the strengths of both modalities, enabling the system to comprehensively grasp visual content and textual questions.
- In the LINS method, the introduced FPAN signifies a noteworthy breakthrough in capturing multi-scale information. This innovative approach adaptively assigns weights to features at different scales, facilitating a more efficient capture of various image details. Furthermore, the incorporation of ALBEF for image feature extraction has been integrated. This thorough and precise analysis of feature characteristics significantly improves the model's accuracy.
- LIUS enables off-the-shelf LLMS without the need for costly end-to-end training, allowing for low-cost, flexible model deployment, and seamless LLM upgrades.

In the upcoming article structure, we will organize the content as follows: In Chapter 2, we will provide a detailed overview of related work. Chapter 3 will delve into the key details of our proposed model. Chapter 4 will focus extensively on our experimental design and results. Finally, Chapter 5 will serve as the conclusion and discussion of this research.

2. RELEVANT WORK

2.1 Transformer-Based Pre-Training

Using the Transformer as a backbone network for pre-training has significantly advanced the state of the art across the domains of natural language processing, computer vision (Vaswani et al., 2017), and vision-language tasks (Han et al., 2022). The self-attention mechanism inherent to the Transformer architecture has demonstrated remarkable proficiency in capturing long-range dependencies and contextual information, thereby enhancing performance across a variety of tasks. This progress has been particularly notable in areas such as natural language understanding, text generation, sentiment analysis, and machine translation (S. Zhang et al., 2022).

In the realm of computer vision, adopting Transformer-based pre-training methods has fundamentally transformed tasks such as image classification, object detection, and segmentation (Scao et al., 2022). The self-attention mechanism allows models to consider the global context of images and relationships between different regions, leading to more accurate and robust feature representations (Jin et al., 2021). This revolution has achieved breakthroughs in image understanding and visual feature extraction, further blurring the boundaries between computer vision and natural language processing (Li et al., 2020). Additionally, in the domain of natural language tasks, Transformer-based pre-training methods have achieved notable accomplishments. For instance, models like GPT-3.5 and GPT-4 efficiently grasp and capture relationships between sentences and paragraphs, thereby generating more coherent and natural text (Li et al., 2020). Furthermore, the availability of diverse pre-trained models enables users to fine-tune the models according to task requirements, significantly alleviating computational costs (Firat, 2023).

2.2 Image-Text Foundation Models

Recent work has introduced image-text foundational models that can encompass visual and visual-linguistic pretraining. For instance, the CLIP model draws inspiration from pretraining methods in the field of natural language processing (NLP) (Shen et al., 2021). Through self-supervised learning on large-scale text corpora, it learns rich language representations. The training process involves joint training of image encoders and text encoders. The model learns to embed images and text into a shared representation space and employs a contrastive loss function to encourage closer embedding distances for the same image-text pairs and farther distances for different pairs. Despite its ability to understand correspondences between images and text, CLIP's semantic understanding of images and text remains limited. It may fail to capture fine-grained relationships or deep semantic meanings, leading to limited performance on certain tasks.

In contrast, ALBEF introduces alignment-guided loss, maximizing mutual information between image and text representations to align vision and language representations (J. Li et al., 2021). This significantly enhances interaction-based learning between vision and language and boosts model performance. To improve learning efficacy on noisy data, ALBEF proposes the momentum distillation method. This method uses a momentum model to generate pseudo-targets, providing additional supervisory signals. Momentum distillation prevents penalization for generating reasonable but different outputs from network annotations, thereby enhancing the model's generalization capacity.

Subsequently, the BLIP model emerges, adopting a unified model architecture known as the Multimodal Mixture of Encoder-Decoder (MED)(Li et al., 2022). This architecture is equally applicable to understanding-based and generation-based tasks. MED comprises an image encoder and a language decoder. The image encoder segments input images into multiple image blocks, encoding them into embedding sequences. Meanwhile, the language decoder generates corresponding language descriptions based on the image encoder's output. During the pretraining phase, BLIP is trained on extensive image-text pairs. The training process includes three objectives: image-text contrastive learning, image-text matching, and image-conditioned language modeling. Through these objectives, the model learns the correlations and correspondences between images and language. These endeavors in the multimodal domain offer novel approaches and methods for integrating images and text, holding the promise to further propel VQA research and applications.

2.3 Enhancing VQA With Language Models

In previous Visual Question Answering (VQA) tasks, a dual-branch architecture was commonly employed to perform feature extraction and fusion between images and questions(Hou et al., 2020). This dual-stream structure enables the encoding of information from images and questions into feature representations. These representations are then combined using specific fusion methods to achieve more accurate question-answering results. An exemplary multimodal model is "VilBERT," a fusion of two prominent models(Lu et al., 2019), "BERT" and "VisualBERT." The text branch

utilizes a structure akin to BERT for encoding input questions, while the image branch adopts a structure similar to VisualBERT, employing Convolutional Neural Networks (CNNs) to encode input images. Subsequently, the features from text and images are fused together, allowing the model to simultaneously consider both question and image information.

While dual-stream networks have achieved considerable success in Visual Question Answering (VQA) tasks, they also face certain limitations and challenges. For instance, dual-stream networks often necessitate separate feature extraction for images and text, which can lead to increased computational costs. Moreover, the increase in model parameters can escalate the complexity of training and inference. Furthermore, discrepancies in modality between images and questions may result in the loss or confusion of information during the feature fusion process. The modal mismatch could potentially impact model performance, especially when significant semantic differences exist between images and questions. Consequently, several studies have adopted language model-based structures for LLM utilization in visual-related tasks (Berrios et al., 2023).

