Multimodal Classification System for Hausa using LLMs and Vision Transformers

Ali Mijiyawa[1] **Fatiha Sadat**[1]

[1]Université du Québec à Montréal (UQAM), Montreal, QC, Canada

mijiyawa.ali@courrier.uqam.ca | sadat.fatiha@uqam.ca

Abstract

This paper presents a classification-based Visual Question Answering (VQA) system for the Hausa language, integrating Large Language Models (LLMs) and vision transformers. By fine-tuning LLMs on monolingual Hausa text and fusing their representations with those of state-of-the-art vision encoders, our system predicts answers from a fixed vocabulary. Experiments conducted on the HaVQA dataset, under offline text–image augmentation regimes, tailored to the specificity of Hausa as a low-resource language, show that this augmentation strategy yields the best performance over the baseline, achieving 35.85% accuracy, 35.89% WuPalmer similarity, and 15.32% F1-score.

1 Introduction

The task of VQA can be approached through three paradigms: *open-ended classification*, selecting an answer from a fixed vocabulary; *multiple-choice (MCQ)*, choosing from given options; and *generative*, producing free-form text responses, sometimes with rationales (Dua et al., 2020). This study adopts a classification-based approach, framing VQA as a closed-vocabulary, multi-class task where the model selects the correct answer from a predefined label set given an image and a question.

Transformer-based language models such as BERT (Devlin et al., 2018) have enabled significant progress in VQA for high-resource languages. However, many African languages, including Hausa, remain underrepresented due to the lack of large annotated multimodal corpora (Kumar et al., 2022; Hedderich et al., 2021), which hinders the development of VQA systems capable of capturing their linguistic and cultural specificities. To the best of our knowledge, no previous study has explored the fine-tuning of Large Language Models (LLMs) combined with vision transformers for classification-based VQA in African low-resource contexts. Furthermore, no prior work has examined data augmentation techniques specifically tailored to VQA systems in African languages.

In this study, we aim to explore the non-generative potential of LLMs in this setting, focusing on the Hausa language. We present a classification-based Hausa VQA system combining fine-tuned LLMs with state-of-the-art vision transformers.

Using the HaVQA dataset (Parida et al., 2023b), we evaluate training paradigms to assess the impact of text and image data augmentation, focusing on offline augmentation adapted to Hausa, where data are expanded before training via text rewriting and image perturbations, and compare this to a baseline without augmentation.

Our main contributions are twofold: (i) We conduct a comprehensive multimodal benchmark of nine LLMs and four vision transformers (36 model variants in total) within a unified fine-tuning framework for Hausa VQA; and (ii) We propose a low-resource data augmentation and multimodal enhancement framework, combining text and image transformations tailored to Hausa linguistic and cultural characteristics. This expands the HaVQA dataset (Parida et al., 2023a) into HaVQA_aug[1] and yields measurable improvements in classification accuracy, Wu-Palmer similarity, and F1-score across models.

The rest of the paper is organized as follows: Section 2 discusses related work; Section 2.4 provides background on the Hausa language; Section 3 formalizes the VQA task and presents the proposed methodology;

Section 4 presents our experiments and evaluations; Section 5 discusses the results; and Section 6 concludes the paper with future directions.

[1]https://github.com/Alimiji/LLM_QRV_Hausa_HaVQA_aug

2 Related Work

2.1 VQA for African Languages

Recent advances in VQA leverage transformer-based models such as BERT (Devlin et al., 2018) and Vision Transformer (Dosovitskiy et al., 2020) to integrate visual and textual information (Tan and Bansal, 2019; Lu et al., 2019). In Africa, only three VQA datasets exist: HaVQA for Hausa using non-large models (Parida et al., 2023a), CVQA covering multiple African languages except Hausa with large multimodal models (Romero et al., 2024), and SwahiliVQA with 10,000 images and 41,448 Q&A pairs, achieving 38.38% accuracy with non-large models (MBWANA and Long Hoang, 2025). Data augmentation improves dataset diversity and model generalization (Chen et al., 2022; Yang et al., 2022). Challenges remain, including limited performance of models like GPT-4o (Olufemi et al., 2025) and cultural biases highlighted by CulturalVQA and WorldCuisines (Nayak et al., 2024; Winata et al.).

