Indonesian Journal of Data and Science

Volume 6 Issue 2
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v6i2.300

Research Article

Information Extraction from Makassar Culinary Images Using Vision Transformers and Cahya GPT-2 (Visual Question Answering Case Study

Tirta Chiantalia Sharief [1]; Hazriani [2]; Syamsul [3]; Anas [4]; Yuyun [5]
[1] Universitas Handayani Makassar, Makassar, 90241, Indonesia, tirtashariff@gmail.com
[2] Universitas Handayani Makassar, Makassar, 90241, Indonesia, hazriani@handayani.ac.id
[3] Universitas Handayani Makassar, Makassar, 90241, Indonesia, syamsulr8@gmail.com
[4] Universitas Handayani Makassar, Makassar, 90241, Indonesia, anas.daeng.pasewang@gmail.com
[5] Badan Riset dan Inovasi Nasional, Bandung 40135, Indonesia, yuyu010@brin.go.id

Correspondence should be addressed to Author; e-mail address

Received 10 October 2025; Accepted 15 December 2025; Published 31 December 2025

Abstract:
This study examines the development of a Visual Question Answering (VQA) system to extract information from images of Makassar culinary specialties by combining the Vision Transformer (ViT) and Cahya_GPT-2 models. The main objective is to integrate visual and natural language understanding so that computers can recognize visual objects (food images) and generate relevant text descriptions. The research method uses an experimental approach with a fine-tuning process of the pre-trained ViT model as a visual encoder and Cahya_GPT-2 as a text decoder. The dataset used includes images of Makassar culinary specialties such as Coto, Konro, Pisang Epe, Barongko, and Jalangkote with question and answer (QnA) annotations. Evaluation is carried out using the ROUGE metric to assess the semantic match between the model's answers and the actual answers. The results show that the developed multimodal model is able to accurately understand the image context with an average ROUGE-L score of 0.63, indicating a good level of closeness between the model's answers and the annotations. In conclusion, the combination of ViT and Cahya_GPT-2 can be an effective approach for natural language-based visual information extraction systems, especially in the Indonesian local culinary domain.

Keywords: Vision Transformer; Cahya_GPT-2; Visual Question Answering; Fine_Tuning; Citra Kuliner Makassar

Dataset link: link "https://drive.google.com/drive/folders/1yQUAJO8kW3EOD6fVf_lwyYsNK1t127C6?usp=sharing"

1. Introduction

The development of artificial intelligence (AI) technology in the current digital era has had a significant impact on various areas of human life, from industry and education to healthcare and even the creative sector. One rapidly developing branch of AI is the ability of computer systems to understand and process multimodal data, namely a combination of visual and textual data [1]. In this context, the integration of natural language processing (NLP) and image recognition (Computer Vision) is an interesting challenge that continues to be studied in the field of intelligent systems. Researchers see that the ability of computers to interpret images and associate them with textual information is a crucial step towards AI systems that are more intuitive and adaptive to real-world contexts [2]. This phenomenon is also evident in modern society, which is increasingly familiar with social media and visual-based content. In Makassar, for example, sharing images of traditional foods such as Coto Makassar, Pisang Epe, Jalangkote, and Barongko has become part of the community's digital culture. This situation presents a significant opportunity for the development of AI-based systems capable of automatically recognizing, classifying, and describing images of typical culinary delights. From the researcher's observation, this kind of technology not only has practical value in tourism

promotion and culinary culture archiving, but also has scientific value because it can strengthen the integration between visual recognition and natural language processing in the local Indonesian domain [3].

Most previous research in image processing has focused on general objects, such as face, vehicle, or landscape detection, and has not examined specific contexts such as traditional Indonesian cuisine. Furthermore, AI-based chatbot systems capable of accepting image input (image-based chatbots) are still relatively rare in Indonesia, especially for local cultural contexts [4]. Therefore, this research aims to develop a Visual Question Answering (VQA) system that combines two different artificial intelligence models: Vision Transformer (ViT) [5] as an image recognition model and Cahya-GPT2 as an Indonesian natural language processing model

Internationally, related research on VQA (Vegetable Quality Assurance) has largely focused on general-purpose datasets in English, such as VQA v2, GQA, and CLEVR. These datasets are designed to cover generic objects and activities, but do not yet represent the specific needs of other languages, particularly Indonesian, as well as local cultural contexts with diverse objects and terms, including those in the realm of traditional culinary arts. Consequently, the development of VQA systems that can understand images and generate factual answers in Indonesian remains very limited [6].

Text-based approaches alone are unable to understand the visual information contained in food images, such as shape, texture, color, presentation, or ingredient composition. In contrast, pure computer vision models can only classify or detect objects without the ability to generate factual textual answers in natural language. This gap demands the integration of modern visual modeling and Indonesian natural language processing to enable VQA [7] systems to operate end-to-end, from visual feature extraction to the generation of relevant linguistic responses.

