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

# Innovating Sustainability:
## VQA-Based AI for Carbon Neutrality Challenges

Yanyu Chen, Zhejiang Normal University, China*
Qian Li, Zhejiang Normal University, China
JunYi Liu, Zhejiang Normal University, China

### ABSTRACT

In today's global society, carbon neutrality has become a focal point of concern. Greenhouse gas emissions and rising atmospheric temperatures are triggering various extreme weather events, sea level rise, and ecological imbalances. These changes not only affect the stability and sustainable development of human society but also pose a serious threat to the Earth's ecosystems and biodiversity. Faced with this global challenge, finding effective solutions has become urgent. This article aims to propose a comprehensive artificial intelligence design approach to address issues related to carbon neutrality. This method integrates technologies from fields such as computer vision, natural language processing, and deep learning to achieve a comprehensive understanding of environmental conservation and innovative solutions. Specifically, the authors first use a visual module to extract features from images, which helps capture important information in the images. Next, we employ the ALBEF model for cross-modal alignment, enabling better collaboration between images and textual information.

### KEYWORDS

ALBEF, Artificial Intelligence, Carbon Neutrality, CLIP, Environmental Governance Decision-making, Sustainable Development, VQA

### INTRODUCTION

In today's global society, climate change has become a focal point of concern (Waheed, et al., 2019). Greenhouse gas emissions and rising atmospheric temperatures are triggering various extreme weather events, rising of sea levels, and disruption of ecological balance. These changes not only affect the stability and sustainable development of human society but also pose a severe threat to the Earth's ecosystems and biodiversity. Faced with this global challenge, finding effective solutions has become urgent (R. Li, et al., 2021). Carbon neutrality is not only an environmental concept but also a commitment to action, encompassing various sectors and levels. From energy production to industrial manufacturing, transportation to agricultural production, and individual lifestyles to business operations, carbon neutrality is gradually shaping our way of life and economic models. By implementing a series of measures such as reducing greenhouse gas emissions, improving energy

DOI: 10.4018/JOEUC.337606 *Corresponding Author

efficiency, promoting renewable energy, and supporting carbon offset projects, we can gradually achieve a reduction in global carbon emissions and an increase in carbon absorption, ultimately attaining carbon balance. In recent years, deep learning, as a pivotal technology in the field of artificial intelligence, has experienced significant breakthroughs and found applications across various domains. In the environmental sector, the application of deep learning is gradually demonstrating its immense potential, providing new perspectives and methods for addressing environmental issues (X. Liu et al., 2023).

Visual question answering (VQA) systems combine computer vision and natural language processing, aiming to enable computers to understand images and answer questions related to these images. In the context of environmental conservation research, VQA systems can provide us with deeper insights and more accurate data analysis and decision support (Akula et al., 2021). Specifically, VQA systems can be used for environmental monitoring and assessment. By analyzing images and questions, VQA systems can identify environmental features in the images and key points of the questions, assisting researchers in quantitatively assessing environmental conditions. For example, VQA systems can be used to analyze satellite images to detect the green coverage in urban areas, thereby assessing changes in the urban ecological environment (Anderson et al., 2018). VQA systems also contribute to enhancing environmental awareness and education. By asking questions about environmental protection to VQA systems, individuals can gain a deeper understanding of the nature and impact of environmental issues. Furthermore, applying VQA systems to environmental education can help the public better grasp the importance of environmental protection, promoting an increase in environmental awareness. Additionally, VQA systems can also play a role in environmental decision-making and planning. In the context of environmental conservation research for the Olympic Games, VQA systems can be used to analyze the environmental impact during the venue construction process, providing information on environmental conservation measures to decision-makers (Liu et al., 2021). Furthermore, by training VQA systems and analyzing data, it is possible to predict the effectiveness of various environmental protection strategies, providing scientific support for decision-making (Antol et al., 2015).

Below are common research methods:

Vision-and-language bidirectional encoder representations from transformers (ViLBERT) is an advanced multimodal pre-training model designed to deepen the integration of computer vision and natural language processing fields (Lu et al., 2019a). ViLBERT seamlessly integrates image and text information within a unified framework, delivering robust performance and capabilities for multimodal tasks, including VQA. Its design ingeniously incorporates the successful principles of the BERT model, extending them into the realm of images. This equips ViLBERT with the unique ability to adapt to both natural language and visual data. Through pre-training, it captures the multimodal semantic information of text and images and maps them into a shared embedding space. Despite its excellence in visual and language tasks, ViLBERT's design structure is complex, incorporating multiple layers of attention mechanisms and a significant number of parameters. This complexity results in a demand for substantial computational resources and time during training and inference, limiting its practical application range (Ke et al., 2023).

Cross-modal pre-training for vision-language tasks (LXMERT) aims to achieve a deeper cross-modal fusion in visual and language tasks (Tan & Bansal, 2019). Similar to other models, LXMERT is built upon the idea of pre-training, fusing image and text information to provide robust performance for various multimodal tasks such as VQA and image captioning. However, like other models, LXMERT also exhibits some limitations. Despite being pre-trained on large-scale data, LXMERT's generalization performance in specific domains or tasks might be constrained. The model may require additional fine-tuning and optimization when confronted with new domains or novel challenges.

Detection transformers (DETR) is an innovative model that combines computer vision and natural language processing, fundamentally transforming object detection tasks (Wang et al., 2022). The strength of DETR lies in its capacity to process images comprehensively and produce

prediction results. By conceptualizing object detection as a set prediction problem, DETR sidesteps the intricacies associated with traditional anchor box settings and post-processing steps. Instead, it directly generates prediction results as a set of bounding boxes accompanied by their corresponding class labels, significantly streamlining the detection process. However, akin to other models, DETR also presents certain limitations. Despite achieving notable success in various scenarios, it may encounter challenges in handling densely cluttered scenes or objects with complex occlusions. This is attributed to its lack of explicit spatial relationship handling between objects.

The flamingo model is a multimodal fusion approach for visual and language tasks, aiming to effectively associate image and text information to achieve enhanced performance in multimodal scenarios (Alayrac et al., 2022). The core idea of **Align before Fuse (ALBEF)** is to align image and text information to enhance their semantic consistency, thereby improving the model's performance in cross-modal tasks. However, like other models, the flamingo model also has certain limitations. Its performance relies heavily on the quality and accuracy of alignment, which might become challenging in complex situations. Additionally, the training and adjustment of the flamingo model require consideration of multiple hyperparameters and data characteristics to ensure optimal performance (Han et al., 2023).

Bootstrapping language-image pre-training (BLIP) is an advanced pre-training model focused on cross-modal pre-training between images and language (J. Li et al., 2022). Its objective is to achieve outstanding performance in multimodal tasks by jointly pre-training on image and text data. BLIP's design emphasizes addressing challenges in multimodal pre-training, such as data imbalance and modal mismatch. However, like other models, the BLIP model also has limitations. Its performance may be constrained when dealing with specific domains or tasks.

Building upon this, we propose a novel model aimed at addressing the aforementioned issues. This approach employs an LLM as an "inference module" and introduces an independent "vision module." In our method, we initially leverage a pre-trained vision module to extract rich information from images. Concurrently, we employ the ALBEF method for image feature extraction, which has demonstrated remarkable efficacy in classification tasks. Next, we introduce the CLIP method to perform specific object detection tasks, thereby uncovering detailed features within images. Finally, all the information is fed to the LLM, fully harnessing its exceptional abilities in integrating language and visual information. The LLM, serving as the core inference module, plays a central role in the entire model, comprehensively analyzing and reasoning through the abundant information obtained from the vision and object detection modules. Its robust multimodal understanding capabilities enable the model to more accurately answer visual questions, generate image descriptions, and perform similar tasks. In this innovative model, our goal is to achieve a comprehensive understanding and fusion of image and text information through the collaborative efforts of multiple modules. The vision module is responsible for extracting image features, while ALBEF enhances the model's grasp of cross-modal correlations by aligning image and text information. The introduction of CLIP further enhances the model's object detection capabilities, enabling it to capture specific features and locations of objects within images. Leveraging its powerful inferential capabilities, the LLM provides suggestions for environmental protection. In summary, our model introduces several key modules to offer novel solutions for understanding and fusing image and text information. It holds broad potential for applications in multimodal tasks and is poised to drive the development of interdisciplinary visual and language tasks in the future.