2.2 LLMs for African Languages

Multilingual LLMs such as mBERT (Devlin et al., 2018) and XLM-R (Conneau et al., 2020) often underperform on African languages due to their scarcity in pretraining corpora (Hedderich et al., 2021; Blasi et al., 2022). Adaptive fine-tuning (e.g., MAFT (Alabi et al., 2022)) and from-scratch models (e.g., AfriBERTa (Ogueji et al., 2021), AfroLM (Doe et al., 2023)) improve alignment with African languages as well as text classification and question answering (Yu et al., 2025). For Hausa, challenges include limited datasets, dialectal variation, and suboptimal tokenization (Muhammad et al., 2025). Community resources like HausaNLP, AfroBench (Ojo et al., 2023), and IrokoBench (Adelani et al., 2025) support NLP development but highlight persistent performance gaps. Integrating these language-specific models with vision transformers and culturally-aware data augmentation is key for effective Hausa VQA.

2.3 Data Augmentation

Recent advances in data augmentation enhance both text and images. Text techniques include synonym replacement, EDA (Wei and Zou, 2019), back-translation (Sennrich et al., 2016), and LLM-based paraphrasing (Ding et al., 2024), while image methods use geometric and photometric transformations, MixUp, and CutOut (Shorten and Khoshgoftaar, 2019a). In VQA, multimodal augmentation combines visual perturbations with question reformulation, QA generation, and adversarial examples (Chen et al., 2022). Such strategies, integrated with adapted LLMs, are crucial for improving classification-based VQA in low-resource African languages like Hausa.

2.4 Hausa Language

Hausa, a Chadic language of the Afroasiatic family, is spoken by over 200 million people across West and Central Africa (Muhammad et al., 2025). It functions as a regional lingua franca and has a rich literary tradition in both *boko* (Latin) and *ajami* (Arabic-derived) scripts. Despite its cultural significance, Hausa remains low-resource in NLP, facing limited datasets, suboptimal tokenization, and dialectal variation (Muhammad et al., 2025). Recent work has advanced text classification, sentiment analysis, machine translation, NER, and POS tagging, while speech datasets like Common Voice and multimodal resources such as HaVQA support VQA research. Pretrained models like AfriBERTa and the HausaNLP catalog further enhance accessibility and development in these tasks.

3 Proposed Methodology

The Hausa VQA task adapts visual question answering to Hausa, aiming to build models that understand Hausa questions about images and provide accurate answers by combining language comprehension with visual reasoning (Parida et al., 2023a; Antol et al., 2015). This work promotes inclusive AI for speakers of low-resource languages (Nekoto et al., 2020; Hedderich et al., 2021), with datasets like HaVQA (Parida et al., 2023a) providing benchmarks for multilingual and multimodal research.

3.1 System Architecture

Our proposed system consists of a text encoder (LLM) that converts a Hausa question q into a dense vector, an image encoder (ViT) producing patch-level visual embeddings, a fusion layer combining both modalities via concatenation or cross-modal attention, and a classification head mapping the fused representation to a global Hausa answer vocabulary $\mathcal{A}$. Given a question–image pair (q, I), the model predicts a unique label $a' \in \mathcal{A}$. The answer space $\mathcal{A}$ comprises 2,991 gold-standard Hausa labels, ensuring each input pair maps to exactly one label.

We adopt a classification-based VQA approach for Hausa, following HaVQA (Parida et al., 2023b).

This simplifies modeling and enables direct supervision with a fixed label set, avoiding the challenges of free-form answer generation in low-resource languages.

3.2 Offline vs. No Augmentation

As shown in Figure 1, the baseline (no augmentation) model is fine-tuned solely on the original HaVQA dataset (Parida et al., 2023a), serving as a performance reference. In contrast, offline augmentation (Figure 2) expands the dataset by duplicating images with geometric transformations (e.g., random flips and rotations) and paraphrasing English question–answer pairs using synonym replacement (Wei and Zou, 2019), which are then translated into Hausa (Parida et al., 2023a). These augmented pairs are linked to the transformed images, creating a larger, fixed dataset. Comparing these two settings enables systematic evaluation of data augmentation's impact on classification accuracy, semantic alignment, and robustness in low-resource scenarios.