To date, there has been no comprehensive research developing an Indonesian-language VQA system by integrating a modern visual encoder based on Vision Transformer (ViT) [8]and an Indonesian language decoder such as Cahya GPT-2, particularly in the traditional culinary domain. This is despite the fact that local culinary delights, such as Makassarese specialties, possess unique visual characteristics and terminology not captured in common international datasets. The lack of curated datasets and trained models in this context is a major obstacle to VQA research focused on local cultural preservation and the application of Indonesian language technology. Based on this gap, this study formulates two main questions:

- How effective is the combination of Vision Transformer (ViT) [9] as a visual encoder and Cahya GPT-2 as a language decoder in answering short factual questions based on images of Makassarese specialties in Indonesian?
- How much improvement does the VQA system performance resulting from the fine-tuning process of the integrated model achieve compared to the baseline model on the same test set?

To answer these research questions, this study makes several key contributions:

- The creation of the first Indonesian-language Makassar culinary VQA dataset, which includes traditional food images and short factual question-answer pairs, as an open resource for further research.
- The implementation of a Vision-Encoder-Decoder architecture that integrates ViT as a visual feature extractor with Cahya GPT-2 as an Indonesian language generative model to generate image-based textual answers.
- A quantitative evaluation of model performance using the ROUGE-L metric to assess the effectiveness of the fine-tuning process in improving answer quality compared to the baseline model.

This research contributes not only to the development of an Indonesian-language VQA system but also to enriching contextual artificial intelligence studies of local cultural richness, particularly traditional Makassar culinary [10].

2. Method

This Research This research uses an applied experimental approach that aims to develop a Visual Question Answering (VQA) system based on a multimodal model to recognize images of Makassar culinary specialties and

generate text descriptions in Indonesian. The method used is a combination of computer vision and natural language processing (NLP) with a Vision-Encoder-Decoder Transformer architecture, which consists of a Vision Transformer (ViT) [11] model as a visual encoder and Cahya-GPT2 as a text decoder. The research design is carried out through six main stages, namely *(a) dataset collection, (b) Annotation and Validation, (c) Dataset Distribution (d) Model Architecture (e) Training Setup (f) Evaluation.*

a) Dataset Collection Stage:

In the initial stage, dataset collection was conducted to build a corpus of Makassar culinary specialty images that represent the visual variety of local foods while also supporting the learning of the Visual Question Answering (VQA) model [12]. The dataset was collected through two main methods: direct image capture in the field and collection from open-license online sources [13].

In the first method, imagery was captured directly at several culinary locations in Makassar City, such as traditional culinary centers, regional food stalls, and local snack centers. All photos were taken using a mobile phone camera with a minimum resolution of 1080x1080 pixels and JPG format. Images were captured using various viewing angles (front, side, and top-view), camera distances, and natural lighting conditions to represent real-world usage conditions. The second method involved collecting additional imagery from online platforms that provide visual content with free use licenses, such as open source image documentation sites and creative repositories, to enrich the variety of settings, food presentation, and visual quality [14]. The types of culinary collected are limited to six typical Makassar foods which have strong visual characteristics and cultural identity, namely:

Table 1. Class Images

No	Class	Number Of Images
1	Coto Makassar	60
2	Konro	60
3	Pisang Epe	60
4	Barongko	60
5	Jalangkote	60
6	Panada	60
	Total	**360**

b) Annotation and Validation

After the image collection and selection process is complete, the next step is dataset annotation in Visual Question Answering (VQA) format, which involves assigning question-answer (QnA) pairs to each image. The goal of this annotation is to establish explicit semantic relationships between visual content and text representations to support the multimodal model learning process [15].

Each image was annotated with three short factual questions that focused on key information that could be directly identified from the food object, including: (1) menu identification (“What is this picture?”), (2) main ingredients (“What is the main ingredient of this food?”), (3) regional origin (“Where does this food come from?”), and (4) general characteristics (“What is barongko?”). Three different types of questions were given for each image to increase the variety of sentence structures and the breadth of language context. With this approach, a total of 1080 question–answer (QnA) pairs were formed from the 360 images collected [16]. Additional validation was carried out through manual evaluation by the principal researcher to ensure:

- Conformity between the question and the visual content, so that the question only refers to information that can be identified through the image.
- K Accuracy of facts, especially regarding food ingredients and region of origin.

- Clarity of language, avoiding ambiguity, informal abbreviations, and spelling variations.

