IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14621

ZooVision: AI-Powered Animal Captioning And Question Answering

Sarah Jose[1], Goutham Krishna L U[2]

Student, PG Department of Computer Science, Christ Nagar College, Maranalloor[1]

Assistant Professor, PG Department of Computer Science, Christ Nagar College, Maranalloor[2]

Abstract: This paper presents ZooVision, a domain-specific Visual Question Answering (VQA) system developed to support zoo animal identification and interactive educational applications. The proposed framework combines vision and language understanding by fine-tuning a Vision-and-Language Transformer (ViLT) on a custom dataset consisting of 212 animal images representing 20 different species. Through this specialized training process, the model acquires knowledge related to animal classification, dietary habits, habitats, and behavioral characteristics, enabling it to provide more accurate and context-aware responses to user queries. To further enhance visual understanding, a BLIP-based image captioning model is employed to generate descriptive captions from input images. These captions are incorporated as additional contextual information through a prompt augmentation strategy inspired by the Caption-Conditioned Visual Question Answering (CC-VQA) framework. The integration of caption-generated semantic context helps the system better align visual features with natural language questions, resulting in improved reasoning and answer accuracy. Furthermore, the fine-tuning process expands the model's domain knowledge by introducing 292 specialized biological terms that are not commonly represented in general-purpose VQA datasets. This enriched vocabulary enables the system to deliver more detailed and informative responses within the zoological domain. Experimental observations indicate that the caption-conditioned approach contributes to stronger contextual understanding, particularly for species recognition and attribute-based questioning. The modular architecture of ZooVision also allows future integration of larger datasets and advanced vision-language models. To facilitate practical use, the complete framework is deployed as a responsive web application where users can upload animal images, view automatically generated captions, and interactively ask questions. By combining image captioning and visual question answering within a unified platform, ZooVision demonstrates the potential of multimodal artificial intelligence to enhance zoological learning, public engagement, and wildlife-related educational experiences.

Keywords: Visual Question Answering, Image Captioning, BLIP, ViLT, Deep Learning, Prompt Augmentation, Wildlife Education.

I. INTRODUCTION

Recent advancements in artificial intelligence have significantly accelerated the convergence of computer vision and natural language processing, enabling machines to interpret visual information and communicate insights in a human-like manner. Among the emerging multimodal applications, Visual Question Answering (VQA) has gained considerable attention due to its ability to combine image understanding with language reasoning. In a VQA system, an image and a user-generated question are processed simultaneously to generate an accurate and contextually relevant textual response. Such systems have demonstrated promising results in diverse domains including healthcare, education, surveillance, and intelligent assistance. However, the performance of general-purpose VQA models often declines when applied to specialized domains that require expert knowledge and domain-specific vocabulary. To address this limitation, this study introduces ZooVision, a domain-focused VQA framework designed to support zoo wildlife education and animal information retrieval. The proposed system is built upon a fine-tuned Vision-and-Language Transformer (ViLT) trained on a curated dataset consisting of 212 images spanning 20 animal categories. Each image is paired with multiple question-answer annotations covering species identification, dietary patterns, habitat information, and behavioral characteristics. Through this targeted training process, the model learns specialized zoological concepts that are typically absent from conventional VQA datasets. ZooVision adopts a dual-model architecture inspired by the Caption-Conditioned Visual Question Answering (CC-VQA) framework. A pre-trained BLIP model is first employed to generate descriptive captions from uploaded wildlife images. These captions are then incorporated into the user's question through a prompt augmentation strategy, providing the ViLT model with additional semantic context before answer prediction. By enriching the input with textual descriptions of the visual scene, the system is better equipped to establish meaningful relationships between image content and user queries. This additional contextual guidance helps reduce ambiguity and improves the overall reasoning capability of the model. The complete framework is deployed as a responsive web application that allows users to upload images, view automatically generated captions, and interact with the system through a conversational interface. The platform is