4 Experiments and Evaluations

4.1 HaVQA dataset and evaluation setup

We use the HaVQA dataset (Parida et al., 2023a), which contains 6,022 English–Hausa question–answer pairs spanning 2,991 unique Hausa labels and 1,555 images from Visual Genome (Krishna et al., 2017). Due to Hausa's low-resource nature, we merged the original development and test sets into a single evaluation set (Kann et al., 2019) to maximize training data. This practice, common in low-resource NLP, increases training robustness while maintaining comprehensive evaluation.

4.2 Training details (Hyperparameters)

All models were trained for 10–20 epochs using HuggingFace's Trainer (final runs: 17 epochs) with a fixed seed of 12345. Training and evaluation used a batch size of 32 per device. We employed the AdamW with $learning\_rate = 1 \times 10^{-5}$ and $weight\_decay = 1 \times 10^{-4}$. Precision was set to bfloat16, and evaluation, logging, and checkpointing occurred every 100 steps, retaining the last three checkpoints. The best model was selected based on WUPS. Data loading used 8 workers and kept all columns. Training was conducted on a single NVIDIA A100 or L4 GPU.

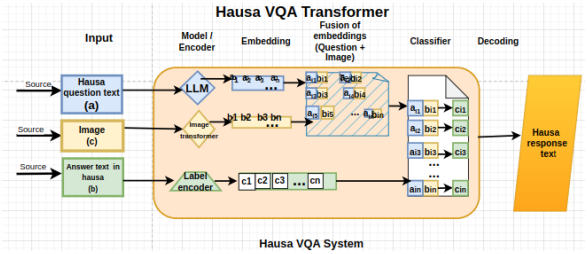

Figure 1: Baseline system: Hausa VQA model combining a large language model (LLM), inspired by Parida et al. (2023b).

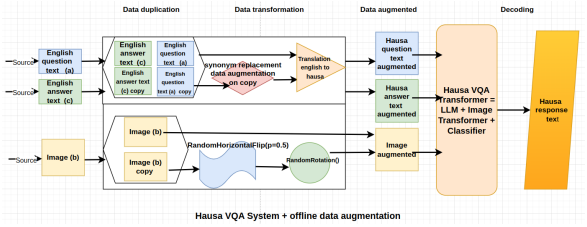

Figure 2: Offline augmentation: Hausa VQA model trained on pre-augmented data before multimodal fusion (Zhang et al., 2015; Wei and Zou, 2019; Parida et al., 2023b).

4.3 Metrics meaning

We evaluate Hausa VQA using Accuracy, F1 Score (Pedregosa et al., 2011), and WuPalmer (WUP) Similarity (Wu and Palmer, 1994). Accuracy measures the proportion of correct predictions, while F1 Score balances precision and recall to account for class imbalance. WUP similarity, computed via wup_similarity in NLTK WordNet[2], assesses semantic relatedness between predicted and reference answers. Together, these metrics provide a comprehensive evaluation, summarized in Table 1.

4.4 Hausa VQA task models training

Tables 2 and 3 list the nine LLMs and four vision transformers evaluated in our Hausa VQA study. Pairing each LLM with every vision encoder yields 36 model variants, all trained as multiclass classifiers over a fixed set of Hausa answer labels, without sequence generation. Each question–image pair from HaVQA (Parida et al., 2023a) is encoded by the LLM and the vision transformer, with fused embeddings fed to a classifier. We compare performance using offline augmentation on the expanded dataset versus the baseline with no augmentation.

4.4.1 No Augmentation training

In this setup, the model is trained on the original HaVQA split: 4,816 training, 1,204 test pairs and covering 1,555 images in total. Each training in-

[2] https://www.nltk.org/api/nltk.corpus.reader.wordnet.html

Metric	Objective	Interpretation	Value Range
Accuracy	Exact match with the reference answer	Answer is entirely correct for high value	0–1
F1 Score	Harmonic mean of precision and recall	Answer is both precise and complete for high value	0–1
WUP Similarity	Semantic similarity via WordNet hierarchy	Answer is semantically close to the reference for high value	0–1

Table 1: Comparison of VQA evaluation metrics. Sources: (Antol et al., 2015; Rajpurkar et al., 2016; Malinowski and Fritz, 2014).