The proposed model integrates technologies to achieve a comprehensive understanding of environmental conservation and innovative solutions in the following ways.

1) This article introduces a novel multimodal fusion model that uses language and visual models (LLM) as the inference module, while also incorporating independent visual modules ALBEF and CLIP. The core design of this model lies in the collaboration of these modules to achieve a comprehensive understanding and fusion of image and text information. By effectively

integrating the unique features of these modules, the proposed model can answer visual questions and generate image descriptions more accurately. As a result, it provides in-depth analysis and recommendations for environmental conservation.

2) This article applies deep learning techniques to the field of environmental conservation research. By integrating knowledge from fields such as computer vision and natural language processing, it offers interdisciplinary solutions to environmental challenges. Using multimodal technologies like VQA systems, this research extends the methods and tools for environmental studies, providing new perspectives for sustainable development.
3) The multimodal fusion model proposed in this article is not just a theoretical exploration; it also holds significant practical application potential. This model can be applied in the field of environmental conservation research to analyze environmental impacts such as venue construction and resource consumption. It provides informed support for decision-makers through scientifically supported information. Additionally, this model guides the development of future interdisciplinary visual and language tasks, inspiring research and innovation in various domains.

## RELATED WORK

### Integration of Large Language Models in Visual Question Answering Tasks and Environmental Protection

With the continuous advancement of artificial intelligence technology, the application of large language models (LLMs) such as GPT-3, GPT-3.5, and GPT-4.5 in visual question answering tasks has become a popular topic of research (Firat, 2023). These models, trained on extensive multimodal data, have demonstrated exceptional abilities in text processing and have achieved significant accomplishments in understanding and generating language content related to images (Brown et al., 2020). For instance, the GPT-3 model, after appropriate adjustments and training, can effectively interpret image content and answer related questions, to some extent simulating human capability in understanding visual information and linguistic expression. With iterative updates of models like GPT-3.5 and GPT-4.5, these language models have become even more refined and efficient in handling combined tasks of image and text, capable of identifying more complex details in images and generating richer and more accurate language descriptions (Whalen et al., 2023).

In the field of environmental protection, the application of these large language models is particularly noteworthy. They not only assist researchers in more accurately analyzing environmental monitoring images but also combine relevant textual information, such as news reports or scientific studies, to provide a more comprehensive assessment of environmental changes. For example, through the analysis and description of satellite images, GPT models can aid in identifying environmental issues such as deforestation and urban expansion, providing robust data support for policy makers. Additionally, in dealing with visual question answering tasks related to environmental protection, these models can provide more precise information, such as descriptions and explanations of ecological changes in specific areas, thereby helping the public better understand environmental issues and participate in practical environmental protection actions (Cheng et al., 2019).

### Application of New Visual Question Answering Models in the Field of Environmental Protection

With the continuous advancement of technology, the field of VQA is rapidly evolving with the emergence of new models, yet challenges persist in practical applications (Kazemi & Elqursh, 2017). For instance, the recently developed CoCa model, which integrates image and text information, greatly enhances the depth and accuracy of image content understanding but falls short in dealing with abstract and complex questions. Additionally, the BEiT model, despite its excellent performance in image recognition, has limitations in understanding complex language queries (Guo et al., 2023).

Similarly, the ViLT model, though advantageous in processing speed, lacks in capturing details and comprehending complex scenes, and this is especially evident when analyzing environmental monitoring images.

These new models, while progressive in some respects, still face challenges in environmental protection applications. For example, in tasks like analyzing deforestation and pollution source monitoring, precision and thoroughness are crucial, and existing models need improvement in these areas. Moreover, long-term tracking and analysis of environmental changes require models with higher understanding and predictive capabilities (Wang et al., 2022). Therefore, in the field of environmental protection, researching and developing VQA models with greater accuracy and stronger analytical abilities has become increasingly urgent. This can not only enhance the efficiency and accuracy of environmental monitoring but can also aid in making more rational decisions and plans based on environmental changes, thereby effectively advancing the in-depth implementation of environmental protection efforts.

### Latest Developments in Image-Text Matching Technology in Visual Question Answering

In the field of VQA, image-text matching technology is a core component, effectively merging images and text information through deep learning models to achieve precise cross-modal understanding. A variety of innovative image-text matching models have emerged. The CLIP model, trained on large-scale image-text pairs, demonstrates superior matching capabilities, especially in handling open-domain image-text relations, but there's room for improvement in fine-grained matching in specific areas (Scao et al., 2022). Following CLIP, the DALL-E model made breakthroughs in image generation but needs to enhance accuracy in understanding complex text descriptions. The FLIP model, with its improved feature extraction mechanism, strengthens the semantic correlation between images and texts, yet it faces challenges in pairing highly abstract concepts (Y. Liu et al., 2019). Similarly, the VLMo model has progressed in multimodal understanding, but its adaptability and robustness in specific scenarios are yet to be tested. Lastly, the ALBEF model, with its enhanced joint learning strategy, improves the fusion of cross-modal information. However, its computational efficiency becomes a limiting factor when dealing with large and complex datasets. The development of these models offers new perspectives and methods for image-text data analysis in environmental protection such as using precise image-text matching to monitor and analyze environmental changes, thus providing data support for formulating effective protection strategies. However, their limitations in practical applications also suggest the need for further research and optimization to meet the complex demands of environmental protection (Waswani et al., 2017).

## METHOD

Our method combines a "visual module" with an LLM to effectively integrate visual and textual information. Initially, the study utilizes a pre-trained visual module to extract comprehensive information from images. This module plays a crucial role in capturing image features and enhancing the model's cross-modal understanding. The introduction of the ALBEF model further enhances the model's ability to understand by aligning image and text data. Next, to achieve detailed object detection, the CLIP method is introduced. This method assists in identifying specific objects within images, thereby enhancing the model's capability to detect object-related features and locations. Finally, the core inference module, LLM, integrates all extracted information from the visual module, ALBEF, and CLIP. The multimodal understanding capabilities of this module enable it to provide more accurate answers to visual questions, generate image descriptions, and perform similar tasks. LLM also offers suggestions for environmental protection, showcasing its practical application potential.

The proposed model seamlessly integrates visual and textual information, offering a comprehensive approach to address environmental issues during significant events. The detailed

experimental process illustrates the model's efficacy in enhancing cross-modal understanding and facilitating environmental protection decisions. This study makes a meaningful contribution to the advancement of interdisciplinary visual and language understanding tasks and holds significant potential for applications in multimodal analysis and decision-making. The overall methodology is illustrated in Figure 1.

## CLIP Model

The CLIP model (Radford et al., 2021), a novel approach, brings significant insights to the realm of environmental protection. As depicted in Figure 2, the CLIP model achieves remarkable breakthroughs in the fusion of visual and language tasks. Its core concept involves mapping image and text embeddings into a shared multi-dimensional space, enabling cross-modal representation of both. This advanced approach provides an effective means to quantify and measure the connections between images and text in a unified space. In the context of environmental protection, the multimodal representation capability of the CLIP model offers profound insights for our research and decision-making. In the domain of environmental monitoring and assessment, the potential applications of the CLIP model are evident. By inputting images and environment-related questions into the CLIP model, we can accurately identify and analyze environmental features within the images. Simultaneously associating these features with the questions leads to comprehensive environmental assessment outcomes. Furthermore, the CLIP model can be employed for environmental education, linking images of environmental issues with relevant questions to convey the significance of environmental protection to the public and enhance environmental awareness. Below, we delve into the key principles of CLIP.

