Journal of Organizational and End User Computing
Volume 36 • Issue 1 • January-December 2024

The Analysis of Smarter Future:
Innovations and Legal Risk in Visual Question Answering Technology in Home Automation

Xiaoyuan Gao
https://orcid.org/0009-0001-8483-4680
School of Law, Tianjin University, China

Juan Li
Continuous Education College, Taiyuan University of Technology, China

Kaiyang Sun
Department of Management, Monash University, Australia

ABSTRACT

With the continuous development of technology, smart home control systems, as one of the applications of artificial intelligence in daily life, are receiving increasing attention. However, smart home control systems face challenges in practical applications, such as complex scenarios and multimodal information. In this context, this paper introduces Visual Question Answering (VQA) technology to enhance the intelligence and user experience of smart home control systems. Despite the potential advantages of VQA technology in smart home control systems, current models still have certain shortcomings in handling local image information and integrating visual-language multimodal features. To address this, this paper proposes an innovative Transformer-based Multimodal Fusion Network (TMFNet) model. TMFNet aims to overcome the limitations of existing models in dealing with complex smart home scenarios by introducing a global-local feature attention mechanism, deep encoding-decoding modules, and multimodal representation modules.

KEYWORDS

Attention Mechanism, Deep Learning, Legal Risk, Legal Risk, Multimodal Fusion Network, Smart Home, Visual Question Answering

INTRODUCTION

In today's era of rapid technological advancement, smart home control systems, an application of AI in daily life, are receiving increasingly more attention (Si et al., 2021). Smart home control systems integrate various smart devices to achieve intelligent management of the home environment, providing users with a more convenient, comfortable, and secure living experience (Hameed et al., 2022; Ning et al., 2022). However, with the widespread adoption of smart home control systems, a series of challenges has emerged. Firstly, the concept of smart home control systems spans multiple domains, including environmental perception, device control, and user interaction, making the design and integration of systems complex and challenging. Additionally, standardization and interoperability issues among various smart devices have become bottlenecks restricting system development (Li et al., 2023; Zhang et al., 2022). Furthermore, smart home control systems face challenges related

DOI: 10.4018/JOEUC.357249

to data privacy and network security, raising concerns among users about the potential leakage of personal information (Chaudhary et al., 2022; Li & Xiao, 2023). To address these challenges, in recent years, deep learning technology has gradually become a research hotspot. Deep learning, with its significant achievements in pattern recognition, natural language processing, and other fields, offers new possibilities for optimizing smart home control systems (Jafarzadegan et al., 2022; Li et al., 2021). Through deep learning, systems can better understand and adapt to user habits, enhancing the level of intelligence in smart home control systems to better meet user needs (Darapaneni et al., 2022; Zhang et al., 2024).

Among the various applications of deep learning, visual question answering (VQA) has garnered significant attention. VQA combines image understanding and natural language processing, enabling computers to answer user questions about images and achieve a profound sense of visual information (Akula et al., 2021; Anderson et al., 2018). The introduction of this technology opens up new possibilities for smart home control systems, allowing users to interact with the system more naturally through voice or text, enhancing the system's user-friendliness and intelligence (Ye & Zhao, 2023; Ye et al., 2023). However, despite the tremendous potential demonstrated by VQA technology in smart home control systems, it still faces challenges (Antol et al., 2015). Firstly, there are limitations in handling complex scenes and multimodal information, making it difficult for the system to provide accurate answers in certain contexts. Secondly, the understanding of user questions and context by VQA systems still requires improvement to enhance the system's intelligence and adaptability (Yang et al., 2016; Luo et al., 2023).

In the optimization research of smart home control systems, researchers' explorations have become a crucial driving force for the advancement of this field. To illustrate this, we introduce the content of four relevant works.

Deep Fusion VQA Model Based on Deep Integration

In this research, a deep fusion VQA model is introduced with the aim of achieving a more comprehensive understanding of the smart home environment. By employing deep fusion, which utilizes convolutional neural networks (CNNs) to extract image features and recurrent neural networks (RNNs) to process semantic information in questions, the model establishes a closer connection between images and questions (Wang et al., 2024). However, despite significant progress in information fusion, this model still requires improvements when handling complex scenes and multimodal information to enhance system generalization and accuracy (Hou et al., 2020).

Multimodal Knowledge Graph VQA Model

Another study introduces a multimodal knowledge graph to enhance the VQA system's understanding of the smart home environment. By modeling knowledge about images, questions, and smart home devices as a graph, this model achieves more global information fusion. However, this approach heavily relies on the construction and updating of knowledge graphs, which may be limited by the timeliness and real-time requirements of the knowledge graph, necessitating further research and optimization for practical applications (Rossa et al., 2013).

Attention-Based VQA Model

Researchers employ an attention-based VQA model to selectively focus on crucial information in smart home scenes. By introducing attention mechanisms, the model dynamically adjusts its focus on images and questions to better answer user queries. However, attention mechanisms may face limitations when handling long-term sequences and diverse user instructions, requiring deeper research to overcome these challenges (Vaswani et al., 2023).

Generative Fusion VQA Model

Another study combines generative models with VQA to achieve a higher-level understanding of the smart home environment (Li, Pang et al., 2024). By generating natural language descriptions related to image content using generative models and integrating them with VQA models, this approach achieves progress in enhancing the system's understanding of complex scenes. However, generative models may encounter challenges related to accuracy in descriptions and diversity in generation, especially when dealing with multimodal information, requiring innovative solutions for improvement (Li et al., 2020; Lu et al., 2023).

Although the aforementioned works have demonstrated some advantages in the application of VQA in smart home control systems, they commonly face two significant issues: insufficient attention to local image information and inadequate fusion of visual-language multimodal features (Jin et al., 2021; Zhang et al., 2023). Firstly, the lack of attention to local image information is evident in the models' inability to adequately focus on the details and local features in complex smart home scenes. For instance, smart home environments may involve numerous devices and subtle variations, and current VQA models may not be sensitive enough when dealing with this local information. This may result in poor performance by the model when answering questions from users about specific devices or particular scenes. Secondly, the issue with the fusion of visual-language multimodal features mainly manifests in the models lacking a sufficiently deep integration strategy when combining image and question information (Zhong & Li, 2024). Although some models adopt deep fusion networks, there still exists a situation where the fusion of information is not closely knit when dealing with multimodal features. This may lead to a lack of sufficient consistency and global understanding in the model's interpretation of the correlation between images and questions, impacting the overall comprehension ability of the model in the smart home environment (Han et al., 2022; Huang et al., 2023).

Based on the aforementioned shortcomings, we propose an innovative transformer-based multimodal fusion network (TMFNet). The design of a TMFNet aims to overcome the issues of insufficient attention to local image information and inadequate fusion of visual-language multimodal features in existing VQA models. Firstly, we introduce a mechanism for attention to local features within global features to simultaneously capture both global and local information in the image. This enables our model to focus more intricately on crucial local regions in the image, thereby enhancing its understanding of complex smart home scenes. To guided attention (GA) from the question to image features, we construct a deep encoding-decoding module. This module effectively integrates question information with image features, enabling the model to answer questions more accurately and comprehensively. With the introduction of this module, we enhance the fusion capability of visual-language features, allowing the model to experience an enhanced understanding of the correlation between images and questions. Next, our multimodal representation module design improves the model's comprehension of critical words in the questions. The introduction of this module not only helps in reducing language noise and improves the accuracy of question expression but also facilitates better extraction of multimodal features, enhancing the model's overall comprehension ability (Sun & Ortiz, 2024). Finally, in the feature aggregation module, we efficiently concatenate multimodal features with question features. This step aims to further strengthen the model's understanding of questions, enabling it to interpret user queries more comprehensively and more accurately generate answers.