LLMs	Pretrained on Hausa?	Fine-tuned on Hausa?	Parameters
mt0-base	Yes	No	580M
mt0-large	Yes	No	1.2B
afriberta_large	Yes	No	126M
afro-xlmr-large	Yes	No	560M
gemini	Yes	No	770M
bloomz560	No	No	560M
bloomz1b7	No	No	1.7B
llama-3.2-1B	No	No	1.23B
deepseek-R1-1.5B	No	No	1.5B

Table 2: Information on Hausa pretraining, fine-tuning, and the number of parameters of the different LLMs used for the Hausa VQA system.

stance is a triplet $(q,\ I,\ a)$: a Hausa question q, its associated image I, and the gold-standard answer a. The question q is tokenised and embedded using a large language model (LLM) encoder, while the image I is patchified and processed by a vision transformer to obtain visual embeddings. The answer space $\mathcal{A}$ is defined as a fixed set of label vectors, each encoding a valid Hausa answer. Textual and visual embeddings are fused, typically via cross-modal attention, into a unified representation. A classification head then projects this fused vector onto the label space, and the top-ranked label is decoded into Hausa text as the final prediction. This baseline serves as the reference configuration to assess the impact of data augmentation strategies.

4.4.2 Offline Augmentation training

To construct the augmented dataset, each English question–answer–image triplet $(q_{en},\ i_{en},\ a_{en})$ from the HaVQA corpus (Parida et al., 2023a) is duplicated: the original instance is preserved, while its copy is transformed. For the textual components, synonym replacement (Wei and Zou, 2019) is applied exclusively to the duplicated English questions q_{en} by using the WordNet module of the NLTK library. This produces a parallel English question–answer set containing both the original questions and their paraphrased variants, each paired with the same gold-standard answers. The resulting English dataset is subsequently translated into Hausa (Parida et al., 2023a) using the Translator module from the googletrans library. For the visual components, only the duplicated images undergo geometric transformations such as random horizontal flips and slight rotations (Shorten and Khoshgoftaar, 2019b).

By combining the original triplets with their augmented counterparts, we obtain the augmented dataset HaVQA_aug[3], whose size is approximately twice that of the original HaVQA corpus. Each training instance is represented as $(q_{ha\text{-}aug},\ i_{ha\text{-}aug},\ a_{ha\text{-}aug})$, where $q_{ha\text{-}aug}$ denotes the paraphrased and translated Hausa question, $i_{ha\text{-}aug}$ the transformed image, and $a_{ha\text{-}aug}$ the corresponding Hausa answer.

The training architecture mirrors the baseline setup: the question is encoded by an LLM, the augmented image is processed by a vision transformer, and the fused multimodal representation is classified into a fixed set of Hausa answer categories. The offline approach enriches both text and image modalities prior to training.

Image Encoder	Parameters
vit-base-patch16-224-in21k	86.4M
clip-vit-base-patch32	149M
mae-base	86M
deit-base-patch16-224	86M

Table 3: Number of parameters of the different image encoders used for the Hausa VQA system.

Dataset	Train	Test	Total Images
HaVQA_aug	9,625	2,407	3,110

Table 4: Details of the HaVQA dataset's offline-augmented partitions for training and evaluation.

5 Results and Discussion

The best Accuracy and WuPalmer scores achieved by each LLM under these regimes are presented in Tables 5 and 6.

Under the no-augmentation baseline, llama-3.2-1B + clip-vit-base-patch32 achieves 19.68% Accuracy and WUP, with the highest F1 among all pairs (Table 5). This remains below the 30.86% WUP reported for DeiT-Base-P-224 + BERT-Base-Hausa (Parida et al., 2023a). Notably, llama-3.2-1B performs robustly across all image encoders despite lacking Hausa-specific pretraining. With offline augmentation, most LLM–image pairs show larger gains, particularly for Hausa-pretrained LLMs. Gemini + vit-base-patch16-224-in21k reaches Wu-

[3] https://github.com/Alimiji/LLM_QRV_Hausa_HaVQA_aug

LLMs	Image Encoder	WuPalmer	Accuracy	F1
mt0-base	deit-base-patch16-224	15.45	15.45	0.35
mt0-large	clip-vit-base-patch32	15.61	15.87	1.03
afriberta_large	vit-base-patch16-224-n21k	18.27	18.42	1.05
afro-xlm-large-7.6b	vit-base-patch16-224-n21k	15.24	15.32	0.49
gemini	vit-base-patch16-224-n21k	35.89	35.85	1.86
bloomz560	deit-base-patch16-224	15.40	15.42	0.34
bloomz1b7	deit-base-patch32	17.62	17.64	0.89
deepseek-R1-1.5B	deit-base-patch32	19.86	19.84	1.73
llama-3.2-1B	deit-base-patch32	18.46	18.36	1.36

Table 5: Best Accuracy, WuPalmer, and F1 score per LLM on HaVQA — *baseline configuration*.