One technique involves training visual encoders to represent each image as a continuous sequence of embeddings, thereby enabling LLM to comprehend(Y. Li et al., 2021). Another approach employs a frozen visual encoder, introducing new layers into the frozen LLM while concurrently conducting contrastive training, followed by retraining from scratch. Additionally, another method suggests employing frozen visual encoders (pre-trained with contrastive learning) and frozen LLM, aligning them using the training of a lightweight Transformer (Whalen & Mouza, 2023). Overall, these language model-based approaches offer promising solutions to overcome challenges faced by the dual-stream structure in visual question answering tasks. By effectively leveraging the application of language models in visual tasks, these methods have the potential to enhance model performance, reduce computational complexity and parameter count, and thus drive further advancements in the field of visual question answering. However, the effectiveness of different methods may vary across distinct tasks and datasets, necessitating further research and experimentation to validate their efficacy and applicability.

3. METHOD

We propose a new framework, called Lius, aimed at enhancing the capabilities of frozen Language and Vision Models (LLMs) by enabling them to handle visual and visual-linguistic tasks beyond their existing natural language understanding abilities. This approach involves introducing an independent "Visual Module" to process image information and then integrating it with the "Inference Module" (LLM) to achieve comprehensive multimodal comprehension.

In the visual module, we adopt two branches to extract feature information from images. The first branch incorporates a pre-trained ResNet-152 model from the standard model library as an image feature extraction network. Through this branch, we are able to obtain a multi-level feature representation of the image, ranging from low-level to high-level, encompassing rich information from edges to objects. To fuse features at different scales, we utilize the FPAN model, which combines features from different layers in a top-down manner. This approach adaptively weights features from different layers, resulting in fused features with enhanced multi-scale expressive capability.

In the other branch, we employ the ALBEF method to match images with text. ALBEF effectively models the correlations between images and questions, thereby facilitating the fusion of visual and language information. By aligning textual information with image information, we gain a better understanding of questions and extract relevant features from images, thus improving the model's performance in multimodal tasks. Following processing through the visual module, we obtain objects, attributes, and captions of images, which serve as content for LLM inference.

In conclusion, the introduction of the Lius framework enhances the capabilities of frozen LLMs in handling visual and multimodal tasks. Through feature extraction and fusion across multiple branches, as well as the application of the ALBEF method, we achieve comprehensive understanding of both

image and text information, leading to improved performance across various multimodal tasks. This innovative approach offers a new solution for cross-disciplinary visual and language tasks, with the potential to drive the development of multimodal intelligent systems in future research. The overall structure of the model is shown in Figure 2.

LIUS executes computer vision and visual reasoning tasks through a frozen LLM and a set of "vision modules". LIUS leverages these vision modules to retrieve a textual description for an image which is used by the "reasoning module" (LLM) to generate a response for a given query.

3.1 Image Tag Mode

In our experiments, a pre-trained residual network is employed for image feature extraction. Since this study utilizes a visual question answering database based on the COCO image dataset for tag extraction, the pre-trained ResNet-152 model available in the standard model library can be utilized as the image feature extraction network (L. Zhang et al., 2022). Simultaneously, the final outputs of the conv3, conv4, and conv5 residual modules are extracted. For the fusion of multi-scale features, we adopt the method of the FPAN model, combining features from different layers in a top-down manner. It is noteworthy that the adaptive weighting strategy of FPAN allows us to better capture image details and semantic information across different scales, thus effectively supporting multi-modal tasks.

This study removes the highly relevant fully connected layers associated with classification tasks. Instead, the output of the final convolutional layer is used as the image feature. Lastly, a 152-layer residual network pre-trained on ImageNet is utilized as the image feature extraction model. Considering computational costs, the experiment does not generate answers for all scale features, but rather explores the optimal fusion mode among features of different scales to obtain a single fused feature, as illustrated in Figure 3. Specifically, the image size in the COCO dataset is standardized to 448×448 as the model input. The output feature size of conv3 is 28×28 with a channel size of 512. The output feature size of conv4 is 14×14 with a channel size of 1024. The output feature size of conv5 is 7×7 with a channel size of 2048, as shown in Figure 4. Referring to the FPAN fusion method and the feature fusion across different scales, we perform feature fusion in a top-down manner. For instance, in the merging of conv5 and conv4, we first process the features extracted by conv5. Subsequently, we employ a 1024-times transposed convolution with a stride of 2 and a dimension of 2048 to extend the features to the same feature size and channel count as conv4. The features are then element-wise added, and the new features are stored as the fused features of conv4 and conv5. A similar method is applied for the fusion of conv5 and conv3.

For the integration of features from these three layers, we further perform transposed convolutions on conv4 and conv5. With 512 transposed convolutions and a stride of 2, the feature size is extended