- We introduce a global-local feature attention mechanism that effectively addresses the insufficient attention to local image information in existing VQA models, thereby enhancing the model's understanding of complex scenarios in smart home environments.
- We introduce a deep encoding-decoding module that effectively integrates question information with image features, enhancing the fusion capability of visual and language features. This improvement makes the model more flexible and adaptive in smart home environments.

- We design a multimodal representation module that lessens language noise, enhances question expression accuracy, and better extracts multimodal features. This leads to a more precise understanding of user queries in smart home control systems.

The advancement of VQA technology undoubtedly amplifies the risk of exposing users' personal information and encroaching upon their privacy rights. Whether it involves augmenting local image information attention or enhancing the fusion capability of visual and language features as well as question expression accuracy, all these efforts aim to improve the precision of intelligent furniture control systems in recognition. However, increasing local image information attention may inadvertently expose sensitive details about users' whereabouts, potential activities, and other identifying data to potential security breaches. Similarly, bolstering the fusion capability of visual and language features, along with improving question expression accuracy, might inadvertently compromise user privacy by enabling more accurate identification of their linguistic patterns within smart home control systems. Therefore, as the actual user of VQA technology, the manufacturer of smart home control systems should prioritize safeguarding users' personal information and privacy during its utilization. Additionally, in order to mitigate potential legal risks and prosecution, it is advisable for the manufacturer to establish prior agreements with users.

This paper is structured as follows. The "Results" section presents the main performance of TMFNet on the VQA-1.0 and VQA-2.0 datasets and includes ablation studies. "Methods" details the main principles of TMFNet. Finally, the "Conclusion" section summarizes the paper, points out the shortcomings of the model, and suggests directions for future work.

RESULTS

In Table 1, we compare the performance of various VQA methods on the Test-dev and Test-std subsets. These methods include traditional approaches such as image-based object-wise importance model (IBOWIM), hierarchical co-attention (HieCoAtt), question guided latent attention block (QLAB), stacked attention network (SAN), Bilateral Attention Network (BAN), adaptive reasoning and answering contexts (ARAC), and dynamic fusion with attentive features (DFAF), as well as our proposed TMFNet. In terms of performance evaluation, we focus on four aspects of questions: yes/no (Y/N), numeric-type (Num), other, and overall (ALL). It is noteworthy that TMFNet demonstrates outstanding performance on the Y/N and ALL aspects of the Test-dev subset, achieving 46.58% and 64.01%, respectively, surpassing other methods significantly. On the Test-std subset, TMFNet also excels in the Y/N and ALL aspects, reaching 44.07% and 63.82%, respectively. Considering the overall performance across all aspects, TMFNet exhibits remarkable performance on the VQA 1.0 dataset, particularly showing significant advantages in handling different types of questions.

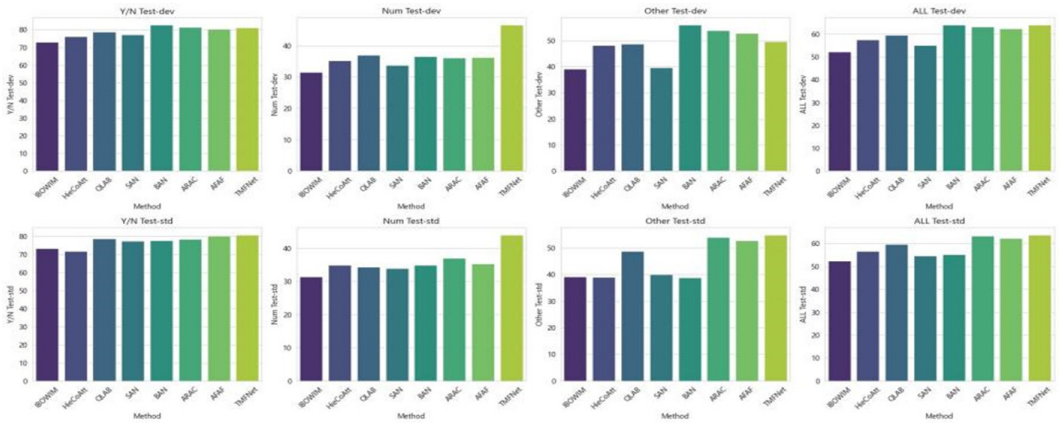
The visual representation of performance in Figure 1 clearly illustrates the outstanding performance of TMFNet in various aspects compared to other methods. This further confirms the promising prospects and application potential of TMFNet in smart home control systems, providing robust support for enhancing user experience and the intelligence level of systems.

Table 2 and Figure 2 represent a similar comprehensive summary of the performance of various VQA methods on the VQA 2.0 dataset. By evaluating the performance across Y/N, Num, Other, and ALL aspects in the Test-dev and Test-std subsets, we gain a thorough understanding of how each method performs across different question types. In the Test-dev subset, TMFNet exhibits superior performance, achieving accuracy rates of 77.59% for Y/N, 42.05% for Num, 45.07% for Other, and 60.48% for ALL, surpassing other methods significantly. This indicates that TMFNet excels in handling various question types, particularly demonstrating notable advantages in Num and ALL performance. On the Test-std subset, TMFNet continues to showcase outstanding performance with accuracy rates across the various categories of 77.24% for Y/N, 39.54% for Num, 51.34% for Other, and 59.29% for ALL. This reinforces TMFNet's excellence on the VQA 2.0 dataset, particularly

Table 1. Performance on VQA 1.0

Method	Test-dev (%)				Test-std (%)			
	Y/N	Num	Other	ALL	Y/N	Num	Other	All
IBOWIM (Feng & Liu, 2022)	73.02	31.50	39.10	52.19	73.23	31.45	39.09	52.36
HieCoAtt (Qiao et al., 2020)	76.17	35.17	48.17	57.47	71.82	34.99	38.99	56.70
QLAB (Feng & Liu, 2022)	78.73	36.98	48.64	59.60	78.71	34.42	48.72	59.59
SAN (Patro et al., 2022)	77.34	33.79	39.59	55.17	77.27	34.00	39.95	54.71
BAN (Kant et al., 2021)	82.71	36.50	55.93	64.00	77.70	34.99	38.84	55.26
ARAC (Li, Gong et al., 2024)	81.47	36.17	53.87	63.27	78.47	36.97	53.87	63.37
DFAF (Niu et al., 2021)	80.47	36.27	52.67	62.37	80.27	35.37	52.77	62.27
TMFNet	81.12	**46.58**	49.60	**64.01**	**80.77**	**44.07**	**54.87**	**63.82**

Figure 1. Performance visualization of various methods on VQA 1.0



showing significant improvements in the Num questions and ALL performance categories. These results, once again, emphasize the outstanding performance of TMFNet as a VQA method, providing robust support for its widespread application in real-world scenarios. In domains such as smart home control systems, the application of TMFNet holds the potential to enhance user experience and elevate the intelligence level of systems.