To maintain consistency in model training and evaluation, cross-set QnA duplication checks were also performed before dividing the dataset into training, validation, and testing sections. Each QnA pair was ensured to only appear in one subset of the data to prevent information leakage that could affect the validity of the performance evaluation results. Through this annotation and validation procedure, the constructed dataset meets the characteristics of high label quality, language consistency, and semantic traceability, making it suitable for use in the learning and evaluation process of Indonesian Visual Question Answering models [17].

c) Dataset Distribution

After the annotation and validation processes were completed, the dataset was further divided into three main subsets: the training set, the validation set, and the testing set. The splitting was done using a 70%:15%:15% ratio, which sequentially resulted in 756 QnA pairs for training, 162 QnA pairs for validation, and 162 QnA pairs for testing out of a total of 1080 question-answer pairs, considering that one Makassarese Food image represents three pairs of questions and answers. The dataset splitting process was carried out using a controlled random split using a fixed random seed value of 42 to ensure that the data division can be reproduced by other researchers in subsequent experiments [18]. The use of this fixed seed is important to maintain the consistency of model evaluation results and allows for fair performance comparisons when repeating experiments or further studies. The class distribution was also kept balanced within each subset, so that each food type (Coto Makassar, Konro, Pisang Epe, Barongko, Jalangkote, and Panada) was proportionally represented in the training, validation, and testing data [19]. This approach was used to prevent training bias due to the dominance of certain classes, which could potentially reduce the reliability of model evaluation [20].

Cross-entropy is used to minimize the difference between the model's answers and the actual answers. ViT's weights are largely frozen to maintain its visual generalization ability, while Cahya-GPT2 focuses on adjusting the weights to allow the model to associate visual representations with the question context. This strategy effectively reduces training time while maintaining stable model performance [21].

d) Annotation and Validation

The Visual Question Answering (VQA) system in this study was built using the Vision-Encoder–Decoder Transformer architecture, which integrates image recognition and natural language models in a single multimodal processing circuit. This architecture consists of two main components: the Vision Transformer (ViT) as the visual encoder and Cahya GPT-2 as the Indonesian language decoder [22].

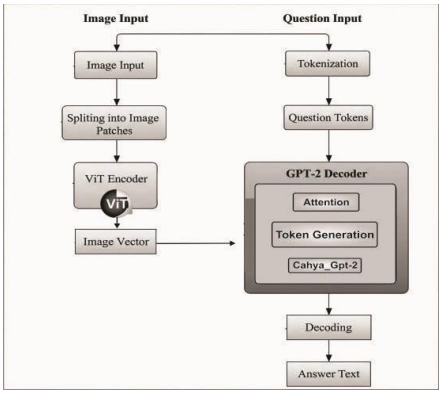

Figure 1. Image to Text Processing Flow

The integration of the two models enables the conversion of visual information into text-based linguistic semantic representations. On the encoder side, the ViT [23] model processes 224×224 pixel input images. The image is divided into 16×16 pixel patches, resulting in 196 patches per image. Each patch is projected into a 768-dimensional embedding space (patch embedding) and enriched with positional embedding to preserve

spatial information. The embedding sequence is then processed through a series of transformer blocks that implement multi-head self-attention mechanisms and feed-forward layers. This mechanism allows the model to learn global relationships between image parts and extract complex visual features such as shape, color, and texture of food objects [24], [25].

The final representation of the encoder is a set of visual embedding vectors, which are then used as context memory for the language decoder. In the multimodal integration stage, the ViT output embeddings are fed to the cross-attention layer of the Cahya GPT-2 decoder, allowing the language model to not only rely on the linguistic context of the question but also directly pay attention to relevant visual features of the input image. This approach enables implicit multimodal fusion at the attention level, without the need for additional fusion modules [26].

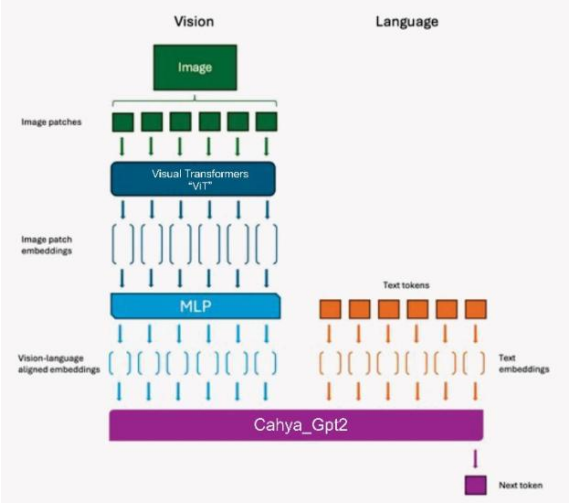

Figure 2. Block diagram flow of the model being built

The Cahya GPT-2 model functions as a text decoder, processing Indonesian input questions and generating answers in the form of token sequences. The tokenization process is carried out using Cahya GPT-2's built-in tokenizer, which is based on subword encoding and optimized for Indonesian vocabulary. Question tokens, along with visual embedding vectors, are processed by the decoder's transformer layer through self-attention and cross-attention mechanisms, then generating a probabilistic representation for each output token [27].