Figure 2. The LIUS framework

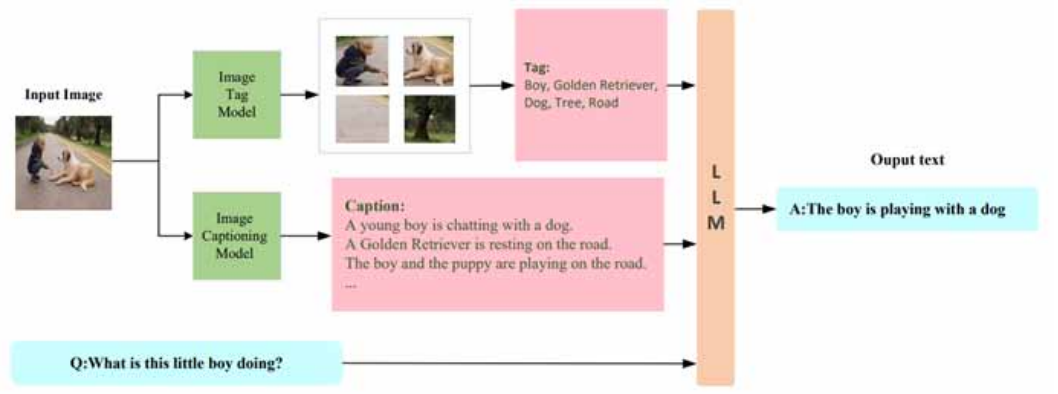


Figure 3. Multi-scale feature fusion in visual question answering system

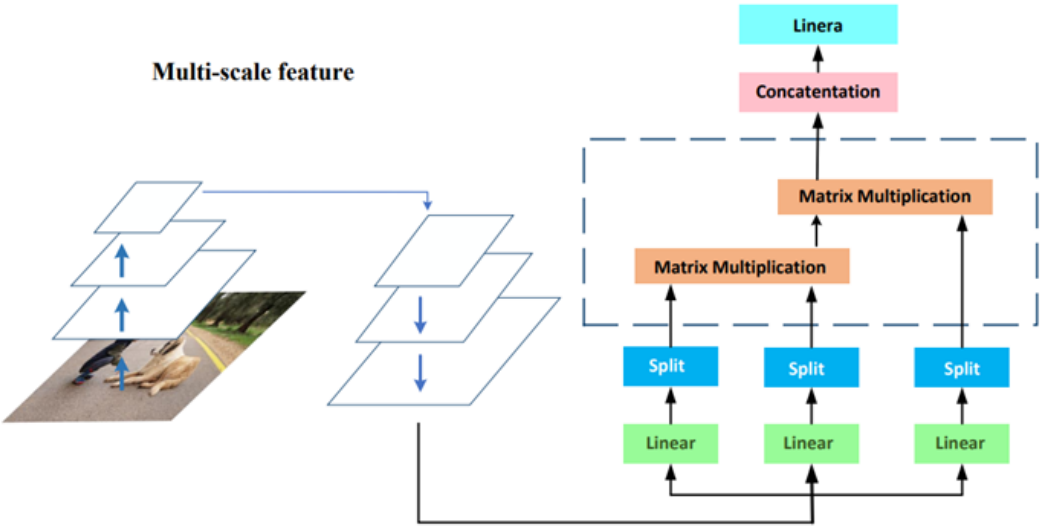
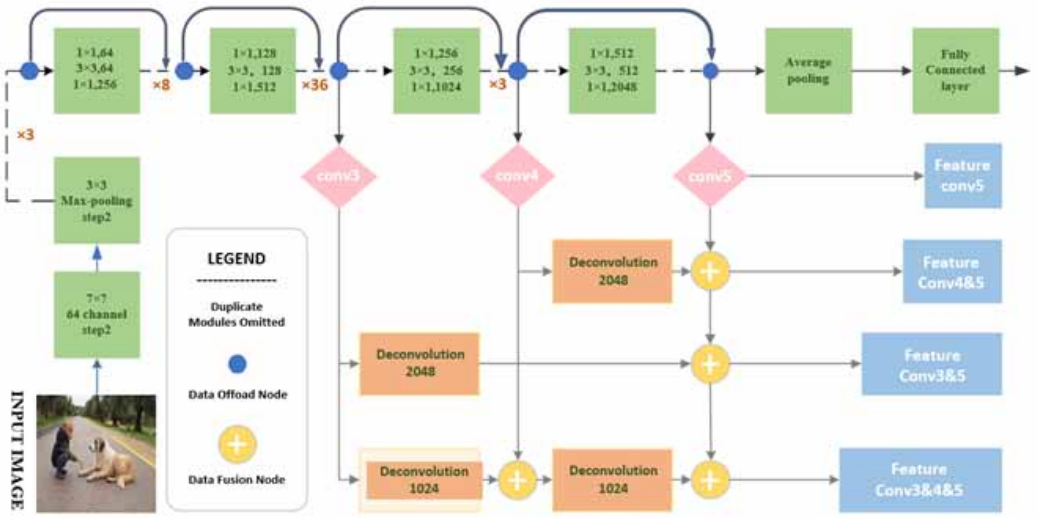


Figure 4. Multi-scale feature extraction model for images in VQA system



to 28 × 28, and the channel count is reduced to 512, matching the feature size and channel count of conv3. Finally, an attention mechanism is introduced to enhance feature expression.

3.2 Image Captioning Model

For LIus, a comprehensive description serves as the bridge that transforms images into textual information, subsequently allowing them to be processed by existing LLMs. Leveraging the ALBEF approach, we establish a close connection between images and text, enabling a deeper cross-modal understanding. ALBEF's alignment-guided loss facilitates effective matching between image features and text representations, enabling the model to accurately comprehend image content and relate it

to natural language queries. Through the guidance of ALBEF, we achieve higher-level semantic connections in LISU, enriching images with more nuanced semantic information and further enhancing the model's performance in visual and visual-linguistic tasks. This fusion of image and text holds significant potential in cross-modal tasks, providing robust support for model intelligence and diversity (J. Li et al., 2021). The network flowchart of ALBEF is illustrated in Figure 5.