As shown in Table 3, we highlight the outstanding performance of TMFNet in zero-shot learning. In the grouped-query attention (GQA) and Visual7W tasks, TMFNet achieved scores of 64.6 and 77.5, respectively, demonstrating its remarkable capabilities in zero-shot learning without prior training data. In comparison, models such as IBOWIM, QLAB, DFAF, and HieCoAtt exhibited relatively lower performance in these two tasks, while TMFNet showcased significant advantages in these zero-shot scenarios. This further validates TMFNet's superior ability to generalize and adapt when handling unknown tasks, providing robust support for its reliability in practical applications. By comprehensively analyzing the data in the table, it is evident that TMFNet excels in zero-shot learning, offering substantial potential for applications in areas such as smart home control systems.

Finally, Figure 3 visually presents the contents of the table, providing a more intuitive representation of TMFNet's zero-shot capabilities across various tasks. This emphasizes the outstanding performance of our approach in addressing unknown tasks, offering compelling support and guidance for the application of TMFNet in intelligent systems.

Table 2. Performance on VQA 2.0

Method	Test-dev (%)			Test-std (%)				
	Y/N	Num	Other	ALL	Y/N	Num	Other	All
IBOWIM	69.49	28.97	35.57	48.66	69.70	28.92	35.56	48.83
HieCoAtt	72.64	31.64	44.64	53.94	68.29	31.46	34.46	53.17
QLAB	75.20	33.45	44.11	55.07	75.18	30.89	44.19	55.06
SAN	73.81	30.26	35.06	50.64	73.74	30.47	35.42	50.18
BAN	79.18	32.97	52.40	60.50	74.17	31.46	34.31	49.05
ARAC	77.94	32.64	50.34	59.74	74.94	32.62	50.34	59.84
DFAF	76.94	32.74	49.14	58.84	76.74	31.84	49.24	58.74
TMFNet	77.59	**42.05**	45.07	**60.48**	**77.24**	**39.54**	**51.34**	**59.29**

Table 3. Unveiling TMFNet zero-shot prowess: Impressive performance across GQA, Visual7W, rendered - SST2, and hateful memes datasets

Models	GQA	Visual7W	Rendered - SST2	Hateful Memes
IBOWIM	60.5	63.2	63.9	53.2
QLAB	51.8	53.6	6..5	57.0
DFAF	62.3	63.2	60.6	56.5
HieCoAtt	64.5	62.5	60.5	65.4
TMFNet	**64.6**	**77.5**	58.0	**66.4**

To validate the effectiveness of global-local attention (GLA), our approach was compared with self-attention (SA) and dynamic self-attention (DSA) methods. The evaluation was conducted on the VQA-1.0 dataset, and all accuracies were measured on the validation set. The specific experimental results are presented in Table 4. Compared to SA, the GLA method exhibits a noticeable improvement in accuracy, particularly in questions related to the Num and Other categories. This enhancement can be attributed to the GLA mechanism capturing crucial local features within the global image context. Num and Other category questions often require attention to local visual objects in the image. Figure 4 visualizes the contents of the table, further confirming the validity of our approach and its feasibility in the context of smart home environments through ablation experiments.

Figure 2. Performance visualization of various methods on VQA 2.0

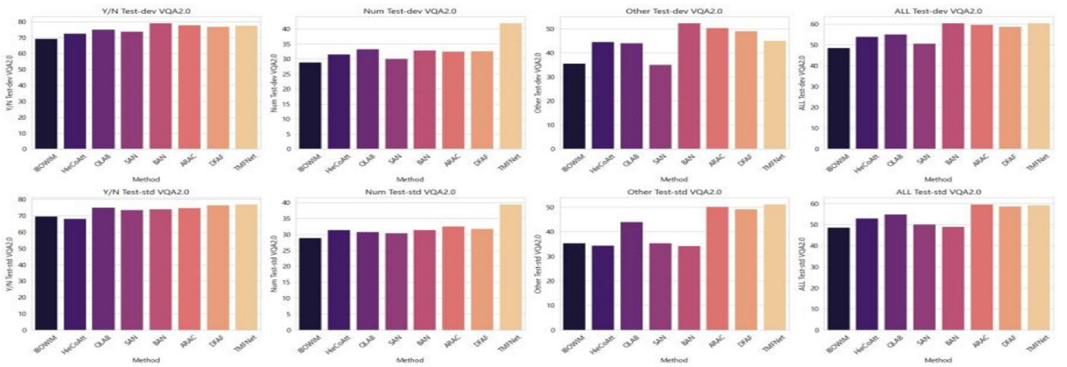


Table 4. The accuracy comparison of three attention modules in VQA 1.0 datasets

Method	ALL	Y/N	Num	Other
SA	61.31	78.42	41.09	52.85
DSA	62.34	79.13	41.31	53.08
GLA	**63.82**	**80.77**	**44.07**	**54.87**

To explore effective strategies for constructing a deep visual-language feature encoding-decoding module that can enhance the performance of VQA models, we devised variations of different CAL modules and conducted comprehensive experiments on these variants, as detailed in Table 5. SA(T)-self-constructing graph attention (SGA)(X,T) served as our validation benchmark, with its network structure illustrated in Figure 5.

The other three variants—GLA(T)-global-guided local attention (GGLA)(X,T), GLA(T)-SGA(X,T), and SA(Y)GGLA(X,T)—correspond to Figure 5(b), Figure 5(c), and our CAL model, respectively, while maintaining the settings of other modules unchanged. Here, X and T represent visual and textual features, respectively. GGLA(X,T) denotes the concurrent utilization of GA and GLA for processing image features. As observed in Table 5, introducing the proposed GLA into the baseline model resulted in an overall improvement in model accuracy. Furthermore, our experiments reveal that when GLA is employed to simultaneously process visual and textual features (GLA(T)-GGLA(X,T)), the model's proficiency in handling number-type questions exhibits significant

Figure 3 Unveiling TMFNet zero-shot prowess: Impressive performance across GQA, Visual7W, rendered - SST2, and hateful memes datasets