LLMs	Image Encoder	WuPalmer	Accuracy	F1
mt0-base	mae-base	32.19	32.16	9.89
mt0-large	mae-base	35.30	35.27	13.45
afriberta_large	clip-vit-base-patch32	32.74	32.70	7.84
afro-xlmr-large-76L	clip-vit-base-patch32	28.45	28.42	4.47
gemini	vit-base-patch16-224-in21k	35.89	35.85	15.32
bloomz560	clip-vit-base-patch32	18.01	17.95	1.24
bloomz1b7	vit-base-patch16-224-in21k	18.36	18.36	0.63
deepseek-R1-1.5B	clip-vit-base-patch32	21.70	21.69	3.42
llama-3.2-1B	clip-vit-base-patch32	17.99	17.99	3.11
gemini	deit-base-patch16-224	33.52	33.49	12.79

Table 6: Best Accuracy, WuPalmer, and F1 score per LLM on HaVQA — *offline* augmentation.

Palmer 35.89%, Accuracy 35.85%, and F1 15.32% (Table 6). mT0-base/large with MAE-Base also benefits substantially (WuPalmer 32.19/35.30%, Accuracy 32.16/35.27%, F1 9.89/13.45%). Non-Hausa-pretrained models such as bloomz and llama-3.2-1B gain modestly; e.g., llama-3.2-1B + clip-vit-base-patch32 shows F1 3.11%, similar to inline augmentation, indicating that its baseline robustness persists but improves little under offline augmentation. Overall, offline augmentation is most effective for Hausa-pretrained backbones.

5.1 Systems Analysis and Limitations

Despite promising results, our Hausa VQA system has several limitations. First, HaVQA is limited in size and diversity, and automatically augmented Hausa texts can introduce noise.bSecond, the system handles only static images, lacking richer multimodal inputs (e.g., video, spatial context), and dialectal or orthographic variation in Hausa is underrepresented, limiting generalization. Model interpretability is also limited, as large models remain opaque without integrated explainability mechanisms.

Our analysis shows that offline augmentation benefits Hausa-pretrained LLMs (mT0, AfriBERTa, Afro-XLM-R, Gemini) considerably, while non-Hausa models (BloomZ, LLaMA, DeepSeek) gain modestly (Table 6). This suggests improvements stem not merely from increased data volume, but from linguistically compatible paraphrases exploited by Hausa-aware tokenizers. However, none of our models is fine-tuned on Hausa (Table 2),

Image Encoder	Text Encoder	WuPalmer
BEiT-L-P224	Bert-Hausa	27.76
ViT-B-P224	Bert-Hausa	28.91
ViT-L-P224	Bert-Hausa	29.67
DeiT-B-P224	**Bert-Hausa**	**30.86**

Table 7: WuPalmer scores of *Bert-Hausa* with four visual encoders on HaVQA(Parida et al., 2023b).

so offline augmentation may act as a proxy for adaptation, potentially overstating gains. Future work includes controlled experiments to separate volume and diversity effects and stricter filtering of synthetic paraphrases.

6 Conclusion and Future Work

This research presents a classification-based Hausa VQA system, combining fine-tuned LLMs with state-of-the-art vision transformers. Experiments on HaVQA show that offline multimodal augmentation, tailored to Hausa linguistic and cultural features, substantially improves performance, achieving 35.85% Accuracy, 35.89% WuPalmer, and 15.32% F1—exceeding prior benchmarks by over 5%. These results highlight the value of language-specific pretraining and multimodal enrichment in low-resource VQA. Future work includes extending HaVQA into a multilingual MT-VQA benchmark, improving interpretability via cross-modal attention analysis, and generalizing the framework to other African languages. We also plan to enhance deployment through model compression, knowledge distillation, and multimodal extensions to speech and video. We will also emphasize ethical alignment, cultural sensitivity, and bias mitigation to foster inclusive and equitable multimodal AI for underrepresented languages.