During the training and fine-tuning process, most parameters in the ViT encoder were frozen to maintain the visual generalization capability of the initial training results on the ImageNet dataset. Weight adjustments focused on all Cahya GPT-2 decoder layers and cross-attention layers so that the model could learn to associate visual features with the semantic structure of Indonesian language that is relevant to the local culinary context. This strategy was chosen to balance computational efficiency with the quality of multimodal learning, while also mitigating the risk of overfitting due to the limited size of the local dataset [6], [5].

In the inference stage, the system processing flow is as follows: (1) the input image is processed by ViT to produce a visual embedding; (2) the embedding is combined with the user's question tokens through a cross-attention mechanism in the decoder; (3) the decoder generates answer tokens gradually using a beam search decoding strategy (beam width = 4) until the final token appears or the maximum limit of 50 tokens is reached. The final output is a short answer sentence that explains relevant information related to the analyzed food image [28]

e) Training Settings

The model training process was conducted on the Google Colab Pro platform with NVIDIA T4 GPU (16 GB VRAM) support for accelerated computation. All experiments were conducted using the HuggingFace

Transformers library and the PyTorch framework. To ensure consistent reproducibility, the entire training process used a fixed random seed of 42 in the data generation module, trainable model weight initialization, and sampling decoding function [29].

In the visual preprocessing stage, all images were resized to 224×224 pixels according to the Vision Transformer input specifications, then normalized using the ImageNet mean and standard deviation scheme. For text, the tokenization process was carried out using the Cahya GPT-2 built-in subword-based tokenizer, with automatic batch padding and truncation up to a maximum sequence length of 512 tokens in the question input process [30]

The model was trained using a supervised learning-based supervised fine-tuning scheme. The loss function used was cross-entropy loss to minimize the difference between the model's predicted output tokens and the reference answer tokens. Given the limited size of the local dataset, most of the ViT encoder parameters were frozen during training, while all Cahya GPT-2 decoder layers and the multimodal cross-attention layer remained trainable [18]. This strategy aims to maintain the encoder's visual generalization capabilities while optimizing associations between visual features and Indonesian language specific to the culinary domain [30]. The following are the training hyperparameter settings:

Table 2. Training hyperparameter settings

Parameter	Nilai
Learning rate	1×10^{-5}
Optimizer	AdamW
Batch size	2
Jumlah epoch	5
Gradient clipping	1.0
Weight decay	0.01
Scheduler	Linear with warm-up
Beam width	4
Maksimum token output	50
Random seed	42

Training was conducted over five epochs, with performance monitoring on the validation set at the end of each epoch. The model with the highest ROUGE-L score on the validation data was kept as the best checkpoint selection. To prevent overfitting, an early stopping mechanism based on validation loss stagnation was used, with a tolerance limit of two consecutive epochs without performance improvement.

f) Evaluation

A performance evaluation of the Visual Question Answering (VQA) system was conducted to measure the model's ability to generate text answers that semantically match human reference answers. The primary evaluation method used in this study is the ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation based on Longest Common Subsequence) metric, which assesses the similarity of the longest word sequence between the model's output sentence and the reference answer. ROUGE-L was chosen for its ability to represent sentence structure alignment and semantic proximity, which are relevant for short question-and-answer tasks in Indonesian. Measurements were conducted on 162 QnA pairs on test data that were never used in the training or validation process. The ROUGE-L score was calculated as a micro-average value across all test samples to obtain an aggregate picture of the model's performance. In addition, a macro-average value was also calculated based on the average score of each food class (Coto Makassar, Konro, Pisang Epe, Barongko, Jalangkote, and Panada) to assess performance stability across culinary categories.

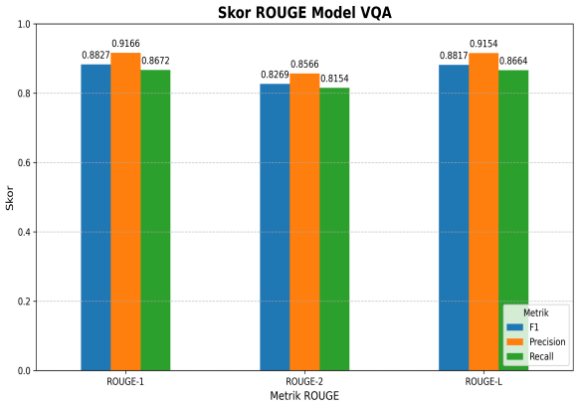

Figure 3. Rouge measurement chart

The evaluation results show that the fine-tuned ViT + Cahya GPT-2 multimodal model achieved an average ROUGE-L score of 0.61 (micro-average) on the test set. This score is an absolute increase from the initial baseline score of 0.42 obtained before fine-tuning with the same model configuration but without adaptation to the local culinary domain. The 0.19 ROUGE-L point increase indicates that the multimodal adaptation process to the Makassar culinary dataset has a significant impact on the quality of the text response output. Macro score analysis shows relatively uniform performance for each food class with small variations between categories. Foods with more complex visual characteristics such as Konro and Coto Makassar show slightly higher scores than sweet snack categories such as Barongko, which tend to have visual similarities to other banana-based foods [31]. This difference indicates that the visual complexity of objects affects the sensitivity of feature representation in Vision Transformer.