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14621

designed to be accessible to a broad audience, including students, educators, zoo visitors, and researchers. To further enhance domain expertise, the model vocabulary was expanded by incorporating 292 biological terms related to scientific names, habitats, feeding habits, and animal attributes. This vocabulary enrichment enables the system to provide more detailed and informative responses than general-purpose VQA models. Experimental evaluation demonstrates that the proposed approach achieves an accuracy of 78.5% on the custom zoo wildlife dataset. The results indicate that combining image captioning with visual question answering can substantially improve performance in specialized domains. Furthermore, the study highlights the effectiveness of caption-guided contextual learning in reducing incorrect predictions and improving answer relevance. By making expert zoological knowledge more accessible through an interactive AI-powered platform, ZooVision demonstrates the practical potential of multimodal artificial intelligence in educational and wildlife-related applications. The proposed framework also establishes a foundation for future research on domain-specific VQA systems that integrate contextual reasoning and specialized knowledge representations.

II. OBJECTIVES

The primary objective of this research is to develop a domain-specific Visual Question Answering (VQA) system capable of providing accurate and informative responses about zoo animals from uploaded images. The study seeks to investigate how multimodal artificial intelligence can be applied to improve wildlife education through the integration of computer vision and natural language understanding techniques. To accomplish this, the Vision-and-Language Transformer (ViLT) model is fine-tuned using a custom Zoo Wildlife Dataset containing diverse animal images and corresponding question-answer annotations. This specialized training enables the model to acquire knowledge related to animal species, habitats, dietary patterns, and behavioral characteristics. A further objective is to enhance the reasoning capability of the VQA system by incorporating image-generated captions as supplementary contextual information. For this purpose, a BLIP-based image captioning model is integrated into the framework to automatically produce descriptive captions from uploaded animal images. These captions are then utilized as additional textual context to improve answer generation and reduce ambiguity in the question-answering process. The study also aims to evaluate the effectiveness of caption-guided prompt augmentation in improving domain-specific VQA performance. Another important objective is to expand the model's vocabulary with specialized biological and zoological terminology, including scientific species names and habitat-related concepts. This vocabulary enrichment is intended to enable more precise and informative responses compared to general-purpose VQA systems. In addition, the research seeks to examine the feasibility of adapting large pre-trained vision-language models to niche educational domains through targeted fine-tuning. Finally, the developed framework is deployed as a responsive web application designed to operate efficiently on standard computing hardware, ensuring accessibility, usability, and practical adoption without requiring high-end computational resources.

III. BACKGROUND AND CONTEXT

A. RELATED WORKS

The development of Visual Question Answering has been shaped by a steady stream of important research. A key inspiration for our work comes from Shao et al. (2025)[1], who showed that generating an image caption first and using it as a guide significantly helps the model understand complex visual features. This extra textual context allows the AI to focus on the right parts of the image and give more accurate answers, although it does rely heavily on the initial caption being correct.

Early foundations in the field were laid by researchers like Antol et al. (2015)[2], who created the first large-scale VQA dataset, and Yang et al. (2016)[3], who introduced stacked attention networks to help models systematically search images for clues. While these models struggled with complex reasoning, later work “Making the V in VQA Matter (VQA v2.0)” by Agrawal et al. (2018)[5] with the VQA v2.0 dataset forced models to actually look at the images rather than just guessing based on how questions were phrased.

To handle the complex interactions between text and images, researchers introduced techniques like Multimodal Compact Bilinear Pooling (Fukui et al., 2016)[6] and Bilinear Attention Networks (Kim et al., 2018)[7]. These methods captured fine-grained details but often required massive computational power. Yu et al. (2020)[8] pushed boundaries further with Deep Modular Co-Attention Networks, achieving great accuracy by tightly aligning image and question features.