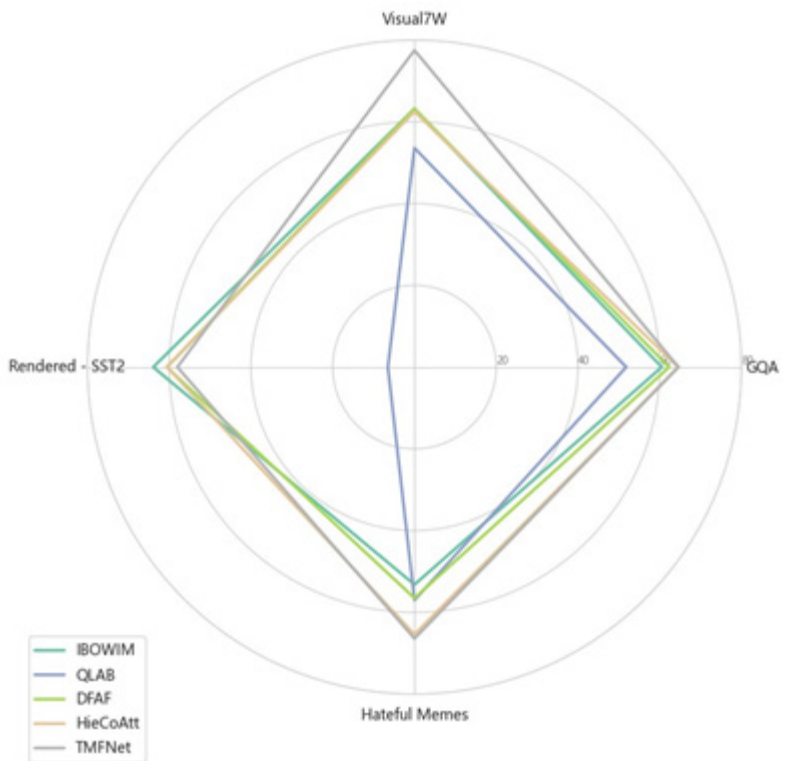


Figure 4. Accuracy comparison of three attention modules in VQA 1.0 datasets

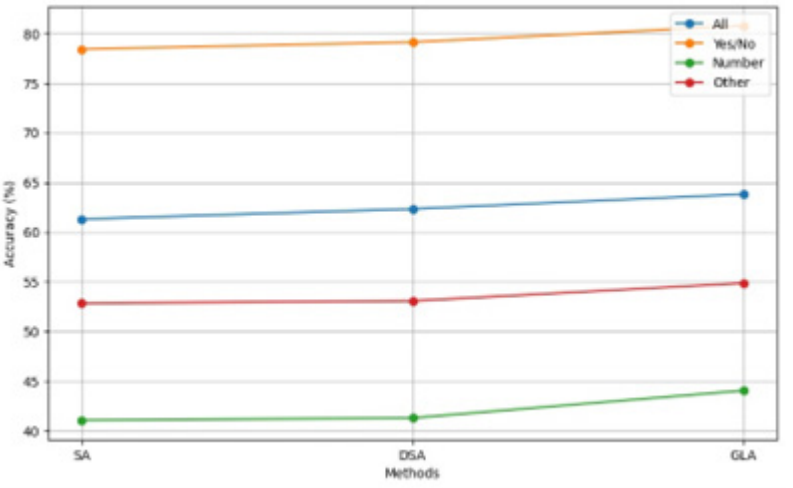


Table 5. Performance comparison of different variants of CAL on the VQA 1.0 dataset

Method	ALL	Y/N	Num	Other
GA(X,T)	62.20	80.10	42.30	52.60
GLA(T)-global-guided local attention (GGLA)(X,T)	62.67	80.24	43.02	54.12
GLA(T)-SGA(X,T)	62.65	80.17	42.15	52.27
SA(T)-GGLA(X,T)	**63.82**	**80.77**	**44.07**	**54.87**

enhancement. This implies that leveraging local key information can assist the model in addressing counting-related challenges to some extent. However, in comparison to the SA(T)-GGLA(X,T) structure, which utilizes only GLA for processing image features, the structure that deploys GLA for both features (GLA(T)-GGLA(X,Y)) introduces additional noise into the text (such as articles and prepositions), thereby impeding the model's performance in handling the other two types of questions. Consequently, we choose the SA(T)-GGLA(X,T) structure, demonstrating superior overall accuracy, to construct the deep visual-language feature encoding-decoding module.

Figure 5. Different variants of CAL

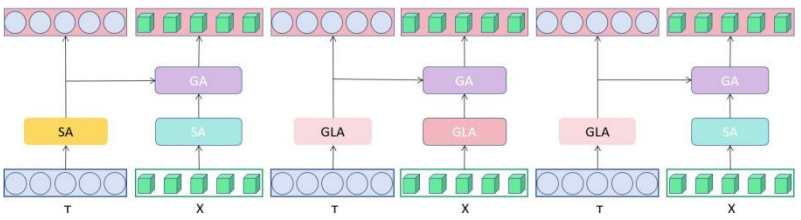
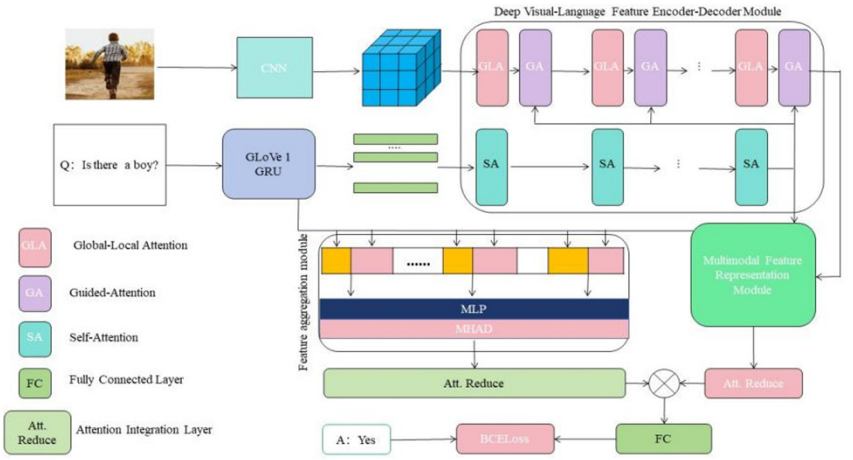


Figure 6. TMFNet structure diagram



METHODS

This study proposes the TMFNet for VQA, leveraging a transformer architecture. To enhance attention to local features, a global-local feature attention mechanism is introduced. Simultaneously, for improved alignment of visual-language features, a deep co-attention module is constructed using a combination of GLA and GA in an encoder-decoder framework, facilitating effective interaction between question and image features. Subsequently, a multimodal representation network is employed to guide the learning of critical question words using visual features as guidance for question features. This approach enables the network to discern the salient aspects of questions and mitigate language noise in VQA tasks. To reinforce question understanding, the feature aggregation module concatenates multimodal features obtained from the multimodal representation network with question features, employing multihead attention for representation learning. This design empowers TMFNet to comprehensively and accurately respond to user queries in complex smart home scenarios, thereby enhancing the system's intelligence and adaptability. The overall network architecture is illustrated in Figure 6.

Deep Visual-Language Feature Encoder-Decoder Module

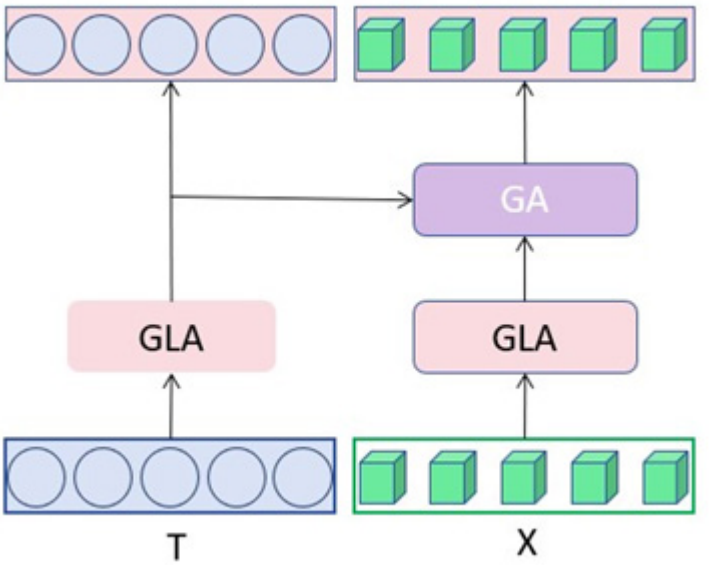
In this section, we mainly introduce the role of the cooperative-attention layer (CAL), also called the co-attention layer, in constructing a deep visual-language encoder-decoder. The network architecture of CAL is illustrated in Figure 7.