References

David Ifeoluwa Adelani and 1 others. 2025. IrokoBench: A new benchmark for african languages in the age of large language models. https://aclanthology.org/2025.naacl-long.139/.

Jesujoba Olabode Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In *Findings of ACL*, pages 1–17.

Stanislaw Antol, Arjun Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering.

In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Long Chen, Yuhang Zheng, and Jun Xiao. 2022. Rethinking data augmentation for robust visual question answering. In *European Conference on Computer Vision (ECCV)*.

Alexis Conneau, Karthik Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Felipe Guzmän, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Luu Anh Tuan, and Shafiq Joty. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1679–1705.

John Doe, Alice Smith, and Rahman Mohammed. 2023. Afrolm: A self-active learning-based multilingual pretrained language model for 23 african languages. *arXiv preprint*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. Beyond answering: Towards multi-step reasoning in machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8849–8863.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.

Katharina Kann, Kyunghyun Cho, and Samuel R. Bowman. 2019. Towards realistic practices in low-resource natural language processing: The development set. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3342–3349, Hong Kong, China. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *International Journal of Computer Vision (IJCV)*, volume 123, pages 32–73. Springer.

Raghavendra Kumar, Michael Hedderich, Bonaventure F Dossou, Chris C Emezue, Krešimir Šojat, Heike Adel, and Iryna Gurevych. 2022. Towards data and benchmarking for african languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3554–3573.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32. Curran Associates, Inc.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in Neural Information Processing Systems*, 27.

Francis Stephen MBWANA and Dang Long Hoang. 2025. Swahilivqa: A dataset for visual question answering in swahili language. In *2025 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6.

Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Idris Abdulmumin, Falalu Ibrahim Lawan, Sukairaj Hafiz Imam, Yusuf Aliyu, Sani Abdullahi Sani, Ali Usman Umar, Tajuddeen Gwadabe, Kenneth Church, and Vukosi Marivate. 2025. HausaNLP: Current status, challenges and future directions for Hausa natural language processing. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 176–191, Vienna, Austria. Association for Computational Linguistics.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Aishwarya Agrawal, and 1 others. 2024. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*.

Wilhelmina Nekoto, Vukosi Marivate, Taiwo Fagbohungbe, and et al. 2020. Participatory research for low-resourced machine translation: A case study

in african languages. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–11.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David Ifeoluwa Adelani. 2023. How good are large language models on african languages? *arXiv preprint arXiv:2311.07978*.

Victor Tolulope Olufemi, Oreoluwa Boluwatife Babatunde, Emmanuel Bolarinwa, and Kausar Yetunde Moshood. 2025. Challenging multimodal LLMs with african standardized exams: A document VQA evaluation. In *CVPR 2025 Workshop Vision Language Models For All*.

Shantipriya Parida, Idris Abdulmumin, Shamsuddeen H. Muhammad, Aneesh Bose, Guneet S. Kohli, Ibrahim S. Ahmad, Ketan Kotwal, Sayan D. Sarkar, Ondrej Bojar, and Habeebah A. Kakudi. 2023a. Havqa: A dataset for visual question answering and multimodal research in hausa language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10162–10183.

Shantipriya Parida, Idris Abdulmumin, Shamsuddeen Hassan Muhammad, Aneesh Bose, Guneet Singh Kohli, Ibrahim Said Ahmad, Ketan Kotwal, Sayan Deb Sarkar, Ondřej Bojar, and Habeebah Kakudi. 2023b. HaVQA: A dataset for visual question answering and multimodal research in Hausa language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10162–10183, Toronto, Canada. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1715–1725.

Connor Shorten and Taghi M. Khoshgoftaar. 2019a. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60.

Connor Shorten and Taghi M. Khoshgoftaar. 2019b. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, and 8 others. WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. ACL '94, page 133–138, USA. Association for Computational Linguistics.

Suorong Yang, Weikang Xiao, Mengchen Zhang, Suhan Guo, Jian Zhao, and Furao Shen. 2022. Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*.

Hao Yu, Jesujoba Oluwadara Alabi, Andiswa Bukula, Jian Yun othersZhuang, and 1 others. 2025. IN-JONGO: A multicultural intent detection and slot-filling dataset for 16 African languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9429–9452, Vienna, Austria. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 649–657.