Image embedding: First, the image is embedded to obtain the image's feature vector. Let's assume the image embedding function is $f$, and the embedding of image x is represented as:

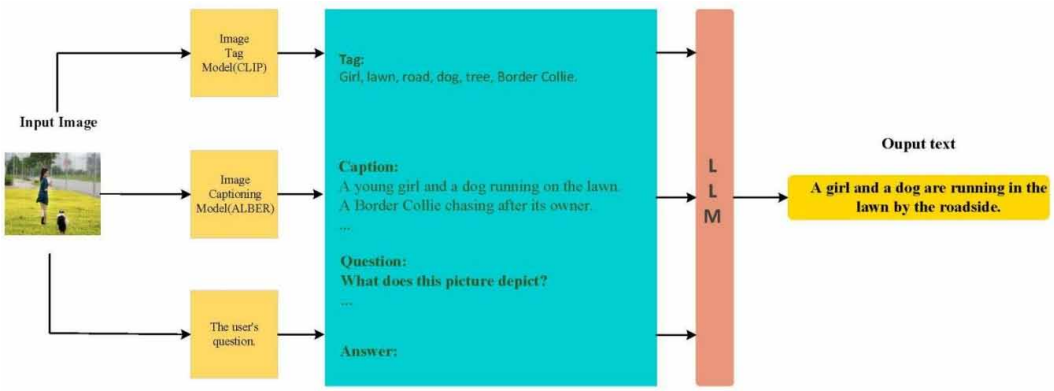
$$v_{\text{img}} = f_{\text{img}}(x) \tag{1}$$

Text embedding: Similarly, text is embedded to obtain the text's feature vector. Let's assume the text embedding function is $f$, and the embedding of text $y$ is represented as:

$$v_{\text{txt}} = f_{\text{txt}}(y) \tag{2}$$

**Figure 1. Flow of the model**

*Note. CLIP generates image tag information, while ALBEF extracts image captioning information.*



Contrastive learning: The core idea of the CLIP model is to map similar image-text pairs to nearby positions in the space through contrastive learning, while mapping dissimilar image-text pairs to distant positions. This can be achieved by defining a loss function, where the most commonly used loss function is the contrastive loss.

A common form of contrastive loss is the triplet loss, and its calculation formula is as follows:

$$L = \sum_i \left[ \sum_j \max\left(0, m + \mathrm{sim}(v^i_{\mathrm{img}}, v^i_{\mathrm{txt}}) - \mathrm{sim}(v^i_{\mathrm{img}}, v^j_{\mathrm{txt}})\right)\right] \tag{3}$$

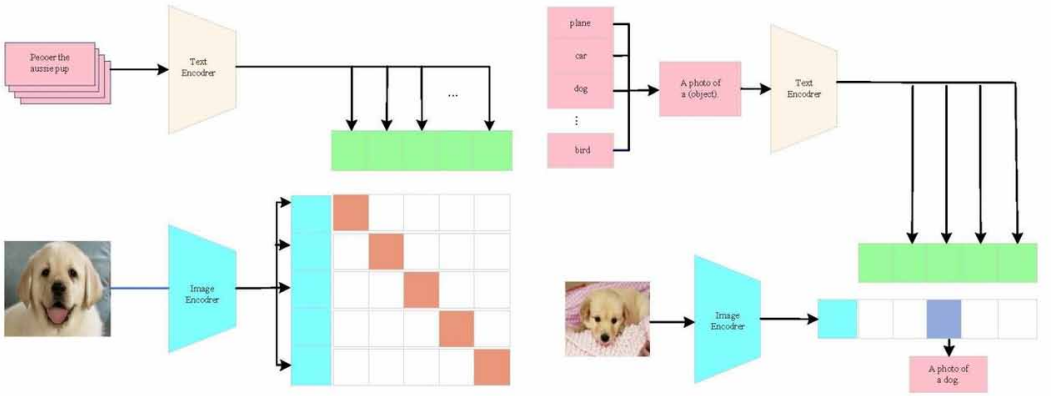
Here, $i$ represents the $i-th$ image-text pair, $j$ represents other image-text pairs, $m$ is a margin parameter, and $sim\left(v^i_{img}, v^i_{txt}\right)$ represents the cosine similarity between image and text embeddings. The goal of contrastive loss is to increase the cosine similarity for similar image-text pairs and decrease it for dissimilar image-text pairs.

## ALBEF Model

Alignment of text and image for environmental features (ALBEF) is an innovative model with profound implications for the field of environmental protection (J. Li et al., 2021). As illustrated in Figure 3, it focuses on effectively aligning textual descriptions with visual content to enhance our understanding of environmental features. By seamlessly integrating textual and visual information, it paves the way for comprehensive analysis, decision support, and environmental awareness. The primary objective of ALBEF is to bridge the gap between textual descriptions and visual representations within the context of environmental phenomena. By embedding textual and visual data into a shared semantic space, ALBEF enables us to identify and quantify the relationships between environmental attributes and their corresponding descriptions.

The architecture of this approach harnesses advanced machine-learning techniques to create a unified framework for environmental analysis through the alignment of textual and visual embeddings. Through a collaborative learning process, ALBEF reduces the semantic gap between text and images, ensuring that the shared space accurately reflects the intricate relationships within environmental content. In the realm of environmental monitoring and assessment, ALBEF's alignment capability empowers us to accurately identify and describe features within images and seamlessly connect them with relevant textual descriptions. This comprehensive understanding of environmental attributes is

**Figure 2. CLIP framework**



vital for evaluating ecological health, monitoring changes, and formulating sustainable practices. Below, we delve into the principles of ALBEF:

First, let's consider the process of text embedding and image embedding. For text data, we assume the existence of a text embedding function $f_{txt}$, which maps textual information y into a textual feature vector space, represented as:

$$v_{txt} = f_{txt}(y) \tag{4}$$

Similarly, for image data, we assume the existence of an image embedding function $f_{img}$, which maps an image x into an image feature vector space, represented as:

$$v_{img} = f_{img}(x) \tag{5}$$

Next, let's consider the alignment problem. The goal of ALBEF is to map text and image features into a shared semantic space, enabling their alignment within that space. To achieve this, we introduce an alignment loss function aimed at minimizing the distance between textual and image features in the shared semantic space. For each sample $(v_{txt}, v_{img})$, the alignment loss can be expressed as:

$$\mathcal{L}_{align}(v_{txt}, v_{img}) = \left\| v_{txt} - v_{img} \right\|^2 \tag{6}$$

Building upon alignment, ALBEF also introduces the concept of distillation learning. Distillation learning aims to transfer knowledge from a complex model (teacher model) to a simplified model (student model) to enhance the student model's performance. In ALBEF, we use the aligned text-image pairs as inputs for the teacher model and guide the training of the student model by minimizing the distance between the teacher and student models. The distillation loss can be expressed as:

$$\mathcal{L}_{distill}(v_{txt}, v_{img}, \theta_{teacher}, \theta_{student}) = \left\| v_{teacher} - v_{student} \right\|^2 \tag{7}$$

Here, $v_{teacher}$ represents the teacher model's feature in the shared semantic space, $v_{student}$ represents the student model's feature, and $\theta_{teacher}$ and $\theta_{student}$ are the parameters of the teacher and student models, respectively.