3. Result and Discussion

This study successfully implemented a Visual Question Answering (VQA) system based on the Vision Transformer (ViT) as a visual encoder and Cahya GPT-2 as an Indonesian language decoder within the Vision-Encoder–Decoder Transformer framework. The system was tested using a test set of 162 question–answer (QnA) pairs that were never used during the training and validation processes.

a) Model Performance

Model performance was evaluated using the ROUGE-L metric, which measures the correspondence between the longest sequence of system responses and a human reference response. The evaluation results showed that the fine-tuned model achieved an average (micro-average) ROUGE-L score of 0.61, an absolute increase from the initial baseline value of 0.42 before the multimodal fine-tuning process. The following table presents a comparison of model performance before and after the training process specifically for the Makassar culinary dataset.

Table 3. Comparison of model performance before and after the process

Configures Model	ROUGE-L
Baseline	0.42
ViT + Cahya GPT-2 (fine-tuned)	0.61

The increase of 0.19 ROUGE-L points reflects that the domain adaptation process specifically has a significant impact on the quality of the model output, especially on the suitability of food ingredient terminology, region of origin, and Indonesian sentence structure.

b) Culinary Class Analysis

Process: In addition to the aggregate evaluation, a macro-average approach was also used to evaluate food categories. The results showed relatively small performance variations between classes, indicating the model's stability in distinguishing food items with different visual characteristics. The Coto Makassar and Konro categories scored slightly higher than sweet snack categories like Barongko, which share a similar visual appearance to other banana dishes. This pattern suggests that the complexity of visual features, such as the presence of sauce, pieces of meat, and complementary elements, provides stronger attentional signals to the Vision Transformer, resulting in more informative visual representations.

c) Quality of Model Answers

Qualitative testing demonstrated that the system was capable of producing factual, concise, and relevant answers to image-based queries. In most samples, the model successfully identified the food type and main ingredients correctly. For example, in the Coto Makassar image, the model responded to the question "What are the main ingredients of this dish?" with the answer "Beef and offal," which fully matched the reference annotation. Furthermore, the model demonstrated the ability to combine visual information and Indonesian linguistic knowledge to answer questions about region of origin, such as providing the answer "Makassar, South Sulawesi" for the Pisang Epe image. This demonstrates the effectiveness of the cross-attention process between visual embeddings and text sequences in Cahya GPT2.

d) Comparison of Previous Studies and Implications.

The findings of this study align with international VQA studies showing that integrating a Vision Transformer-based visual encoder with a GPT-based language decoder can improve visual-text reasoning performance compared to unimodal approaches. Previous studies in non-local domains reported similar improvements when fine-tuning domain-specific datasets. Therefore, this study extends these findings to the context of Indonesian language and local cuisine, an area that has been underexplored previously.

From a practical perspective, the developed system has the potential to be applied as a visual culinary chatbot capable of automatically answering user questions about ingredients, dish types, and regions of origin. In addition to supporting digital tourism promotion, this technology can also be used as a medium for culinary culture education and documentation of regional food knowledge in an interactive format.

4. Conclusion

This research successfully developed a Visual Question Answering (VQA) system based on a multimodal transformer model that combines two different architectures: the Vision Transformer (ViT) for visual processing and Cahya-GPT2 for natural language processing in Indonesian. The primary objective of this research was to build a system capable of understanding visual information from images of Makassarese culinary specialties and transforming it into informative, relevant, and contextual text output. Through a targeted fine-tuning process, the system demonstrated significant improvements in its ability to understand the relationship between image and text modalities, while also demonstrating the effectiveness of integrating the transformer encoder-decoder architecture in a local research context

Technically, the test results showed that the fine-tuned combined ViT and Cahya-GPT2 model was able to produce answers with a better semantic match than the pre-trained model. The ROUGE-L metric increased from 0.42 to 0.61 after the fine-tuning process, indicating an improvement in the model's ability to understand the contextual meaning between words and sentences. The developed system is capable of recognizing various images of Makassarese specialties, such as Coto Makassar, Pisang Epe, Barongko, Jalangkote, Konro, and Panada, and providing appropriate descriptions and answers based on user queries. This improved performance demonstrates the effectiveness of a multimodal approach in bridging the gap between image and text processing in specific domains such as traditional culinary arts.

