Assistive Visual Question Answering for Visually Impaired Users

Mounika Kottapalli
Department of Computer Science and Engineering
The University of Texas at Arlington
mxk5510@mavs.uta.edu

***Abstract*—In order to improve the accessibility and independence of visually impaired people, this project introduces an Assistive Visual Question Answering (VQA) system that allows them to ask questions about their environment and get insightful responses. The system is constructed utilizing two complementary methods using the VizWiz dataset, which is a real-world collection of image-question pairings from blind users. PaliGemma-2 is used to extract features, which are then put into a lightweight classifier for effective prediction across a fixed response space in the first method, which is based on classification and inspired by the "Less Is More" paradigm. The second is a generation-based method that allows for the development of open-ended answers by combining different vision-language models (CLIP-ViT-GPT2, SigLIP-GPT2, ViT-GPT2). According to experimental data, generation-based models provide more flexibility but demand more processing resources and have trouble handling ambiguous circumstances, whereas the classification-based model achieves better accuracy and functions dependably on noisy, unanswerable questions. This work advances useful AI solutions for visually impaired users by illustrating the trade-offs between expressiveness and efficiency in VQA systems.**

Index Terms—VizWiz, PaliGemma-2, CLIP-ViT-GPT2, SigLIP-GPT2, ViT-GPT2

I. Introduction

Visual Question Answering (VQA) is an interdisciplinary activity that combines computer vision and natural language processing. It involves a system answering natural language questions based on visual material. By answering questions about their surroundings that they cannot see, VQA technology has the potential to greatly increase everyday accessibility and independence for those with visual impairments.

The goal of this research is to develop an assistive VQA system specifically for people with visual impairments. This research make use of the VizWiz dataset, which comprises more than 30,000 image-question pairs gathered from actual blind user use cases. Poor image quality, a significant percentage of "unanswerable" situations, and a wide variety of frequently unclear queries are some of the particular difficulties this dataset presents.

In order to tackle these issues, two different methods were employed:

1) Inspired by the philosophy "Less Is More," classification-based VQA uses a lightweight neural network to classify textual and visual elements into a predetermined set of 6,503 frequent answers.
2) Generation-based VQA allows the system to go beyond the preset answer set by combining GPT-2 with cutting-edge visual encoders (CLIP, SigLIP, and ViT) to produce free-form responses.

Each strategy has advantages and disadvantages. The classification method is reliable and effective, particularly for answers that appear frequently and unanswerable questions, whereas the generation method permits more expressive, open-ended responses but necessitates a substantial increase in training and processing power. By bridging the gap between current VQA models and practical accessibility requirements, this research hopes to contribute to AI systems that are more inclusive.

A vast collection of visual question answering (VQA) samples that have been carefully chosen to represent the requirements and difficulties encountered by visually impaired users is known as the VizWiz dataset. About 31,000 image-question-answer triplets make up this collection; the spoken questions and images were initially recorded by blind users using a mobile application. Many artificial or crowdsourced VQA datasets lack the authenticity and richness that these real-world samples bring. The dataset is particularly difficult because of problems like uneven framing, ambiguous or context-dependent questions, and poor image quality (blurry focus, low illumination), which closely resemble the unpredictability of actual assisting events.

An 80-10-10 ratio was used to re-split the initial VizWiz training and validation splits into fresh training, validation, and test sets after they had been merged into a single pool of about 25,000 samples for this project. In VizWiz, each question has up to ten human-annotated responses; however, only the most common (top-ranked) response was chosen for each sample in order to maintain simplicity and conformity to a classification framework. Image pathways, matching natural language queries, and one ground truth response are all included in the final dataset. The fact that out of 6,503 unique answers, nearly 75% only appear a few times, and a sizable portion (4,843) are marked as "unanswerable" highlights the difficulty of this task and emphasizes the significance of the dataset for creating approachable, practical VQA systems. This is one of the main challenges in using VizWiz.