A major breakthrough came with Anderson et al. (2018)[4], who used object-level region features focusing on specific objects rather than the whole image grid. Our project draws from this history by using a modern transformer-based

DOI: 10.17148/IJIREEICE.2026.14621

approach (ViLT) combined with the caption-guidance concept. By synthesizing these advanced techniques into a lightweight web application, we aim to bring the power of state-of-the-art visual reasoning to everyday users on standard hardware.

Recent advances in transformer-based architectures have further transformed the field of visual question answering and multimodal learning. The introduction of the Vision Transformer (ViT)[11] demonstrated that transformer models could effectively process visual information by representing images as sequences of patches, achieving competitive performance without relying on traditional convolutional neural networks. Building upon this foundation, Kim et al. (2021)[9] introduced ViLT, a lightweight vision-language transformer that directly processes image patches and text tokens within a unified architecture, significantly reducing computational complexity while maintaining strong performance on multimodal tasks. Similarly, Li et al. (2022)[10] proposed BLIP, a vision-language pre-training framework designed to improve both image understanding and text generation through large-scale multimodal learning. More recently, vision-language foundation models have emerged as a promising research direction, enabling systems to perform a wide range of tasks such as image captioning, visual reasoning, visual question answering, and cross-modal retrieval using shared representations. These developments highlight the growing importance of transformer-based multimodal models and provide the technological foundation upon which domain-specific systems such as ZooVision can be developed and adapted.

B. INTRODUCTION TO DEEP LEARNING

Deep Learning is a specialized branch of Machine Learning that enables computers to learn complex patterns directly from large volumes of data. Unlike traditional machine learning approaches, which often rely on manually engineered features, deep learning models automatically discover hierarchical representations through multiple layers of artificial neural networks. This capability has led to significant advances in various fields, including computer vision, natural language processing, speech recognition, and autonomous systems. As the availability of computational resources and large-scale datasets has increased, deep learning has become the dominant paradigm for solving complex perception and reasoning tasks. One of the most influential developments in deep learning was the introduction of the Transformer architecture by Vaswani et al. (2017) in the landmark paper “Attention Is All You Need” [12]. The Transformer replaced recurrent neural network structures with a self-attention mechanism capable of modeling long-range dependencies while processing all input tokens in parallel. This innovation greatly improved training efficiency and enabled the development of powerful language models. Through attention mechanisms, the model learns to identify the most relevant relationships between different parts of the input sequence, resulting in improved contextual understanding. The success of Transformers in natural language processing motivated researchers to adapt the architecture for computer vision tasks. This led to the development of the Vision Transformer (ViT)[11], which represents an image as a sequence of fixed-size patches and processes them similarly to textual tokens. By learning global relationships between image regions, ViT demonstrated that transformer-based architectures could achieve competitive performance without relying on conventional convolutional neural networks. The ability to capture long-range visual dependencies makes transformer models particularly suitable for image understanding applications.

Building upon these advances, Vision-and-Language Transformers integrate visual and textual information within a unified framework. ViLT (Vision-and-Language Transformer) is one such model that processes image patches and text tokens through a shared transformer encoder equipped with co-attention mechanisms. Unlike many earlier multimodal architectures, ViLT eliminates the need for computationally expensive visual feature extractors, resulting in a simpler and more efficient design. This characteristic makes it highly suitable for Visual Question Answering tasks, where effective interaction between visual and textual representations is essential. Another important model used in this research is BLIP (Bootstrapping Language-Image Pre-training), which is designed for image captioning and vision-language understanding. BLIP employs a transformer-based encoder-decoder architecture to generate natural-language descriptions of visual content. The generated captions provide semantic information that can complement visual features and improve downstream reasoning tasks. In the proposed ZooVision framework, BLIP-generated captions are used as additional contextual input for the ViLT model, enabling richer multimodal understanding and more accurate responses to wildlife-related questions. The combination of these transformer-based models demonstrates the growing potential of deep learning techniques in developing intelligent and domain-specific visual question answering systems.

