Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3241~3252
e-ISSN: 2548-6861

Indonesia–English Bilingual Visual Question Answering Using Partial Fine-Tuning on a ViT-GPT2 Architecture

Anas [1]*, Hazriani [2], Yuyun [3]***, Syamsul Rijal [4]**, Tirta Chiantalia Sharief [5]****
* Operations and Service Control (POP), Regional VI, PT Pos Indonesia
** Computer Systems, Handayani University Makassar
*** National Research and Innovation Agency (BRIN), Republic of Indonesia
anas.daeng.pasewang@gmail.com[1], hazriani@handayani.ac.id [2], yuyu010@brin.go.id [3], syamsulr8@gmail.com [4], tirtashariff@gmail.com [5]

Article Info

Article history:

Received 2026-06-13
Revised 2026-07-02
Accepted 2026-08-07

Keyword:

Visual Question Answering,
Vision-Language Model,
Vision Transformer,
GPT-2,
Fine-Tuning.

ABSTRACT

Visual Question Answering (VQA) is a multimodal task that integrates visual understanding and natural language processing to generate answers based on information contained in an image. Most existing VQA research focuses on the English language and general-domain datasets, limiting its applicability to bilingual environments and domain-specific scenarios. This study proposes an Indonesia–English bilingual VQA model based on the VisionEncoderDecoderModel architecture, which combines a Vision Transformer (ViT) as the visual encoder and an Indonesian GPT-2 model as the language decoder. Bilingual capability is achieved through the introduction of special language tokens, <id> and <en>. The model is trained using a combination of the bilingual VQAv2 and bilingual LosariVQAv1 datasets, representing general-domain and local tourism-domain knowledge, respectively. Four fine-tuning strategies are evaluated: Encoder Freeze, Full Fine-Tuning, Partial-4, and Partial-6. Experimental results show that the Partial-6 strategy achieves the best performance on LosariVQAv1, obtaining an Exact Match score of 26.67%, a BLEU score of 26.80%, and a CIDEr score of 292.67, while maintaining competitive performance on VQAv2 with an Exact Match score of 41.26%. Cross-language evaluation reveals only a small performance gap between Indonesian and English. The findings indicate that partial fine-tuning provides a better balance between generalization capability and domain adaptation than the other fine-tuning strategies evaluated in this study.

I. INTRODUCTION

Visual Question Answering (VQA) is a multimodal task that integrates visual understanding and natural language processing to generate answers based on information contained in an image. This task requires a model not only to recognize visual objects but also to understand the context of a question and the semantic relationships between visual and textual information. The advancement of Transformer architectures has led to significant progress in various multimodal tasks, including image captioning, visual reasoning, and Visual Question Answering [1]. The ability of Transformers to model long-range dependencies through the self-attention mechanism has made them the foundation of many modern Vision-Language Models (VLMs).

In recent years, numerous Vision-Language Models have been developed to enhance multimodal understanding capabilities. Models such as LXMERT [7], ViLT [8], BLIP [9], OFA [10], BLIP-2 [11], InstructBLIP [12], LLaVA [13], and Qwen-VL [14] have demonstrated that integrating visual and textual representations can significantly improve performance across various multimodal tasks. These developments indicate a paradigm shift from task-specific models toward more general-purpose multimodal models

capable of handling multiple tasks within a unified architecture [15], [16].

Despite the rapid advancement of Vision-Language Models, most Visual Question Answering research remains focused on the English language due to the widespread use of benchmark datasets such as VQAv2 [4], GQA, and Visual Genome, which predominantly contain English question–answer pairs. This situation limits the applicability of VQA technology in multilingual environments, particularly for the Indonesian language. In practice, Indonesian users primarily interact in Bahasa Indonesia, making reliance on English a significant barrier to the adoption of VQA systems in various artificial intelligence applications.

In addition to language-related challenges, most VQA studies utilize general-domain datasets dominated by everyday objects such as people, vehicles, animals, and household items [4]. While these datasets are effective for evaluating model generalization, the learned visual representations may not adequately capture the characteristics of specific domains. Consequently, models that perform well on general benchmarks may not achieve similar performance when applied to domain-specific environments with distinct visual characteristics.

One domain that could significantly benefit from VQA technology is tourism. Indonesia possesses a wide variety of tourist destinations rich in visual and contextual information; however, the application of Vision-Language Models to support tourism information services remains limited. VQA systems can function as digital tourism assistants capable of answering tourists' questions based on photographs of tourist attractions captured in real time. Nevertheless, research on VQA for local Indonesian tourism domains remains scarce in the existing literature.

To address these challenges, this study develops an Indonesia–English bilingual Visual Question Answering model using the VisionEncoderDecoderModel architecture, which combines a Vision Transformer (ViT) [2] as the visual encoder and an Indonesian GPT-2 model [3] as the language decoder. Unlike most previous VQA studies that focus on a single language, this research introduces special language tokens, <id> and <en>, to control the output language of the model. This approach enables a single model to generate answers in either Indonesian or English without requiring separate models for each language.

In addition to supporting bilingual capabilities, this study integrates two datasets with distinct characteristics. The first dataset is a bilingual version of VQAv2, obtained by translating the original VQAv2 dataset using the multilingual machine translation model mBART-50 [17]. The second dataset is LosariVQAv1, a local tourism-domain Visual Question Answering dataset developed in this research using images of tourist attractions located at Losari Beach in Makassar City. The combination of these datasets enables the model to learn both general visual knowledge and domain-specific visual characteristics related to local tourism.

This study also focuses on analyzing fine-tuning strategies for Vision-Language Models. Previous studies have shown that fine-tuning strategies significantly influence model performance when adapting to new domains [9], [11]. Therefore, four training strategies are evaluated: Encoder Freeze, Full Fine-Tuning, Partial Fine-Tuning of the last four encoder layers (Partial-4), and Partial Fine-Tuning of the last six encoder layers (Partial-6). The analysis aims to identify the configuration that provides the best balance between generalization capability on general-domain data and adaptation capability to the local tourism domain.