Methodologically, this study emphasizes the importance of a selective fine-tuning approach, where the learning process focuses on the language model (Cahya-GPT2) while the parameters of the visual model (ViT)

are largely frozen. This strategy not only streamlines training time and computing resources but also maintains ViT's visual generalization capabilities to new objects. This approach can serve as a reference for further research that combines models with different architectures in a single multimodal pipeline, particularly in the context of local Indonesian data, which is still limited in quantity and diversity.

From an application perspective, the results of this study make a significant contribution to the development of artificial intelligence systems in the fields of computer systems, computer vision, and natural language processing (NLP). The resulting system has the potential to be used in various applications, such as culinary chatbots, travel recommendation systems, and digital education platforms about traditional culture and food. Implementation of this system can also be part of efforts to preserve local culture through the digitization of regional culinary knowledge. Thus, this research is not only academically valuable but also has broad social impact in supporting culturally context-based digital transformation.

However, this study still has several limitations. First, the number and variety of Makassar culinary datasets used are relatively limited, so the model's generalization ability to images from different sources still needs to be improved. Second, the model has not been tested on complex or abstract questions that require deeper semantic reasoning. Third, the system still relies on an internet connection and a cloud-based environment for GPU computing, which at production scale can pose challenges to efficiency and data security. These limitations open up opportunities for further research to expand the dataset, increase model capacity, and optimize the architecture for on-device inference.

Conceptually, this research demonstrates that the integration of the Vision Transformer and a local Generative Pre-trained Transformer (GPT) such as Cahya-GPT2 can provide a strong foundation for developing multimodal artificial intelligence systems in Indonesia. This approach not only strengthens the machine's ability to understand visual and linguistic contexts but also provides a new direction for AI research that is more adaptive to Indonesia's social, cultural, and linguistic contexts. With this foundation, this research is expected to be an initial contribution to building an AI research ecosystem that is more oriented towards local, contextual, and socially impactful data.

Suggestion

Based on the results obtained, several development recommendations can be put forward. First, the dataset needs to be expanded with a larger number and variety of culinary imagery, involving various lighting conditions, shooting angles, and diverse backgrounds to improve the model's robustness to visual variations. Second, further research is recommended to explore other multimodal architectures such as CLIP (Contrastive Language–Image Pre-training) or BLIP (Bootstrapping Language–Image Pre-training), which have better contrastive learning capabilities in linking images and text. Third, testing new instruction-tuned language models such as IndoBERT, LLaMA-ID, or Gemma-2B-ID can also be conducted to improve semantic understanding of complex questions in Indonesian.

Furthermore, integrating the system with a more interactive web- or mobile-based user interface would be very helpful in testing the model's performance in real-world environments. This application development could be directed towards an interactive culinary chatbot system that not only answers user questions but also provides culinary recommendations, ingredient composition, and nutritional information. In an academic context, the results of this research can serve as the basis for curriculum development or further research projects in the field of Multimodal Artificial Intelligence, which combines computer vision, machine learning, and natural language processing to solve local data-driven problems.

With all its results, contributions, and development directions, this research confirms that innovation based on local multimodal transformer models can be a crucial pillar in strengthening Indonesia's AI ecosystem. Through a combination of technological power and cultural richness, such systems are expected to play an active role in encouraging cultural digitalization, supporting smart tourism, and expanding the role of AI as a means of education and promotion of Indonesia's culinary heritage in the future.

References

[1] R. Pakpahan, "Analisa Pengaruh Implementasi Artificial," *J. Inf. Syst. Informatics Comput.*, vol. 5, no. 2, pp. 506–513, 2021, doi: 10.52362/jisicom.v5i2.616.

[2] I. G. Made, W. Anditya, G. Ayu, and V. Mastrika, "Klasifikasi Kematangan Tomat pada Citra Digital Menggunakan DeiT (Data-efficient Image Transformer)," vol. 3, pp. 737–744, 2025, doi: https://doi.org/10.24843/JNATIA.2025.v03.i04.p03.

[3] S. J. Grace and D. Gunawan, "Perbandingan Cnn, Resnet50, Dan Vision Transformer Untuk Klasifikasi Kanker Payudara Berbasis Web," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 10, no. 2, pp. 945–956, 2025, doi: 10.36341/rabit.v10i2.6420.

[4] M. Farwati, I. Talitha Salsabila, K. Raihanun Navira, T. Sutabri, and U. Bina Darma Palembang, "Analisa Pengaruh Teknologi Artificial Intelligence (Ai) Dalam Kehidupan Sehari-Hari," *Jursima*, vol. 11, no. 1, pp. 39–45, 2023, doi: https://doi.org/10.47024/js.v11i1.563.