II. Literature Review

Yan et al. [1] focus on the VizWiz dataset, an accessibility-focused collection of photos taken by blind users, and offer a simple yet powerful method for Visual Question Answering (VQA). The authors use pre-trained CLIP (Contrastive Language–Image Pre-training) encoders for both picture and text modalities in place of deep, computationally costly models. A lightweight linear classifier is used to process these concatenated features. The study also presents answer-type prediction, an auxiliary task that serves as a gating mechanism to improve answer selection. The model circumvents the huge output spaces typical of VQA challenges by employing a curated vocabulary of 5,726 answers chosen using a greedy approach. Despite its simplicity, the model earned competitive results in the VizWiz 2022 Challenge, including 60.15% accuracy on Task 1 and 83.78% average precision on Task 2. The effectiveness of simplicity in VQA tasks, particularly for accessible applications, is demonstrated by this work.

Gurari et al. [2] present the VizWiz Grand Challenge, a benchmark dataset and challenge created to investigate Visual Question Answering (VQA) in terms of accessibility for blind users. More than 31,000 photos and natural language queries were gathered directly from blind people via a smartphone app to create the dataset. VizWiz data shows real-world issues like low image quality, unclear framing, and a variety of question intent, including unanswerable queries, in contrast to standard VQA datasets where images and questions are frequently chosen or crowdsourced. The study offers a comprehensive examination of these difficulties, draws attention to the discrepancy between the requirements of blind users and conventional VQA models, and establishes a framework for assessing answerability as a crucial performance indicator.

III. Dataset Description

The VizWiz dataset consists of over 30,000 image-question pairs collected from visually impaired users via a mobile application. Each pair includes a user-taken photo, a spoken question (transcribed), and up to ten human-annotated answers.

To maintain consistency and simplify the model design, only the most frequent (top-ranked) answer from the ten annotator responses is selected as the ground truth. This approach supports supervised learning and reduces noise caused by ambiguity in multiple responses.

Despite simplifying the answer space, the dataset remains challenging due to:

- Poor image quality (e.g., blurriness, low lighting)
- Unclear or ambiguous questions
- A high proportion of “unanswerable” cases

Out of 6,503 unique answers in the dataset, nearly 75% appear only a few times. Approximately 4,843 answers are marked as *“unanswerable”*, further emphasizing the need for robust handling of noisy inputs.

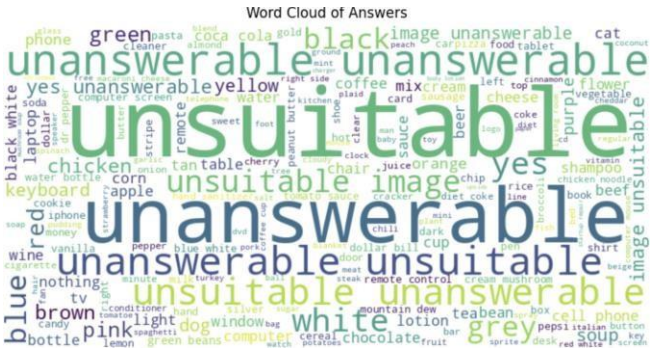

Fig. 1: Word Cloud of Answers in VizWiz Dataset

A representative image from the VizWiz dataset, along with its associated question and ground truth answer, is shown in Fig. 2 to illustrate the nature of the input data.

Example:

Fig. 2: Sample image from the VizWiz dataset with question and answer.

IV. Implementation

To tackle the intricate problems of Visual Question Answering (VQA) in authentic assistive environments, this study investigates two complementary strategies: a classification-based approach and a generation-based approach. While both aim to process image-question pairs and produce natural language responses, they differ significantly in design and functionality.

The classification-based method, inspired by the “Less Is More” paradigm, emphasizes efficiency by employing a lightweight classifier built on PaliGemma-2 to map multimodal features to a fixed set of answers. In contrast, the generation-based method leverages transformer-based vision-language models (CLIP-ViT-GPT2, SigLIP-GPT2, ViT-GPT2) to generate expressive, free-form answers. This dual-strategy framework enables comparison between expressiveness and efficiency in assistive VQA for visually impaired users.