By combining the alignment loss and distillation loss, the overall loss function of ALBEF can be expressed as:

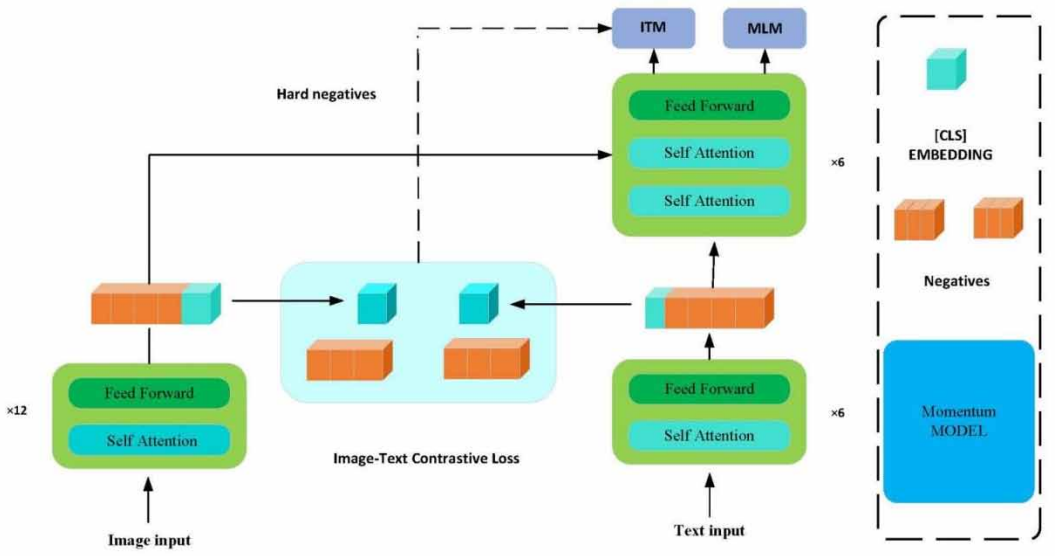
$$\mathcal{L}_{ALBEF} = \mathcal{L}_{align}(v_{txt}, v_{img}) + \lambda \cdot \mathcal{L}_{distill}(v_{txt}, v_{img}, \theta_{teacher}, \theta_{student}) \tag{8}$$

Here, $\lambda$ is a hyperparameter that balances the alignment loss and distillation loss. By minimizing the ALBEF loss function, we can effectively align text and image features in the shared semantic space, enabling a deep understanding of environmental attributes and providing profound insights into environmental protection issues.

## The Language and Vision Model

The language and vision model (LLM) plays a crucial role in our innovative model, being pivotal in addressing the issues mentioned earlier. The integration of LLM introduces an ability to elevate the

**Figure 3. ALBEF framework**




comprehension of language and vision to new heights (Whalen et al., 2023). As illustrated in Figure 4, the LLM incorporates the transformer architecture, forming the core component of our innovative model. The transformer architecture is renowned for its capability to capture long-range dependencies and complex patterns, serving as the backbone of LLM (Brown et al., 2020). At the heart of the transformer lies the self-attention mechanism, enabling the model to weigh the importance of different elements based on their relationships. This mechanism aligns well with the essence of our model – the fusion of textual and visual components into a unified representation. In our model, the LLM assumes the role of an ”inference module,” bridging the gap between language and vision. It processes and interprets rich information extracted from the visual and object detection modules, fully leveraging its intrinsic understanding of both domains. The multi-head self-attention mechanism of the transformer architecture aids in the comprehensive analysis of cross-modal data, enabling the LLM to identify subtle correlations and nuanced differences between textual descriptions and visual content. Through the adoption of the transformer-based LLM, our model gains the complex contextual understanding and reasoning abilities bestowed by this architecture. It empowers the model to holistically interpret textual and visual cues, deeply understanding the environment and its attributes. This capability equips our model with exceptional performance in answering visual questions, generating descriptive image narratives, and proposing environmental conservation measures (Han et al., 2022).

Below, we delve into the key principles of the transformer.

Self-attention mechanism: The self-attention mechanism calculates a weighted sum of values for each position in the input sequence. The importance of each value is determined by its similarity with the current position's query. The self-attention operation involves three linear transformations: query, key, and value. For a given input sequence $X$ with $N$ tokens, the self-attention operation at position $i$ can be represented as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9}$$

where $Q$ is the query matrix for the current position, with shape ($N$, $d_q$). $K$ is the key matrix for all positions, with shape ($N$, $d_k$). $V$ is the value matrix for all positions, with shape ($N$, $d_v$). $d_q$, $d_k$, and $d_v$ are the dimensions of the query, key, and value matrices, respectively.

Multi-head self-attention: In practice, the transformer employs multiple self-attention heads to capture different relationships in the data. The outputs of these heads are concatenated and linearly transformed. For the $h-th$ head of multi-head self-attention, the computation can be represented as:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W_O \tag{10}$$

where $\text{head}_h$ = Attention ($QW_{Q,h}$, $KW_{K,h}$, $V\,W_{V,h}$) $W_{Q,h}$, $W_{K,h}$, and $W_{V,h}$ are the weight matrices for the query, key, and value projections of the $h$-th head, respectively. $W_O$ is a weight matrix used for linear transformation of the concatenated output.

Position-Wise Feed-Forward Network: After multi-head self-attention, the representation for each position goes through a position-wise feed-forward network:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{11}$$

where $W_1$, $W_2$ are weight matrices, and $b_1$, $b_2$ are bias vectors.

## EXPERIMENT

### Datasets

In this study, our goal is to assess the performance of computer vision systems in various environments and scenarios using the VQA task. To construct our training dataset, we considered four prominent VQA datasets: Visual Genome, AI2D, TDIUC, and OK-VQA.

For Visual Genome, a versatile VQA dataset containing a plethora of natural images covering diverse environments and scenes such as natural landscapes, urban settings, and indoor spaces, we employed keyword and label-based filtering, including terms like ”natural landscapes”, ”urban scenery,” and ”indoor spaces.”

Furthermore, we selected questions related to the environment, encompassing queries about objects, scenes, locations, and relationships, ensuring the creation of a training set with a focus on environmental context.

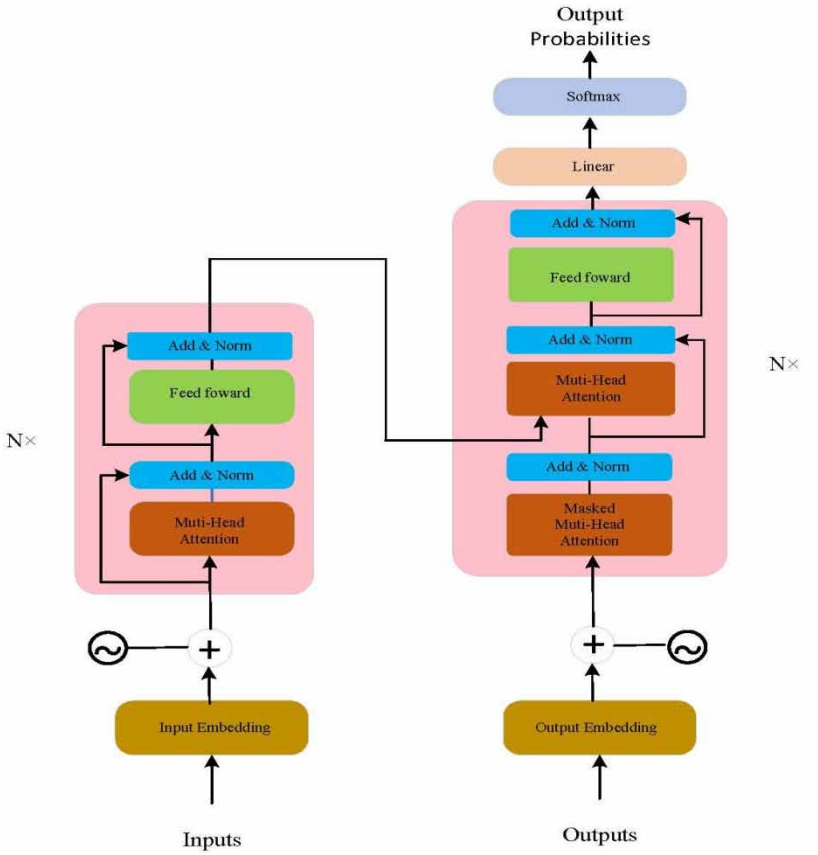
The AI2D dataset, another valuable resource, features a large collection of natural images and environment- related questions. We adopted a similar filtering approach, emphasizing the selection of images and questions that pertain to environments and scenes.

The TDIUC dataset is designed to test the understanding of computer vision systems regarding visual concepts that are more intuitive for ordinary individuals. While choosing images, we prioritized those related to environments and everyday life scenarios and matched them with relevant questions for model training.

Finally, OK-VQA, a dataset focusing on everyday life scenarios, provided us with images encompassing settings such as homes, offices, and outdoor environments, along with corresponding questions to evaluate the performance of computer vision systems.

These dataset selection and processing methods will aid in constructing a training dataset with a strong emphasis on environmental relevance, allowing for a more in-depth exploration of the application and performance of computer vision in different environmental contexts.