4. EXPERIMENTS

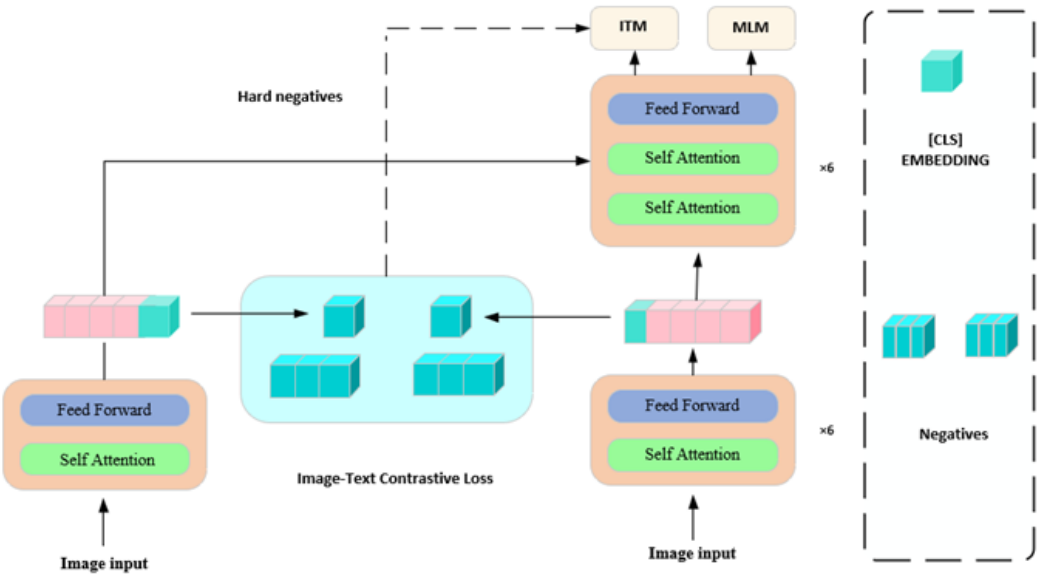
4.1 Environment Setup

4.1.1 Datasets

VQAv2 Dataset (Zhang et al., 2023): VQAv2 is a widely used dataset in visual question answering research, designed to assess models' ability to understand images and answer related questions. The dataset comprises over one million question-image pairs, encompassing images from diverse scenes and objects. Each question is associated with an image and may pertain to various aspects of the image, including its content, scenes, objects, locations, and more. The primary objective of this dataset is to assist models in comprehending images and providing accurate answers to questions related to them, thereby advancing research in the field of computer vision.

OK-VQA Dataset (Marino et al., 2019): OK-VQA is a dataset specifically designed for the task of visual question answering, with the aim of evaluating models' understanding and reasoning abilities when processing multimodal information, including both images and text. Compared to other visual question answering datasets, OK-VQA presents greater challenges as it includes a broader range of multimodal questions, requiring models to reason about different objects, relationships, and contextual information within images. To further elevate the challenge, A-OKVQA extends the OK-VQA dataset by introducing adversarial samples. These adversarial samples are created by injecting disruptive text or images into the questions and introducing redundant or confusing answers. This extension amplifies the complexity of the task, demanding that models possess stronger reasoning

Figure 5. Illustration of ALBEF. It consists of an image encoder, a text encoder, and a multimodal encoder.



and comprehension skills to provide accurate answers even in the presence of disruptions, thereby promoting research in multimodal understanding.

Rendered - SST2 Dataset (Huang et al., 2023): Rendered - SST2 is a dataset used for sentiment analysis, evaluating models' ability to analyze the sentiment polarity (positive, negative, or neutral) of given text. The dataset encompasses various types of sentences, ranging from news headlines to comments. Models are required to determine the sentiment inclination of each sentence and categorize it into positive, negative, or neutral sentiment categories. Rendered - SST2 contributes to the advancement of research in sentiment analysis and text understanding, providing valuable resources for natural language processing tasks.

Hateful Memes Dataset (Hamza et al., 2023): The Hateful Memes dataset aims to detect malicious or offensive content within memes, assessing models' capability to recognize potential malicious or offensive information within both image and text content. The dataset consists of a collection of images accompanied by corresponding text. Models are tasked with identifying the presence of objectionable content and classifying it accordingly. The Hateful Memes dataset holds significant importance in research related to computer vision and natural language processing, particularly in the evaluation of models' performance in malicious content detection.

4.1.2 Experimental Environment

In my experimental configuration, I employ the following hardware components: an Intel i7-13650 CPU as the processor, an NVIDIA GTX 4060 graphics card, and a generous 32 GB of memory. My software environment is structured as follows: it operates on a CUDA 11.6 general-purpose computing architecture, utilizes CUDNN 9.0 as the GPU acceleration library, and relies on the Python deep learning framework.

4.2 Implementation Details

Our model employs an innovative multimodal processing methodology, which includes the introduction of a distinct "visual module" seamlessly integrated with the "inference module" (LLM) to achieve a comprehensive grasp of image information. To generate contextually relevant captions for questions, we harness the capabilities of ALBEF [30] for caption generation and execute image-question matching. For the precise localization of image regions pertinent to questions, we utilize ResNet-152. Images are resized to dimensions of 448×448 to serve as input for the model. In the course of processing, conv3 generates feature maps measuring 28×28 with a complement of 512 channels. Conv4 yields feature maps sized at 14×14, housing 1024 channels, while conv5 produces feature maps at 7×7 with a capacity of 2048 channels.

Moreover, our approach incorporates a top-down feature fusion technique. Our model represents an innovative multimodal framework meticulously crafted to harness the synergies of linguistic and visual information, enabling a deeper understanding of the intricate interplay between images and textual content, and fostering advanced reasoning capabilities. Through the harmonious amalgamation of the visual module, FPAN, and ALBEF components, our model exhibits outstanding performance across a diverse array of tasks, offering pioneering solutions for cross-domain image and language challenges.