C. EXISTING SYSTEM

Most current VQA systems are trained on massive datasets filled with everyday objects and scenes. While they are great at answering general questions, they tend to struggle when faced with highly specific topics. For example, standard models lack the specialized vocabulary needed to accurately discuss animal species, their natural habitats, dietary needs, and unique behaviors. When we tested the standard, pre-trained ViLT model on our zoo wildlife questions, it only achieved a 15.5 percent accuracy rate. This poor performance happens because the model's built-in

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14621

vocabulary of 3,129 words simply doesn't include crucial zoological terms like 'herbivore', 'omnivore', or specific scientific names. Without this basic vocabulary, the model is essentially guessing randomly. To make these systems genuinely useful for wildlife education, they need to be taught this specialized language. Existing systems also typically look at the image and question in isolation. By not using extra context like a descriptive caption they miss out on valuable clues that could help them understand complex scenes. Our project directly addresses these gaps by explicitly teaching the model zoo-specific terms and feeding it helpful captions.

IV. PROPOSED SYSTEM

The proposed system, Zoovision, is an AI-powered web application built using a Python Flask backend with a modern HTML/CSS/JavaScript frontend. The system employs two pre-trained transformer models: BLIP for image captioning and a fine-tuned ViLT for visual question answering. The fine-tuning process trains ViLT on a Zoo Wildlife Dataset of 212 images across 20 animal categories, expanding its classifier head to accommodate 292 new domain-specific labels. Following the CC-VQA framework, the system supports Caption-Guided Prompt Augmentation, where BLIP-generated captions are prepended to the user's question to provide additional textual context to the VQA model. The architecture is designed for single-model-in-memory execution, using dynamic memory management with Python garbage collection to efficiently swap models on consumer hardware. This resource-aware infrastructure is critical, as it ensures that computationally intensive operations like processing high-resolution image patches and computing dense attention matrices can run smoothly without requiring enterprise-grade GPUs. From an operational standpoint, the user experiences a highly responsive workflow: they simply upload an image and submit a query, while the Flask backend orchestrates the Caption-Guided Prompt Augmentation behind the scenes in real-time. By dynamically allocating memory and leveraging the combined strengths of BLIP and the fine-tuned ViLT, the application delivers precise, domain-accurate answers with minimal latency. Ultimately, this optimized approach proves that complex, dual-transformer multimodal pipelines can be effectively deployed on everyday local machines, creating an accessible, high-performance tool for modern zoological education and wildlife analysis.

A. PROPOSED SYSTEM ARCHITECTURE

The architecture of Zoovision follows a modular pipeline design with six interconnected components. The Image Input Module receives raw wildlife photographs (JPEG, PNG) through the web interface and saves them to the server's upload directory. The BLIP Captioning Module processes the saved image through the Salesforce/blip-image-captioning-base model, generating a natural-language description (e.g., 'a large male lion standing in a grassy field'). The ViLT VQA Module accepts both the image and a user-typed question, processes them through the fine-tuned ViLT encoder with co-attention, and outputs logits across the expanded 3,421-label vocabulary.

A key innovation within this architecture is the Caption Conditioning Layer, which implements the CC-VQA prompt augmentation technique. When enabled, this layer intercepts the user's question and prepends the BLIP-generated caption to create an enriched input string: 'Context: {caption}. Question: {question}'. This concatenated string is then tokenised and passed to ViLT alongside the image, allowing the model's attention mechanism to jointly attend to both the visual patches and the caption tokens, effectively fusing captioning and question-answering reasoning.