GLA

In response to the problem of exclusively modeling global features and thereby overlooking essential local features in SA mechanisms, we introduce a local key feature attention module embedded within global features (Waswani et al., 2017). This module first identifies relationships among global features, extracting overarching features. Subsequently, it employs a locally attentive mechanism with path-selective functionality to capture local key features. Finally, by generating pertinent weights from local key features and applying them to the global features extracted by TMFNet, the module achieves a targeted emphasis on crucial local features.

As shown in Figure 8, the global feature perceptron represents SA without residual connections and LayerNorm, while the local feature perceptron signifies a module with the capability of extracting crucial visual features. The input features are linearly mapped to obtain four vectors: Q, K, V1, and

Figure 7. CAL network structure diagram



V2. In the Global Feature Perception module, Q, K, and V1 are input into multihead attention to obtain the global feature V', as shown in Equation 1.

$$V' = MHAtt\left(Q, K, V_1\right), \tag{1}$$

On the other hand, V2 is input into local attention to obtain local key features L'. To emphasize the focus on local features, the local key features L' are passed through a sigmoid function to obtain a weight matrix. Subsequently, the global features V' are multiplied element-wise with this weight matrix through the Hadamard product, resulting in the computation of local key features F, as shown in Equation 2.

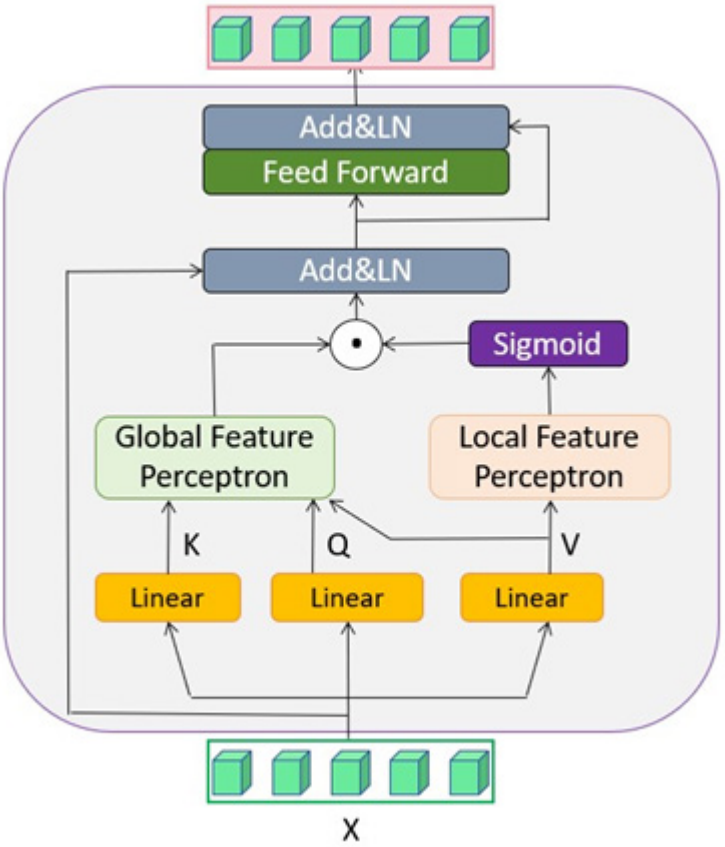
$$F = V' \odot \sigma(L'), \tag{2}$$

where V' represents the global features obtained from the global feature perceptron module, L' is the locally significant feature derived by inputting V2 into the local attention module, σ is the sigmoid function, generating a weight matrix with its input being the local key feature L′, ⊙ representing the Hadamard product, representing element-wise multiplication, and F is the resulting local key features, obtained by element-wise multiplying the global features V′ with the weight matrix generated by the sigmoid function. The overall formula is shown in Equation 3.

$$F' = LN(f + FF(f)),$$

$$f = LN(X + F). \tag{3}$$

Figure 8. Local attention structure diagram in global relations



where f represents the features obtained after normalizing the local relational features with the input features.

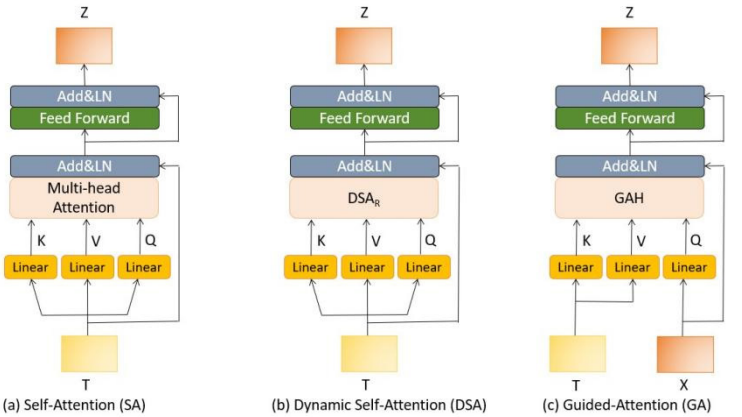
CAL

The CAL comprises three fundamental attention units: SA, DSA, and GA. Each attention unit collaboratively captures information at different levels and types. SA is employed to capture internal correlations within the input sequence, DSA emphasizes key information that dynamically changes within the input sequence, and GA concentrates on specific regions through a guiding mechanism. This collaborative design enables the CAL to comprehensively and accurately capture relevant features within the input sequence, providing a robust foundation for higher-level information extraction. Figure 9 illustrates the architectural diagram of the three attention mechanisms.

SA is a mechanism that computes the attention weights for each element in the input sequence X, considering the relationships between all elements. It is defined as shown in Equation 4.

$$SA\left(X\right) = softmax\left(\frac{KW_q\left(XW_k\right)^T}{\sqrt{d_k}}\right)XW_v \tag{4}$$

Figure 9. Three basic attention structures



where X represents the input sequence, Wq, Wk, Wv, which are learned weight matrices, and dk is the dimensionality of the attention heads. The softmax function generates attention weights, and the output is obtained by multiplying the input sequence X by these weights.

The DSA extends the SA mechanism by introducing a dynamic adjustment matrix M. It emphasizes key information that dynamically changes within the input sequence, as shown in Equation 5.

$$DSA\left(X\right) = softmax\left(\frac{KW_q\left(X\odot M\right)\left(XW_k\right)^T}{\sqrt{d_k}}\right)\left(X\odot M\right)W_v \tag{5}$$

where X is the input sequence, Wq, Wk, Wv are learned weight matrices, ⊙ denotes element-wise multiplication, and M is the dynamic adjustment matrix. This formulation introduces adaptability by dynamically adjusting the attention weights based on the dynamic features captured by M. GA is shown in Equation 6.

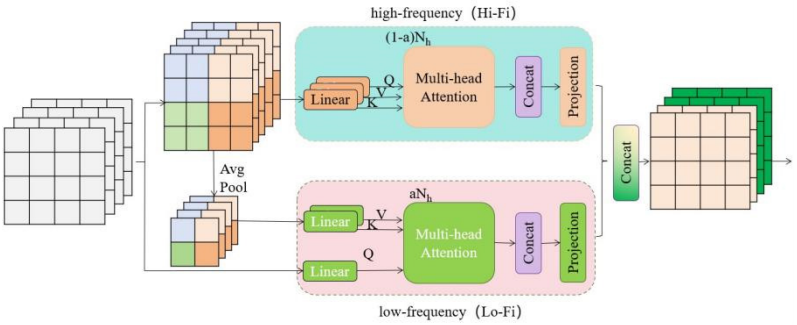
$$GA\left(X,G\right) = softmax\left(\frac{KW_q\left(GW_k\right)^T}{\sqrt{d_k}}\right)XW_v \tag{6}$$

where X is the input sequence, Wq,Wk,Wv are learned weight matrices, and G is the vector generated by the guiding mechanism. GA incorporates a guiding mechanism, represented by the vector G, to direct the attention mechanism towards specific regions of interest. The softmax function generates these adjusted attention weights, and the final output is computed by multiplying the input sequence X by these weights.