4.3 Main Results

Comparing State-of-the-Art Results with End-to-End Approaches:

Table 1 presents the performance on the VQAv2, OK-VQA, and A-OKVQA datasets, while also indicating whether each model underwent end-to-end training. It is worth emphasizing that our model achieved outstanding performance across all these tasks and datasets without the need for complex end-to-end training.Firstly, for models that underwent end-to-end training, they exhibited remarkable performance on the VQAv2 and OK-VQA datasets. However, their performance on the A-OKVQA dataset was comparatively lower. Furthermore, for models that did not undergo end-to-end training,

Table 1. Performance on VQAv2, OK-VQA, and A-OKVQA

Methods	End-to-End	VQAv2		A-OKVQA		OK-VQA
	Training?	Val	test	Val	Test	Test
$FewVLM_{base}$ (Alayrac et al., 2022)	√	43.4	42.5	33.5	32.5	11.6
$FewVLM_{large}$	√	47.7	47.6	35.8	34.5	16.6
$VLKD_{ViT-B/16}$ (Li et al., 2023)	√	38.6	39.7	34.2	34.6	10.5
$VLKD_{ViT-L/14}$	√	42.6	44.5	37.1	36.0	13.3
$Flamingo_{3b}$ (Barr et al.)	√	42.5	49.2	35.5	34.5	41.2
$Flamingo_{9b}$	√	50.3	51.8	37.1	38.1	44.7
$PICa_{175b}$ (Yang et al., 2022)	×	53.5	54.5	-	-	17.7
$LENS_{6.7b}$ (Gupta et al., 2022)	×	58.7	58.7	35.5	33.3	40.2
$LENS_{30b}$	×	59.5	59.8	37.8	35.7	40.5
$Img2LLM_{6.7b}$ (Guo et al., 2023)	×	57.6	57.0	33.3	32.2	38.2
$Img2LLM_{30b}$	×	59.5	60.4	36.9	36.0	41.8
$VLM2LLM_{6.7b}$	×	57.8	59.8	35.5	38.4	38.7
$VLM2LLM_{30b}$	×	59.9	60.0	38.7	38.5	42.3
$VLM2LLM_{175b}$	×	60.5	60.3	40.7	40.5	45.3
LIUS	×	61.5	61.3	41.2	41.5	45.6

their performance on the A-OKVQA dataset was relatively subpar. In contrast, our model achieved an outstanding accuracy of 45.6% on the A-OKVQA dataset, highlighting the exceptional performance and versatility of our model in multimodal tasks. It provides efficient solutions for various tasks without the need for cumbersome end-to-end training.

Testing the Ability to Handle Open-ended Questions:

Table 2 presents the performance of various methods on two different test sets (Test-dev and Test-std), as well as their accuracy in answering different types of questions. In this series of experiments, our model stands out with outstanding performance. Firstly, for "Yes/No" type questions, our model achieved superior accuracy on both test sets, reaching 84.68% and 84.31%, respectively, closely matching or surpassing other methods. This demonstrates the excellent ability of our model to understand and answer binary questions.

Table 2. Performance on VQA 1.0

Method	Test-dev (%)				Test-std (%)			
	Y/N	Num	Other	ALL	Y/N	Num	Other	ALL
QLAB (Ravi et al.)	82.33	39.55	52.15	55.71	76.71	34.95	46.63	55.89
SAN(Chen et al., 2022)	80.85	37.35	43.15	59.32	80.85	37.53	43.59	58.26
ASST-LSTM (Aishwarya et al., 2022)	86.87	37.56	44.35	68.53	81.53	38.45	44.52	59.62
Img2LLM (Guo et al., 2023)	85.32	39.42	58.52	68.50	82.23	39.52	57.41	66.59
LIVS (Gupta et al., 2022)	84.32	39.85	56.23	65.28	83.26	38.91	56.32	65.82
Ours	84.68	51.52	53.13	68.58	84.31	47.62	58.54	68.34

Secondly, for "Number" type questions, our model also performed exceptionally well, achieving accuracies of 51.52% and 47.62% on the Test-dev and Test-std test sets, respectively, placing it in a leading position among the various methods. This highlights our model's outstanding performance in handling quantity-related questions. Finally, for "Other" type questions, our model achieved excellent accuracy on both test sets, reaching 53.13% and 58.54%, respectively. This further confirms our model's superiority in handling diverse questions. Our model demonstrated outstanding performance on various types of questions and two test sets, showcasing its versatility and superiority across multiple tasks and datasets. These results not only emphasize our model's excellent performance in various challenging tasks but also highlight its unique advantages in zero-shot reasoning and cross-domain tasks.

Zero-shot Testing on Other Datasets:

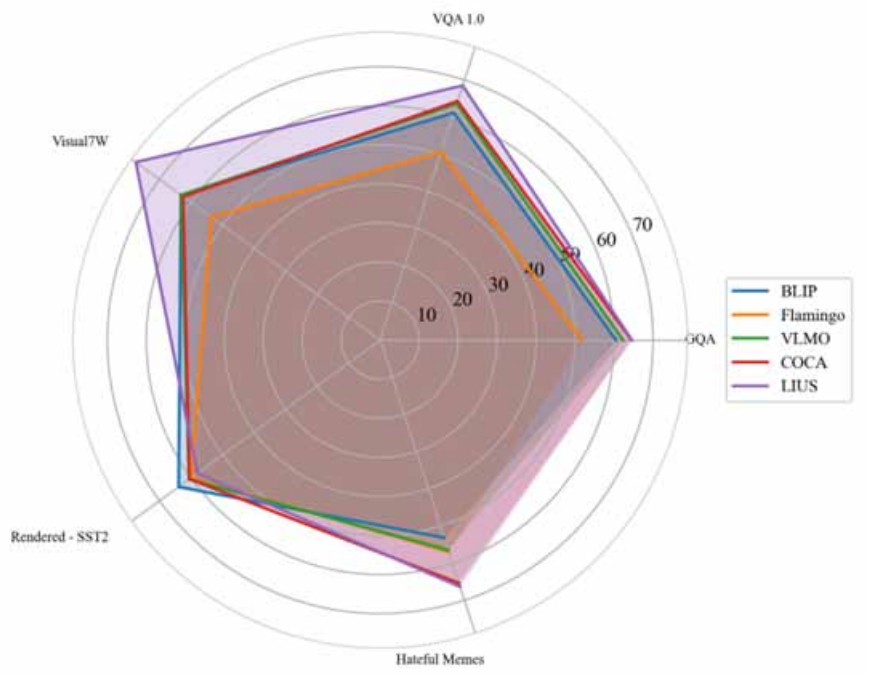
Our model has demonstrated outstanding performance across multiple tasks and datasets, as shown in Table 3 and Figure 6. On the GQA, VQA 1.0, Visual7W, Rendered - SST2, and Hateful Memes datasets, our model achieved accuracies of 64.54, 68.53, 77.65, 58.04, and 66.45, respectively, surpassing the performance of other models (BLIP, Flamingo, VLMO, and COCA) by a wide margin.