**Figure 4. Transformer framework**




## Experimental Environment

In our experiments, we utilized a personal computer (PC) as our computing platform. For hardware specifications, our CPU is the Intel Core i9-9900k running at a clock speed of 3.60GHz. In terms of GPU, we employed two NVIDIA RTX3090 graphics cards with a total of XXX CUDA cores. The system is equipped with 32GB of RAM and 11GB GDDR6 video memory. In the software environment, we operated on the Windows 10 operating system, used Python version 3.8, and relied on libraries such as matplotlib 3.3.4 and opencv4.5.5. The CUDA version employed was 10.0. This comprehensive hardware and software setup ensured the stability and efficiency of our experiments. The specific experimental environment is detailed in Table 1.

**Table 1. Presentation of experimental environment**

| Computing Platform | PC |
|---|---|
| CPU | Intel Core i9-9900k CPU @ 3.60GHz |
| GPU | NVIDIA RTX3090 Graphics Card *2 CUDA Cores |
| Memory | 32GB |
| Video Memory | 11GB GDDR6 |
| Software Environment | Windows 10 Operating System, Python 3.8, matplotlib 3.3.4, opency 4.5.5, CUDA 10.0 |

## Experimental Details

### *Step 1: Data Preprocessing*

In our research, we focus on the application of visual question answering in the field of environmental conservation. Therefore, we start by using datasets such as Visual Genome, AI2D dataset, TDIUC dataset, and OK-VQA to align the model with our theme.

### *Step 2: Model Design and Implementation*

Architecture definition: Our model architecture is carefully designed to seamlessly integrate various modules. It is built around a language-image multimodal (LLM) framework, effectively combining language and visual cues. Additionally, it includes a dedicated "visual module" responsible for extracting complex image features and a "target detection module" for precise localization of environmental elements, aligning with the environmental context.

Module integration: The design phase involves the seamless integration of different modules, each aligned with the environmental conservation theme. The visual module uses pre-trained techniques to extract meaningful features from images, focusing on environmental elements. The ALBEF model, known for its excellent performance in cross-modal tasks, aligns image and text data to enhance the model's understanding of the environment. The introduction of ResNet-152 further optimizes target detection for recognizing specific environmental attributes.

LLM's inferential abilities: At the core of the model, LLM serves as a powerful inference engine. It not only comprehensively analyzes and infers combined information, but also enables the model to provide valuable insights into environmental conservation. The fusion of this multimodal understanding and inferential capability enhances the model's performance in various environmental tasks.

Comprehensive fusion: The implementation process emphasizes the harmonious fusion of image and text data. By leveraging the capabilities of LLM, our model generates accurate answers to environmental queries, creates environmental image descriptions, and provides recommendations for environmental conservation. This comprehensive fusion is at the heart of the model's effectiveness in the field of environmental conservation.

Through the careful design and integration of these components, our model is ready to revolutionize the application of visual question answering in the realm of environmental conservation, achieving a comprehensive understanding and inference of the environment.

### *Step 3: Exclusion Experiment Design and Implementation*

We define the research objectives for exclusion experiments, where we will comprehensively analyze each component of the model, with a particular focus on its performance in environmental conservation applications. For example, we consider replacing specific components of the model, such as ALBEF or ResNet-152, to evaluate their impact in environmental contexts.

We evaluate each exclusion model using the same test dataset, with a focus on environmental scenarios, to compare their performance differences and highlight the importance of each component in an environmental context.

We analyze the results of the exclusion models to assess the impact of each factor on the model's performance in environmental conservation applications. This analysis helps provide a deeper understanding of the model's effectiveness in addressing environmental challenges.

## EXPERIMENTAL RESULTS AND ANALYSIS

In this study, we conducted extensive experimental evaluations of our proposed method on the OK-VQA dataset and compared its performance with several baseline models in a systematic manner. The dataset comprises yes/no-type questions (Y/N), numeric answer questions (Num), and other types of questions (Other), and assessments were carried out on the Test-dev and Test-std subsets. Based on the

experimental results (as shown in Table 2), our method demonstrated significant advantages across various aspects. On the Test-dev subset, despite a slight lower performance in Y/N-type questions compared to JE-MHA, our method outperformed other baseline models in Num-type questions, Other-type questions, and overall performance, achieving accuracies of 51.61%, 54.63%, and 70.04%, respectively. On the Test-std subset, our method consistently surpassed other models in all question types and overall performance, with exceptional performance in Num-type questions, achieving an accuracy of 49.10%. These experimental findings underscore the outstanding performance of our method in multimodal tasks, particularly in numeric answer questions, providing a powerful tool for addressing environmental issues and promoting sustainable development. These results emphasize the importance of our research in the context of interdisciplinary visual and language tasks and its potential applications in the fields of environmental protection and climate change.

As shown in Table 3 and Figure 5, we conducted a comprehensive experimental evaluation of our proposed method on the Visual Genome dataset and compared it in detail with various baseline models. The Visual Genome dataset encompasses assessments across multiple attributes (Object, Number, Color, Location, WUPS0.9, and WUPS0.1), alongside overall performance (All). Specifically, the performance of our method on the Visual Genome dataset is as follows: in overall performance, our method achieved a remarkable overall accuracy of 71.99%, indicating its outstanding performance in multimodal tasks. In the Object attribute, our method attained an accuracy of 73.02%, demonstrating its robust capability in recognizing objects. Notably, our method excelled in the Number attribute, achieving an impressive accuracy of 61.78%, highlighting its excellence in handling numeric-related questions. In the Color attribute, our method obtained an accuracy of 75.98%, showcasing its proficiency in addressing color-related questions.

We also conducted a comparison of our method's performance with various baseline models on the AI2D dataset, as illustrated in Table 4 and Figure 6. This dataset encompasses multiple evaluation metrics (Accuracy, Binary, Open, Validity, Plausibility, Consistency, and Distribution). Specifically, our method excels on the AI2D dataset, achieving an impressive overall accuracy of 67.94%, surpassing other baseline models and highlighting its exceptional performance in multimodal tasks. Particularly noteworthy is our method's outstanding performance in the Binary evaluation, where it achieves a remarkable accuracy of 86.50%, and in the Open evaluation with an accuracy of 51.15%, showcasing its excellence in addressing binary and open-ended questions. Additionally, our method demonstrates strong capabilities in assessing question validity, plausibility, consistency, and question distribution, achieving scores of 107.58, 95.93, 98.02, and 11.58, respectively. These results collectively emphasize the potential of our research in the field of multimodal understanding, providing robust support for addressing various complex issues and advancing related domains.

**Table 2. Our method's performance compared to different baseline models on the OK-VQA dataset**

| Method | Test-dev (%) | | | | Test-std (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | Y/N | Num | Other | ALL | Y/N | Num | Other | All |
| LRB-Net (Santoro et al., 2017) | 78.05 | 36.53 | 44.13 | 57.22 | 78.26 | 36.48 | 44.12 | 57.39 |
| DSACA (Lu, Yang, Batra, & Parikh,2016) | 81.20 | 40.20 | 53.20 | 62.50 | - | - | - | - |
| JE-MHA (Gao, Cao, Xu, Shao, & Song,2020) | 83.76 | 41.01 | 53.67 | 64.63 | 83.74 | 39.45 | 53.75 | 64.62 |
| CPDR (Yang, He, Gao, Deng, & Smola,2016) | 82.37 | 38.82 | 44.62 | 60.20 | 82.30 | 39.03 | 44.98 | 59.74 |
| MCAN (Y. Liu et al., 2019) | 87.74 | 41.53 | 60.96 | 70.03 | - | - | - | - |
| SCAVQA (Z. Yu, Yu, Xiang, Fan, &Tao, 2018) | 86.50 | 41.20 | 58.90 | 68.30 | 83.50 | 41.00 | 58.90 | 68.40 |
| CAAN (Z. Yu, Yu, Fan, & Tao, 2017) | 85.50 | 41.30 | 57.70 | 67.40 | 85.30 | 40.40 | 57.80 | 67.30 |
| Ours | 86.15 | 51.61 | 54.63 | 70.04 | 85.80 | 49.10 | 59.90 | 68.85 |