Multimodal Feature Representation Module

At this stage, the deep encoder-decoder further encodes the obtained visual-language features, primarily utilizing the mechanisms of GA and high-low (HiLo) frequency attention. The GA mechanism concentrates attention on specific regions of interest by introducing a guiding vector G, enhancing the model's focus on crucial information. This mechanism adjusts attention distribution more selectively by incorporating the guiding vector G into attention calculations, allowing the model to capture relevant information more precisely within the input sequence. The HiLo mechanism focuses on adjusting attention across different ranges of the spectrum to better handle various

Figure 10. High- and low-frequency attention structure map



frequency information present in the input. The introduction of this mechanism helps the model gain a more comprehensive understanding of input features and process information at different frequencies more finely. The collaborative operation of these two attention mechanisms enables the deep encoder-decoder to effectively encode visual-language features, providing richer and more accurate representations for subsequent tasks.

High-Low Frequency Attention

HiLo frequency attention refers to the attention given to high-frequency (Hi-Fi) and low-frequency (Lo-Fi) signals separately in the feature map, as illustrated in Figure 10. Essentially, the Lo-Fi attention branch design captures the overall dependency relationships of features. It does not require high-resolution feature maps but demands global attention. On the other hand, the Hi-Fi attention branch is dedicated to capturing fine-grained local dependency relationships. This requires high-resolution feature maps but can be achieved through local attention.

Multimodal Feature Representation

The multimodal feature representation module primarily employs GA (learning for image-guided problems) and HiLo frequency attention to further learn the captured visual-language features. The mechanism of GA is specifically designed to address the learning challenges posed by image-guided problems. By introducing a guiding vector, attention is concentrated on specific regions, facilitating a better understanding of the visual-language correlations. This mechanism enhances the model's ability to accurately capture key information in the images, providing more targeted learning for the representation of visual-language features. Simultaneously, the incorporation of HiLo frequency attention allows the module to process Hi-Fi and Lo-Fi signals differently based on the frequency information in the feature maps. The Lo-Fi attention branch focuses on overall dependency relationships, while the Hi-Fi attention branch specializes in local dependency relationships. This design enables the module to comprehensively and intricately learn and represent the captured visual-language features.

Feature Aggregation Module

This module plays a crucial role in computing the associations between text features and the bimodal embedding of visual-language information. The bimodal embedding inherently encapsulates the intricate relationships between textual and visual features. The primary objective of this module is to comprehensively capture the nuanced connections between the given problem and the bimodal representation. Subsequently, it aims to extract the most pertinent information from the bimodal features to facilitate further analysis (Han et al., 2022).

The architecture of the feature aggregation module is visually represented in Figure 4. The design of this module is meticulously crafted to enhance the model's ability to amalgamate and leverage both textual and visual information effectively. The feature aggregation process involves intricate computations to distill the key insights embedded in the bimodal embedding, enabling the model to make more informed and contextually relevant predictions.

This module serves as a critical component in the overall architecture, contributing to the model's capacity to holistically understand and interpret the relationships between textual and visual elements. Its effectiveness is pivotal in achieving accurate and meaningful results in tasks involving visual-language comprehension and analysis.

EXPERIMENT

Dataset

In this experiment, we utilized two datasets, namely VQA-1.0 and VQA-2.0.

VQA-1.0 (Guo et al., 2022): The VQA-1.0 dataset serves as a benchmark dataset for evaluating models' capabilities in understanding and answering questions related to images. Developed collaboratively by researchers from academia and industry, VQA-1.0 comprises a diverse set of images from the Microsoft Common Objects in Context dataset, each accompanied by human-generated questions and corresponding answers. Each image is associated with multiple questions, providing models with a comprehensive and challenging task. As a foundational dataset in the field of VQA, VQA-1.0 has facilitated the development and evaluation of algorithms for image-based question-answering tasks.

VQA-2.0 (Marino et al., 2019): Building upon the success of VQA-1.0, the VQA dataset VQA-2.0 introduces additional complexity to further challenge models. Like its predecessor, VQA-2.0 is constructed from the Microsoft Common Objects in Context dataset, encompassing a large number of images covering various scenarios and backgrounds. The dataset includes diverse questions, spanning different language structures and requiring varying levels of reasoning. Importantly, VQA-2.0 addresses certain limitations of VQA-1.0 by introducing a balanced dataset, ensuring a fair distribution of answer types for a more rigorous evaluation of model performance. This enhancement aims to encourage the development of models that demonstrate a nuanced understanding of both images and questions, making VQA-2.0 a valuable resource for driving the latest technological advancements in VQA research.

EXPERIMENTAL ENVIRONMENT

Implementation Details

Data Processing

In the data preprocessing phase of our experiment involving the VQA datasets, VQA-1.0 and VQA-2.0, we implemented several crucial steps to ensure the datasets were optimally prepared for the analysis. Firstly, data normalization procedures were applied. For image data, this entailed scaling pixel values to a range between 0 and 1, achieved by dividing each pixel value by 255, the maximum pixel value. Textual content, comprising questions and answers, underwent a tokenization process. Each piece of text was converted to lowercase, tokens were mapped to indices in a predefined vocabulary, and these indices were transformed into dense vectors via embedding layers. These vectors were then normalized to have a mean of zero and a standard deviation of one.

Following normalization, the datasets were partitioned to facilitate training, validation, and testing. As shown in Table 6, the VQA-1.0 dataset was divided into a training set of 50,213 images, a validation set of 19,876 images, and a test set consisting of 29,947 images. In terms of annotated images, the VQA-1.0 dataset included 249,732 images for training, 100,214 images for validation,

Table 6. Dataset partitioning on the VQA 1.0 and VQA 2.0 datasets

	VQA-1.0			VQA-2.0		
	Training Set	Validation Set	Test Set	Training Set	Validation Set	Test Set
Data per Image	50,213	19,876	29,947	82,783	40,504	81,434
Data per Annotated Image	249,732	100,214	150,089	443,757	214,354	447,793

Table 7. Model parameter settings

Parameter	Value	Description
batch size	64	Batch size
learning rate	2.5×10^{-5}	Initial learning rate
LR decay re	0.2	Learning rate decay rate
epoch	13	Number of epochs
hidden size	512	Hidden layer size
max length	16	Maximum sequence length
word embed size	300	Word embedding dimension
LAYER	6	Number of layers
head	8	Number of heads

and 150,089 images for testing. Similarly, the VQA-2.0 dataset was categorized into a training set with 82,783 images, a validation set of 40,504 images, and a test set comprising 81,434 images. This partitioning was designed to ensure a comprehensive evaluation of models across a diverse array of scenarios and questions, thereby providing a robust assessment of their visual and textual understanding capabilities.