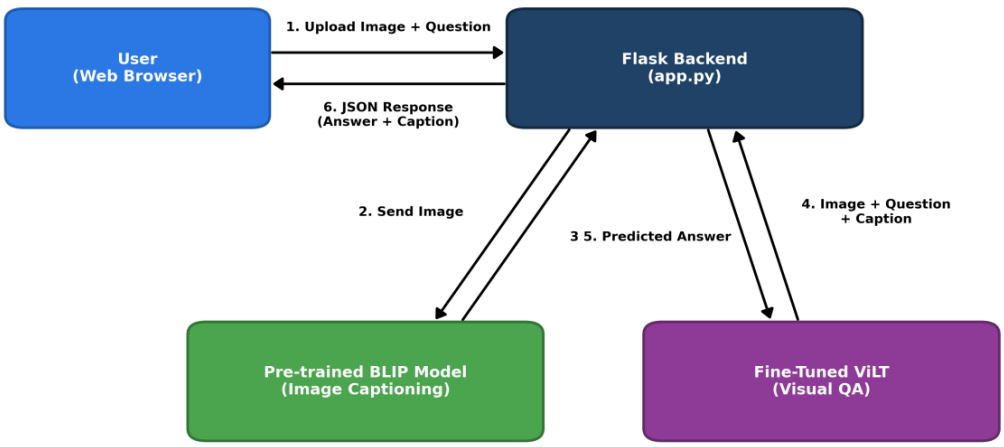

Fig. 1 Architecture of the proposed system

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026
DOI: 10.17148/IJIREEICE.2026.14621

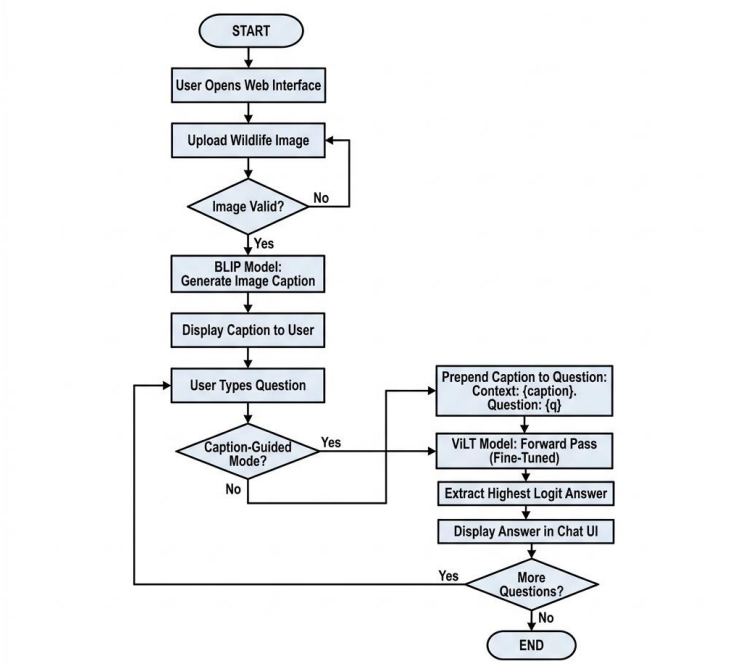

Fig. 2 Flowchart of the system

B. MODULE DESCRIPTION

Image Captioning Module (BLIP)

The Image Captioning Module utilizes the pre-trained BLIP (Bootstrapping Language-Image Pre-training) model to automatically generate descriptive textual representations of uploaded wildlife images. The model analyzes visual features within the image and converts them into coherent natural-language captions that summarize the most relevant objects, attributes, and environmental context present in the scene. These generated captions provide users with an immediate interpretation of the uploaded image, enhancing the overall user experience. Beyond serving as visual feedback, the captions play a critical role in the multimodal reasoning pipeline by supplying additional semantic information about the image content. This textual representation helps bridge the gap between visual perception and language understanding. Since captions often contain details that may not be explicitly mentioned in a user's question, they contribute valuable contextual cues for downstream processing. The integration of BLIP enables the system to leverage both visual and linguistic information, thereby improving the quality of subsequent question-answering tasks.