Based on the aforementioned background, this study seeks to answer three primary research questions. First, how can an Indonesia–English bilingual Vision-Language Model be developed using a combination of ViT and Indonesian GPT-2 within the VisionEncoderDecoderModel architecture? Second, what is the impact of integrating general-domain and local tourism-domain datasets on Visual Question Answering performance? Third, which fine-tuning strategy provides the best balance between generalization and domain adaptation in a bilingual environment?

The main contributions of this study are threefold. First, this study develops an Indonesia–English bilingual Visual Question Answering model based on the VisionEncoderDecoderModel architecture by integrating a Vision Transformer as the visual encoder and an Indonesian GPT-2 model as the language decoder, enabling bilingual answer generation within a single unified framework through the introduction of special language tokens. Second, this work introduces LosariVQAv1, a bilingual Visual Question Answering dataset representing the Indonesian local tourism domain, which complements the general-domain bilingual VQAv2 dataset and supports multi-domain learning. Third, this study provides a comprehensive comparative analysis of four fine-tuning strategies—Encoder Freeze, Full Fine-Tuning, Partial-4, and Partial-6—to investigate the trade-off between preserving pretrained visual knowledge and adapting the model to a domain-specific bilingual environment. These contributions distinguish the proposed approach from previous Vision-Language Model studies that primarily focus on English-only datasets and general-domain applications.

Compared with previous studies that mainly evaluate Vision-Language Models on English benchmark datasets, this work emphasizes bilingual capability, adaptation to an Indonesian local tourism domain, and systematic evaluation of partial fine-tuning strategies within a unified VisionEncoderDecoderModel architecture. Therefore, the novelty of this research lies not only in the proposed bilingual framework but also in the integration of a newly developed local dataset and the comprehensive analysis of fine-tuning strategies for multilingual and multi-domain Visual Question Answering.

The findings of this study are expected to contribute to the advancement of Vision-Language Models for the Indonesian language and serve as a foundation for developing Visual Question Answering systems applicable to tourism and other

multimodal domains. Furthermore, this research is expected to provide a useful reference for developing bilingual VQA models for low-resource languages through transfer learning and fine-tuning approaches based on modern Vision-Language Models.

II. METHODS

This study develops an Indonesia–English bilingual Visual Question Answering (VQA) model using the VisionEncoderDecoderModel architecture, which integrates a Vision Transformer (ViT) as the visual encoder and an Indonesian GPT-2 model as the language decoder. The proposed approach enables the model to understand visual information and generate answers in two languages within a unified architecture. The research methodology consists of bilingual model construction, dataset preparation, fine-tuning strategy design, model training, and performance evaluation.

A. Transformer

Transformer is a deep learning architecture introduced by Vaswani et al. [1] and has become the foundation of many modern models in Natural Language Processing (NLP), Computer Vision (CV), and Vision-Language Models (VLMs). Unlike Recurrent Neural Networks (RNNs), which process data sequentially, Transformers employ a self-attention mechanism that enables the model to learn relationships among data elements in parallel, resulting in higher computational efficiency.

The core component of the Transformer architecture is the Scaled Dot-Product Attention mechanism, which is formulated as:

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

dengan:

- Q= Query matrix,
- K= Key matrix,
- V= Value matrix,
- d_k= dimensi vektor Key.

To enhance representational capacity, the Transformer employs a Multi-Head Attention mechanism that enables the model to learn multiple relationships simultaneously:

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O$$

dengan:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

The Transformer architecture serves as the backbone of many state-of-the-art models, such as BERT, GPT, ViT, BLIP, and Qwen-VL [1], [15], [16]. In the proposed model, Transformer architectures are utilized in both the visual encoder and the language decoder, facilitating integrated multimodal representation learning.

B. Vision Transformer (ViT)

The visual encoder used in this study is ViT-Base Patch16-224 (google/vit-base-patch16-224-in21k), which was introduced by Dosovitskiy et al. [2]. Vision Transformer (ViT) is an adaptation of the Transformer architecture for image processing tasks, replacing conventional convolutional operations with a self-attention mechanism.

The ViT model employed in this research has the specifications presented in Table I.

TABLE I
VISION TRANSFORMER SPECIFICATIONS

Parameter	Value
Model	ViT-Base Patch16-224
Number of Parameters	±86 Million
Hidden Size	768
Transformer Layers	12
Attention Heads	12
Patch Size	16×16
Input Resolution	224×224
Pretraining Dataset	ImageNet-21K

The model was pretrained on the ImageNet-21K dataset, which contains more than 14 million images spanning approximately 21,000 object categories [2].

Let the input image be represented as:

$$I \in \mathbb{R}^{H \times W \times C}$$

Given a patch size of $P \times P$, the number of visual tokens generated is calculated as:

$$N = \frac{HW}{P^2}$$

Each patch is subsequently projected into an embedding vector and processed by the Transformer encoder to generate a global visual representation. This representation is then passed to the decoder through a cross-attention mechanism.

ViT was selected due to its ability to learn powerful visual representations across a wide range of vision-language tasks, as well as its compatibility with the Transformer-based architecture employed by the decoder [2], [9], [11].

C. Indonesian GPT-2

The language decoder employed in this study is the indonesian-nlp/gpt2 model, which is based on the GPT-2 architecture introduced by Radford et al. [3]. GPT-2 is an autoregressive language model designed to predict the next token based on all previously generated tokens.

The specifications of the GPT-2 model used in this study are presented in Table II.

TABLE II
INDONESIAN GPT-2 SPECIFICATIONS

Parameter	Value
Architecture	GPT-2 Small
Number of Parameters	±124 Million
Transformer Layers	12
Hidden Size	768
Attention Heads	12
Initial Vocabulary Size	50,257
Language	Indonesian

The model was pretrained on a large-scale Indonesian corpus comprising news articles, web documents, and diverse general-purpose text sources. Consequently, it is capable of learning rich linguistic representations that reflect the syntactic and semantic characteristics of the Indonesian language.

The probability of a token sequence in GPT-2 is defined as:

$$P(x) = \prod_{t=1}^{T} P(x_t \mid x_{<t})$$

where:

- x_t represents the token generated at time step t
- $(x_1, x_2, \ldots, x_{t-1})$ represents all previously generated tokens.
- T denotes the length of the token sequence.