Network Parameter Setting

In the experiment, we utilized the pre-trained ResNext152 model on Visual Genome for image feature extraction. By applying max-pooling to the obtained grid features, we obtained visual features with dimensions of 256×2,048. For computational convenience, these features were further mapped to a 256×512 format. Regarding text features, the question length was set to 16, and feature extraction was performed using the global vector model with a dimensionality of 300. The resulting language features were formed using a gated recurrent network, resulting in dimensions of 16×512.

In the design of the attention module, we employed eight multihead attentions, each with a dimensionality of 64, and implemented a deep encoder-decoder structure with a total of six layers. During the training process, we set the batch size to 64, and the learning rate was adjusted according to the formula $\min(2.5\times10^{-5},10^{-4}\cdot t)$, where t represents the current epoch, starting from epoch 1. After 10 epochs of training, the learning rate decayed by 1/5 every two epochs. We chose Adam as the optimizer, with $\beta1 = 0.9$ and $\beta2 = 0.98$. Overall, we completed training for 13 epochs, and specific parameter settings and their meanings are referenced in Table 7.

Baseline

IBOWIM: In the IBOWIM model, image-based information is leveraged to answer questions by evaluating the importance of individual objects within the image. Utilizing a CNN, image features are extracted, and a GLA mechanism is employed to focus on different objects within the image. The interaction between questions and image features is then processed using an RNN. The model

excels at enhancing question understanding by attending to various objects in the image. However, challenges may arise when dealing with complex scenes involving multiple objects, potentially limiting its effectiveness on large-scale datasets.

HieCoAtt: HieCoAtt employs a hierarchical co-attention mechanism, facilitating a comprehensive understanding by simultaneously focusing on multiple levels of both the image and the question. The model utilizes a CNN to extract image features and employs an attention-based RNN to interact with both question and image features. The layered attention mechanism proves advantageous in handling complex relationships between images and questions, although it comes with increased training complexity and potential computational resource requirements.

QLAB: QLAB introduces a question-guided latent attention block to guide attention toward image details relevant to the posed question. By extracting image features using a CNN and incorporating a question-guided latent attention block, the model achieves more targeted interactions between questions and images. While improving question understanding, the model may still face challenges in comprehending complex scenes, particularly when relying on the expressive capability of questions.

SAN: SAN utilizes a stacked attention network structure to incrementally enhance the understanding of both images and questions. Image features are extracted using a CNN, and stacked SA modules are employed to process interactions between questions and image features. The SAN mechanism aids in establishing deeper relationships between images and questions, although training deeper models may necessitate additional data and computational resources.

BAN: BAN introduces a bilateral attention mechanism, considering both local and global information in images and various parts of the question. Employing a CNN for image feature extraction and bilateral attention blocks for question-image interaction, the model achieves a more holistic understanding. While bilateral attention proves effective in global comprehension, handling large-scale datasets may pose challenges.

ARAC: ARAC incorporates adaptive reasoning and answering contexts to enhance question understanding. The model uses a CNN to extract image features and adapts reasoning and answering contexts to process interactions between questions and image features. The adaptive contexts contribute to adapting to distinctive features in different questions, yet the model's performance may depend on the availability of sufficient contextual information.

DFAF: DFAF employs dynamic fusion and attentive feature mechanisms to dynamically integrate question and image features for improved question understanding. Utilizing a CNN for image feature extraction, the model employs dynamic fusion and attentive feature modules to enhance question interactions. DFAF affords flexibility in handling diverse question types, but additional training data may be required for optimal performance, especially in complex contexts.

Evaluation Metric

The VQA dataset treats the visual question-answering task as a classification task and introduces a novel evaluation metric for VQA. This metric has gained wide acceptance and become a consensus in the field of VQA. Specifically, the evaluation metric for VQA is expressed in Equation 7.

$$Accuracy\left(a\right) = min\left\{\frac{n}{3}, 1\right\} \tag{7}$$

where n represents the number of model-predicted answers that match the manually annotated correct answers, a denotes an answer, and min{·} is the minimum value operation. Under this evaluation metric, an answer is considered correct as long as the number of occurrences of the model-generated answer is greater than or equal to 3 among the 10 human-annotated answers corresponding to each question.

CONCLUSION

In this paper, we propose the TMFNet, which brings substantial benefits to optimizing smart home control systems. Firstly, the global-local feature attention mechanism effectively addresses existing VQA models' shortcomings in paying insufficient attention to local image information, thereby enhancing the model's understanding of complex smart home scenarios. Secondly, the deep encoding-decoding module ensures effective integration of question information with image features, improving the fusion capability of visual and language features, and making the model more flexible and adaptive. Lastly, the multimodal representation module helps mitigate language noise, enhances question expression accuracy, and better extracts multimodal features, thus improving the model's overall comprehension of smart home control systems.

Research on optimizing smart home control systems presented in this paper holds significant importance. TMFNet provides an innovative solution to the deficiencies in attention and feature fusion present in existing systems. Moreover, the optimized model enhances system flexibility in smart home environments, better meets user needs, and improves user friendliness and intelligence. This advancement positively contributes to the progress of smart home technology and enhances the user experience.

Despite these advancements, the model has some limitations. Handling long-term sequences and diverse user instructions may pose challenges for the attention mechanism, requiring further research to overcome these issues. Furthermore, generative models may face challenges related to accuracy in descriptions and diversity in generation, especially with multimodal information, necessitating innovative approaches for improvement.

Future plans include further enhancing the TMFNet model to address its current limitations. Exploring advanced deep learning technologies to improve the performance of smart home control systems is also a priority. Additionally, future research will focus on data privacy and network security to ensure the continued safety and reliability of smart home systems in the evolving technological landscape. These efforts aim to make significant contributions to research and applications in the field of smart homes.

AUTHOR NOTE

Xiaoyuan Gao (https://orcid.org/0009-0001-8483-4680)

The authors of this publication declare there are no competing interests.

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Funding for this research was covered by the authors of this article.

PROCESS DATES

September 20, 2024
Received: July 4, 2024, Revision: August 13, 2024, Accepted: September 8, 2024

CORRESPONDING AUTHOR

Correspondence should be addressed to Xiaoyuan Gao (China, 18931753506@163.com)

REFERENCES

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*, 2148–2166. DOI: 10.18653/v1/2021.emnlp-main.164

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. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6077–6086. 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. *2015 IEEE International Conference on Computer Vision (ICCV)*, 2425–2433. DOI: 10.1109/ICCV.2015.279

Chaudhary, M., Vashistha, S., & Bansal, D. (2022). Automated detection of anti-national textual response to terroristic events on online media. *Cybernetics and Systems*, *53*(8), 702–715. DOI: 10.1080/01969722.2022.2044596

Darapaneni, N., Sunilkumar, C. M., Paroha, M., Paduri, A. R., Mathew, R. G., Maroli, N., & Sawant, R. E. (2022). Object detection of furniture and home goods using advanced computer vision. *2022 Interdisciplinary Research in Technology and Management (IRTM)*, 1–5. . Conference: 2022 Interdisciplinary Research in Technology and Management (IRTM).DOI: 10.1109/IRTM54583.2022.9791508

Feng, J., & Liu, R. (2022). LRB-Net: Improving VQA via division of labor strategy and multimodal classifiers. *Displays*, *75*, 102329. DOI: 10.1016/j.displa.2022.102329

Guo, J., Li, J., Li, D., Meng, A., Li, B., Tao, D., & Steven. (2022). From images to textual prompts: Zero-shot VQA with frozen large language models. *ArrXiv Preprint*. DOI: 10.48550/arxiv.2212.10846

