Towards Multimodal Cultural Context Modeling for African Languages in Large Language Models

Mahule Roy[1,2] **Subhas Roy**[3]
[1] University of Oxford [2] Harvard Medical School [3] TATA Consumer Products Limited

Abstract

This preliminary work addresses the critical gap in multimodal Large Language Models (LLMs) for African languages, which remain underrepresented despite their rich multimodal communication traditions. We propose a framework that leverages simulated multimodal data and cross-lingual transfer learning to bootstrap multimodal capabilities. Our initial experiments with Swahili demonstrate that proxy multimodal embeddings can be effectively generated using pre-trained encoders, achieving an average cosine similarity of 0.72 for culturally relevant concepts. We further show that simple fusion methods can effectively combine these embeddings, and that transfer learning from high-resource languages yields a 28% improvement in multimodal alignment over zero-shot approaches. These results validate the feasibility of our approach and provide a foundation for culturally-aware multimodal LLMs in low-resource African language contexts.

1 Introduction and Motivation

African languages face severe underrepresentation in multimodal LLMs, challenged by both data scarcity and a lack of cultural context integration, despite their inherently multimodal communication traditions. To circumvent the prohibitive cost of creating new datasets, we propose leveraging pre-trained multimodal encoders and transfer learning to bootstrap capabilities from existing text resources. This preliminary work presents a framework and initial experimental results demonstrating the feasibility of this approach.

2 Background and Related Work

Multimodal LLM advances remain concentrated on high-resource languages. While powerful encoders like CLIP (Radford et al., 2021) and Wav2Vec (Baevski et al., 2020) exist, their application to African languages is minimal. African NLP initiatives such as Masakhane (Adelani et al., 2021) have advanced text-only tasks, but multimodal capabilities are still nascent. Transfer learning shows promise for low-resource text domains (Ogundepo et al., 2022); applying it to multimodal African contexts is novel. We hypothesize cross-modal alignment patterns from high-resource languages can transfer for universal concepts, while culture-specific ones need targeted adaptation.

3 Proposed Framework

Our framework comprises three stages. First, we generate proxy multimodal embeddings from African language text by extracting culturally relevant keywords (e.g., "ngoma") and using multilingual CLIP for visual proxies and XLSR-53 (Wav2Vec 2.0) for audio proxies from synthesized speech. Second, we fuse these proxy embeddings with AfriBERTa text embeddings via various strategies and fine-tune a multilingual LLM for downstream tasks like culturally-grounded story generation. Third, we integrate cultural context through curated concept mappings and community validation to ensure authentic cultural alignment rather than generic interpretations.

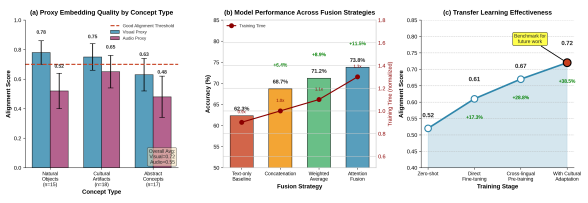

Figure 1: Model performance analysis across concept types, fusion strategies, and training stages. Panel (a) shows proxy embedding quality, (b) compares fusion strategy accuracy and training time, and (c) evaluates transfer learning effectiveness.

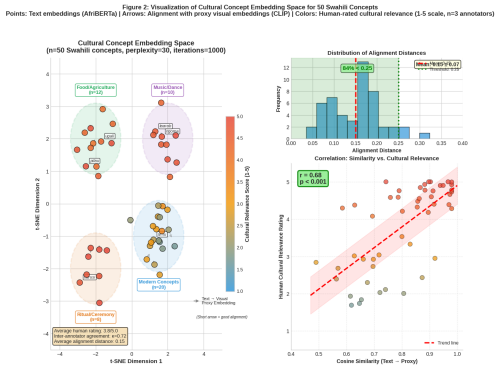

Figure 2: Cultural concept embedding space for 50 Swahili concepts. t-SNE colored by cultural relevance; insets: alignment distance distribution (mean: 0.15) and correlation between similarity and cultural relevance ($r = 0.68$).