To enable GPT-2 to function as a multimodal decoder, the model configuration was modified by activating the cross-attention mechanism through the parameters config_decoder.is_decoder = True and config_decoder.add_cross_attention = True. This configuration allows the decoder to receive visual representations from the ViT encoder during the answer generation process.

D. Bilingual VisionEncoderDecoderModel Architecture

The proposed model is built using the VisionEncoderDecoderModel framework, which combines a Vision Transformer (ViT-Base Patch16-224) as the visual encoder and an Indonesian GPT-2 model as the language decoder in a unified architecture.

The input image is resized to 224 × 224 pixels and normalized using the ImageNet mean and standard deviation before being processed by the ViT encoder. The encoder divides the image into 16 × 16 patches and converts them into visual embeddings with a hidden dimension of 768. These visual embeddings are then passed to the decoder through the cross-attention mechanism.

The Indonesian GPT-2 decoder also uses a hidden dimension of 768, with 12 Transformer layers and 12 attention heads. Since both the encoder and decoder have the same hidden dimension, the visual embeddings produced by the encoder can be directly utilized by the decoder without requiring additional projection layers.

To support bilingual answer generation, each question is prefixed with a special language token, <id> for Indonesian or <en> for English. These tokens are added to the tokenizer vocabulary, and the decoder embedding layer is resized accordingly. During training, Indonesian and English samples are combined in the same training process (joint training), allowing a single model to generate answers in both languages based on the language token.

Unlike the standard VisionEncoderDecoderModel, the proposed model extends the decoder by introducing bilingual language tokens while maintaining a single encoder–decoder architecture. The overall architecture and information flow of the proposed model are illustrated in Figure 1.

TABLE III
SPECIAL TOKENS USED IN THE TOKENIZER

Token	Function
<id>	Indonesian Language Identifier
<en>	English Language Identifier
<bos>	Beginning of Sequence
<pad>	Padding Token

The bilingual token mechanism was implemented using the following code:

```
tokenizer.add_special_tokens({
  "additional_special_tokens": ["<id>", "<en>"],
  "bos_token": "<bos>",
  "pad_token": "<pad>"
})
```

Following the introduction of the special tokens, the vocabulary size was expanded, requiring the GPT-2 decoder's token embedding layer to be resized accordingly. The overall architecture and processing workflow of the proposed model are conceptually illustrated in Figure 1.

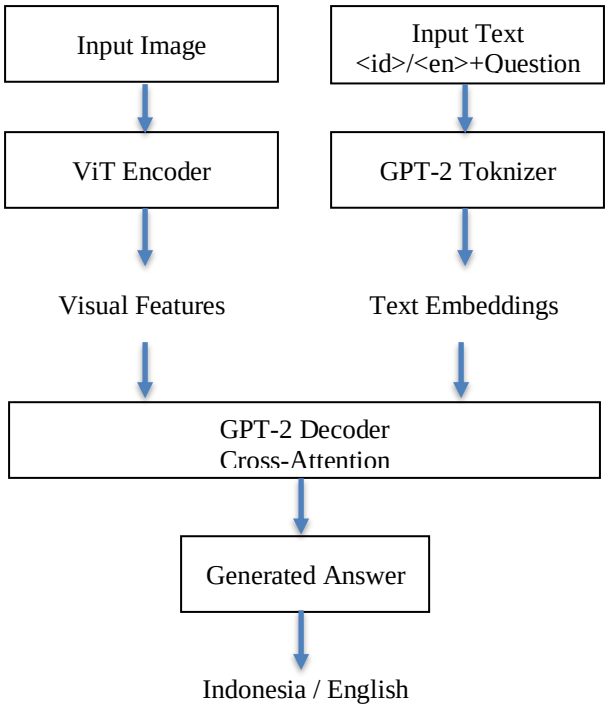

Figure 1. ViT-GPT2 Architecture with Cross-Attention for Bilingual Visual Question Answering

E. Datasets

This study utilizes two datasets with distinct characteristics: a general-domain dataset and a local tourism-domain dataset.

1) *Bilingual VQAv2*: The VQAv2 dataset [4] was used to represent the general domain. The original English question–answer pairs were translated into Indonesian using

the mBART-50 multilingual machine translation model [17], resulting in a bilingual Indonesia–English dataset. Consequently, each image is associated with corresponding Indonesian and English question–answer pairs, enabling balanced bilingual training while preserving the original visual content. The translated question–answer pairs were verified to ensure consistency between the Indonesian and English versions before being used for model training.

TABLE IV
STATISTICS OF THE BILINGUAL VQAV2 DATASET

Subset	Sampels	Number of Images
Train	200,004	18,661
Validation	10,002	883
Test	10,006	973

2) Bilingual LosariVQAv1: LosariVQAv1 is a bilingual tourism-domain Visual Question Answering dataset developed in this study using photographs collected from several tourist attractions around Losari Beach, Makassar, Indonesia. The dataset was designed to complement the general-domain VQAv2 dataset by providing images and question–answer pairs representing local tourism objects, including landmarks, monuments, public facilities, traditional architecture, environmental conditions, and cultural attractions.

The annotation process was conducted using Label Studio, where each image was manually annotated to generate question–answer pairs in Indonesian. The Indonesian annotations were subsequently translated into English using the mBART-5**0** multilingual machine translation model to create bilingual question–answer pairs. Each image therefore has corresponding Indonesian and English annotations, enabling balanced bilingual training within a single model. The bilingual annotations were verified to ensure consistency between the Indonesian and English versions before being used for model training.

TABLE V
STATISTICS OF THE LOSARIVQAV1 DATASET

Subset	Samples	Number of Images
Train	19,140	365
Validation	2,402	61
Test	2,310	45

Overall, the dataset contains 19,140 training samples, 2,402 validation samples, and 2,310 test samples obtained from 471 images (365 training, 61 validation, and 45 test images). The bilingual LosariVQAv1 dataset is publicly available to support reproducibility and future research on Indonesian Vision-Language Models.