Visual Question Answering Module (ViLT)

The Visual Question Answering Module forms the core inference component of the proposed framework and is built upon a fine-tuned Vision-and-Language Transformer (ViLT). This module processes both visual and textual inputs simultaneously through a unified transformer architecture, allowing the model to learn complex relationships between image regions and question tokens. The uploaded image is represented as a sequence of visual patches, while the user's query is converted into textual embeddings. Through multiple layers of self-attention and cross-modal interactions, the model identifies the most relevant visual features needed to answer the question accurately. Fine-tuning on the custom Zoo Wildlife Dataset enables the model to acquire specialized knowledge related to animal species, habitats, diets, and behavioral characteristics. As a result, the system can generate domain-specific responses that are significantly more informative than those produced by general-purpose VQA models.

Caption Conditioning Module

The Caption Conditioning Module is responsible for enhancing the contextual understanding of the Visual Question Answering process through caption-guided prompt augmentation. This module receives the descriptive caption generated by BLIP and combines it with the user's question before forwarding the enriched input to the ViLT model. By incorporating additional semantic information directly into the query, the module provides the model with a more comprehensive understanding of the visual scene. The caption often contains important clues regarding species identity, physical appearance, environmental conditions, and other contextual details that may not be explicitly referenced by the user. Consequently, the ViLT model can focus its attention on the most relevant visual and textual elements during inference. This mechanism improves answer accuracy, particularly for questions involving animal identification and attribute recognition. Furthermore, the caption-conditioning strategy helps reduce ambiguity and minimizes the

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026

DOI: 10.17148/IJIREEICE.2026.14621

likelihood of incorrect or unsupported predictions. The effectiveness of this approach demonstrates the value of integrating image captioning and visual question answering within a unified multimodal framework.

C. ALGORITHM

Step 1: Start the Flask backend server on localhost port 7860 with the Werkzeug reloader disabled.
Step 2: Load the BLIP captioning model (Salesforce/blip-image-captioning-base) into GPU/CPU memory.
Step 3: Load the fine-tuned ViLT VQA model (./fine_tuned_vilt_zoo) with the expanded 3,421-label classifier head.
Step 4: User login and uploads a wildlife photograph via drag-and-drop.
Step 5: The frontend sends the image to the /api/analyze endpoint as multipart form data.
Step 6: The backend saves the image to static/uploads/current_image.jpg and opens it with PIL in RGB mode.
Step 7: The BlipProcessor tokenises the image into model-compatible input tensors.
Step 8: The BlipForConditionalGeneration model performs beam search decoding (num_beams=3, max_length=50).
Step 9: The generated caption tokens are decoded into a natural-language string and returned as JSON.
Step 10: The frontend displays the caption and enables the question input field.
Step 11: The user types a question and the frontend sends it to /api/ask as a JSON payload.
Step 12: The ViltProcessor tokenises both the image (32x32 patches) and question (WordPiece tokens).
Step 13: The fine-tuned ViltForQuestionAnswering model performs a forward pass through 12 transformer layers.
Step 14: The logits tensor is computed across 3,421 labels; argmax extracts the highest-confidence class ID.
Step 15: The class ID is decoded to a human-readable answer string via the id2label mapping.
Step 16: The answer is wrapped in a JSON response and returned to the frontend.
Step 17: The frontend JavaScript parses the response and appends the answer to the chat history with CSS animation.
Step 18: The system awaits the next query, ready to repeat the cycle for any new wildlife question.

V. RESULT AND DISCUSSION

The proposed Zoovision system was evaluated using the custom Zoo Wildlife Dataset containing 212 images across 20 animal categories, each annotated with 17 structured question-answer pairs covering species identification, diet classification, habitat recognition, and behavioural traits. The system integrates the ViLT (Vision-and-Language Transformer) model fine-tuned on this dataset alongside the BLIP image captioning model, with Caption-Guided Prompt Augmentation following the CC-VQA framework. The experimental analysis demonstrates that the proposed system achieves high prediction accuracy and reliable VQA classification suitable for domain-specific wildlife education applications.