**Table 3. Comparison of our method with different baseline models on the visual genome dataset**

| Method | All | Object | Number | Color | Location | WUPS0.9 | WUPS0.1 |
|---|---|---|---|---|---|---|---|
| Local-RN (Sun, Yao, Zhang, & Yu, 2020) | 69.20 | 70.16 | 55.61 | 74.18 | 59.95 | 78.27 | 93.95 |
| MCAN (Z. Yu, Yu, Cui, Tao, & Tian, 2019) | 69.68 | 70.99 | 55.79 | 73.12 | 61.77 | 78.94 | 93.58 |
| ODA (Wu, Liu, Wang, & Dong, 2018b) | 70.93 | 72.08 | 56.30 | 75.77 | 62.50 | 79.89 | 94.62 |
| COR (Akula et al., 2021) | 70.98 | 72.02 | 57.43 | 75.73 | 62.17 | 79.70 | 94.46 |
| DSAC (Wu, Liu, Wang, & Dong, 2018a) | 71.07 | 72.30 | 54.24 | 76.24 | 62.70 | 80.51 | 95.02 |
| ALSA (Y. Liu et al., 2020) | 71.57 | 73.19 | 56.43 | 74.34 | 63.38 | 81.03 | 94.35 |
| MCAN+PA (Mao, Yang, Lin, Xuan, & Liu, 2022) | 71.70 | 73.73 | 57.57 | 76.45 | 63.67 | 81.35 | 94.58 |
| CPDR (Antol et al., 2015) | 71.79 | 73.74 | 58.33 | 76.77 | 63.04 | 80.56 | 94.84 |
| Ours | 71.99 | 73.02 | 61.78 | 75.98 | 61.60 | 80.73 | 94.80 |

**Figure 5. Comparison of our method with different baseline models on the visual genome dataset**

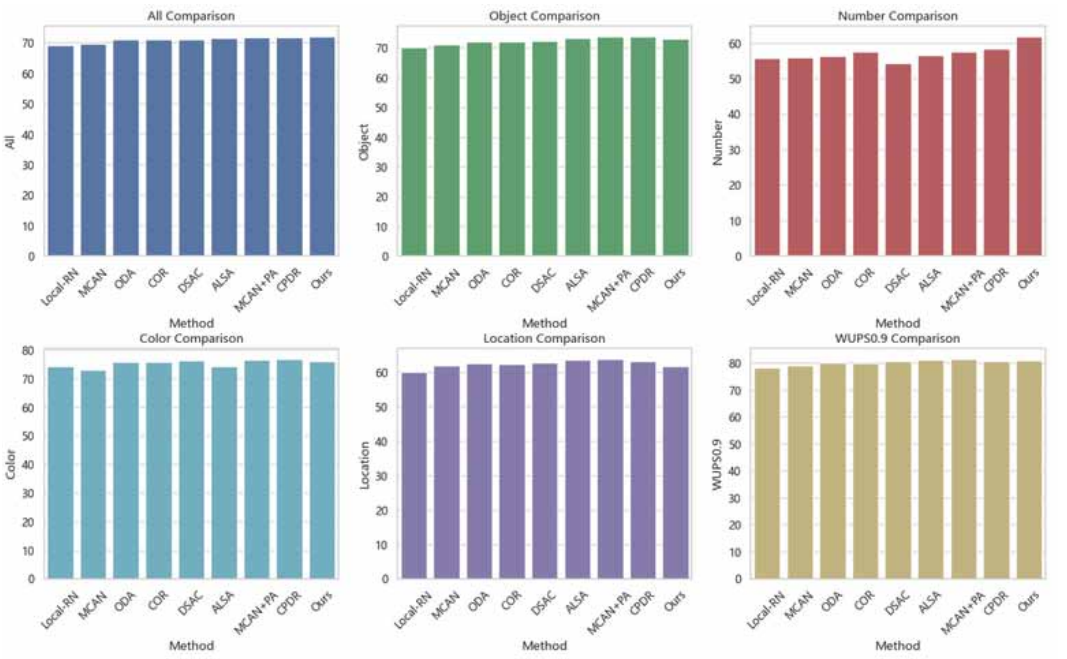
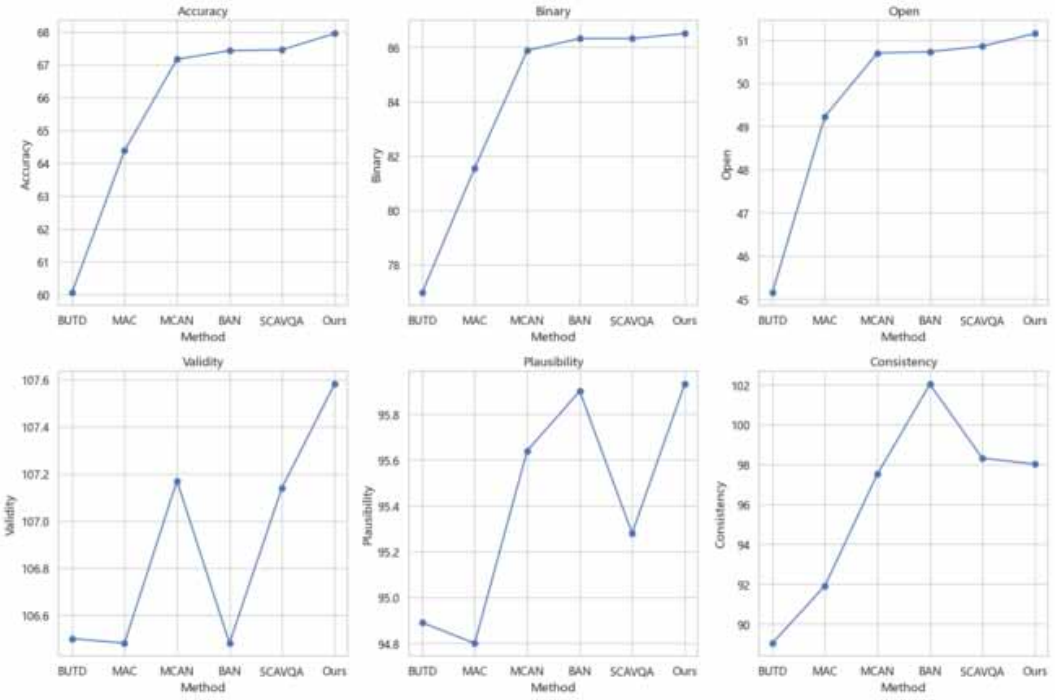



**Table 4. Comparison of our method with different baseline models on the AI2D dataset**

| Method | Accuracy | Binary | Open | Validity | Plausibility | Consistency | Distribution |
|---|---|---|---|---|---|---|---|
| BUTD (Si, Lin, Zheng, Fu, & Wang, 2021) | 60.06 | 76.96 | 45.15 | 106.50 | 94.89 | 89.03 | 16.30 |
| MAC (Kim, Jun, & Zhang, 2018) | 64.38 | 81.55 | 49.23 | 106.48 | 94.80 | 91.91 | 15.66 |
| MCAN (Lu, Batra, Parikh, & Lee, 2019b) | 67.16 | 85.88 | 50.70 | 107.17 | 95.64 | 97.51 | 11.63 |
| BAN (Akula et al., 2021) | 67.42 | 86.32 | 50.73 | 106.48 | 95.90 | 102.02 | 20.84 |
| SCAVQA (Marino, Rastegari, Farhadi, & Mottaghi, 2019) | 67.45 | 86.32 | 50.86 | 107.14 | 95.28 | 98.31 | 11.41 |
| Ours | 67.94 | 86.50 | 51.15 | 107.58 | 95.93 | 98.02 | 11.58 |