A. *Classification-Based VQA Using PaliGemma-2*

On top of feature embeddings taken from the PaliGemma 2 model, a lightweight neural classifier is built to facilitate quick and precise prediction within a fixed answer vocabulary. The model learns to map multimodal input features to one of 6,503 predefined answer classes in the VizWiz dataset using this classification-based method, which turns the Visual Question Answering (VQA) job into a supervised multi-class classification problem.

The original training and validation sets were combined to yield a total of about 25,000 samples. This model used a stratified split based on the answer label to guarantee a representative and equitable distribution of answer classes across subsets. Three subsets of the dataset were created test (10%), validation (10%), and training (80%). To preserve the proportionate dis- tribution of common and uncommon answers—a crucial fea- ture in a highly unbalanced dataset like VizWiz—stratification was implemented. Initially, a temporary pool was created by separating 20% of the samples. Once more, stratified sampling was used to divide this pool equally into test and validation sets (10% each). The splits' final sizes were as follows :

- Training Set: 80% (˜20,000 samples)
- Validation Set: 10% (˜2,500 samples)
- Test Set: 10% (˜2,500 samples)

For each image-question pair, the PaliGemma-2 model extracted fixed-length feature vectors. These vectors were passed to a fully connected neural network (VQAClassifier), which consisted of:

- A linear layer to reduce feature dimensions,
- A ReLU activation function,
- A dropout layer for regularization,
- A final linear projection to the 6,503 answer classes.

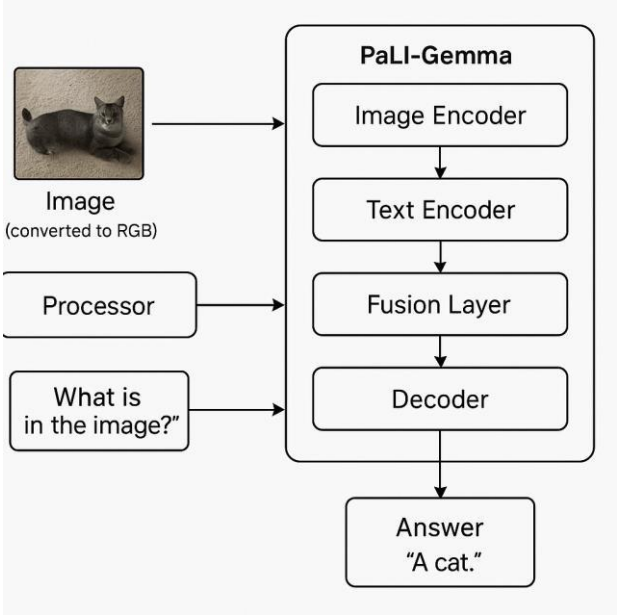

Fig. 3: Classification based Model Architecture

The model was trained over 20 epochs using the Adam optimizer and cross-entropy loss. The best model (based on validation accuracy) was saved for final testing.

Performance Metrics:

- **Train Accuracy:** 0.7740, **Loss:** 0.6387
- **Validation Accuracy:** 0.7041, **Loss:** 1.0675

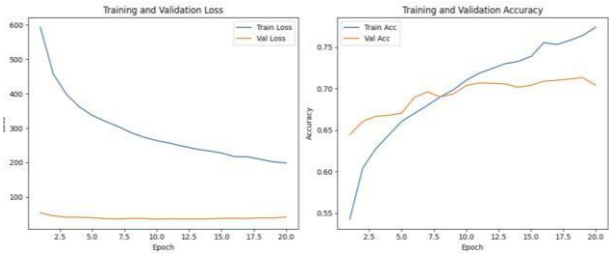

Fig. 4: Accuracies and Losses for Paligemma2 VQA

These results suggest the classifier effectively balances accuracy and efficiency, making it suitable for real-time, resource-constrained assistive applications.