This series of remarkable achievements highlights the exceptional versatility of our model across different tasks and datasets. Of particular note is our model's accuracy of 77.65 on the Rendered - SST2 dataset, which requires the model to perform inference in a zero-shot scenario. This underscores our model's outstanding generalization ability when dealing with new tasks and domains. This outstanding zero-shot performance further emphasizes the unique advantages of our model.

Table 3. Unveiling LIUS's zero-shot prowess: Impressive performance across GQA, VQA 1.0, Visual7W, Rendered - SST2, and Hateful Memes datasets

Models	GQA	VQA 1.0	Visual7W	Rendered - SST2	Hateful Memes
BLIP (Li et al., 2022)	60.43	61.25	63.22	63.52	53.44
Flamingo (Barr et al.)	51.82	50.62	53.68	62.43	57.42
VLMO (Bao et al., 2022)	62.53	63.23	63.52	60.53	56.58
COCA (Yu et al., 2022)	64.52	64.42	64.23	65.23	65.68
Ours	64.54	68.53	77.65	58.04	66.45

Figure 6. Performance comparison of different models on various datasets. Different colors represent different methods.



4.4 Ablation Experiment

In Table 4 and Figure 7, LIUS underwent four ablation experiments on the OK-VQA dataset.

Ablation of image captioning model (ALBEF):

In the ablation study of the image captioning model (ALBEF), we assessed the performance of different multimodal models in the absence of the ALBEF model. The results demonstrate that the ALBEF model excels in this multimodal task, achieving an accuracy of 45.59, serving as the baseline model for comparison. In comparison, other multimodal models perform less effectively in the absence of image captioning capabilities provided by the ALBEF model. Specifically, the BLIP model achieves an accuracy of 43.55, VILBERT model reaches 44.52, CLIP model achieves 43.52, and COCA model attains 44.56. These results underscore the crucial role of the ALBEF model in multimodal tasks, with its image and text understanding capabilities significantly enhancing overall performance. Consequently, the ALBEF model emerges as the top-performing model in this multimodal task, showcasing its distinctive advantage in profound image-text understanding and reasoning.

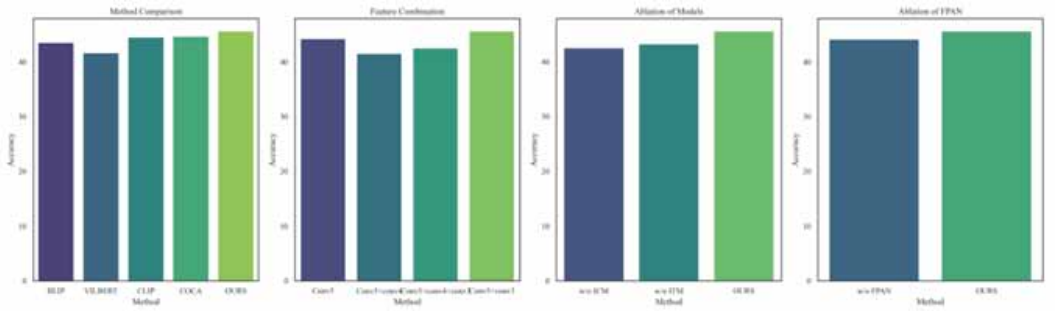
Ablation of different image feature combinations in ResNet-152:

In our comprehensive ablation study examining various combinations of image features within the ResNet-15 model, we systematically evaluated their influence on the model's performance in a multimodal task. The study encompassed four distinct image feature combinations, each yielding varying levels of accuracy. When utilizing only the image features from the Conv5 layer, the model achieved an accuracy of 40.52, indicating that relying solely on the deepest layer's image features resulted in relatively lower performance. Incorporating image features from both Conv5 and Conv4 layers in the Conv5+conv4 configuration led to a modest accuracy improvement, reaching 41.53. However, the addition of Conv4 features had a limited impact. When considering image features from Conv5, Conv4, and Conv3 layers in the Conv5+conv4+conv3 configuration, the accuracy showed a slight decline, measuring 40.56. This suggests that introducing more image features did not significantly enhance the model's performance. Finally, by focusing on image features from Conv5

Table 4.

Method	Accuracy
Ours (full model)	45.59
Ablation of image captioning model(ALBEF)	
BLIP	43.55
VILBERT	44.52
CLIP	43.52
COCA	44.56
ALBEF	45.59
Ablation of different image feature combinations in ResNet-15	
Conv5	40.52
Conv5+conv4	41.53
Conv5+conv4+conv3	40.56
Conv5+conv3	44.59
Ablation of image captioning model and image tags model	
w/o Image Captioning Model	40.52
w/o Image Tags Model	41.53
Ablation of multi-feature fusion	
w/o FPAN	43.52

Figure 7. The selected examples from LIUS employ the tags and captions modules, with the Resnet 152 and ALBEF as the visual encoders, and the language module is LLM



and Conv3 layers alone in the Conv5+conv3 configuration, the model achieved the highest accuracy of 44.59. These findings underscore the critical role of Conv3 features in improving performance in multimodal tasks and highlight the nuanced impact of different image feature combinations on model accuracy.