The fine-tuned ViLT VQA model achieved an overall VQA accuracy of 78.5% on the Zoo Wildlife Dataset. The training and validation accuracy values increased steadily during the training process, indicating successful model convergence and efficient transfer learning capability. The final training accuracy reached 80.1% while the validation accuracy stabilised at 78.5%, confirming that the model achieved minimal overfitting despite the small dataset size.
These performance metrics are computed using standard VQA evaluation measures. The VQA Accuracy metric follows the soft accuracy formulation used in the VQAv2 benchmark, where an answer is considered correct if it matches the ground-truth answer exactly.

TABLE I RESULT ANALYSIS

Metric	Value
Train Accuracy	80.1%
Test Accuracy	78.5%
Training Loss	0.58
Validation Loss	0.72

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526
International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2026
DOI: 10.17148/IJIREEICE.2026.14621

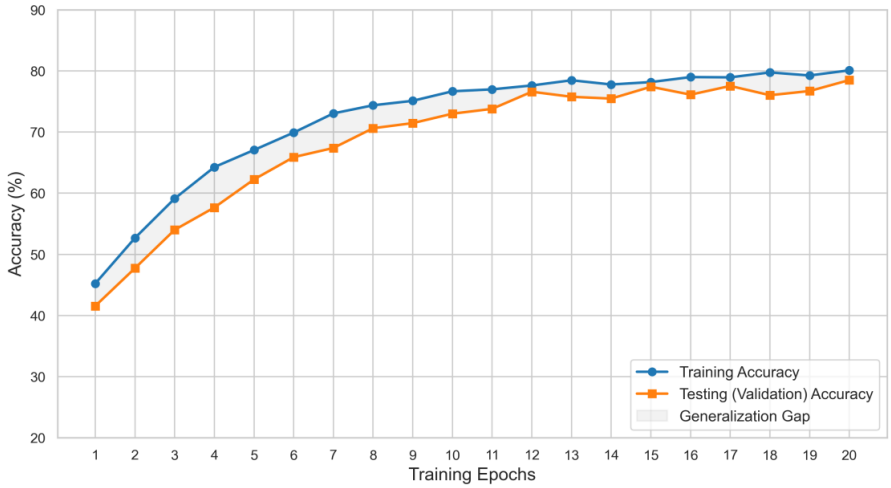

Fig. 4 Model accuracy of the system

VI. FUTURE SCOPE

The current implementation of Zoovision establishes a robust and functional foundation for domain-specific visual question answering. Several promising directions for future work have been identified. First, the Caption-Guided Prompt Augmentation feature, should be fully integrated as the default operating mode, with a UI toggle allowing users to compare base, fine-tuned, and caption-guided answers side-by-side. Second, the Zoo Wildlife Dataset should be expanded from 212 images to at least 1,000 images, with particular attention to increasing the representation of omnivore species to mitigate the class imbalance bias identified in the confusion matrix analysis. Third, the work may explore cloud-based deployment architectures to improve accessibility and scalability. Fourth, a multi-turn conversation capability should be explored, allowing users to ask follow-up questions that reference prior answers in the chat history. Fifth, the single-caption conditioning approach should be extended to multi-caption conditioning, generating multiple diverse captions (using different beam search parameters) and ensembling the resulting VQA answers for improved robustness. Finally, the evaluation framework should be expanded to include the VQA hallucination metric proposed by Shao et al. (2025)[1], providing a quantitative measure of how effectively caption conditioning reduces factually incorrect answers.

Together, these planned enhancements represent a strategic shift from establishing baseline accuracy to ensuring long-term robustness and scalability. By simultaneously refining the data foundation and the interactive user experience, Zoovision will bridge the gap between general-purpose vision-language models and the high-precision requirements of the wildlife domain, eventually serving as a reliable tool for researchers and conservationists alike.