B. *Generation-Based VQA Approach*

The second option considered for the Visual Question Answering (VQA) system is the generation-based approach, which uses advanced vision-language models to generate free form written responses to queries based on image content. Several transformer-based architectures, including CLIP-ViT GPT2, SigLIP-GPT2, and ViT-GPT2, were used to execute this strategy. For the purpose of producing natural language responses, these models integrate the potent language model GPT-2 with cutting-edge vision transformers (ViTs).

1) *Architecture Overview:*

- **Vision Encoder:** To process and extract useful charac teristics from the input image, the models first employ a Vision Transformer (ViT) or more advanced variants such as SigLIP
- **Textual Input:**The transformer-based language model receives the multimodal input, which is created by con-catenating the question with the image features.
- **Answer Generation:** Using the contextual relationship between the image and the question, the GPT-2 model creates a textual response based on the combined image question representation.

The Visual Question Answering (VQA) system presented in this work employs a transformer-based multimodal architec-ture that combines a vision encoder with a language decoder (GPT-2). The architecture is designed to process both visual and textual inputs in a unified framework to generate accurate, natural language answers to image-based questions.

Image Encoding – The input image is passed through a pre-trained vision encoder to extract a dense feature vector representation.

Projection – This visual representation is then projected into

a 768-dimensional embedding space to match the hidden size expected by the GPT-2 decoder.

Text Formatting – The input question is tokenized using GPT-2's tokenizer after being formatted as `"Question: [question] Answer:"`.

Fusion – The projected image embedding is prepended to the tokenized question to form a combined multimodal input sequence.

Generation – GPT-2 generates the answer autoregressively, conditioned on both the image and the textual prompt. The

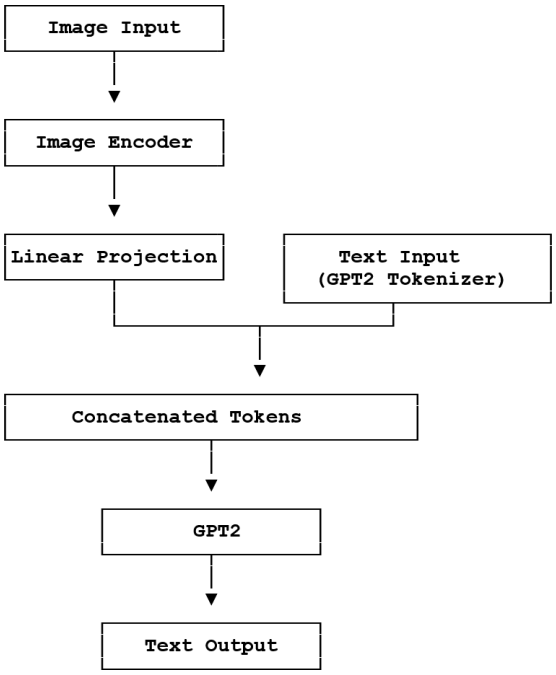

Fig. 5: Generation based Model Architecture

answer generation process is controlled using parameters such as `max_new_tokens` and `max_length`, which limit the number of tokens generated during decoding. These parameters ensure that the responses are concise and relevant to the question context.

During training, the model is optimized using cross-entropy loss, computed only over the answer tokens. The question portion of the input is masked to prevent the model from learning to reproduce it. Padding and masked tokens are excluded from the loss computation to ensure focused and efficient learning.

For evaluation, answer accuracy is computed using an exact match between the predicted output and ground truth references. If multiple ground truth answers are available, the prediction is considered correct if it matches any one of them. This evaluation strategy follows standard VQA protocols and effectively captures the model's ability to generate semantically accurate responses.