[5] V. No, "Klasifikasi Motif Batik Nusantara Menggunakan Vision Transformer (ViT) Berbasis Deep Learning Imam," *J. Inform. dan Teknol.*, vol. 8, no. 2, pp. 511–522, 2025, doi: https://doi.org/10.29408/jit.v8i2.31108.

[6] A. Riswanto and R. E. Rachmadi, "Artificial Intelegence Dalam Sistem Informasi Manajemen Dan Kinerja Berkelanjutan," *J. Lentera Bisnis*, vol. 12, no. 1, p. 124, 2023, doi: 10.34127/jrlab.v12i1.754.

[7] T. Febriyanto and S. Syofian, "Implementasi Deep Learning Menggunakan Vision Transformer Untuk Klasifikasi Penyakit Daun Padi," *J. TIFDA (Technology Inf. Data Anal.*, vol. 1, no. 2, pp. 34–39, 2024, doi: 10.70491/tifda.v1i2.47.

[8] R. E. Prasetya, M. A. Soeleman, F. Al Zami, A. Affandy, A. Marjuni, and M. I. S. Assaqty, "Enhancing Vision Transformer Performance with Rotation Based Augmentation for Classifying Images of Colon Cancer Pathology," *INTENSIF J. Ilm. Penelit. dan Penerapan Teknol. Sist. Inf.*, vol. 9, no. 2, pp. 235–249, 2025, doi: 10.29407/intensif.v9i2.24918.

[9] R. R. Ar, Agusriyati, and S. Moka, "Pendeteksian Dini Stunting Pada Balita Menggunakan Vision Transformer (VIT) Berbasis Citra Tubuh," *J. Inform. dan Tek. Elektro Terap.*, vol. 13, no. 3S1, pp. 896–902, 2025, doi: 10.23960/jitet.v13i3S1.7888.

[10] K. Umam, H. Khotimah, S. E. Purwanto, E. Azhar, A. Fatayan, and I. Nuriadin, "Augmented Reality dan Artificial Intelegence untuk Pemeblajaran dalam Persepektif Guru Matematika," *JagoMIPA J. Pendidik. Mat. dan IPA*, vol. 4, no. 2, pp. 273–279, 2024, doi: 10.53299/jagomipa.v4i2.595.

[11] Mohamad Asyqari Anugrah, Yaya Wihardi, and Rani Megasari, "Unsupervised Clustering of Handwritten Essay Answer Images Using Vision Transformer," *J. Komput. Teknol. Inf. Sist. Inf.*, vol. 4, no. 2, pp. 845–854, 2025, doi: 10.62712/juktisi.v4i2.517.

[12] K. A. Pradani and L. H. Suadaa, "Automated Essay Scoring Menggunakan Semantic Textual Similarity Berbasis Transformer Untuk Penilaian Ujian Esai," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 6, pp. 1177–1184, 2023, doi: 10.25126/jtiik.2023107338.

[13] K. Leonardi Kohsasih and J. Darwin, "Penerapan Algoritma Transformer Dalam Aplikasi Parafrase Teks Otomatis," *TAMIKA J. Tugas Akhir Manaj. Inform. Komputerisasi Akunt.*, vol. 5, no. 1, pp. 103–109, 2025, doi: 10.46880/tamika.Vol5No1.pp103-109.

[14] M. N. Achmadiah *et al.*, "Deteksi Kepadatan Penumpang di Stasiun Kereta Api Menggunakan Vision Transformer pada Jetson Orin Nano," *J. Elkolind*, vol. 12, no. 1, 2025, doi: DOI: http://dx.doi.org/10.33795/elkolind.v12i1/7495 153.

[15] A. H. Pradhana and E. Daniati, "Benchmarking Vision Transformer Klasifikasi Visual Masakan Padang Dengan Robustness Melalui Augmentasi Data," *Pros. Semin. Nas. Teknol. dan Sist. Inf.*, vol. 5, no. 1, pp. 152–164, 2025,

doi: 10.33005/sitasi.v5i1.2527.

[16] A. Indrasetianingsih, E. M. P. Hermanto, M. Ilham, N. Rahmawati, and I. A. Hariyanti, "Implementasi Model Deit Untuk Membedakan *Figure* Buatan Ai Dan Manusia Pada Ilustrasi Animasi 2D," *LPPM Nusa Mandiri*, vol. 19, no. 2, pp. 146–153, 2025, doi: https://doi.org/10.33480/inti.v19i2.6306 IMPLEMENTASI.

[17] Mas Nurul Achmadiah, Novendra Setiawan, and A. D. Risdhayanti, "Perbandingan Efisiensi Vision Transformer dan MobileNet untuk Optimasi Deteksi Objek di Edge Device," *J. Elektron. dan Otomasi Ind.*, vol. 12, no. 2, pp. 346–353, 2025, doi: 10.33795/elkolind.v12i2.8730.