The bilingual LosariVQAv1 dataset is publicly available for research purposes at the following repository: https://drive.google.com/drive/folders/1aa78KidB68HuinEZeJ7lD_S4npNwH4C_?usp=sharing

F. Fine-Tuning Strategies

This study evaluates four fine-tuning strategies to investigate the trade-off between preserving pretrained visual knowledge and adapting the model to the local tourism domain. Four configurations were evaluated:

*1) Encoder Freeze***:** All parameters of the ViT encoder are frozen, while all GPT-2 decoder parameters are updated during training. This strategy evaluates the transferability of pretrained visual features without modifying the visual encoder.

2) Full Fine-Tuning: All parameters of both the ViT encoder and GPT-2 decoder are updated during training. Although this strategy provides maximum model adaptability, it may increase the risk of catastrophic forgetting.

3) Partial-4: Only the last four encoder layers (Layers 9–12) are unfrozen, while the remaining encoder layers remain frozen. The GPT-2 decoder is fully trainable.

4) Partial-6: Only the last six encoder layers (Layers 7–12) are unfrozen, while the first six encoder layers remain frozen. The GPT-2 decoder is fully trainable. This strategy is expected to provide a better balance between preserving pretrained visual representations and learning domain-specific features.

TABLE VI

FINE-TUNING CONFIGURATIONS

Strategy	ViT Encoder	GPT-2 Decoder	Objective
Encoder Freeze	All layers frozen	Fully trainable	Evaluate pretrained visual features
Full Fine-Tuning	Layers 1–12	Fully trainable	Maximum adaptation
Partial-4	Layers 9–12	Fully trainable	Moderate domain adaptation
Partial-6	Layers 7–12	Fully trainable	Balance generalization and adaptation

These strategies were selected because previous studies have shown that updating only the higher layers of the vision encoder is often sufficient to adapt pretrained visual representations to a new domain while reducing the risk of catastrophic forgetting, particularly when the target dataset is relatively small [9], [11].

G. Training Configuration

The model was trained using the Hugging Face Transformers framework on the Google Colab Pro environment equipped with NVIDIA GPUs. Before being processed by the Vision Transformer, all input images were resized to 224 × 224 pixels and normalized using the ImageNet mean and standard deviation. These preprocessing steps are consistent with the pretrained ViT-Base Patch16-224 model.

TABLE VII
TRAINING CONFIGURATION

Parameter	Value
Epoch	3
Batch Size	4
Gradient Accumulation	4
Optimizer	AdamW
Scheduler	Cosine
Warmup Steps	1,000
Beam Size	3
Max New Tokens	24
Weight Decay	0.01
Decoder Learning Rate	3×10^{-4}
Encoder Learning Rate	1×10^{-5} (default); 5×10^{-5} (Partial-6)

Table VII summarizes the training configuration used in all experiments. The model was trained for three epochs using the AdamW optimizer with a cosine learning rate scheduler and 1,000 warm-up steps. A batch size of four with gradient accumulation of four was employed to overcome GPU memory limitations. Different learning rates were assigned to the encoder and decoder because the decoder required greater adaptation to the bilingual VQA task, whereas the pretrained ViT encoder already contained rich visual representations.

During inference, beam search with a beam size of three was used to improve answer generation quality. The maximum number of generated tokens was limited to 24, which was sufficient for the short-answer characteristics of the VQA task while avoiding unnecessarily long responses. The same training configuration was applied to all fine-tuning strategies to ensure a fair comparison, except for the encoder learning rate adjustment evaluated in the Partial-6 experiment.

H. Evaluation Metrics

Before evaluation, the output token IDs generated by the GPT-2 decoder were converted back into text using the tokenizer's decoding function (detokenization). Special tokens such as <id>, <en>, <bos>, <pad>, and other padding tokens were removed before computing the evaluation metrics. The decoded answers were then compared with the corresponding reference answers using Exact Match (EM), BLEU, and CIDEr. All evaluation metrics were computed on the final decoded text rather than on token IDs.

The same evaluation procedure was applied to both Indonesian and English outputs to ensure a fair comparison between the two languages.

Model performance was evaluated using three widely adopted metrics for generative tasks and Vision-Language Models.

1) *Exact Match (EM)*: Exact Match measures the proportion of predicted answers that exactly match the reference answers.

$$EM = \frac{1}{N}\sum_{i=1}^{N} \delta(y_i, \hat{y}_i)$$

where:

$$\delta(y_i, \hat{y}_i) = \begin{cases} 1, & y_i = \hat{y}_i \\ 0, & other \end{cases}$$

This metric is commonly used in Visual Question Answering research [4].

2) *BLEU*: *Bilingual Evaluation Understudy* introduced by Papineni et al. [5], measures the similarity of n-grams between the predicted answer and the reference answer.

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where BP *denotes the brevity penalty and* p_n denotes the modified n-gram precision.

3) *CIDEr*: *Consensus-based Image Description Evaluation*, introduced by Vedantam et al. [6], is widely used in image captioning and Vision-Language Model tasks.

CIDEr employs TF-IDF weighting over n-grams to measure the semantic similarity between predicted and reference answers:

$$CIDEr(c_i, S_i) = \frac{1}{m}\sum_{j=1}^{m}\sum_{n=1}^{4} w_n \frac{g_n(c_i) \cdot g_n(s_{ij})}{\| g_n(c_i) \| \| g_n(s_{ij}) \|}$$

This metric was selected because it is more sensitive to semantic information quality than BLEU [6].

Based on the proposed methodology, the next stage is to evaluate the different fine-tuning strategies in order to identify the configuration that provides the best balance between generalization capability on the general domain and adaptation capability to the local tourism domain.

III. RESULTS AND DISCUSSION

This section presents the evaluation results of the proposed Indonesia–English bilingual Visual Question Answering (VQA) model built using a VisionEncoderDecoderModel architecture that integrates a Vision Transformer (ViT) and an Indonesian GPT-2 model. The evaluation was conducted to assess the impact of different fine-tuning strategies on the model's ability to answer questions in both general-domain and local tourism-domain scenarios. In addition to quantitative evaluation using Exact Match (EM), BLEU, and CIDEr metrics, this study also provides cross-language analysis, training convergence analysis, and qualitative analysis to examine the characteristics and limitations of the proposed model.