2) ***ViT-GPT2****:* The ViT-GPT2 model uses `google/vit-base-patch16-224-in21k` as the image encoder and GPT-2 as the text generator. To process the original VizWiz dataset, each image is matched with a corresponding question and ground-truth response. The images are transformed into pixel values using a vision feature extractor, while the question-answer pairs are tokenized into model-compatible input sequences. The question and answer are concatenated to form the complete input for each data point. Labels are then generated by tokenizing the full text while masking the question tokens to prevent the model from simply copying them during training. This pre-processing ensures efficient data loading and enables the model to learn natural language responses grounded in both the image and the question.

For the Visual Question Answering (VQA) task, the ViT-GPT2 model serves as a hybrid architecture that combines vision and language understanding. It consists of two main components: a Vision Transformer (ViT) for extracting visual features from the image, and GPT-2 for generating textual responses. The ViT encoder processes the input image and produces a dense feature vector, which is then projected into a 768-dimensional embedding space to match GPT-2's hidden size. The input question is tokenized and embedded using GPT-2's tokenizer. The projected image embedding and the text embeddings are concatenated to form a single multimodal input. This combined representation enables the model to generate contextually relevant answers that reflect both the visual content and the question intent.

The model was trained for 3 epochs on the VizWiz dataset.

Results:

- **Train Accuracy:** 0.6819, **Loss:** 0.0752
- **Validation Accuracy:** 0.5605, **Loss:** 0.1373

Sample Predictions (ViT-GPT2):

Question: What does this can say?
Answer: classic roast
Predicted: unanswerable

Question: What color is this shirt?
Answer: grey
Predicted: white

Question: What breed of dog is this?
Answer: unsuitable
Predicted: unsuitable

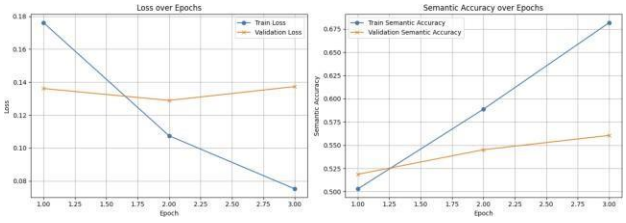

Fig. 6: Accuracies and Losses for VIT-GPT2 Model

3) ***SigLIP-GPT2****:* The SigLIP-GPT2 model integrates `google/siglip-base-patch16-224` as the image encoder and GPT-2 as the language generator. Similar to the ViT-GPT2 setup, each image in the VizWiz dataset is paired with a corresponding question and answer. The images are processed using `SiglipProcessor` to produce normalized pixel tensors, while the question and answer are formatted and tokenized in the form `"Question: [question] Answer:"`. During training, the question tokens are masked to focus the model's learning on generating accurate answers.

SigLIP (Sigmoid Language-Image Pretraining) improves upon CLIP by replacing the cosine similarity bottleneck with a sigmoid-based alignment objective, enabling more flexible feature learning. In this architecture, the visual features extracted by the SigLIP encoder are projected into a 768-dimensional embedding space and concatenated with the GPT-2 token embeddings. This fused sequence is passed through GPT-2, which generates the answer autoregressively. The model leverages the strength of SigLIP's visual understanding and GPT-2's language modeling to handle diverse image-question pairs in a natural and context-aware manner.

The model was trained for 3 epochs on the VizWiz dataset.
Results:

- **Train Accuracy:**0.5342, **Loss:** 0.0957
- **Validation Accuracy:** 0.4800 , **Loss:** 0.1546

Sample Predictions (SigLIP-GPT2):