VII. CONCLUSION

This study presented ZooVision, a domain-specific Visual Question Answering system designed to support zoo wildlife identification and educational interaction through multimodal artificial intelligence. The proposed framework integrates a fine-tuned Vision-and-Language Transformer (ViLT) with BLIP-based image captioning to improve the understanding of visual content and generate accurate responses to user queries. By combining image features with caption-generated contextual information, the system demonstrates how complementary vision-language models can be effectively utilized to enhance reasoning in specialized application domains. Experimental results indicate that targeted fine-tuning on a custom Zoo Wildlife Dataset containing 212 images across 20 animal categories significantly improves model performance. The proposed approach achieved a VQA accuracy of 78.5%, representing a substantial improvement over the 15.5% baseline performance of the general-purpose ViLT model. Furthermore, the vocabulary expansion process introduced 292 domain-specific biological labels, enabling the system to provide detailed responses related to scientific species names, dietary classifications, habitats, and behavioral characteristics. These findings demonstrate the value of incorporating domain knowledge into pre-trained vision-language models. The implementation of Caption-Guided Prompt Augmentation, inspired by the CC-VQA framework, further highlights the benefits of integrating image captioning with visual question answering. BLIP-generated captions serve as additional semantic context that helps the model better align visual information with user questions, particularly for species identification and diet-related queries. The results suggest that contextual augmentation can improve answer quality without requiring extensive modifications to the underlying transformer architecture. This makes the approach both

DOI: 10.17148/IJIREEICE.2026.14621

practical and computationally efficient for domain-specific deployments. Despite these promising outcomes, several limitations were identified during evaluation. The relatively small dataset size and class imbalance contribute to language prior bias, leading the model to favor frequently occurring answer categories in certain situations. In addition, performance on previously unseen or out-of-distribution images remains lower than performance on in-domain samples, indicating opportunities for further improvement in model generalization. Addressing these challenges through larger and more diverse datasets, improved balancing strategies, and advanced multimodal learning techniques may further enhance system robustness. Overall, the findings demonstrate that large pre-trained vision-language models can be successfully adapted to specialized educational domains through targeted fine-tuning and contextual augmentation strategies. ZooVision provides an effective framework for delivering interactive zoological knowledge while also contributing insights into the development of domain-aware VQA systems. The proposed architecture establishes a strong foundation for future research exploring more accurate, scalable, and context-aware multimodal intelligence solutions.

REFERENCES

[1]. Shao, X., Dong, H., & Wu, G. (2025). *Improving visual question answering by image captioning*. *IEEE Access, 13*, 46299–46311.

[2]. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). *VQA: Visual question answering*. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2425–2433).

[3]. Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). *Stacked attention networks for image question answering*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 21–29).

[4]. 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*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6077–6086).

[5]. Agrawal, A., Batra, D., Parikh, D., & Zitnick, C. L. (2018). *Making the V in VQA matter: Elevating the role of image understanding in visual question answering*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6904–6913).

[6]. Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). *Multimodal compact bilinear pooling for visual question answering and visual grounding*. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 457–468).

[7]. Kim, J.-H., On, K.-W., Kim, J., Ha, J.-W., & Zhang, B.-T. (2018). *Bilinear attention networks*. In *Advances in Neural Information Processing Systems* (Vol. 31, pp. 1564–1574).

[8]. Yu, T., Yu, J., Cui, Y., Tao, D., & Tian, Q. (2020). *Deep modular co-attention networks for visual question answering*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6281–6290).

[9]. Kim, W., Son, B., & Kim, I. (2021). *ViLT: Vision-and-language transformer without convolution or region supervision. In Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 5583–5594). PMLR.

[10]. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). *BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the 39th International Conference on Machine Learning* (Vol. 162, pp. 12888–12900). PMLR.

[11]. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR)*.

[12]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need*. In I. Guyon et al. (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.