A. Comparative Analysis of Fine-Tuning Strategies

Experiments were conducted to evaluate the impact of different fine-tuning strategies on model performance using the LosariVQAv1 dataset as a representative local tourism domain. Four configurations were evaluated: Encoder Freeze, Full Fine-Tuning, Partial Fine-Tuning of the last four encoder layers (Partial-4), and Partial Fine-Tuning of the last six encoder layers (Partial-6).

TABLE VIII
MODEL PERFORMANCE COMPARISON ON LOSARIVQAV1

Strategi	EM (%)	BLEU (%)	CIDEr
Freeze	23.07	24.97	265.67
Full Fine-Tuning	24.76	25.77	279.64
Partial-4	24.33	25.42	275.44
Partial-6	**26.67**	**26.80**	**292.67**

As shown in Table VIII, the Partial-6 strategy achieved the best performance across all evaluation metrics. Compared with the Freeze configuration as the baseline, Exact Match improved by 15.61%, BLEU improved by 7.33%, and CIDEr improved by 10.16%. These results indicate that partially unfreezing the encoder provides better adaptation to the visual characteristics of the local tourism domain.

Interestingly, Full Fine-Tuning did not outperform Partial-6 despite updating all encoder parameters during training. This finding suggests that updating the entire encoder may lead to the loss of some general visual representations acquired during pretraining, a phenomenon commonly referred to as catastrophic forgetting. In contrast, Partial-6 preserves most of the pretrained visual knowledge while providing sufficient flexibility to learn new domain-specific visual characteristics.

These findings are consistent with previous transfer learning studies on Vision-Language Models, which have shown that partial fine-tuning is often more effective than updating all parameters when the target dataset is relatively limited in size [9], [11], [20].

Although the Partial-6 strategy achieved the best overall performance, the Exact Match score on LosariVQAv1 remained relatively modest (26.67%). This result reflects the challenging nature of generative Visual Question Answering, where semantically correct answers may still be considered incorrect because Exact Match requires an identical textual response. Therefore, the Exact Match score should be interpreted together with BLEU and CIDEr, which also consider lexical overlap and semantic similarity.

To evaluate model generalization on a general domain, the same experiment was conducted using the bilingual VQAv2 dataset.

TABLE IX
MODEL PERFORMANCE COMPARISON ON VQAV2

Strategi	EM (%)	BLEU (%)	CIDEr
Freeze	40.26	11.51	108.63
Full Fine-Tuning	40.91	15.30	110.81
Partial-4	40.46	11.72	109.64
Partial-6	40.72	12.41	110.28
Partial-6 (LR Encoder = 5×10^{-5})	**41.26**	**16.49**	**111.46**

Unlike LosariVQAv1, performance differences among the evaluated strategies on VQAv2 were relatively small. This indicates that all configurations were able to maintain generalization capability on the general domain. Nevertheless, the Partial-6 configuration with an encoder learning rate of 5×10^{-5} achieved the best performance across all evaluation metrics.

These results suggest that increasing the encoder learning rate provides additional adaptation capability without degrading performance on the general domain. Therefore, Partial-6 can be considered the most balanced strategy for preserving general visual knowledge while enhancing domain adaptation capability.

B. Cross-Language Analysis and Domain Adaptation

One of the primary objectives of this study is to develop a model capable of generating answers in both Indonesian and English using a single architecture. Therefore, evaluations were conducted separately for each language using the best-performing configuration, namely Partial-6. For the cross-language evaluation, the Indonesian and English test sets were evaluated separately using the same trained bilingual model. Each evaluation sample consisted of the same image paired with language-specific question–answer pairs. Since both language versions were derived from identical visual content, only the textual question and answer differed, allowing a fair comparison of bilingual performance without introducing additional visual variations.

TABLE X
MODEL PERFORMANCE BY LANGUAGE

Dataset	Indonesia EM (%)	Inggris EM (%)
LosariVQAv1	27.27	26.06
VQAv2	41.64	39.80

Based on Table X, the performance gap between Indonesian and English is relatively small. On LosariVQAv1, the Exact Match difference is only 1.21 percentage points, while on VQAv2 the difference is 1.84 percentage points. These results indicate that the language tokens <id> and <en> effectively guide the decoder in generating responses in the intended language.

The relatively small performance gap suggests that the proposed language-token mechanism effectively enables bilingual answer generation within a single decoder. Most prediction errors were observed consistently in both languages, indicating that the primary limitations originated from visual reasoning rather than language translation. In several cases, English predictions generated semantically correct answers using alternative wording, which reduced the Exact Match score despite preserving the intended meaning. Therefore, lexical variation contributed to part of the observed performance difference between Indonesian and English.

This bilingual capability represents one of the main contributions of this study, as it enables a single model to support two languages without requiring a separate translation process during inference. From a deployment perspective, this approach is more efficient than translation-based pipelines because it reduces both system complexity and processing time.

When comparing domains, performance on VQAv2 remains higher than that on LosariVQAv1. This difference can be attributed to the substantially larger dataset size and greater visual diversity of VQAv2. The VQAv2 dataset contains more than 200,000 training samples, whereas LosariVQAv1 contains approximately 19,000 question–answer pairs.

Nevertheless, the observed performance improvement on LosariVQAv1 demonstrates that the fine-tuning process successfully transferred general visual knowledge into domain-specific knowledge related to the Losari Beach tourism environment. These findings suggest that integrating general-domain and domain-specific datasets is an effective strategy for developing multi-domain Vision-Language Models.

C. Interpretation of Training and Validation Loss Curves

To evaluate the stability of the training process, the progression of training loss and validation loss was monitored throughout the fine-tuning process using the best-performing configuration, namely Partial-6.

TABLE XI
TRAINING AND VALIDATION LOSS PROGRESSION

Step	Training Loss	Validation Loss
2500	7.45	1.47
5000	5.69	1.39
7500	5.40	1.34
10000	5.24	1.32
12500	5.11	1.30
15000	4.72	1.30
17500	4.43	1.29
20000	4.38	1.28
22500	4.35	1.26
25000	4.29	1.26

As shown in Table XI, the training loss decreased consistently throughout the training process, from 7.45 at step 2,500 to 4.29 at step 25,000. This 42.42% reduction indicates that the model gradually learned the relationship between the visual features extracted by the encoder and the textual representations utilized by the decoder to generate answers. The steady decline in training loss suggests that the optimization process proceeded effectively without significant training instability.