Question: What Color?
Answer: multi color
Predicted: grey

Question: What color is this?
Answer: white
Predicted: grey

Question: What's in this box?
Answer: vicks vaporub
Predicted: unanswerable

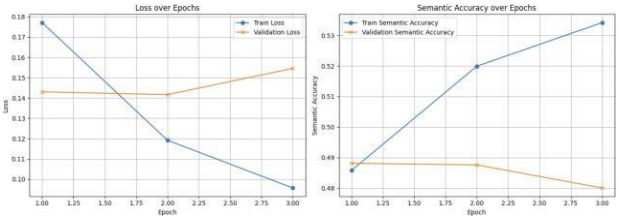

Fig. 7: Accuracies and Losses for SigLip-GPT2 Model

4) ***CLIP-ViT-GPT2****:* The CLIP-ViT-GPT2 model combines `openai/clip-vit-large-patch14` as the vision encoder with GPT-2 for text generation. Each image from the VizWiz dataset is paired with a relevant question and answer. The image is processed using the `CLIPProcessor` to extract dense visual embeddings, while the question and answer are formatted into a structured text prompt. The tokenized input follows the form `"Question: [question] Answer:"`, with masking applied to the question segment during training to prevent data leakage.

CLIP (Contrastive Language–Image Pretraining) uses a contrastive objective to align visual and textual representations in a shared latent space. In this model, CLIP's image embeddings are linearly projected to match GPT-2's 768-dimensional hidden state. These projected embeddings are then prepended to the GPT-2 input sequence. The model decodes this fused representation to generate natural language responses that incorporate both the visual context and the question semantics. This integration enables the model to generalize well across a wide range of visual and linguistic inputs, producing answers that are contextually and semantically aligned.

The model was trained for 3 epochs on the VizWiz dataset.
Results:

- **Train Accuracy:**0.5282, **Loss:** 0.0956
- **Validation Accuracy:** 0.4789 , **Loss:** 0.1513

Sample Predictions (CLIP-ViT-GPT2):

Question: What Color?
Answer: white
Predicted: white

Question: What color is this?
Answer: brown
Predicted: blue

Question: What's in this box?
Answer: vicks vaporub
Predicted: unanswerable

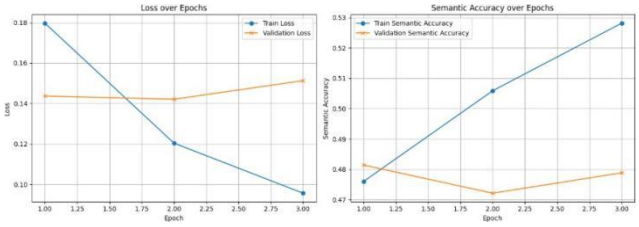

Fig. 8: Accuracies and Losses for CLIP-ViT-GPT2 Model

While generation-based models allow for open-ended and expressive answers, they demand significantly more computational resources and are more sensitive to poor image quality and ambiguous questions in the VizWiz dataset.

V. MODEL EVALUATION

This section presents the evaluation outcomes of the Visual Question Answering (VQA) models tested on the VizWiz dataset. The models were assessed based on their ability to generate accurate answers to image-based questions. Performance is reported using metrics such as accuracy, loss, precision, recall, and F1 score. Both quantitative metrics and qualitative examples are included to provide a comprehensive understanding of model behavior on real-world data.

A. *Classification-Based Model: PaLI-Gemma2 (MLP)*

The classification-based model adopts a fixed answer vocabulary and predicts the most likely answer class given an image-question pair. It utilizes image embeddings from a pre-trained PaLI-Gemma2 vision encoder, followed by a Multi-Layer Perceptron (MLP) for classification.

Test Set Performance:

- **Accuracy:** 0.7171
- **Precision:** 0.6880
- **Recall:** 0.7171
- **F1 Score:** 0.6963
- **Top-1 Accuracy:** 0.7171
- **Top-5 Accuracy:** 0.9843

Inference: The model demonstrates strong overall performance, particularly in Top-5 predictions, indicating that even when the most probable answer is not ranked first, it frequently appears among the top predicted candidates. This suggests effective semantic understanding, though minor ambiguities persist in exact answer selection.