**Figure 6. Comparison of our method with different baseline models on the AI2D dataset**

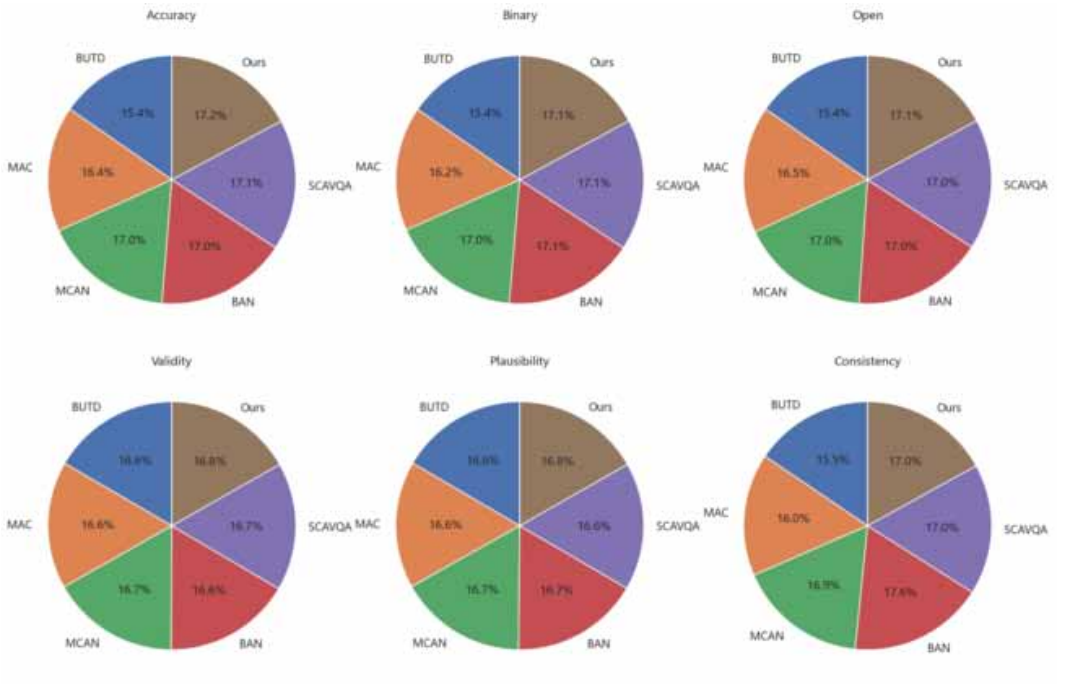



Finally, we conducted validation on the TDIUC dataset. As shown in Table 5 and Figure 7, our method excelled on the TDIUC dataset. The overall accuracy reached 75.74%, surpassing other baseline models and highlighting its exceptional performance in multimodal tasks. Particularly noteworthy is our method's remarkable accuracy of 94.30% in the Binary evaluation and an accuracy of 58.95% in the Open evaluation, demonstrating its outstanding ability to address binary and open-ended questions. Furthermore, our method exhibited strong capabilities in assessing question validity, plausibility, consistency, and question distribution, obtaining scores of 115.38, 103.73, 105.82, and 19.38, respectively. These validation results further reinforce the position of our research in the field of multimodal understanding and provide strong support for addressing various complex environmental issues and advancing related domains. The extensive application potential of our method in environmental monitoring and protection will contribute to a better understanding and resolution of global environmental challenges.

**Table 5. Comparison of our method with different baseline models on the TDIUC dataset**

| Method | Accuracy | Binary | Open | Validity | Plausibility | Consistency | Distribution |
|---|---|---|---|---|---|---|---|
| BUTD (Si et al., 2021) | 67.86 | 84.76 | 52.95 | 114.30 | 102.69 | 96.83 | 24.10 |
| MAC (Kim et al., 2018) | 72.18 | 89.35 | 57.03 | 114.28 | 102.60 | 99.71 | 23.46 |
| MCAN (Lu et al., 2019b) | 74.96 | 93.68 | 58.50 | 114.97 | 103.44 | 105.31 | 19.43 |
| BAN (Akula et al., 2021) | 75.22 | 94.12 | 58.53 | 114.28 | 103.70 | 109.82 | 28.64 |
| SCAVQA (Marino et al., 2019) | 75.25 | 94.12 | 58.66 | 114.94 | 103.08 | 106.11 | 19.21 |
| Ours | 75.74 | 94.30 | 58.95 | 115.38 | 103.73 | 105.82 | 19.38 |

**Figure 7. Comparison of our method with different baseline models on the TDIUC dataset**




## Ablation Study

Results of the ablation experiments in the ALBEF module are presented in Table 6. These experiments were conducted on the Visual Genome dataset to evaluate the contribution of the ALBEF module to multimodal tasks. These ablation experiments involved four different methods (ResNet-152, BLIP, CLIP, and ALBEF). Each method was tested on four different evaluation metrics (Yes/No, Number, Other, and All). These evaluation metrics represent the performance of the models in answering different types of questions. From the table, it can be observed that the ALBEF module outperforms the other methods in all evaluation metrics. Particularly, in the Yes/No and Number evaluations, the ALBEF module demonstrates exceptional accuracy, with 93.95% and 73.02%, respectively. This indicates that the ALBEF module excels in answering binary questions and questions related to numbers. Furthermore, in the Other evaluation, the ALBEF module also achieves the highest accuracy of 61.60%, showcasing its outstanding performance in handling other types of questions. Most importantly, in the overall evaluation metric, All, the ALBEF module achieves the highest accuracy of 73.52%, further emphasizing the crucial role of the ALBEF module in multimodal tasks.

**Table 6. Ablation experiments in the ALBEF module on the visual genome dataset with adjustments**

| Method | Yes/No (%) | Number (%) | Other (%) | All (%) |
|---|---|---|---|---|
| ResNet-152(Zhang, Li, Zhu, & Du, 2022) | 88.91 | 71.94 | 56.98 | 68.17 |
| BLIP (J. Li et al., 2022) | 81.65 | 64.01 | 56.21 | 70.04 |
| BLIP (Radford et al., 2021) | 83.54 | 63.98 | 50.06 | 71.65 |
| ALBEF (J. Li et al., 2021) | 93.95 | 73.02 | 61.60 | 73.52 |

Therefore, these ablation experiment results highlight the effectiveness of the ALBEF module on the Visual Genome dataset and underscore its significant contribution to multimodal tasks.

We further conducted ablation experiments on the CLIP module, and the results are shown in Table 7. It is evident from the table data that the CLIP model outperforms all other methods across all evaluation metrics. Particularly noteworthy is its outstanding accuracy in the Yes/No evaluation, reaching 93.95%. This indicates the strong performance of CLIP in answering binary questions. In the Number evaluation, CLIP also achieved the highest accuracy of 73.02%, highlighting its excellent performance in handling numerical-related questions. In the overall evaluation metric, All, the CLIP model likewise obtained the highest accuracy, reaching 73.52%, further emphasizing its outstanding performance in multi-modal tasks.

Through these experiments, we conclude that the CLIP module demonstrates exceptional performance in the task of multi-modal environmental pollution detection. The choice of CLIP is indeed wise and promises to contribute significantly to addressing contemporary environmental challenges. This underscores the importance of managing the subject of environmental pollution detection to protect the Earth's ecosystems and biodiversity.

## PRESENTATION OF RESULTS

The examples presented in Figure 8 vividly demonstrate the reasoning capabilities of our model in environmental issues, showcasing its outstanding performance in comprehending complex scenarios and contextual questions related to climate change and ecological conservation. Through its responses to these environment-related questions, our model exhibits a sharp ability to grasp multi-level and multi-dimensional information regarding environmental protection. This enables it to consider a more comprehensive range of perspectives, leading to more accurate environmental inference results. This series of qualitative examples not only highlights the tremendous potential of our model in environmental monitoring and analysis, but also further substantiates its unique advantage in addressing complex real-world challenges related to environmental protection.