At the same time, the validation loss also decreased from 1.47 to 1.26. Although the reduction was smaller than that observed for the training loss, the downward trend indicates that the performance improvement was not limited to the training data but also generalized to the validation data that were not used during model learning. This finding suggests that the model was able to generalize reasonably well to unseen samples.

Furthermore, the rate of validation loss reduction began to slow after step 12,500. Between steps 12,500 and 25,000, the validation loss decreased by only 0.04. This behavior indicates that the model was approaching convergence and that subsequent iterations yielded progressively smaller performance gains. Therefore, the number of training steps employed in this study can be considered sufficient to achieve stable model performance.

To provide a clearer illustration of the training dynamics, the progression of training loss and validation loss is visualized in Figure 2.

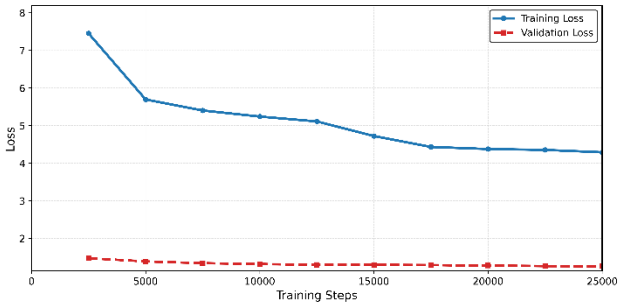

Figure 2. Training and Validation Loss Curves of the Partial-6 Model

Based on Figure 2, no significant increase in validation loss was observed throughout the training process. This indicates that the model did not exhibit severe overfitting. Moreover, the gap between the training loss and validation loss remained relatively stable, suggesting that the fine-tuning process converged effectively and resulted in good generalization capability. These findings are consistent with the results presented in Tables VIII and IX, which demonstrate that the Partial-6 configuration achieved the best performance across both evaluation datasets.

Overall, the combination of Partial Fine-Tuning, the Cosine Learning Rate Scheduler, the Warmup Strategy, and different learning rates for the encoder and decoder resulted in a stable, convergent, and effective optimization process for developing an Indonesia–English bilingual Vision-Language Model.

D. Qualitative Analysis and Error Analysis

In addition to quantitative evaluation, qualitative analysis was conducted to better understand the characteristics of the answers generated by the model and to identify the types of errors that still occur.

1) Analysis of Correct Predictions: Table XII presents several examples of correctly predicted answers generated by the model across various question categories, including object identification, visual attribute recognition, location recognition, and object counting.

TABLE XII
EXAMPLES OF CORRECTLY PREDICTED ANSWERS GENERATED BY THE MODEL

Image	Question (Ground Truth)	Prediction
	Apa yang terlihat dalam gambar ini? GT: Masjid Amirul Mukminin	Masjid Amirul Mukminin

Image	Question (Ground Truth)	Prediction
	Berapa jumlah menaranya? GT: Dua	Dua menara tinggi.
	What is seen in this picture? GT: Amirul Mukminin Mosque	Amirul Mukminin Mosque
	Apa yang nampak dalam gambar ini? GT: Patung Kapal Phinisi	Patung Kapal Phinisi
	How's the weather? GT: Bright	It appears sunny with a clear sky.
	What do you see in this picture? GT: Phinisi Ship Statue	Kapal Phinisi Statue
	Bangunan apa ini? GT: Rumah tongkonan	Rumah tongkonan
	What is this building? GT: Tongkonan house	A traditional tongkonan house
	Di mana lokasi gambar ini? GT: Pantai Losari	Di area Anjungan Bugis di Pantai Losari
	Where is this image located? GT: Losari Beach	It is taken at the Bugis area of Losari Beach in Makassar
	Warna apa tulisannya? GT: Merah	Tulisan tersebut berwarna merah.
	Ada berapa patung orang? GT: Tiga patung	Tiga patung
	How many statues of people are there? GT: Three statues	Three statures.
	What color is the sky? GT: Blue	Blue.

Based on Table XII, the model successfully answered questions related to object identification, visual attributes, color recognition, and tourist location recognition with a high degree of accuracy. These results indicate that the visual representations generated by the ViT encoder were effectively utilized by the GPT-2 decoder to produce appropriate answers in both Indonesian and English.

The model's ability to recognize specific tourist locations further demonstrates that fine-tuning on the LosariVQAv1 dataset enhanced its understanding of local tourism-related objects and environments. In addition, the model was capable of generating semantically correct answer variations, as illustrated by the prediction of *"A traditional tongkonan house"* for the reference answer *"Tongkonan house"* and the identification of Losari Beach using a more descriptive response. These examples suggest that the model integrates both visual information extracted by the ViT encoder and linguistic knowledge learned by the GPT-2 decoder to generate contextually appropriate answers. Nevertheless, some incorrect predictions indicate that the decoder may still rely on linguistic priors when visual evidence is ambiguous or insufficient.

2) Error Analysis: Table XIII presents several examples of predictions that do not match the ground-truth answers. These cases were analyzed to identify the limitations of the proposed model in visual understanding and image-based reasoning.

TABLE XIII
EXAMPLES OF INCORRECT MODEL PREDICTIONS

Image	Question (Ground Truth)	Prediction
	Berapa jumlah patung komodo yang atas di atas bola dunia? GT: Satu	Tiga patung
	How many Komodo dragon statues are there on top of the globe? GT: One	Two visible near the top and right side
	Menghadap kemana patung tersebut? GT: Menghadap ke laut	Menghadap ke arah kamera.
	Which way is the statue facing? GT: Facing the sea	Facing forward toward the camera.
	Tulisan besar apa yang ada pada gambar? GT: BUGIS	Tulisan 'CITY OF MAKASSAR'.
	Apa yang sedang dilakukan patung tersebut? GT: Menari	Menari di atas sebuah bola besar
	What is the statue doing? GT: Dancing	It appears to be sitting on a large pedestal
	Apakah ini patung kendaraan tradisional? GT: Ya, becak tradisional	Ya, terlihat seperti becak tradisional

Based on Table XIII, the observed errors can be categorized into three main types. First, Counting Errors, which occur when the model incorrectly estimates the number

of objects present in an image, as demonstrated by the Komodo statue counting example. Second, Spatial Reasoning Errors, which involve incorrect interpretation of spatial relationships or object orientations, such as determining the direction in which a statue is facing. Third, Activity Recognition Errors, which arise when the model misidentifies actions or activities performed by an object, such as confusing a dancing pose with a stationary position.