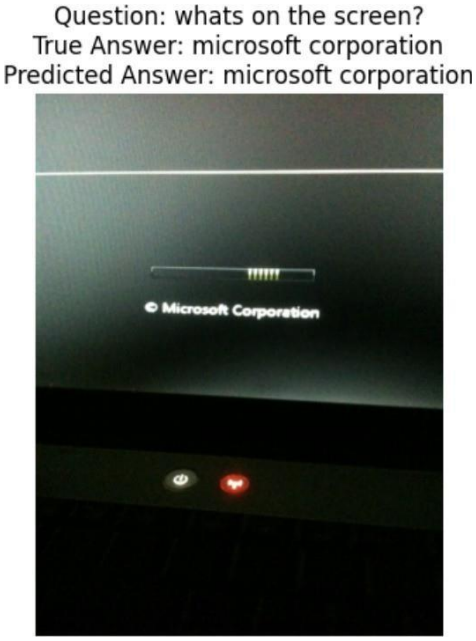

Fig. 9: Sample output from the classification-based model showing predicted answer from fixed vocabulary.

The model's performance across training, validation, and test sets is summarized in Table I, highlighting consistent generalization with a high Top-5 test accuracy despite moderate validation loss.

TABLE I: Training, Validation, and Test Performance of PaLI-Gemma2 Classification Model

Dataset	Accuracy	Loss
Train	0.7740	0.6387
Validation	0.7041	1.0675
Test	0.7171	–

B. *Generation-Based Models*

The generation-based models adopt an autoregressive approach where GPT-2 generates free-form answers conditioned on both the image and question. Unlike the classification-based approach, these models are not restricted to a fixed vocabulary and are capable of producing more flexible and descriptive outputs. Each model uses a different vision encoder to extract image embeddings, which are then fused with the question tokens and passed to GPT-2 for decoding.

1) ***ViT-GPT2****:* The ViT-GPT2 model was evaluated on the VizWiz test dataset to assess its generative capabilities. The model produces free-form answers by conditioning both visual features and textual input.

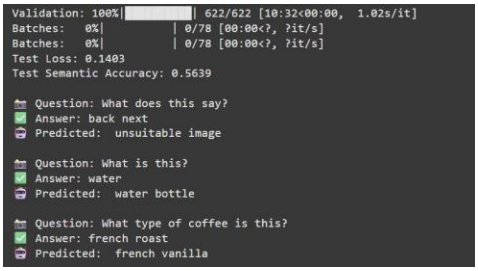

Fig. 10: test predictions generated by the ViT-GPT2 model..

The training, validation, and test results demonstrate consistent model behavior, with minimal drop in accuracy and a moderate increase in loss across datasets, as summarized in Table II.

TABLE II: Training, Validation, and Test Performance of ViT-GPT2 Model

Dataset	Accuracy	Loss
Train	0.6819	0.0752
Validation	0.5605	0.1373
Test	0.5639	0.1403

2) ***SigLIP-GPT2****:* The SigLIP-GPT2 model was evaluated on the VizWiz test dataset to measure its ability to generate natural language responses by using visual features that are closely aligned with their related text through contrastive training.

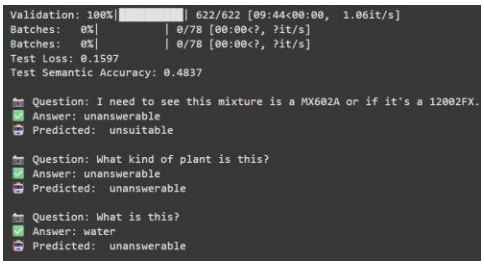

Fig. 11: Test predictions generated by the SigLIP-GPT2 model.

The training, validation, and test performance metrics show a moderate decline across datasets, with semantic accuracy

remaining consistent under varying conditions, as shown in Table III.

TABLE III: Training, Validation, and Test Performance of SigLIP-GPT2 Model

Dataset	Accuracy	Loss
Train	0.5342	0.0957
Validation	0.4800	0.1546
Test	0.4837	0.1597

3) ***CLIP-ViT-GPT2****:* The CLIP-ViT-GPT2 model leverages the CLIP ViT-B/32 image encoder and GPT-2 to generate open-ended answers for visual questions on the VizWiz dataset.