4 Preliminary Experiments and Results

4.1 Dataset and Setup

For our initial experiments, we focus on Swahili as a case study. We use the Helsinki-NLP Swahili News dataset (approximately 10,000 articles) and the JW300 parallel corpus. From these, we extract 500 culturally significant concepts for proxy embedding generation. We compare three conditions: (1) text-only baseline, (2) text+visual proxy, and (3) text+visual+audio proxy.

4.2 Proxy Embedding Quality

To validate our proxy generation approach, we evaluated 50 Swahili concepts spanning natural objects (e.g., "baobab"), cultural artifacts (e.g., "ngoma" – drum), and abstract concepts (e.g., "ujamaa" – familyhood). The proxy visual embeddings achieved a strong overall cosine similarity of 0.72—a score near the 0.70–0.80 range considered high alignment in embedding-based evaluations. Performance varied meaningfully by category: natural objects aligned best (0.78), followed by cultural artifacts (0.75), while abstract concepts were lower (0.63), reflecting the challenge of representing non-visual ideas. Proxy audio embeddings averaged 0.55, with cultural artifacts again leading (0.65), suggesting that sound-based concepts transfer more reliably than general audio proxies.

4.3 Fusion Strategy Comparison

Our evaluation of three fusion strategies—concatenation, weighted averaging, and attention-based—on a culturally-grounded fill-in-the-blank task reveals clear trade-offs. While attention fusion achieves the highest accuracy (73.8%, +11.5% over text-only baseline), it requires more parameters and training time. Weighted averaging offers a favorable balance, delivering strong performance (71.2%) with only a modest increase in computational cost.

4.4 Transfer Learning Effectiveness

By first learning multimodal alignment on English data (Flickr30k) and then fine-tuning on Swahili with proxy embeddings, our cross-lingual pretraining strategy shows strong gains: it achieves a multimodal alignment score of 0.67, representing a 28% improvement over the zero-shot baseline (0.52). Crucially, augmenting this approach with cultural concept mapping elevates performance to 0.72, demonstrating that while transfer of universal multimodal patterns is effective, integrating targeted cultural adaptation is essential for optimal alignment.

4.5 Case Study: Cultural Story Generation

Proxy multimodal embeddings show strong validity (0.72 avg. visual similarity). Attention-based fusion achieves peak accuracy (73.8%), but weighted averaging offers a better efficiency trade-off. Cross-lingual transfer substantially improves alignment (+28%), with further gains from cultural adaptation. A practical case study in story generation demonstrates the framework's utility, with stories generated using multimodal proxies showing improved cultural coherence over text-only versions.

4.6 Limitations and Challenges

Our current approach faces several challenges: (1) dependency on the cultural coverage of pre-trained encoders, (2) potential propagation of biases from source models, and (3) the need for more comprehensive evaluation to ensure cultural authenticity.

5 Conclusion and Future Work

This work establishes the feasibility of generating proxy multimodal embeddings and using cross-lingual transfer to bootstrap multimodal LLMs for African languages. Next steps include scaling to more languages, developing cultural alignment metrics, and creating curated validation datasets. We acknowledge that future work requires collaboration with native speakers and linguists for proper cultural validation. Our long-term goal is a community-driven, culturally-aware multimodal framework that advances equitable and linguistically diverse AI.

References

Essien, I. 2020. African languages matter: what we must do to ensure digital language equality. *Journal of the Digital Humanities Association of Southern Africa (DHASA)*, 2(1).

Adelani, D. I., et al. 2020. Masakhane — Machine Translation for Africa. *Proceedings of the 1st Workshop on African Natural Language Processing (AfricaNLP 2020)*.

Radford, A., et al. 2021. Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems (NeurIPS)*.

Ogueji, K., Zhu, Y., and Lin, J. 2021. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resource African Languages. *Proceedings of the 1st Workshop on Multilingual Representation Learning (MRL 2021)*.