The observed errors also indicate several limitations of the proposed model. The relatively small size of the LosariVQAv1 dataset, together with the complexity of free-text answer generation, limits the model's ability to perform higher-level visual reasoning. Consequently, tasks involving counting, spatial reasoning, and activity recognition remain challenging despite improvements achieved through the Partial-6 strategy.

Errors were also observed in questions requiring fine-grained visual interpretation, particularly when distinguishing object attributes or subtle visual details. Variations in lighting conditions, viewing angles, and image quality occasionally made accurate recognition of object appearance more difficult.

These findings are consistent with previous studies reporting that object counting, spatial reasoning, and activity understanding remain challenging tasks for many modern Vision-Language Models [11], [13], [20]. Although the proposed model demonstrates strong performance in object recognition, visual attribute identification, and tourist location recognition, the error analysis indicates that higher-level visual reasoning capabilities still require further improvement. Future enhancements may be achieved through the incorporation of more diverse training data, more effective fine-tuning strategies, or the adoption of more advanced Vision-Language Model architectures.

Furthermore, the proposed model was evaluated using a tourism-domain dataset (LosariVQAv1). Therefore, its generalization capability to other domain-specific environments, such as healthcare, education, or industrial applications, remains an important direction for future research.

E. Summary of Findings

Based on the experimental results, the Partial-6 Fine-Tuning strategy proved to be the most effective configuration for bilingual multi-domain Visual Question Answering. This strategy achieved the highest performance on the LosariVQAv1 dataset, obtaining an Exact Match score of 26.67%, a BLEU score of 26.80%, and a CIDEr score of 292.67, while simultaneously maintaining competitive performance on the VQAv2 dataset with an Exact Match score of 41.26%.

Furthermore, the proposed model demonstrated consistent bilingual capabilities, with only a small performance gap observed between Indonesian and English. These results indicate that the combination of Vision Transformer, Indonesian GPT-2, special language tokens, and the Partial Fine-Tuning strategy provides an effective approach for developing Vision-Language Models in multilingual and multi-domain environments. The findings support the hypothesis that partially unfreezing the encoder layers offers a better balance between generalization capability and domain adaptation than either the Encoder Freeze or Full Fine-Tuning approaches.

IV. CONCLUSION

A. Conclusion

This study successfully developed an Indonesia–English bilingual Visual Question Answering (VQA) model based on the VisionEncoderDecoderModel architecture, integrating a Vision Transformer (ViT) as the visual encoder and an Indonesian GPT-2 model as the language decoder. Bilingual capability was achieved through the introduction of special language tokens, <id> and <en>, enabling a single model to generate answers in both Indonesian and English without requiring separate models for each language.

Experimental results demonstrate that the Partial Fine-Tuning strategy outperformed both Encoder Freeze and Full Fine-Tuning approaches. On the LosariVQAv1 local tourism dataset, the Partial-6 configuration achieved the best performance, obtaining an Exact Match score of 26.67%, a BLEU score of 26.80%, and a CIDEr score of 292.67. These results indicate that, within the scope of the evaluated datasets, unfreezing the last six encoder layers provides an effective trade-off between preserving pretrained visual representations and adapting to the visual characteristics of the target tourism domain.

On the bilingual VQAv2 dataset, the best-performing configuration achieved an Exact Match score of 41.26%, a BLEU score of 16.49%, and a CIDEr score of 111.46. These results demonstrate that the model maintained its generalization capability on a general domain despite being adapted to a local tourism domain. Therefore, the experimental results suggest that the Partial-6 strategy offers a favorable trade-off between maintaining general-domain performance and adapting to the evaluated tourism-domain dataset.

Cross-language evaluation revealed that the model was able to generate answers in both Indonesian and English with only a small performance gap. On LosariVQAv1, the Exact Match difference between the two languages was only 1.21 percentage points, while on VQAv2 the difference was 1.84 percentage points. These findings indicate that the proposed language-token mechanism effectively supports bilingual answer generation within a unified architecture.

The training loss and validation loss analyses showed stable convergence without significant signs of overfitting. Furthermore, the qualitative analysis demonstrated that the model performs well in object recognition, visual attribute identification, color recognition, and tourist location recognition. However, challenges remain in tasks requiring object counting, activity recognition, and spatial reasoning.

Overall, the results indicate that the combination of Vision Transformer, Indonesian GPT-2, special language tokens, and the Partial-6 Fine-Tuning strategy constitutes an effective approach for developing bilingual Visual Question Answering systems across both general and local tourism domains. These findings suggest that, for the datasets evaluated in this study, the Partial-6 fine-tuning strategy achieved a more favorable trade-off between preserving pretrained visual knowledge and adapting to the target domain than the Encoder Freeze and Full Fine-Tuning strategies. Further validation on additional datasets and architectures is required before generalizing this conclusion to broader Vision-Language Model applications.

B. Research Contributions

The main contributions of this study can be summarized as follows:

1. Development of an Indonesia–English bilingual Visual Question Answering model based on the VisionEncoderDecoderModel architecture integrating Vision Transformer and Indonesian GPT-2.
2. Introduction of special language tokens, <id> and <en>, to support bilingual answer generation within a unified multimodal model.
3. Development and utilization of the bilingual LosariVQAv1 dataset as a VQA dataset representing the Indonesian local tourism domain.
4. Integration of the bilingual VQAv2 and LosariVQAv1 datasets to support multimodal learning across both general and domain-specific environments.
5. Comprehensive analysis of Encoder Freeze, Full Fine-Tuning, Partial-4, and Partial-6 strategies to identify the most effective fine-tuning approach.
6. Demonstration that the Partial-6 Fine-Tuning strategy achieved the best overall performance across the bilingual VQAv2 and LosariVQAv1 datasets evaluated in this study.