Ablation of image captioning model and image tags model

In this ablation study, we systematically assessed the impact of removing the image captioning model and the image tags model on the model's performance in a multimodal task, emphasizing their complementary roles. Firstly, we considered the scenario where the image captioning model was removed (w/o Image Captioning Model), resulting in a reduced accuracy of 40.52. This observation

underscores the critical importance of the image captioning model in the context of multimodal tasks, as its absence led to a notable performance decline. Secondly, we examined the case where the image tags model was removed (w/o Image Tags Model), yielding an accuracy of 41.53. Comparatively, the removal of the image tags model had a relatively smaller impact on performance, highlighting its secondary role. Collectively, these experimental findings emphasize the interdependency of the image captioning and image tags models in multimodal tasks. While both models contribute to overall performance, the absence of the image captioning model had a more pronounced effect, underscoring its indispensable role in comprehensively understanding and reasoning with combined image and text information. Therefore, this study highlights that the presence of both modules is essential for achieving optimal performance in multimodal tasks.

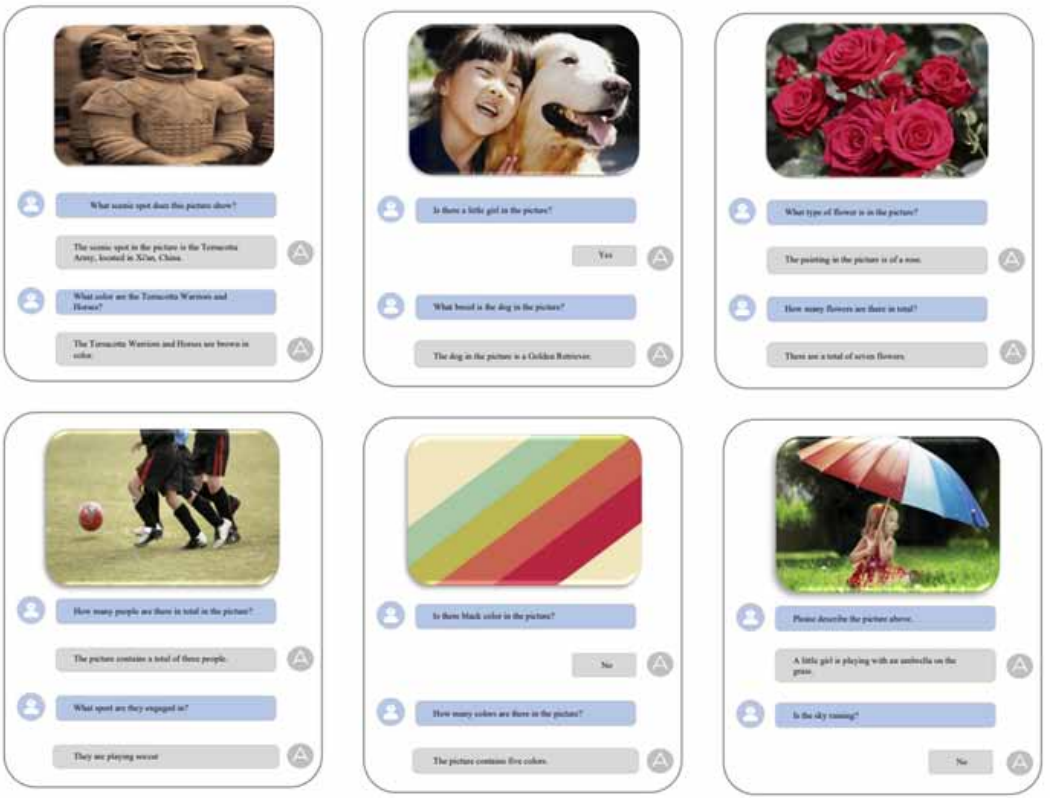
Ablation of multi-feature fusion:

In our final ablation experiment, we investigated the impact of removing the Feature Pyramid Attention Network (FPAN) from the multi-feature fusion process, denoted as "w/o FPAN," and examined its influence on the model's performance in a multimodal task. Upon the removal of FPAN, the model's accuracy recorded a value of 43.52. This experimental result underscores the pivotal role played by FPAN in the context of multimodal tasks, as its absence led to a notable decrease in model performance.

4.5 Presentation of Results

These examples vividly presented in Figure 8 show case the exceptional performance of our model in the realm of reasoning. These instances encompass complex scenes and contextual queries, and

Figure 8. Illustration of LIUS



through their responses to these questions, our model demonstrates a profound understanding of information, enabling it to consider multiple perspectives comprehensively across various levels and dimensions, thereby achieving more precise inference outcomes. These specific cases not only highlight the potential of our model but also further substantiate its unique advantage in tackling complex real-world challenges. These demonstrations provide strong support for the outstanding performance of our model and offer compelling evidence for its application in various multimodal tasks.

5. CONCLUSION

We introduce LIUS, a versatile and computationally efficient approach that effectively harnesses frozen Language and Vision Models (LLMs) to synergize with visual modules, achieving competitive performance even when compared to larger multimodal pretraining systems. It offers adaptability to a variety of open-source or black-box language models, regardless of their pretraining or multimodal data, providing flexibility and scalability for future performance enhancements within the community. By capitalizing on the strengths of LLMs and our modular approach, LIUS makes significant strides in solving certain tasks without the need for additional pretraining. Its seamless integration with various visual tasks showcases its versatility and potential for widespread applications. In future work, an intriguing direction to explore would be expanding LIUS's applicability, incorporating it into tasks involving different modalities. For instance, integrating LIUS into tasks like audio classification or video action reasoning could yield valuable insights. Such extensions would involve orchestrating the role of LLMs and integrating them with complementary modules.