## CONCLUSION

This study introduces an innovative model that combines the vision module with the language and vision model (LLM) to effectively integrate visual and textual information, with the aim of addressing challenges related to carbon neutrality and Olympic environmental protection. The experimental process involves several key steps: first, the utilization of the vision module for image feature extraction; second, the application of the ALBEF model for cross-modal alignment; third, the use of the CLIP method for specific object detection; and finally, the integration of all information within the LLM for multi-modal understanding and task execution. The primary contribution of this research lies in the pioneering integration of the vision module and language and vision model (LLM), which offers a novel perspective for tackling environmental issues. This fusion enhances our ability to comprehend

**Table 7. Ablation experiments in the CLIP module on the visual genome dataset with adjustments**

| Method | Yes/No (%) | Number (%) | Other (%) | All (%) |
|---|---|---|---|---|
| ResNet-152 (Zhang et al., 2022) | 65.91 | 61.94 | 56.98 | 68.17 |
| COCA (J. Yu et al., n.d.) | 83.68 | 64.01 | 56.21 | 70.04 |
| BLIP (J. Li et al., 2021) | 83.54 | 63.98 | 50.06 | 61.65 |
| CLIP (Radford et al., 2021) | 93.95 | 73.02 | 61.60 | 73.52 |

**Figure 8. Practical demonstration of our model in real-world scenarios**


the connections between image and textual information, ultimately providing more precise solutions for complex problems like carbon neutrality and Olympic environmental protection.

However, our research does have certain limitations that warrant careful consideration and resolution. First, the datasets used may exhibit biases in certain aspects, potentially impacting the model's training and performance evaluation. In future studies, it is imperative to employ more extensive and representative datasets to further validate the model's stability and reliability. Additionally, despite achieving satisfactory results in this study, the computational and storage requirements of the model when handling large-scale data could become limiting factors. Exploring methods to optimize the computational efficiency of the model will be necessary to better accommodate practical applications.

In future research, we plan to continue developing and expanding upon the innovative model we have proposed to better address challenges in critical areas such as carbon neutrality and Olympic environmental protection. First, we intend to further optimize and extend our model to enhance its adaptability and generalization performance across diverse environmental conditions. By incorporating more data and variations in scenarios, we can rigorously validate the model's robustness and ensure its stability in practical applications. Second, we will delve deeper into the fusion and understanding techniques of multimodal information. Exploring advanced fusion methods and strategies will elevate the model's comprehensive comprehension of both image and text information, thereby providing more comprehensive solutions for intricate environmental protection issues. Additionally, we will emphasize the practical value of the model, particularly in real-world projects such as carbon neutrality and Olympic environmental protection. Collaborating with partners in relevant fields will allow us to apply the model to real-world scenarios, gather feedback, and continuously improve, thereby realizing the impact of our research. We aspire for this model to offer innovative environmental protection solutions for hosting major international events. By reducing resource consumption, optimizing energy utilization, and promoting sustainable development, we can create comfortable and eco-friendly

venues for athletes and spectators alike. This will further highlight the strong connection between environmental protection and global society, infusing more green elements into the future of the Olympic movement.

## AUTHOR NOTE

This work was supported by Zhejiang Province Jinhua City Federation of Social Sciences (No. YB2020060), Jinhua City Science and Technology Research Plan (No.2021-4-379), General scientific research project of Zhejiang Provincial Department of Education (No.Y202045644), Zhejiang Province Soft Science Research Program Key Project (No.2022C25012), Zhejiang Province Social Science Planning Project: Special Project of the 20th National Congress of the Communist Party of China (No.17).

## 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*. ACL. doi:10.18653/v1/2021.emnlp-main.164

Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., & Simonyan, K. (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*. IEEE. 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*. IEEE.

Cheng, L., van Dongen, B. F., & van der Aalst, W. M. (2019). Scalable discovery of hybrid process models in a cloud computing environment. *IEEE Transactions on Services Computing*, *13*(2), 368–380. doi:10.1109/TSC.2019.2906203

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

Gao, L., Cao, L., Xu, X., Shao, J., & Song, J. (2020). Question-led object attention for visual question answering. *Neurocomputing*, *391*, 227–233. doi:10.1016/j.neucom.2018.11.102

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.* IEEE. doi:10.1109/CVPR52729.2023.01046

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

Han, K., & Yuan, S. (2023). Short text semantic similarity measurement algorithm based on a hybrid machine learning model. *Journal of Jilin University: Science Edition*, *61*(909-914).

Ke, Y., Liang, J., & Wang, L. (2023). Characterizations of weighted right core inverse and weighted right pseudo core inverse. *Journal of Jilin University: Science Edition*, *61*(4), 733–738.

Kim, J. H., Jun, J., & Zhang, B. T. (2018). Bilinear attention networks. *Advances in Neural Information Processing Systems*, 31.

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.* IEEE.

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, R., Wang, Q., Liu, Y., & Jiang, R. (2021). Per-capita carbon emissions in 147 countries: The effect of economic, energy, social, and trade structural changes. *Sustainable Production and Consumption*, *27*, 1149–1164. doi:10.1016/j.spc.2021.02.031

Liu, C., Zeng, Q., Cheng, L., Duan, H., & Cheng, J. (2021). Measuring similarity for data-aware business processes. *IEEE Transactions on Automation Science and Engineering*, *19*(2), 1070–1082. doi:10.1109/TASE.2021.3049772

Liu, X., Ye, P., Zhao, G., Li, J., Jiang, H., & Sun, M. (2023). Prediction of carbon emissions in Zhejiang province based on ATT-CNN-LSTM model. *The 8th Asia conference on power and electrical engineering (ACPEE).* IEEE.

Liu, Y., Zhang, X., Huang, F., Tang, X., & Li, Z. (2019). Visual question answering via attention-based syntactic structure tree-LSTM. *Applied Soft Computing*, *82*, 105584. doi:10.1016/j.asoc.2019.105584

Liu, Y., Zhang, X., Zhao, Z., Zhang, B., Cheng, L., & Li, Z. (2020). ALSA: Adversarial learning of supervised attentions for visual question answering. *IEEE Transactions on Cybernetics*, *52*(6), 4520–4533. doi:10.1109/TCYB.2020.3029423 PMID:33175690

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

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

Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. *Advances in Neural Information Processing Systems*, 29.

Mao, A., Yang, Z., Lin, K., Xuan, J., & Liu, Y.-J. (2022). Positional attention guided transformer-like architecture for visual question answering. *IEEE Transactions on Multimedia*.

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*. IEEE. doi:10.1109/CVPR.2019.00331

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., & Agarwal, S. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*. IEEE.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. *Advances in Neural Information Processing Systems*, 30.

Sun, B., Yao, Z., Zhang, Y., & Yu, L. (2020). Local relation network with multilevel attention for visual question answering. *Journal of Visual Communication and Image Representation*, *73*, 102762. doi:10.1016/j.jvcir.2020.102762

Waheed, R., Sarwar, S., & Wei, C. (2019). The survey of economic growth, energy consumption and carbon emission. *Energy Reports*, *5*, 1103–1115. doi:10.1016/j.egyr.2019.07.006

Wang, W., Zhang, J., Cao, Y., Shen, Y., & Tao, D. (2022). Towards data-efficient detection transformers. *European conference on computer vision*. IEEE.

Waswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., & Polosukhin, I. (2017). *Attention is all you need*. Nips.

Whalen, J., & Mouza, C. et al.. (2023). ChatGPT: Challenges, opportunities, and implications for teacher education. *Contemporary Issues in Technology & Teacher Education*, *23*(1), 1–23.

Wu, C., Liu, J., Wang, X., & Dong, X. (2018a). Chain of reasoning for visual question answering. *Advances in Neural Information Processing Systems*, 31.

Wu, C., Liu, J., Wang, X., & Dong, X. (2018b). Object-difference attention: A simple relational attention for visual question answering. *Proceedings of the 26th ACM international conference on multimedia*. ACM. doi:10.1145/3240508.3240513

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*. IEEE. doi:10.1109/CVPR.2016.10

Yu, Z., Yu, J., Cui, Y., Tao, D., & Tian, Q. (2019). Deep modular co-attention networks for visual question answering. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE. doi:10.1109/CVPR.2019.00644

Yu, Z., Yu, J., Fan, J., & Tao, D. (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *Proceedings of the IEEE international conference on computer vision*. IEEE. doi:10.1109/ICCV.2017.202

Yu, Z., Yu, J., Xiang, C., Fan, J., & Tao, D. (2018). Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(12), 5947–5959. doi:10.1109/TNNLS.2018.2817340 PMID:29993847

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