C. Research Limitations

Despite the promising results, this study has several limitations:

1. The LosariVQAv1 dataset focuses on a single tourist destination, namely Losari Beach, and therefore does not fully represent the visual diversity of Indonesian tourism.
2. The language decoder is based on GPT-2 Small, containing approximately 124 million parameters, which limits its multimodal reasoning capability compared with modern Vision-Language Models built upon Large Language Models.
3. The evaluation focuses on fine-tuning strategies within a single architecture and does not include direct comparisons with state-of-the-art Vision-Language Models such as BLIP-2, InstructBLIP, LLaVA, and Qwen-VL.
4. The bilingual VQAv2 dataset was generated through automatic translation using mBART-50, which may introduce translation-related inconsistencies and quality variations.
5. The model still exhibits limitations in tasks requiring counting, activity recognition, and spatial reasoning, as identified in the error analysis presented in Section III.
6. Consequently, further validation on datasets from other application domains is required before the proposed model can be generalized to broader real-world scenarios.
7. This study compares several fine-tuning strategies within the VisionEncoderDecoderModel architecture but does not include direct comparisons with recent Vision-Language Models such as BLIP-2, InstructBLIP, LLaVA, or Qwen-VL under the same experimental settings.
8. The contribution of individual components, such as bilingual language tokens and partial fine-tuning, was evaluated through comparative experiments but was not further investigated using a dedicated ablation study.
9. The experimental results were obtained from a single training run for each configuration. Therefore, performance variability due to different random initialization or random seeds was not investigated.
10. Statistical significance testing was not performed because the objective of this study was to compare the relative performance of different fine-tuning strategies rather than to establish statistical significance between model configurations.

D. Future Work

Based on the findings of this study, several directions for future research can be explored:

1. Expanding the local tourism dataset by incorporating images from various tourist destinations across Indonesia to create a more representative and visually diverse tourism-oriented VQA dataset.
2. Developing multilingual datasets that include additional regional and international languages to broaden the applicability of VQA systems in multilingual environments.
3. Integrating Large Language Model-based decoders such as Gemma, Qwen, Llama, or other advanced multilingual language models to improve answer generation and multimodal reasoning capabilities.
4. Conducting direct comparisons with modern Vision-Language Models, including BLIP-2, InstructBLIP, LLaVA, and Qwen-VL, to provide a more comprehensive assessment of the proposed model's performance.
5. Exploring instruction tuning, parameter-efficient fine-tuning, and advanced multimodal reasoning approaches to improve performance on tasks requiring higher-level visual reasoning.
6. Deploying the proposed model in real-world applications, such as image-based digital tourism

assistants capable of providing interactive information in both Indonesian and English.
7. Conduct comprehensive comparisons with recent Vision-Language Models, including BLIP-2, InstructBLIP, LLaVA, and Qwen-VL, using identical datasets and evaluation protocols.
8. Perform ablation studies to quantify the individual contributions of bilingual language tokens, partial fine-tuning strategies, and dataset combinations to the overall model performance.
9. Evaluate model robustness through multiple experimental runs with different random seeds and analyze performance stability across training sessions.
10. Apply statistical significance tests to determine whether performance differences between fine-tuning strategies are statistically meaningful.

E. Research Implication

The findings of this study demonstrate that Vision-Language Models can be effectively adapted not only to English benchmark datasets but also to Indonesia–English bilingual environments through appropriate fine-tuning strategies. The integration of general-domain and domain-specific datasets further improves the model's capability to learn diverse visual concepts while preserving general visual knowledge acquired during pretraining.

From a practical perspective, the proposed bilingual VQA model has potential applications in several real-world scenarios. In the tourism sector, it can support intelligent digital tourism assistants capable of answering visitors' questions based on photographs of tourist attractions. In education, the model can assist interactive learning systems by providing image-based explanations in both Indonesian and English. Furthermore, the proposed approach may also support public information services, such as smart city applications, digital museums, cultural heritage information systems, and multilingual public service platforms requiring visual question answering capabilities.

Although this study focuses on a tourism-domain dataset, the proposed bilingual VisionEncoderDecoderModel framework can potentially be adapted to other domain-specific applications, including healthcare, industrial inspection, agriculture, and environmental monitoring, through domain-specific fine-tuning using appropriate datasets. Therefore, the proposed approach provides a flexible foundation for developing multilingual Vision-Language Models in low-resource domains beyond tourism.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[2] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," *International Conference on Learning Representations (ICLR)*, 2021.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI Technical Report, 2019.

[4] A. Goyal, Y. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017.

[5] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, 2002.

[6] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015.

[7] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," *Proc. EMNLP-IJCNLP*, pp. 5100–5111, 2019.

[8] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," *International Conference on Machine Learning (ICML)*, 2021.

[9] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," *International Conference on Machine Learning (ICML)*, 2022.

[10] P. Wang et al., "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," *International Conference on Machine Learning (ICML)*, 2022.

[11] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *International Conference on Machine Learning (ICML)*, 2023.

[12] D. Dai et al., "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[13] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[14] B. Bai et al., "Qwen-VL: A frontier large vision-language model with versatile abilities," 2023.

[15] Z. Wang et al., "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5475–5498, 2024.

[16] P. Xu et al., "Multimodal foundation models: From specialists to general-purpose assistants," *ACM Computing Surveys*, vol. 57, no. 1, pp. 1–44, 2024.

[17] Y. Tang et al., "Multilingual translation with extensible multilingual pretraining and finetuning," *Proc. 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 863–884, 2021

[18] T. Wolf et al., "Transformers: State-of-the-art natural language processing," *Proc. EMNLP System Demonstrations*, pp. 38–45, 2020.

[19] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov, "Label Studio: Data labeling software," HumanSignal, 2020.

[20] Y. Du et al., "Vision-language pre-training: Basics, recent advances, and future trends," *Foundations and Trends in Computer Graphics and Vision*, vol. 18, no. 1–2, pp. 1–214, 2024.