Like any research endeavor, LIUS also comes with its own limitations. Firstly, the visual capabilities of LIUS heavily rely on its underlying visual components, namely ALBEF and Resnet152. While these models exhibit significant performance improvements, there is still room for further enhancement by effectively leveraging their strengths and combining them with LLMs. Future research should explore methods for integrating these models efficiently and harnessing the synergies between visual and language components to achieve better performance across various tasks. Secondly, it should be acknowledged that evaluating experiments using the LIUS model requires substantial computational resources. This might pose challenges for smaller or mid-sized labs as well as communities with limited access to such resources. Future efforts should strive to make computational resources more accessible and explore methods to alleviate the computational burden while maintaining the effectiveness of the approach.

In conclusion, LIUS offers an innovative solution for cross-domain visual and language tasks, leveraging the existing Language and Vision Models (LLMs) and modular design principles to achieve competitive performance. It demonstrates remarkable performance across various tasks, particularly evident in the VQA domain. This work establishes a foundation for further collaboration and innovation in the fields of vision and language, laying a robust groundwork for achieving deeper and broader multi-modal understanding. LIUS is more than just a model; it represents a significant endeavor in the exploration of multi-modal intelligence, showcasing the immense potential for advancements in intelligent systems.

REFERENCES

Aishwarya, R., Sarath, P., Sneha, U., & Manmadhan, S. (2022). Stacked Attention based Textbook Visual Question Answering with BERT. *2022 IEEE 19th India Council International Conference (INDICON)*.

Akula, A., Changpinyo, S., Gong, B., Sharma, P., Zhu, S.-C., & Soricut, R. (2021). Crossvqa: Scalably generating benchmarks for systematically testing vqa generalization. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., & Reynolds, M. (2022). Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems, 35*, 23716-23736.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2018.00636

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. *Proceedings of the IEEE International Conference on Computer Vision*. doi:10.1109/CVPR.2018.00636

Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O. K., Aggarwal, K., Som, S., Piao, S., & Wei, F. (2022). Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems, 35*, 32897-32912. doi:10.1109/CVPR.2018.00636

Chen, C., Han, D., & Chang, C.-C. (2022). CAAN: Context-Aware attention network for visual question answering. *Pattern Recognition*, *132*, 108980. doi:10.1016/j.patcog.2022.108980

Firat, M. (2023). What ChatGPT means for universities: Perceptions of scholars and students. *Journal of Applied Learning and Teaching*, *6*(1).

Guo, J., Li, J., Li, D., Tiong, A. M. H., Li, B., Tao, D., & Hoi, S. (2023). From Images to Textual Prompts: Zero-shot Visual Question Answering with Frozen Large Language Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Gupta, V., Li, Z., Kortylewski, A., Zhang, C., Li, Y., & Yuille, A. (2022). Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Hamza, A., Javed, A. R., Iqbal, F., Yasin, A., Srivastava, G., Połap, D., Gadekallu, T. R., & Jalil, Z. (2023). Multimodal Religiously Hateful Social Media Memes Classification based on Textual and Image Data. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., & Tao, D. (2022). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 87–110. doi:10.1109/TPAMI.2022.3152247 PMID:35180075

Hou, R., Zhao, Y., Hu, Y., & Liu, H. (2020). No-reference video quality evaluation by a deep transfer CNN architecture. *Signal Processing Image Communication*, *83*, 115782. doi:10.1016/j.image.2020.115782

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *International Conference on Machine Learning*.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems, 34*, 9694-9705.

Li, X., Fang, Y., Liu, M., Ling, Z., Tu, Z., & Su, H. (2023). Distilling large vision-language model with out-of-distribution generalizability. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., & Wei, F. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings*.

Li, Y., Yang, Z., & Hao, T. (2021). *TAM at VQA-Med 2021: A Hybrid Model with Feature Extraction and Fusion for Medical Visual Question Answering*. CLEF (Working Notes).

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems, 32.

Lu, S., Ding, Y., Liu, M., Yin, Z., Yin, L., & Zheng, W. (2023). Multiscale feature extraction and fusion of image and text in VQA. *International Journal of Computational Intelligence Systems*, *16*(1), 54. doi:10.1007/s44196-023-00233-6

Marino, K., Rastegari, M., Farhadi, A., & Mottaghi, R. (2019). Ok-vqa: A visual question answering benchmark requiring external knowledge. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2019.00331

Ravi, S., Chinchure, A., Sigal, L., Liao, R., & Shwartz, V. VLC-BERT: Visual Question Answering with Contextualized Commonsense Knowledge-Supplementary Material. *Science and Technology, 37*, 38.57. doi:10.1109/CVPR.2019.00331

Schwenk, D., Khandelwal, A., Clark, C., Marino, K., & Mottaghi, R. (2022). A-okvqa: A benchmark for visual question answering using world knowledge. *European Conference on Computer Vision*.

Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., & Keutzer, K. (2021). How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Whalen, J., & Mouza, C. (2023). ChatGPT: Challenges, Opportunities, and Implications for Teacher Education. *Contemporary Issues in Technology & Teacher Education*, *23*(1), 1–23.

Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., & Wang, L. (2022). An empirical study of gpt-3 for few-shot knowledge-based vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*. doi:10.1609/aaai.v36i3.20215

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* doi:10.1609/aaai.v36i3.20215

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*. doi:10.1609/aaai.v36i3.20215

Zhang, L., Li, H., Zhu, R., & Du, P. (2022). An infrared and visible image fusion algorithm based on ResNet-152. *Multimedia Tools and Applications*, *81*(7), 9277–9287. doi:10.1007/s11042-021-11549-w

Zhang, L., Zhai, X., Zhao, Z., Wen, X., & Zhao, B. (2023). What If the TV Was Off? Examining Counterfactual Reasoning Abilities of Multi-modal Language Models. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. doi:10.1109/ICCVW60793.2023.00497

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., & Lin, X. V. (2022). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*. doi:10.1109/ICCVW60793.2023.00497