[18] D. Rohim and M. Fauziah2, "Desain Dan Kelayakan Bahan Ajar Berbasis Contextual Learning Untuk Keterampilan Menulis Permulaan Di Sekolah Dasar," *Pendas J. Ilm. Pendidik. Dasar*, vol. 9, no. 2, p. 746, 2024, doi: 10.31004/abdidas.v5i5.1012.

[19] S. R. Arifin, "Sintesis Teks Ke *Figure*: Tinjauan Atas Dataset," *J. EEICT (Electric Electron. Instrum. Control Telecommun.*, vol. 7, no. 1, 2024, doi: 10.31602/eeict.v7i1.13066.

[20] N. W. S. Y. Ari Cahyani, N. N. Ganing, and I. K. A. Putra, "Pengaruh Model Pembelajaran Consept Sentence Berbantuan Media Audio Visual Terhadap Keterampilan Menulis Bahasa Indonesia," *J. Pedagog. dan Pembelajaran*, vol. 2, no. 2, p. 203, 2019, doi: 10.23887/jp2.v2i2.17909.

[21] Primanto and A. M. Simarmata, "Implementation of a Deep Learning Model Using Teachable Machine for Early Pneumonia Detection from X-Ray Images," *INOVTEK Polbeng - Seri Inform.*, vol. 10, no. 3, pp. 1444–1451, 2025, doi: 10.35314/7mjewe22.

[22] Muhammad Fakhri Fadhlurrahman, Munir, and Yaya Wihardi, "Pengenalan Ekspresi Wajah Peserta Didik di Ruang Kelas Menggunakan Vision Transformer (ViT)," *J. Komput. Teknol. Inf. Sist. Inf.*, vol. 4, no. 2, pp. 1047–1058, 2025, doi: 10.62712/juktisi.v4i2.531.

[23] A. N. Salsabila, M. Liebenlito, and D. U. Zulkifli, "Perbandingan Deteksi Alzheimer: ViT, CNN dan ViT dengan Bobot pada Citra Medis," *Indones. J. Comput. Sci.*, vol. 13, no. 1, pp. 1401–1412, 2024, doi: 10.33022/ijcs.v13i1.3765.

[24] Eva Putriany and Dhani Ariatmanto, "Literatur Reviu Sistematis: Identifikasi Jenis Ular Berbasis Computer Vision," *Jnanaloka*, p. 43, 2024, doi: 10.36802/jnanaloka.2024.v5-no01-43-50.

[25] Z. A. Annisa, R. S. Perdana, and P. P. Adikara, "Kombinasi Intent Classification dan Named Entity Recognition pada Data Berbahasa Indonesia dengan Metode Dual Intent and Entity Transformer," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 5, pp. 1017–1024, 2024, doi: 10.25126/jtiik.2024117985.

[26] M. S. A. Aria, C. Slamet, and M. D. Firdaus, "Klasifikasi Fake dan Real Menggunakan Vision Transformer dan EfficientNet-B0 pada *Figure* Asli dan Generatif AI," *Smatika J.*, vol. 15, no. 01, pp. 179–192, 2025, doi: 10.32664/smatika.v15i01.1531.

[27] Rismayani, S. Wahyuni, N. S. Layuk, R. H. Loly, and A. N. Daud, "Desain Sistem Speech Recognition Penerjemah Bahasa Toraja Menggunakan Hidden Markov Model," *J. Penelit. Pos dan Inform.*, vol. 11, no. 2, pp. 107–119, 2021, doi: 10.17933/jppi.v11i2.286.

[28] G. Ekayanda and M. Rahardi, "Analysis of Deep Learning Algorithms Using ConvNeXt and Vision Transformer for Brain Tumor Disease," *J. Inform. dan Komputasi Terap.*, vol. 9, no. 6, 2025, doi: https://doi.org/10.17933/jppi.v11i2.286.

[29] M. D. Noverta Effendi*1, Witrihan Ramadhani2, Fitri Farida3, "Jurnal Computer Science and Information Technology (CoSciTech) things," *J. Comput. Sci. Inf. Technol.*, vol. 5, no. 2, pp. 358–366, 2024, doi: https://doi.org/10.37859/coscitech.v6i2.10026.

[30] Y. P. Salsabil, F. L. Nisa, and Marseto, "Jurnal Pengabdian Masyarakat IPTEK Jurnal Pengabdian Masyarakat IPTEK," *J. Pengabdi. Masy. Dharma Andalas*, vol. 2, no. 2, pp. 62–66, 2022, doi:

https://doi.org/10.62335/mgjxst31.

[31] P. W. Cahyo and L. Sudarmana, “Klasterisasi Penjawab Berdasar Kualitas Jawaban pada Platform Brainly Menggunakan K-Means,” *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 11, no. 2, pp. 148–153, 2022, doi: 10.32736/sisfokom.v11i2.1314.