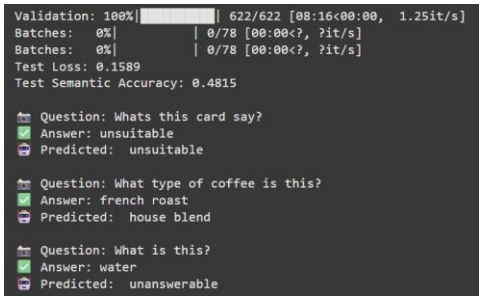

Fig. 12: Test predictions generated by the CLIP-ViT-GPT2 model.

The model demonstrated stable generative behavior across datasets with consistent loss values. Table IV summarizes the final performance across training, validation, and test phases.

TABLE IV: Training, Validation, and Test Performance of CLIP-ViT-GPT2 Model

Dataset	Accuracy	Loss
Train	0.5282	0.0956
Validation	0.4789	0.1513
Test	0.4815	0.1589

VI. INFERENCE

Inference was performed by manually providing input images and corresponding questions. Images were uploaded from the local system, and questions were entered through a prompt interface. The models then returned predicted answers based on these inputs.

For classification, the model selected an answer from a fixed set of classes. In generation-based models, GPT-2 produced a natural language response. This interactive setup allowed for evaluating the models' ability to generalize to real-world, unseen inputs beyond the structured test dataset.

(a) Inference Example 1

(b) Inference Example 2

Fig. 13: Sample outputs during manual inference with custom images and questions.

VII. Conclusion

This study investigates the application of Vision and Language Models (VLMs) for Visual Question Answering (VQA), focusing on advanced architectures such as GPT-2, ViT, CLIP, and their combinations to enhance image-text understanding. Several models were trained—ViT-GPT2, SigLIP-GPT2, and CLIP-ViT-GPT2—each integrating visual and textual modalities to answer image-based queries.

Results indicate varying levels of success: the ViT-GPT2 model, combining ViT embeddings with GPT-2's language capabilities, achieved the highest accuracy on both training and test sets. SigLIP-GPT2 and CLIP-ViT-GPT2 showed moderate outcomes, with training-validation performance gaps indicating potential overfitting. CLIP-based models, despite their promise in multimodal learning, struggled to generalize to unseen data. Additionally, the integration of PaLI-Gemma-2 in the classification pipeline demonstrated enhanced robustness, particularly in handling noisy visual inputs typical of accessibility-focused applications.

Key Contributions and Innovations:

- **Efficient Pipeline:** Adoption of the "Less Is More" strategy using PaLI-Gemma-2, achieving a balance between performance and computational efficiency.
- **Fine-Tuning on VizWiz:** A novel application of PaLI-Gemma-2, improving resilience to noise and question diversity.
- **Dual-Model Approach:** Comparative analysis of classification (PaLI-Gemma-2) and generation (CLIP/SigLIP/ViT + GPT-2) strategies, highlighting trade-offs.
- **Real-World Impact:** System tailored for visually impaired users, addressing challenges such as noisy data and limited device resources.

Challenges and Future Directions:

- **Classification Limitations:** Fixed answer space (6,503 answers), limiting flexibility for out-of-vocabulary responses.
- **Generation Issues:** Difficulty handling unanswerable questions and high computational costs.
- **Dataset Complexity:** VizWiz presents ongoing challenges due to image noise and answer diversity.
- **Future Work:** Explore hybrid pipelines that classify unanswerable questions before generating responses; optimize for edge deployment via model quantization; and integrate larger generative models (e.g., GPT-3) to expand vocabulary coverage.

References

[1] F. Deuser, et al., "Less Is More: Linear Layers on CLIP Features as Powerful VizWiz Model," *arXiv preprint arXiv:2206.05281*, 2022.

[2] D. Gurari, et al., "VizWiz Grand Challenge: Answering Visual Questions from Blind People," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.