Hameed, A., Violos, J., & Leivadeas, A. (2022). A deep learning approach for IoT traffic multi-classification in a smart-city scenario. *IEEE Access: Practical Innovations, Open Solutions*, *10*, 21193–21210. DOI: 10.1109/ACCESS.2022.3153331

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, 87-110*(1), 1. DOI: 10.1109/TPAMI.2022.3152247

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

Huang, X., Zhu, S., & Ren, Y. (2023). A Semantic Matching Method of E-Government Information Resources Knowledge Fusion Service Driven by User Decisions. *Journal of Organizational and End User Computing*, *35*(1), 1–17. DOI: 10.4018/JOEUC.317082

Jafarzadegan, M., Safi-Esfahani, F., & Beheshti, Z. (2022). An Agglomerative Hierarchical Clustering Framework for Improving the Ensemble Clustering Process. *Cybernetics and Systems*, *53*(8), 679–701. DOI: 10.1080/01969722.2022.2042917

Jin, W., Cheng, Y., Shen, Y., Chen, W., & Ren, X. (2021). A Good Prompt Is Worth Millions of Parameters: Low-resource Prompt-based Learning for Vision-Language Models. *ArXiv Preprint, arxiv.2110.08484*. https://doi.org//arxiv.2110.08484DOI: 10.48550

Kant, Y., Moudgil, A., Batra, D., Parikh, D., & Agrawal, H. (2021). Contrast and classify: Training robust VQA models. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1604–1161. DOI: 10.1109/ICCV48922.2021.00163

Li, G., Wang, X., & Zhu, W. (2020). Boosting visual question answering with context-aware knowledge aggregation. *MM '20: Proceedings of the 28th ACM International Conference on Multimedia*, 1227–1235. DOI: 10.1145/3394171.3413943

Li, M., & Xiao, W. (2023). Research on the effect of e-leadership on employee innovation behavior in the context of "self" and "relationship.". *Journal of Organizational and End User Computing*, *35*(1), 1–20. DOI: 10.4018/JOEUC.317090

Ri, Zhao, S., & Yang, B. (2023). Research on the application status of machine vision technology in furniture manufacturing process. *Applied Sciences (Basel, Switzerland)*, *13*(4), 2434. DOI: 10.3390/app13042434 PMID: 39354955

Li, S., Gong, C., Zhu, Y., Luo, C., Hong, Y., & Lv, X. (2024). Context-aware multi-level question embedding fusion for visual question answering. *Information Fusion*, *102*, 102000. DOI: 10.1016/j.inffus.2023.102000

Li, T., Pang, G., Bai, X., Zheng, J., Zhou, L., & Ning, X. (2024). Learning adversarial semantic embeddings for zero-shot recognition in open worlds. *Pattern Recognition*, *149*, 110258. DOI: 10.1016/j.patcog.2024.110258

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. *Conference and Labs of the Evaluation Forum (CLEF)*. https://api.semanticscholar.org/CorpusID:237299018

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

Luo, M., Du, B., Zhang, W., Song, T., Li, K., Zhu, H., Birkin, M., & Wen, H. (2023). Fleet rebalancing for expanding shared e-Mobility systems: A multi-agent deep reinforcement learning approach. *IEEE Transactions on Intelligent Transportation Systems*, *24*(4), 3868–3881. DOI: 10.1109/TITS.2022.3233422

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*, 3195–3204. DOI: 10.1109/CVPR.2019.00331

Ning, X., Tian, W., Yu, Z., Li, W., Bai, X., & Wang, Y. (2022). HCFNN: High-order coverage function neural network for image classification. *Pattern Recognition*, *131*, 108873. DOI: 10.1016/j.patcog.2022.108873

Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.-S., & Wen, J.-R. (2021). Counterfactual VQA: A cause-effect look at language bias. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12695–12705. DOI: 10.1109/CVPR46437.2021.01251

Patro, B. N., Anupriy, , & Namboodiri, V. P. (2022). Explanation vs. attention: A two-player game to obtain attention for VQA and visual dialog. *Pattern Recognition*, *132*, 108898. DOI: 10.1016/j.patcog.2022.108898

Qiao, Y., Yu, Z., & Liu, J. (2020). VC-VQA: Visual calibration mechanism for visual question answering. *IEEE International Conference on Image Processing (ICIP)*. DOI: 10.1109/ICIP40778.2020.9190828

Rossa, F. D., Dercole, F., & Piccardi, C. (2013). Profiling core-periphery network structure by random walkers. *Scientific Reports*, *3*(1), 1467. DOI: 10.1038/srep01467 PMID: 23507984

Si, Q., Lin, Z., & Zheng, M. yu, Fu, P., & Wang, W. (2021). Check it again: Progressive visual question answering via visual entailment. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Vol. 1, Issue Long Papers, pp. 4101–4110). Association for Computational Linguistics. DOI: 10.18653/v1/2021.acl-long.317

Sun, Y., & Ortiz, J. (2024). An AI-based system utilizing IoT-enabled ambient sensors and LLMs for complex activity tracking. *Academic Journal of Science and Technology*, *11*(3), 277–281. DOI: 10.54097/dj2pt496

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need. *Advances in Neural Information Processing Systems*. Advance online publication. DOI: 10.48550/arXiv.1706.03762

Wang, J., Li, F., An, Y., Zhang, X., & Sun, H. (2024). Towards robust LiDAR-Camera fusion in BEV space via mutual deformable attention and temporal aggregation. *IEEE Transactions on Circuits and Systems for Video Technology*, *1*(7), 5753–5764. Advance online publication. DOI: 10.1109/TCSVT.2024.3366664

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2015). Stacked attention networks for image question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 21–29. DOI: 10.48550/arxiv.1511.02274

Ye, Yao, K., & Xue, J. (2023). Leveraging empowering leadership to improve employees' improvisational behavior: The role of promotion focus and willingness to take risks. *Psychological Reports*, *332941231172707*. Advance online publication. DOI: 10.1177/00332941231172707 PMID: 37092876

Ye, S., & Zhao, T. (2023). Team knowledge management: How leaders' expertise recognition influences expertise utilization. *Management Decision*, *61*(1), 77–96. DOI: 10.1108/MD-09-2021-1166

Zhang, H., Wang, C., Tian, S., Lu, B., Zhang, L., Ning, X., & Bai, X. (2023). Deep learning-based 3D point cloud classification: A systematic survey and outlook. *Displays*, *79*, 102456. DOI: 10.1016/j.displa.2023.102456

Zhang, H., Wang, C., Yu, L., Tian, S., Ning, X., & Rodrigues, J. (2024). PointGT: A method for point-cloud classification and segmentation based on local geometric transformation. *IEEE Transactions on Multimedia*, *26*, 8052–8062. DOI: 10.1109/TMM.2024.3374580

Zhang, R., Tan, Y., Wang, Y., Wang, H., Zhang, M., Liu, J., & Xiong, J. (2022). Predicting the concentrations of VOCs in a controlled chamber and an occupied classroom via a deep learning approach. *Building and Environment*, *207*, 108525. DOI: 10.1016/j.buildenv.2021.108525

Zhong, Z., & Li, X. (2024). Re-Visiting the green puzzle: The effect of eco-positioning on inertial consumers. SSRN Electronic Journal. DOI: 10.2139/ssrn.4138686